



Molecular Modeling

ACS SYMPOSIUM SERIES 576

Molecular Modeling

From Virtual Tools to Real Problems

Thomas F. Kumosinski, EDITOR
U.S. Department of Agriculture

Michael N. Liebman, EDITOR
Amoco Technology Company

Developed from a symposium sponsored
by the Division of Agricultural and Food Chemistry
at the 205th National Meeting
of the American Chemical Society,
Denver, Colorado,
March 28–April 2, 1993



American Chemical Society, Washington, DC 1994

Molecular modeling



Library of Congress Cataloging-in-Publication Data

Molecular modeling: from virtual tools to real problems / Thomas F. Kumosinski, editor, Michael N. Liebman, editor.

p. cm.—(ACS symposium series, ISSN 0097-6156; 576)

“Developed from a symposium sponsored by the Division of Agriculture and Food Chemistry at the 205th National Meeting of the American Chemical Society, Denver, Colorado, March 28–April 2, 1993.”

Includes bibliographical references and indexes.


ISBN 0-8412-3042-0

1. Biomolecules—Computer simulation—Congresses.
2. Biomolecules—Structure—Mathematical models—Congresses.
3. Proteins—Structure—Computer simulation—Congresses.
4. Proteins—Structure—Mathematical models—Congresses.

I. Kumosinski, Thomas F. II. Liebman, Michael N., 1947—
III. American Chemical Society. Division of Agricultural and Food Chemistry. IV. American Chemical Society. Meeting (205th: 1993: Denver, Colo.) V. Series.

QP517.M3M65 1994
547.7'0442'011—dc20

94-38705
CIP

The paper used in this publication meets the minimum requirements of American National Standard for Information Sciences—Permanence of Paper for Printed Library Materials, ANSI Z39.48-1984. 

Copyright © 1994

American Chemical Society

All Rights Reserved. The appearance of the code at the bottom of the first page of each chapter in this volume indicates the copyright owner's consent that reprographic copies of the chapter may be made for personal or internal use or for the personal or internal use of specific clients. This consent is given on the condition, however, that the copier pay the stated per-copy fee through the Copyright Clearance Center, Inc., 27 Congress Street, Salem, MA 01970, for copying beyond that permitted by Sections 107 or 108 of the U.S. Copyright Law. This consent does not extend to copying or transmission by any means—graphic or electronic—for any other purpose, such as for general distribution, for advertising or promotional purposes, for creating a new collective work, for resale, or for information storage and retrieval systems. The copying fee for each chapter is indicated in the code at the bottom of the first page of the chapter.

The citation of trade names and/or names of manufacturers in this publication is not to be construed as an endorsement or as approval by ACS of the commercial products or services referenced herein; nor should the mere reference herein to any drawing, specification, chemical process, or other data be regarded as a license or as a conveyance of any right or permission to the holder, reader, or any other person or corporation, to manufacture, reproduce, use, or sell any patented invention or copyrighted work that may in any way be related thereto. Registered names, trademarks, etc., used in this publication, even without specific indication thereof, are not to be considered unprotected by law.

PRINTED IN THE UNITED STATES OF AMERICA

American Chemical Society
Library
1155 16th St., N.W.
Washington, D.C. 20036

In Molecular Modeling: Kumosinski, et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 1994.

1994 Advisory Board

ACS Symposium Series

M. Joan Comstock, *Series Editor*

- | | |
|---|---|
| Robert J. Alaimo
Procter & Gamble Pharmaceuticals | Douglas R. Lloyd
The University of Texas at Austin |
| Mark Arnold
University of Iowa | Cynthia A. Maryanoff
R. W. Johnson Pharmaceutical
Research Institute |
| David Baker
University of Tennessee | Julius J. Menn
Western Cotton Research Laboratory,
U.S. Department of Agriculture |
| Arindam Bose
Pfizer Central Research | Roger A. Minear
University of Illinois
at Urbana–Champaign |
| Robert F. Brady, Jr.
Naval Research Laboratory | Vincent Pecoraro
University of Michigan |
| Margaret A. Cavanaugh
National Science Foundation | Marshall Phillips
Delmont Laboratories |
| Arthur B. Ellis
University of Wisconsin at Madison | George W. Roberts
North Carolina State University |
| Dennis W. Hess
Lehigh University | A. Truman Schwartz
Macalaster College |
| Hiroshi Ito
IBM Almaden Research Center | John R. Shapley
University of Illinois
at Urbana–Champaign |
| Madeleine M. Joullie
University of Pennsylvania | L. Somasundaram
DuPont |
| Lawrence P. Klemann
Nabisco Foods Group | Michael D. Taylor
Parke-Davis Pharmaceutical Research |
| Gretchen S. Kohl
Dow-Corning Corporation | Peter Willett
University of Sheffield (England) |
| Bonnie Lawlor
Institute for Scientific Information | |

Foreword

THE ACS SYMPOSIUM SERIES was first published in 1974 to provide a mechanism for publishing symposia quickly in book form. The purpose of this series is to publish comprehensive books developed from symposia, which are usually “snapshots in time” of the current research being done on a topic, plus some review material on the topic. For this reason, it is necessary that the papers be published as quickly as possible.

Before a symposium-based book is put under contract, the proposed table of contents is reviewed for appropriateness to the topic and for comprehensiveness of the collection. Some papers are excluded at this point, and others are added to round out the scope of the volume. In addition, a draft of each paper is peer-reviewed prior to final acceptance or rejection. This anonymous review process is supervised by the organizer(s) of the symposium, who become the editor(s) of the book. The authors then revise their papers according to the recommendations of both the reviewers and the editors, prepare camera-ready copy, and submit the final papers to the editors, who check that all necessary revisions have been made.

As a rule, only original research papers and original review papers are included in the volumes. Verbatim reproductions of previously published papers are not accepted.

M. Joan Comstock
Series Editor

Preface

THE IMPORTANCE OF PROTEINS in the field of food and medicinal sciences has never been of more importance than it is in today's global marketplace. In the future, biotechnology is expected to lead to the development of designer-type products for food and nonfood uses with tailor-made functionalities, for example, products that have pesticide resistance or transport stability as well as nutritional and therapeutic products. The current thinking is that modification of molecules as well as genetic engineering will provide the vehicle for success. Ultimately, genetic engineering may even allow the possibility of the creation of transgenic animals and plants whose byproducts will need only a small amount of processing to achieve a desired product. However, the historical problem of developing quantitative measures of animal and plant components for structure–function relationships with high predictability will ultimately plague the researcher. Without the knowledge of these relationships, the researcher or developer will be limited to costly hit-or-miss experiments that have limited success rates. Many proteins that have great economic impact on the food industry are noncrystallizable and, therefore, their three-dimensional structures cannot be determined by X-ray crystallography, which is currently the best technique for structural determination.

In recent years, the emergence of molecular modeling as a technique for refining existing three-dimensional molecular structures and for building new predicted models has yielded a methodology capable of developing a molecular basis for structure–function relationships.

Now, not only food proteins but preservatives, emulsifiers, stabilizers, and so on can be modeled for their effectiveness in a food system. Similarly, new peptides, carbohydrates, polysaccharides, and small molecules can be tested for their potential effectiveness.

For these reasons, we organized the symposium upon which this book is based. This symposium brought together a group of 30 internationally recognized experts to address the issue of molecular modeling from experiment to computation. The emphasis of the work discussed included the needs of the food and agriculture industries, which may appear on the surface to be different from the more traditional areas of modeling in medicinal chemistry, but is warranted by the new developments in plant and animal biotechnology and genome analysis. One only has to look at

the most widely read scientific journals that publish such molecular modeling results to see the presence and importance of molecular modeling in their publications.

The overall organization of this book follows that of the symposium. The intent was to bring together experimentalists and theoreticians to discuss new approaches for biological and food systems. The broad sections include experimental methods for molecular structure, prediction of molecular structure, analysis at the molecular level for function at system level, methods for study of molecular interaction—both experimental and computational—and recognition of molecular interactions. Because molecular structure determination is an essential part of molecular biology, and molecular biology is concerned with explaining functions at the molecular level, this volume will be useful as a reference book not only to food chemists and food scientists, but also to molecular biologists, physical biochemists, physical chemists, biophysicists, and biotechnologists. This book will also have worldwide appeal for scientists in the fields of drug design, environmental science, and plant and animal physiology.

THOMAS F. KUMOSINSKI
Eastern Regional Research Center
Agricultural Research Service
U.S. Department of Agriculture
600 East Mermaid Lane
Philadelphia, PA 19118

MICHAEL N. LIEBMAN
Bioinformatics Program
Amoco Technology Company
Mail Code F-2
150 West Warrenville Road
Naperville, IL 60563-8460

October 27, 1994

Chapter 1

Molecular Modeling Transferring Technology to Solutions

Michael N. Liebman

**Bioinformatics Program, Amoco Technology Company, Mail Code F-2,
150 West Warrenville Road, Naperville, IL 60563-8460**

The changes which are being realized in the food and agricultural product areas suggest that the time is appropriate to consider transfer of technology to assist in the rational design of new products and assessment of potential benefit and risk. In this overview we present an introduction to where molecular modeling is today and where it's evolution is headed in terms of hardware and software capabilities. The difficult issues involved in complex problem-solving are identified in terms of relevant research goals to reveal value in identifying and solving the correct problem to reach these goals. The value in looking to new methodologies and applications is also described.

Man has long attempted to effect control over the plants and animals which have shared his environment and provided food, fuel, clothing and early forms of transportation. This control has involved development and utilization of crop science, soil science and animal husbandry to effect enhanced food quality and production, resistance to disease and enhancement of other consumer-driven characteristics, e.g. color, taste, texture. Historically, an advantage in applying these technologies to agricultural products, whether as chemical additives or modulators or cross-breeding in plants or animals, has been the ease of access to the experimental system. Recent focus by both

0097-6156/94/0576-0001\$08.00/0
© 1994 American Chemical Society

In Molecular Modeling; Kumosinski, T., et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 1994.

regulatory and consumer groups indicate the need to improve methods for characterizing experimental systems and to establish more accurate risk assessment capabilities. Access to information concerning the genetics, physiology and the structure and function of molecular targets in agriculture have added to the complexity of resolving these issues but also provide the opportunity to develop applications using more rational approaches.

The requirements for developing rational approaches to agricultural applications include both the integration of computational and experimental methodologies and the informatics and database technologies necessary to support the integration. Many of these component technologies have been developed for use in pharmaceutical research, where the target organism is not suitable for experimental development and model organisms only serve as candidate screens during drug development. It is thus of benefit to agricultural researchers to learn what technologies in rational drug design and protein engineering might be suitable for transfer to agricultural applications.

Common Paths, Common Needs and Technology Transfer

Product Development Path- The proposal for the potential transfer of technology can be better understood when one examines the analogy between agricultural and pharmaceutical needs. This is described in terms of the customers which comprise each area's respective product development path (Figures 1 and 2). Although the commonly held view might be that drug development proceeds along a different path and with different priorities, the analogous representation of pharmaceutical product development can also be viewed in terms of a sequence of producers and consumers.

Product Development Needs-While these paths may appear to involve different priorities, both have a fundamental need to develop an accurate analysis of the relationship between structure and function of the (macro)molecules which serve as targets for the products or as the products, themselves. Within the pharmaceutical industry, regulatory and economic considerations have resulted in stronger emphasis on the development and implementation of computational and experimental tools to aid in the determination of structural identities and features relevant to product development. The focus of the papers presented here is on providing an overview of the experimental and computational tools which represent current state-of-the-art, along with examples of their application to actual problems, to serve as a vehicle to support the transfer of appropriate technologies to problem solving in the agriculture domain.

Potential for Technology Transfer-The methods and

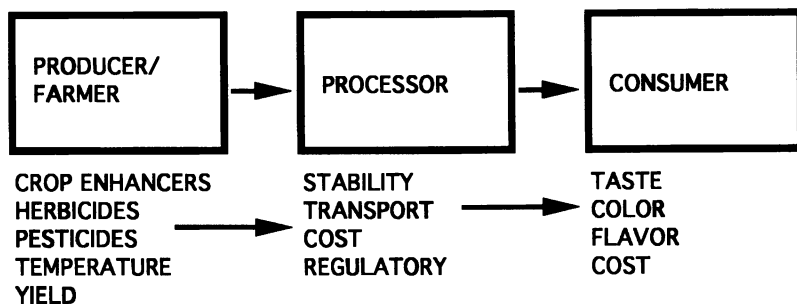


Figure 1. Agricultural Product Development Pathway showing the progression from source to processor to consumer, with their associated priorities and goals. This is constructed analogous to the Pharmaceutical Development Path in Figure 2.

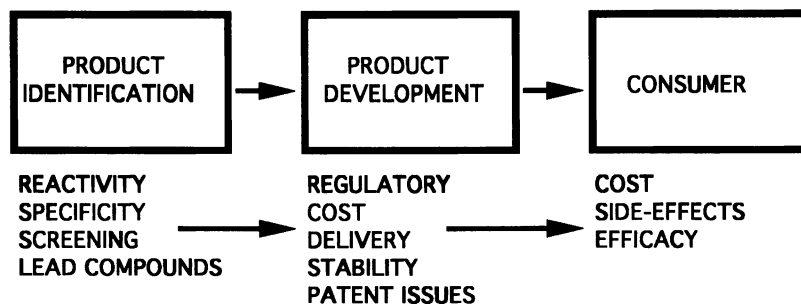


Figure 2. Pharmaceutical Development Path showing the progression from product or lead identification to product development to consumer, with their associated priorities and goals.

problems which are presented represent most of the continuum of problem areas which confront biological research, incorporating both today's existing information as well as in anticipation of access to genome sequences. This continuum includes gene location (either for target selection or as desired product), structural organization of the gene product (i.e. protein or enzyme), structural response to environmental factors (i.e. conformational), definition of specificity (in vivo versus in vitro), participation in higher order organizations (e.g. biological pathways), secondary effects produced by inter-pathway interaction (i.e. side-effects), and clinical observations (e.g. existing disease or risk assessment). This extension, from gene location to clinical observation, is viewed as the basis for defining the "new" medicine which will follow access to the sequence of the human genome and high speed sequencing tools. It can be readily seen to have a parallel in agriculture, e.g. development or selection of crop for particular disease resistance, etc. The commonality of these two broad applications resides in their utilization of the hierarchical processing of biological information from the genome through the gene product (Figure 3) to its biological function or dysfunction (Figure 4). The key to understanding the impact of this information processing is in the analysis of the structure and resulting function of the protein.

Common Goals

Sequence to Structure- A longstanding goal of structural biology has been the accurate prediction of the secondary and tertiary structure of a protein from its constituent amino acid sequence. The tacit assumption has been that adequate information exists within the amino acid sequence to enable the correct three-dimensional structure to be assembled during protein synthesis. This interpretation is based on the observed fidelity of protein conformation associated with a specified polypeptide sequence, and the ability of the sequence to return to its native conformation following denaturation in an environment devoid of other protein synthesis components, e.g. ribosomes. Significant computational and experimental research efforts have examined both the correlation of the amino acid sequence, or sequence-derived properties, with the resultant protein structure, and the mechanism by which the folding occurs. Most notable in the more than 20 year period of activity in this area has been the growth of the database of atomic resolution protein structures, initially observed by x-ray crystallography and more recently by NMR; the evolution of the methods used to analyze this data; and evolution of those questions which scientists hope to answer through such studies.

Since the first determination of the three-dimensional structure of a protein in a crystalline environment, the

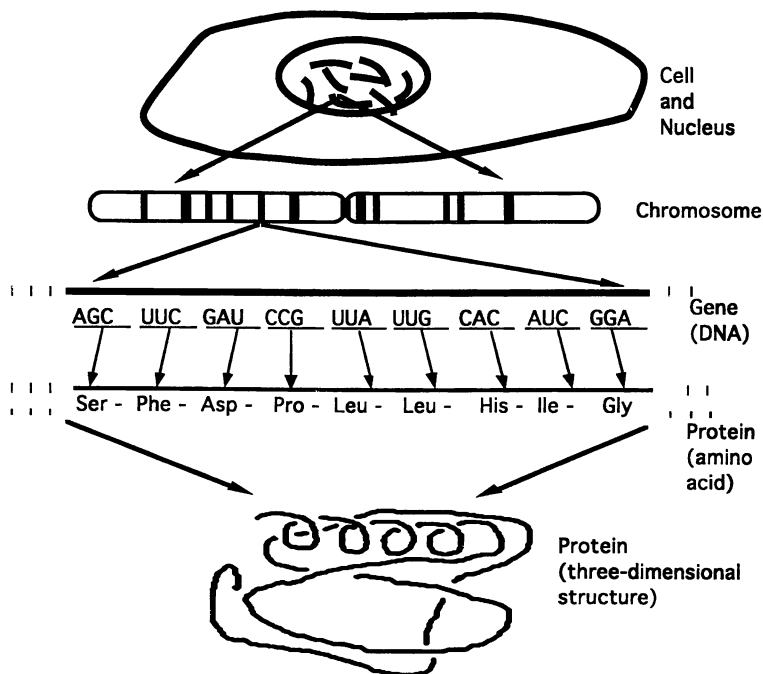


Figure 3. The “Holy Grail” of molecular biology as represented by the transformation of information from nucleus to chromosome to gene to gene sequence to amino acid sequence to three-dimensional structure of the gene product (protein).

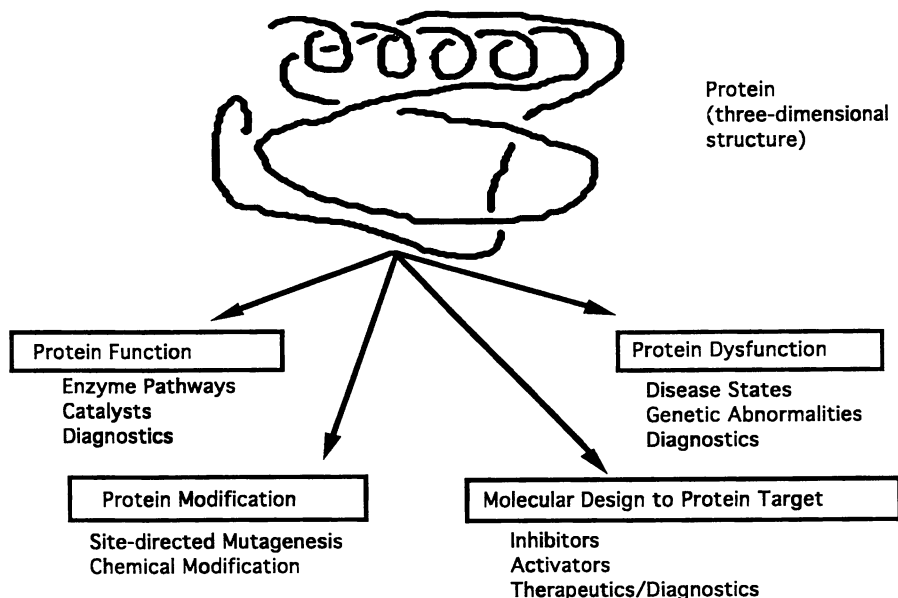


Figure 4. The targeted values associated with understanding the relationship between the three-dimensional structure of a protein and its biological function.

Downloaded by 89.163.34.136 on October 17, 2012 | http://pubs.acs.org
Publication Date: December 14, 1994 | doi: 10.1021/bk-1994-0576.ch001

information base has grown to more than 1000 structures, most of which are publicly available through the Protein Data Bank. This growth has been nonlinear and is indicative of technological advances which permit the higher resolution, greater than 2.0Å resolution, of many of the new entrants. Structures, or sets of conformations which are consistent with observed distance constraints, as revealed by two-dimensional NMR, have become recent additions to this database. As this database has grown, the appropriateness for application of statistical methods has increased, and studies have ranged from the correlation of local, secondary structural features, to longer range, tertiary structural interactions. This structural distinction is typically maintained through the selection of a viewing window onto the amino acid sequence, where the window boundaries are defined in terms of distance apart along the polypeptide chain. Such correlations have attempted to utilize both the first order data, i.e. amino acid sequence, and physicochemical parameters that may be derived from the sequence, e.g. hydrophobicity. More recent approaches have utilized the methodologies developed in computer science to explore knowledge bases and pattern recognition, as well as the connectionist approach. The most striking evolution in the process of protein structure prediction is reflected in the underlying questions and goals which continue to drive the conversion of information to knowledge, and reflects the more rapid technological advances and needs for the application of protein engineering and biotechnology. The difficult conversion of information into knowledge both derives benefit and suffers from increased access to computers and databases.

The goal of determining the three-dimensional structure of a protein from its constituent amino acid sequence stems from the universal acceptance that the function, in vivo and in vitro, of an enzyme (catalyst protein) is a direct consequence of its folded shape and its physical properties. The capability of describing and understanding the relationship between structure and function in these biological macromolecules is essential to success in modern medical applications including therapeutics and diagnostics, as well as in developing biotechnology-based chemical synthesis and dealing with impending environmental issues (Figure 4). Research in this area has been a significant focus of both the biotechnology and pharmaceutical industries. Public domain sequence databases are growing almost exponentially, fueled in part by the increasing interest in the Human Genome Sequencing Project at the international level. Structural information lags behind the sequence information by orders of magnitude with no obvious path to reduce this gap. The problem, from a structure prediction basis, can be viewed as the transfer of information along a pathway:

Gene Sequence -->	Protein Sequence -->	Protein Structure -->	Protein Function
----------------------	-------------------------	--------------------------	---------------------

Structure to Function- This sequence-to-structure pathway actually crosses only the first major barrier confronting researchers in this area. The most significant utilization of the structural information for proteins and enzymes derives from determination of their functional characteristics. This is of particular relevance to the sequence information provided in the Human Genome Initiative (or similarly from plant genomes) depicting several areas of near-term and long-term commercial potential. As in much of research, we ask questions which may not always address the information we need to know, but rather are bounded by our perception of what we believe can be answered. Although frequently unstated, we would like to accurately predict the structural features of an enzyme necessary to

1. understand the structural and evolutionary source of its specificity and reactivity;
2. understand adaptation through tissue, organism or even phylum differences;
3. design more specific inhibitors, substrates or activity modulators;
4. design and perform site-directed mutations which can predictably enhance thermal stability, response to pH, alter specificity or reactivity parameters;
5. identify/ classify a protein as to its functionality based on analysis of its amino acid/nucleic acid sequence alone;
6. address potential issues of patentability, patent protection or patent avoidance in defining new materials through these processes; and
7. acquire and maintain such knowledge in a form which can be applied in a potentially enhanced and proprietary manner, i.e. commercialize the product, the process and/or the knowledge.

We can summarize the value of the analysis of protein structure and its relation to function by representing the role of a protein in function and dysfunction, and as a target for designing a molecule, or as the molecule to be produced in a modified or unmodified form as shown in Figure 4.

Virtual Tools of Molecular Modeling

The key to successfully analyzing the complex relationship between structure and function requires data and information from both experimental and computational technologies and the integration of fundamental principles from a range of disciplines. The rapidly evolving (and improving) cost-performance curve is placing high

performance computer technology on the desktop. Access to high resolution, interactive computer graphics, conformational energy minimization, molecular dynamics, quantum mechanical calculations, homologous protein structure generation and sequence database searching are examples of some tools which can be routinely accessed through commercial molecular modeling and molecular biology software packages on these desktop workstations. The issue is no longer what can we integrate, but rather what should we integrate to solve a problem. In Figure 5, we show the range of tools which are commonly explored, both computationally and experimentally, when starting at the top with information about the gene product in terms of sequence, structure and/or concerning inhibitors or modulators of its function. The rational progression towards the application goals at the bottom can follow a multitude of paths, with no single path presently able to guarantee success, even for a specific class of problem, e.g. designing an in vivo inhibitor for a known enzyme structure. Access to the component tools has been significantly enhanced through use of commercial packages for molecular modeling, e.g. Biosym, Cambridge Scientific, ChemDesign, MDL Info Systems, Molecular Simulations, Oxford Molecular and Tripos, as well as for molecular biology, e.g. DNASTar, GCG, IntelliGenetics, MacVector, etc. These provide access to both the computational tools and user-friendly interfaces as well as linkages to the rapidly expanding databases of structure and sequence data. The problem is not how to access relevant data and information, but rather how to extract knowledge to make the information useful. Years of evolution have enabled biology to develop the appropriate rules and biological/biochemical/biophysical tools essential to process most of this information successfully, although mistakes do occur. The challenge to our achieving some degree of success lies in developing the ability to identify the information necessary to solve the problem from the wide band of information which is available, and develop the tools necessary to identify, select and apply that information efficiently. This challenge generalizes to any of the information-intense areas which confront us.

Defining Real Problems

The key to addressing this involves 1) learning how to identify the right question; 2) being able to evaluate if adequate information exists to answer that question and what that information is; and 3) evaluating the impact of the inherent bias present in either the answer or method we used to find the answer.

Asking the Right Question- The most difficult, and yet the most significant step in solving a problem involves establishing what are the actual goals to accomplish to

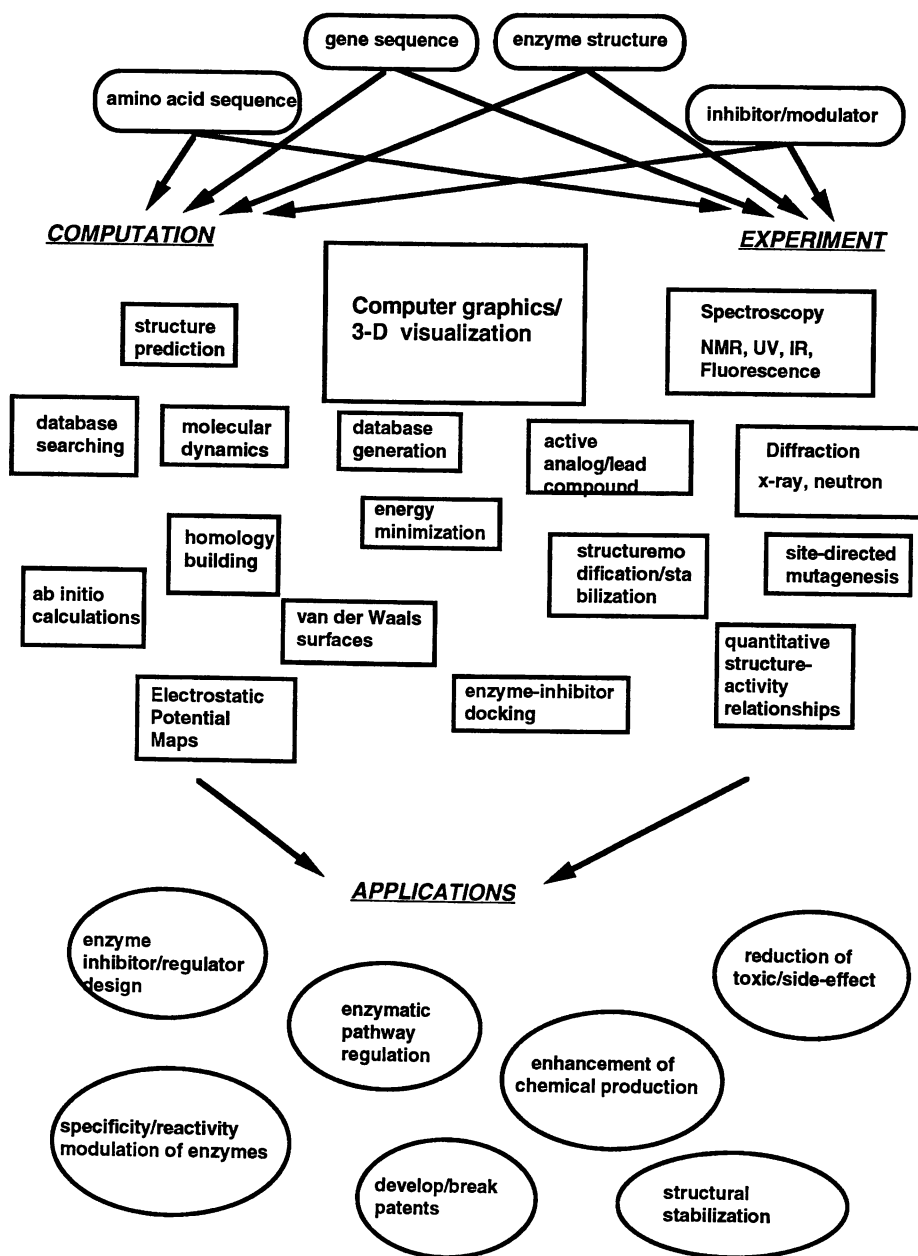


Figure 5. The “real-world” situation involving taking proprietary information and selecting the correct “pathway” (i.e. combination of experimental and computational molecular modeling approaches) to produce the desired knowledge or application.

complete the task. A general observation is that many questions are frequently stated which do not adequately reflect these goals, but rather reflect an anticipation of the solution or approach to solving the problem based on the existing knowledge of the person presenting the problem to be solved. As a result, the same question may be presented by different investigators, yet their actual needs may be very different. We can examine this distinction by expanding the question to examine underlying issues, and readily recognize that this is not a unique characteristic of problem-solving in the pharmaceutical or biological domains. Thus the first stage of problem-solving involves an identification and analysis of the true goals.

An example of a question, and one frequently asked, concerns the ability to predict the three-dimensional structure of a protein from its amino acid sequence. Can we predict the structure of a protein from its amino acid sequence? Initially confronted with this, some of the underlying questions include:

- do we want to predict protein structure or identify protein family?
- can we predict tertiary structure or, at least, secondary structure composition?
- how accurate a prediction is needed to be of value?
- how do we assess the accuracy of the predicted structure? by an overall comparison or by correlation of structural or functional features?
- an accuracy of 90-95% is a significant improvement over the current 65-70% barrier, but is adequate to enable accurate determination of structure and function?
- what accuracy do we need for rational design of an active site inhibitor, given that a full x-ray structure, i.e. 100% accuracy, does not guarantee success
- what accuracy do we need to rationally design a site directed mutation?
- can we understand differences in physiological function among homologous enzymes?
- is it necessary to predict structure from computational approaches without the benefit of using available experimental data, given that first principle calculations are not possible?
- do we want to inhibit an enzyme or the physiological process in which it functions?
- do we really need to predict structure, or predict and identify function within a molecule?

Do We Have Enough Information- In addition, it is necessary to evaluate whether sufficient information is available to answer the question presented, or the underlying questions:

- does an amino acid sequence correlate with a single structure, or a family of structures?
- is this correlation based on the amino acid identity, or

- on the physicochemical properties that an amino acid is capable of providing in a given environment?
- does experimental data, e.g. spectroscopy, physicochemical properties, provide additional information?
 - are there homologous proteins, either in the traditional sense of evolutionary homology, or as structural or functional homologues?
 - are the specificity profiles developed for substrates and inhibitors using in vitro assays for a given enzyme adequate to establish the in vivo structure-function role which governed its evolution?
 - if the goal is inhibition of a process, is the enzyme being studied the best target, or the target of convenience?
 - can information about natural mutations and their functional differences provide insight?
 - do we adequately understand how the molecule works, either in vitro or in vivo, even if we can accurately predict describe its structure?
 - can we use the modeling to rationally design experimental approaches which can assist in enhancing the prediction method?

Evaluating Bias in Data or Methodology- Also of importance is the need to examine the possibility of introducing or overlooking biases which may be present in either the experimental or computational methods of analysis, or in selecting the data used for the analysis:

- can any basis set of proteins adequately be constructed as a model for the world of proteins without introducing some bias?
- do specificities towards active-site directed inhibitors and\substrates adequately assay the natural specificities of enzymes which interact specifically with macromolecules?
- are environmental biases present, e.g. pH, temperature, compartmentalization?
- does enzyme mechanism adequately provide an objective means for classification, e.g. Enzyme commission numbers (E.C.)?
- do methods of analysis, which commonly emphasize similarity, enable detection of discrete but potentially significant structural and functional differences?

Dealing With Bias- By addressing each of the three challenges noted above, we can optimize the potential success of our approach to problem-solving. The most difficult challenge involves identifying and overcoming inherent bias in either the available or selected data to be analyzed, or the method of analysis, itself. The effect of these biases can be readily seen in their impact on the analysis of the structure-function relationship in proteins. If we use the definition that all proteins can

be classified as either soluble (i.e. water soluble) or membrane bound (i.e. non-water soluble), a representation of our current knowledge can be compared to the complete universe of proteins. The bias present in our existing knowledge is then the difference between those proteins which are known and the remainder of the universe of all proteins (Figure 6). Because we cannot quantitatively or qualitatively evaluate this bias, it has the potential to impact our ability to produce fully generalizable results from the data set being analyzed, and in a manner which remains unknown.

As shown in Figure 6, the size and therefore the potential impact of the unknown bias can be significant in defining the limit of our existing knowledge. As an example of how this bias might exist, if the soluble proteins which we studied included only hemoglobins, myoglobins, cytochrome b562, hemerythrins and erythrocurins, and the non-soluble proteins included rhodopsin and other seven-bundle membrane spanning proteins, we might assume that all proteins contained alpha helices as their main secondary structural feature. Because we know that other structural classes of proteins do exist, we can readily see the bias in this data set, but in examining the set of proteins whose structures we do know, we are not able to identify any bias which might similarly exist. A similar observation of bias can exist in any set of proteins which we select, not only prejudiced by structure, but also function, e.g. enzymes may not adequately represent structural proteins, as well as physico-chemical properties, e.g. molecular weight, subunit organization, etc. Biases may also have an historical origin. Serine proteases were first isolated from digestive organs because of convenience of access to material. Their proteolytic activity classified them as digestive enzymes with limited specificity towards macromolecular substrates rather than their highly refined role in limited proteolysis as observed in blood coagulation, complement activation, etc. Even today, newly discovered serine proteases are described as being chymotrypsin-like (i.e. hydrophobic residue substrate) or trypsin-like (i.e. charged residue substrate) independent of the physiological process in which they participate, and the greater range of substrate reactivity which has been observed. Amino acid sequences of serine proteases are numbered according to an alignment with the sequence in chymotrypsin, not because of its established role in the evolution of this family but because of the chronology of its observation.

Bias can also exist in the methodologies used or developed for problem-solving and limit our ability to adequately describe the systems under analysis. Interest in examining the secondary structure of a protein can be significantly limited by the common bias which focuses on the existence of regular structures including alpha helices and beta sheets. Both the ease of recognizing structural

regularity and the ability to focus on locating a limited set of structural features result in an observer, using a computer graphics display, to "see" the helices and sheets at the expense of other structural features which may appear less regular. In addition, the molecule will appear differently as it is viewed from the different orientations which result from rotation and translation of the atomic coordinates, thus hiding or exposing different structural regions during its graphical manipulation. By contrast, the information content of the molecular conformation remains constant throughout its graphical manipulation, meaning that the secondary structural features are not changing, only the ability to perceive them through this form of representation. It is also difficult to visually process structural regions of non-regular conformation for comparison with potentially similar regions in the same or other proteins, this further limits the ability to "see" other similar structural features in different proteins. These biases are compounded further when attempting to evaluate the similarity between two complete protein molecules, either visually using computer graphics as noted above, or using computer algorithms such as root-mean-squared superposition.

Compensation for bias in data selection or in methodology is most difficult when the bias can not be adequately measured. This limitation bears directly on the ability to generalize the results of the analysis and we cannot expect to completely eliminate these biases in either the data or methodologies. Thus our inability to develop general rules for protein structure prediction from amino acid sequence may be indicative of such a bias because critical information is both missing and not assessable from the existing data. Biases can be used constructively when they are used to focus on the particular aspects of a specific problem and generalization is not the overall goal. We can intentionally include a particular bias and evaluate the significance of the results and the limitation in their generalizability by careful analysis and interpretation of each stage of the analysis. We can construct, for example, a set of structurally and/or functionally related proteins using characteristics which might relate homologous families, proteins from the same tissue within an organism, operating at the same pH maximum, etc. We should expect that analysis of such sets would be able to best predict characteristics of the next member of this class, and failure to be accurate in predictions about the next member of a set would suggest an inability to generalize. Thus analysis of the serine proteases with an inability to accurately predict structure and function of the next member of this class strongly indicates that the potential for accurate prediction of non-serine proteases is significantly limited. The characteristics which are actually learned from such protein set analyses are clearly

class-associated, although some subset of these characteristics may be generally applicable. The difficulty and the value comes in developing the ability to recognize and separate these characteristics. The analysis is carried out in a parallel manner for multiple protein classes as depicted in Figure 7. This is conceptually analogous to the computational approach termed "divide and conquer".

Evolving New Methods for Problem Solving

As we have noted, conventional molecular modeling has progressed from the computational tools maintained and applied by specialists on high-performance computers and graphics workstations, to the delivery of hands-on toolsets on personal computers and workstations. These technological advances have helped to integrate modeling into the general problem-solving process by lowering the barriers to entry and significantly enlarging the community of its practitioners. A primary benefit of this evolution has been the enhanced communication between those who need to solve a particular problem and those who conceive of and develop the modeling tools. A secondary benefit, which is only beginning to be realized, is that this broader community of those using modeling has led to the introduction of new approaches based on the transfer of technology from other disciplines. Pattern recognition, neural networks, parallel processing, array processing, database architecture and searching, distributed computing, Petri Net analysis, artificial intelligence techniques and fuzzy logic are some of the newly incorporated methodologies, many of which rely on the improvement in computing access on the desktop.

The opportunity for the future resides in evolving the techniques for problem-solving beyond conventional boundaries to incorporate the wide-range of available technologies and disciplines. Examination of the functional requirements for defining and solving a specific problem can lead to the development of methodologies which can address a wide-range of seemingly disparate problems. Our interest in analyzing the blood coagulation pathway has led to development of tools for pathway representation, comparison and simulation, determination of potential control points for direct pathway intervention, e.g. enzyme inhibitor or drug, analogy to evolutionarily related pathways, e.g. complement or fibrinolysis, genetic origin of pathway evolution, risk assessment of secondary effects through linked pathways, assessment of genetic differences, e.g. mutated but non-lethal enzymes, in terms of functional differences in pathway function and response and identification of genetic control elements to potentially express pathways in alternative organisms. These tools are neither limited in their application to the coagulation pathway, nor to biological processes. Thus the potential

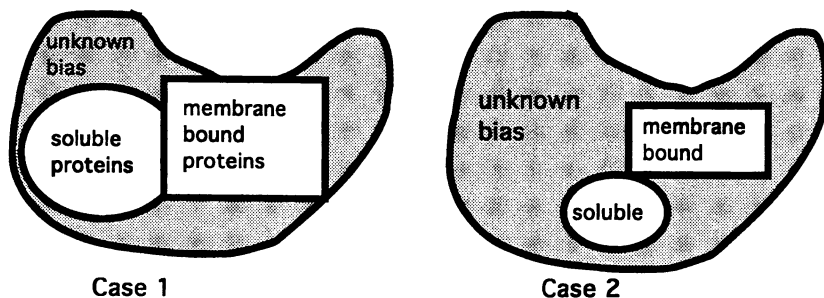


Figure 6. The bias which represents the influence of unknown or unattainable information remains unassessable in terms of that information which is known and presents a formidable barrier to modeling the whole world.

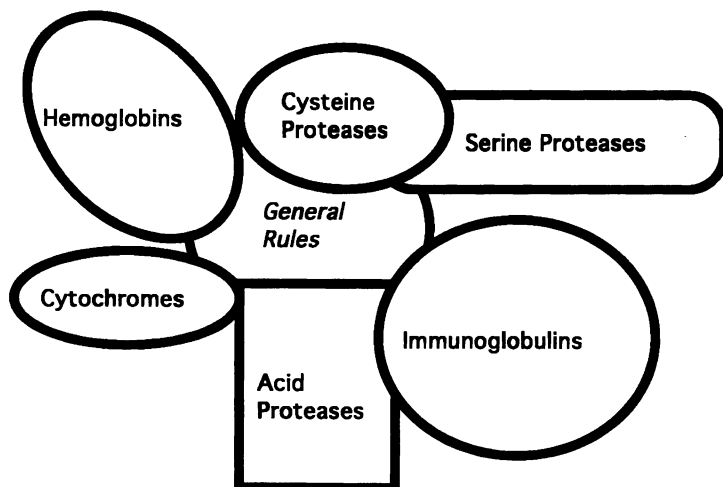


Figure 7. Utilizing a known bias to create parallel data models which can reveal information which contains the predetermined bias, and information which may be of general applicability.

for technology transfer out as well as in to the application should not be overlooked.

Conclusion

The overview presented here addresses the nature of problem-solving with a focus on issues arising in the prediction and analysis of protein structure and function. The goal has been to show the potential value in applying the methodologies developed and deployed within the pharmaceutical area to functionally analogous problems in agricultural and food chemistry. This discussion has been

directed at identifying underlying problems and questions which are present but not always addressed. This approach was used to show the potential which can still be realized by the rational development and utilization of computational modeling in conjunction with experimental verification to tackle the complex problems of today and tomorrow. The papers which follow in this volume serve as examples of moving conventional approaches to their successful application.

References

1. *Pharmaceutical R&D: Costs, Risks and Rewards*, Office of Technology Assessment, **1993**
2. *Mapping Our Genes*, Office of Technology Assessment, Johns Hopkins University Press, **1988**
3. *Annual Reports in Medicinal Chemistry*, Division of Medicinal Chemistry, J. Bristol, Editor, Academic Press
4. Bernstein, F.C., T.F. Koetzle, G.J.B. Williams, E.F. Meyer, Jr., M.D. Brice, J.R. Rodgers, O. Kennard, T. Shimanouchi and M. Tasumi, *J. Mol. Biol.* 112, 535, **1977**
5. Wilcox, G.L, Poliac M. and Liebman, M.N., *Tetra. Comp. Let.*, 191 **1990**
6. Liebman, M. N., and Brugge, A. L., *Santa Fe Institute Studies in the Sciences of Complexity, Volume VII*, eds. G. Bell and T. Marr, Addison-Wesley Longman Publishing Group, 183, **1989**;
7. Liebman, M. N., *J. Comp.-Aided Molecular Design* 1, 323 **1987**
8. Liebman, M. N., *J. Indus. Microbiology* 3, 127 **1988**
9. Liebman, M. N. *Enzyme* 36, 115, **1986**.
10. Prestrelski, S. J., Williams, A. L. and Liebman, M. N. *Proteins*, 14, 430 **1992**
11. Prestrelski, S. J., Byler, D. M. and Liebman, M.N., *Proteins*, 14, 440 **1992**
12. Liebman, M.N., Venanzi, C.A., Weinstein, H., *Biopolymers* 24, 1721, **1985**.
13. Liebman, M. N., *Application of Neural Networks to the Analysis of Structure and Function in Proteins, Second International Conference on Bioinformatics, Supercomputing and Complex Genome Analysis*, Ed. Lim, Fickett, Cantor and Robbins, World Scientific, p331 **1993**
14. Reddy, V. N., Mavrovouniotis, M. and Liebman, M. N., *Petri Net Representations in Metabolic Pathways*, ISMB, **1993** in press

RECEIVED August 2, 1994

Chapter 2

Reliability of X-ray Crystallographic Structures

Richard Bott

Department of Enzymology, Genencor International, 180 Kimball Way,
South San Francisco, CA 94080

The process of X-ray crystallographic structure determination requires growing crystals and visualizing these structures in electron density maps. This process introduces some limitations in the reliability of these structures. The resolution limit, crystallographic R-factor and atomic temperature factors provide important clues in assessing the confidence a researcher can have in the coordinates of any particular segment in the protein structure. The structure of subtilisin determined independently in a number of laboratories from crystals grown in different conditions provides a means to obtain an empirical estimate of error. In the case of subtilisin BPN', the structure of which has been determined at resolutions ranging from 1.8-1.6Å resolution with R-factors ranging from 0.18-0.14, there is very good agreement between structures determined from different crystal forms. This observation suggests that the individual models are fair representations of the structure of the enzyme in solution.

The three-dimensional structures of macromolecules, predominantly proteins, determined using X-ray crystallography now figure prominently in all biochemical textbooks. The number of crystallographic structures available in the Brookhaven Protein Data Bank is increasing exponentially, driven by commercial as well as academic interests to determine protein structures to serve as the basis for rational drug design and protein engineering. This growth is also a consequence of the increasing number of active crystallography laboratories and the continuous improvement in crystallographic techniques and hardware.

All biochemists are familiar with the quaternary changes that occur to effect the allosteric regulation of oxygen uptake by haemoglobin in the bloodstream and the proposed mechanism of action for hydrolysis of peptide bonds in serine proteases. Both are derived from numerous crystallographic studies of the protein structures. Knowledge of the three-dimensional structure of a biological

0097-6156/94/0576-0018\$08.00/0
© 1994 American Chemical Society

macromolecule is an essential component in dissecting the relationship between the structure of a particular molecule and its functionality in medical or chemical applications. A better understanding of the limitations placed on the structure by the process of X-ray crystallographic structure determination should be of value to all who use the results.

It is not possible to reduce all of X-ray crystallography into a short article. It is not my intention to duplicate many far more thorough developments of the theory and formula of X-ray crystallography (1-3). Instead I will attempt to present an overview of the process of structure determination namely crystallization, data collection, interpretation of electron density maps and refinement of model coordinates. I will also show how each of these factors provide clues regarding the overall confidence levels for the coordinates. The aim is to provide a general audience with a better insight into the evaluation of the reliability of coordinates based on the experimental results of the structure determination. I will rely heavily on the work done by a number of laboratories on subtilisin where data on how the coordinates might compare with the molecule in "solution" is available.

Crystallographic Structure Determination

The first step is to grow crystals of the protein or macromolecule of interest. The necessity for growing crystals is based on the radiation employed, X-rays, having a wavelength comparable to the interatomic distance of covalently bonded atom (0.15nm). X-rays are ionizing radiation, creating free radicals that will randomly break covalent bonds throughout the molecule, degrading the protein sample. The crystals serve as diffraction gratings, where x-rays scattered from all molecules in the crystal positively interfere. This gives diffraction patterns such as the one presented in Figure 1. The spots in this figure represent as much as a 10 billion-fold amplification of the x-rays scattered from a single molecule, depending of the number of repeats or "unit cells" present in the crystal. The diffraction will be limited by the degree of long-range periodicity that exists from molecule to molecule throughout the crystal. This limit is referred to as the "resolution".

The time required for growing crystals of sufficient size for data collection varies from, in the best cases, a matter of hours for the most pure and facile proteins to many months, although crystals grown for a year or more have in some cases been used. The conditions giving the crystals are empirically selected often with pH near the isoelectric point of the enzyme. The selection is opportunistic and may not coincide with the conditions under which the enzyme is optimally active. In some cases, the crystals can induce some interesting alterations such as the elevation of the pK of the catalytic histidine in α -lytic protease, giving rise to the suggestion that there was not a hydrogen bond between the catalytic serine and histidine in the active site of that enzyme (4). This result can now be reconciled with NMR studies where the finding is that the hydrogen bond does form in solution but under the conditions of the crystallization the histidine is in fact protonated at pH 7.9.

Once the crystals are grown, diffraction data can be collected. The diffraction data arises from the coherent interference of all molecules in the crystal.

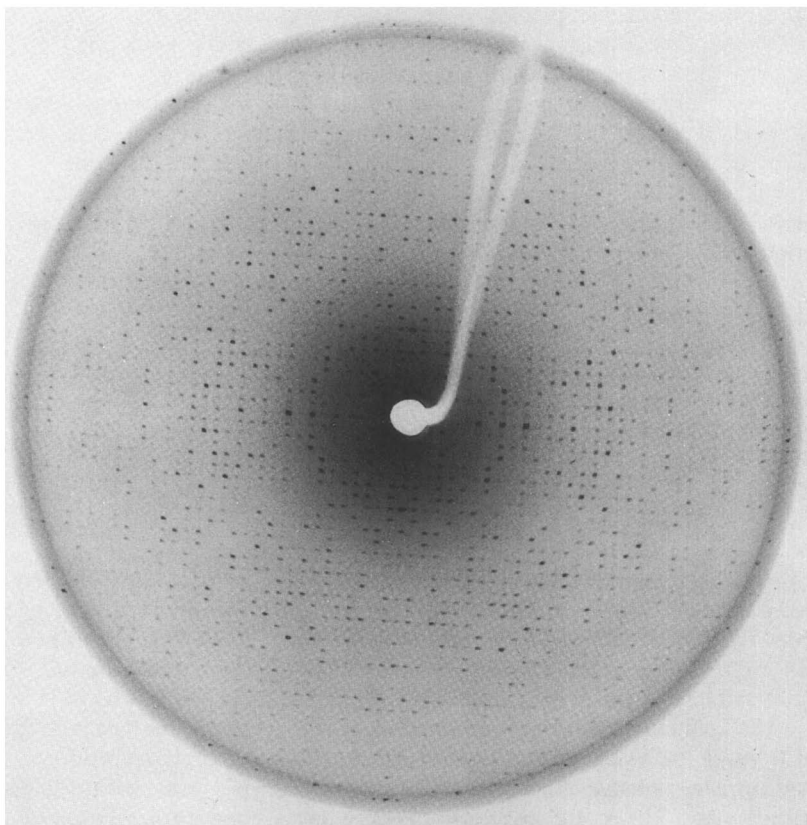


Figure 1. X-ray diffraction pattern showing the 0h0l projection of data from subtilisin BPN' grown at pH 6.0

The diffraction data is a radially distributed array of diffraction maxima whose positions are determined by the lattice repeat and any internal symmetry found in the crystal. This data can most easily be visualized as a series of spots that are uniformly spaced but have different intensities. The variation of the intensities of the spots (diffraction maxima) in Figure 1 arises from the interference of scattered X-rays from atoms within a single molecule and as a consequence contains information on the three-dimensional arrangement of the atoms within the protein molecule.

This information can then be used to visualize the scattering matter, electrons, by calculating an electron density map. The "data" used in this calculation includes the intensities measured directly from the crystal along with "phases" that go into a Fourier summation. The crystal is a periodic function of matter in three-dimensions. Fourier summations can define any periodic function as the summation of an large number of wave functions. The amplitude of each wave in the summation is proportional to a particular diffraction maxima, modulated by a phase displacement. A much more thorough discussion can be found in Ref 1. While this is NOT intuitive, unlike the correspondence of a particular atom to a peak in NMR, the net effect is an electron density map that has quite detailed information about the protein in the crystal. The "detail" is dependent on the resolution limit regulating the fineness of waves that are included in the Fourier summation.

With single counter diffractometers, "high" resolution usually meant 2.8 Å while, with the appearance of area detectors capable of collecting data much faster and with greater sensitivity, high resolution now means 2.0-1.6 Å or better resolution. Even at 2.0-1.6 Å resolution it is not possible to differentiate covalently bonded atoms, but at this resolution, side chain and main chain moieties can often be recognized as shown in Figure 2. This figure looks at the electron density of a tyrosine side chain. The molecular orbital that forms a doughnut shaped ring can be seen with the expected vacant central cavity for this side chain. Resolution limits provide useful information regarding the relative rigidity of a particular molecule. Structures determined from crystals that diffract to high resolution will have better overall reliability limits.

The diffraction data represents the ensemble diffraction from all molecules in the crystal. It follows that the resulting electron density map from this data will represent an "averaged" electron density of all molecules in the crystal. If some residues vary from molecule to molecule within the crystal then these residues or portions thereof may have an average electron density falling below the cutoff level for "noise". These segments will not be easily fitted and will remain ambiguous. In fact there is a continual gradation of relative rigidity within the molecule in general the atoms in the interior are more rigid than those on the surface.

The electron density map serves as a guide to fit an atomic model having the expected amino acid sequence to best match this electron density. Once a model has been fitted, it is possible to use the coordinates of the model to calculate the diffraction pattern arising from this model. This calculated diffraction pattern having calculated intensities for each diffraction maxima can be compared with the observed diffraction pattern from the crystal. The model is imperfect, a single

model representing an averaged structure and incomplete, lacking the disordered bulk solvent molecules and hydrogen atoms (due to the restricted resolution). With the exception of some of the most high resolution structures, corresponding to 1.0 Å or better, it is usually not possible to "resolve" hydrogen atoms. Despite these limitations there is good overall agreement between the relative magnitudes of the calculated diffraction intensity and the observed intensities that are observed from the crystal, suggesting that the model still fairly represents the molecule in the crystal.

It is possible to refine the model by adjusting the position of the atomic coordinates to minimize the difference between calculated and observed diffraction patterns. In the case of high resolution data the number of observations/variables can be 2-3:1 for reasonably complete data sets at 2.0-1.6 Å resolution. Added to these are the quasi observations or additional restraints placed on the molecules by the stereochemistry of bond lengths, angles, planarity and stereochemistry of amino acids and polypeptide linkages. The refinement procedure not only gives a better agreement between the data calculated from the model and the observed data, but when the refined model is used to calculate a new electron density map the resulting map usually indicates new information relating to errors in the model and additional features such as tightly bound solvent molecules.

There are algorithms to estimate the overall mean error (5,6) and these rely on the same agreement between the observed and calculated structure factors that are refined. The structure factors are proportional to the square root of the intensity of the diffraction maxima. The agreement between the observed and calculated diffraction intensities is measured by an R factor defined by equation 1.

$$R = \sum_h (|F_o(h)| - |F_c(h)|) / |F_o(h)| \quad (1)$$

In this equation, $F_o(h)$ and $F_c(h)$ are the absolute values of the observed and calculated structures factors for the spot having indices (h) corresponding to h,k,l , each index in turn representing the integral number of spacings along the axial repeats from the center of the diffraction pattern in Figure 1. The estimated mean error is directly related to the value of the R-factor for the higher resolution data. The lower the R-factor the lower the estimated mean error will be. If more than one potential crystallographic model is available, a researcher would then be usually better off choosing the one determined at high resolution and giving the lowest R-factor for high resolution data. For high resolution (1.8- 1.6Å) X-ray crystallographic structures, giving R-factors in the range of 0.18-0.14, the methods give estimates of the mean error on the order of 0.2 Å.

In the course of the refinement an atomic temperature factor is also refined for each atom in the model. This atomic temperature factor is a measure of the relative vibrational motion different atoms have in the molecule. Internal atoms in the rigid, center of the molecule will have lower temperature factors than atoms on the surface. It would be expected that in any coordinate set, the electron density for atoms having high temperature factors will be more diffuse and fitted with lower confidence than the well ordered atoms having low temperature factors.

Accuracy and Reliability of Crystallographic Models: An Example

It is important to consider how well this crystal structure matches the same molecule in solution. To address this question one would ideally need to have the crystal structure of the same molecule determined independently. Ideally the different structures would be determined from crystals grown under different conditions. By comparing the crystal structures of the same enzyme determined from different crystal forms, grown under different conditions we can infer how each might differ from the solution structure of the enzyme. The different crystal lattice interactions would be expected to distort the crystal structure in different ways. Thus the divergence we see between crystal structures determined in different crystal forms should diverge more from each other than the solution structure.

The results from subtilisin meet most of these conditions and offer a chance to answer this question. In part because of the high commercial interest in subtilisin engineering, the number of independent crystallographic models is larger with most now available in the protein data bank. The coordinate sets have been determined from crystallographic data collected from crystals that differ in space group and the conditions for crystal growth ranging in pH from 6.0 to 9.5 and in precipitant from ammonium sulfate to acetone. What is compared here is the native enzyme grown at pH 6.0 from ammonium sulfate and the three-dimensional structure of a variant enzyme having six site-specific substitutions; Met 50 replaced by Phe, Asn 76 replaced by Asp, Gly 169 replaced by Ala, Gln 206 replaced by Cys, Tyr 217 replaced by Lys and Asn 218 replaced by Ser (7). We have seen that site-specific substitutions produce limited and often, very subtle conformational changes that are localized at the site of substitution (8), and thus we expect that the perturbations resulting from these additional differences to be minimal.

The structure of native and variant subtilisin BPN' determined under the different extremes of pH 6.0 versus 9.0 and different precipitating agents, ammonium sulfate versus acetone, are still, very similar in their overall folding (Figure 3) and in conformation of the side chain atoms. An example of this is presented for atoms in the active site (Figure 4). The overall rms variation from Ca atoms is 0.38 Å. The algorithms to estimate the overall mean error give estimates of the mean error of 0.2 Å.

While knowing the mean error is quite useful it would be more useful to establish the variation about this mean with the aim of establishing criteria for what constitutes statistically significant differences between these structures. These criteria would also correspond to a confidence level for scientists using these models to adjust the thresholds for docking algorithms that take into account the overall deformation that models might be allowed to undergo before a significant disruption in the structure occurred. We have employed an empirical method for estimating the mean error for atoms having the same degree of thermal mobility as measured by the refined crystallographic temperature factor. This method relies on the finding that a linear relationship exists between the logarithm of the distance between equivalent atoms and the temperature factors of those atoms (9). The crystallographic temperature factors reflect the relative reliability of any particular

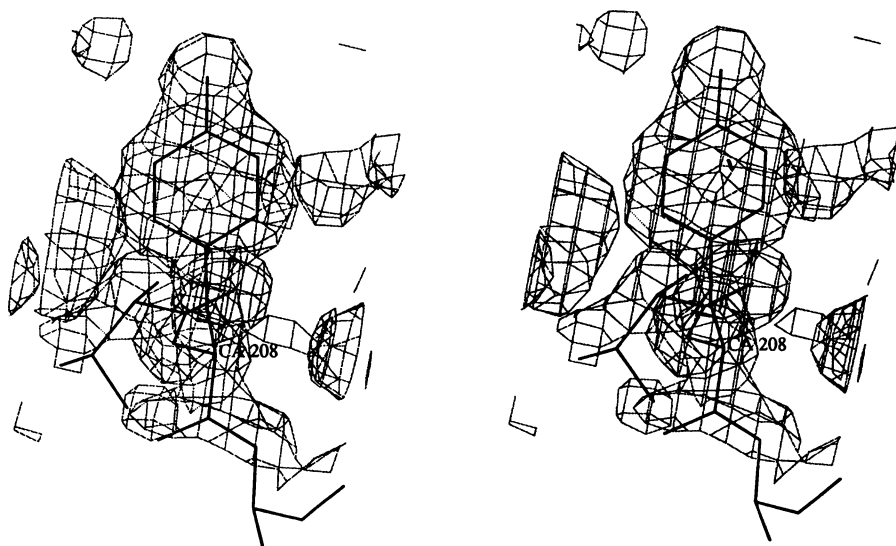


Figure 2. Stereographic representation of electron density map at 1.6 Å resolution. The side chain of Tyr 208 of subtilisin from *Bacillus lentus* is superposed on the electron density.

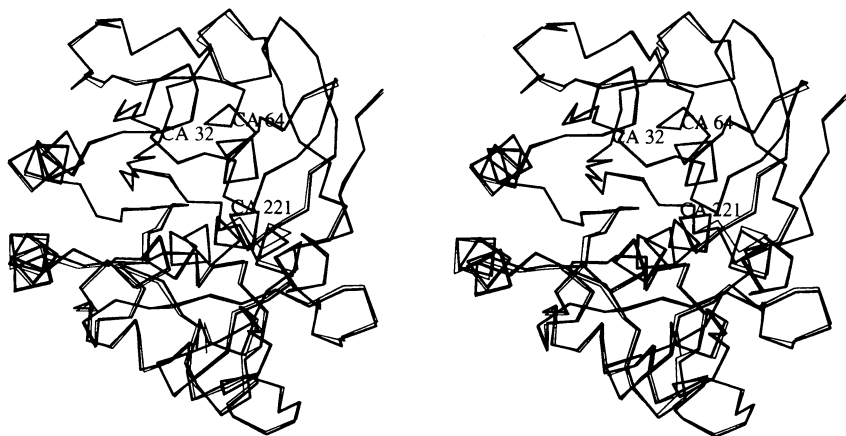


Figure 3. Stereographic view comparing C α trace of native subtilisin BPN' determined at pH 6.0, 40% sat. ammonium sulfate (thick lines) and subtilisin BPN' variant (M50F/ N76D/ G169A/ Q209C/ Y217K and N218S) at pH 9.0 55 % acetone (thin lines).

segment of a protein. An equation can then be determined for the line (Figure 5) representing the mean error and the root mean squared deviation from this line. Residues in the two structures compared having significant departures from the mean error can be identified.

This method determines, by linear least squares fit, the equations for mean error as a function of the crystallographic temperature factor B and the root mean squared deviation. This is analogous to the standard analysis of variance conducted for the mean error between equivalent atoms in two crystallographic coordinate sets. Using this method we have identified regions that represent potentially significant differences between different site-specific variants and the native enzyme (10).

In this paper the interest is not in the particular differences but rather with the threshold of difference between equivalent atoms that represent significant departures from random variation. This method provides an empirical estimate of the error that would be found in the structure in the same or different crystal forms. The error boundaries should highlight the confidence levels appropriate under these circumstances.

The values from the equations obtained from comparing crystal structures in the same and different forms are compared in Table 1. In this table, we compare the values of the mean error and the variation of the error about the mean in two pairwise comparisons. The first comparison is between the native enzyme and a variant having a single site-specific substitution, phenylalanine replacing methionine at position 222 (M222F). The second is between the native enzyme determined from a crystal grown at pH 6.0 from ammonium sulfate and a variant with six site-specific substitutions (M50F/ N76D/ G169A/ Q206C/ Y217K/ N218S) determined from a crystal grown at pH 9.0 from acetone. All structures in these comparisons have a mean value of 10. The table presents the estimates of the mean error, along with the standard deviation from the mean for atoms having temperatures factors of 5, 10, 15 and 20 representing atoms that are relatively more rigid than the average atoms, the mean atoms and atoms that are more or very much more, disordered than the "average" atoms in these pairwise comparisons.

The error between structures in the same crystal lattice might represent the error expected if the structures were re-determined using independent data sets that would be representative of experimental error. While the error between structures determined in different crystal lattices and grown under different conditions might reflect structures that independently diverge from the solution structure. The divergence between these structures would in fact overestimate the deviation that either of the structure might have with the structure in solution. The values for the threshold of a significant difference between equivalent atoms having the mean temperature factor of 10.0 range from 0.23 to 0.59 Å depending on whether the coordinate sets were determined in the same crystal form, under similar conditions of pH and precipitating agent or under significantly different crystallization conditions. In the second case, when comparing subtilisin structures determined at different extremes of pH and precipitant, coordinates for "average", well ordered atoms should vary by no more than 0.6 Å before differences above this threshold may be regarded as significant. This also defines a confidence level for these structures relative to a "solution" structure. It should be noted that for the most

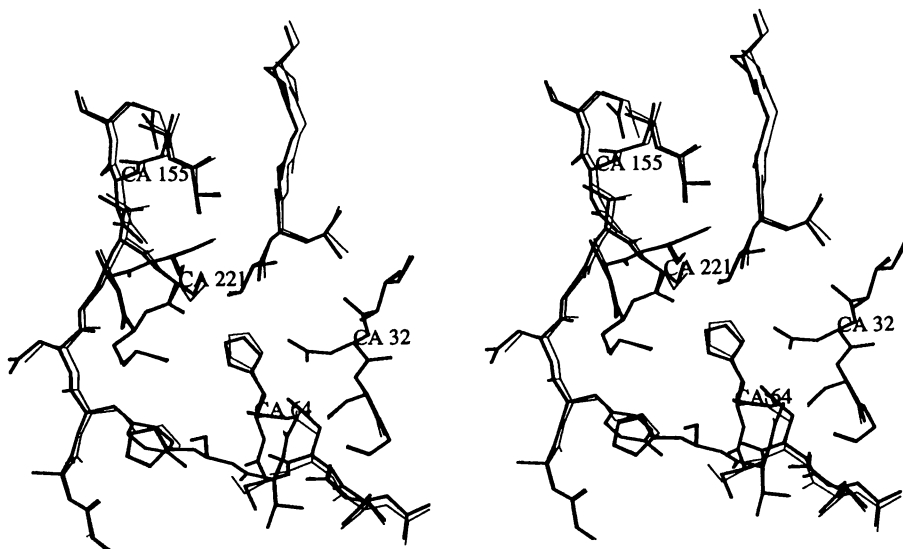


Figure 4. Stereographic comparison of the active site of native subtilisin BPN' determined at pH 6.0, 40% sat. ammonium sulfate (thick lines) and subtilisin BPN' variant (M50F/ N76D/ G169A/ Q209C/ Y217K and N218S) at pH 9.0 55 % acetone (thin lines).

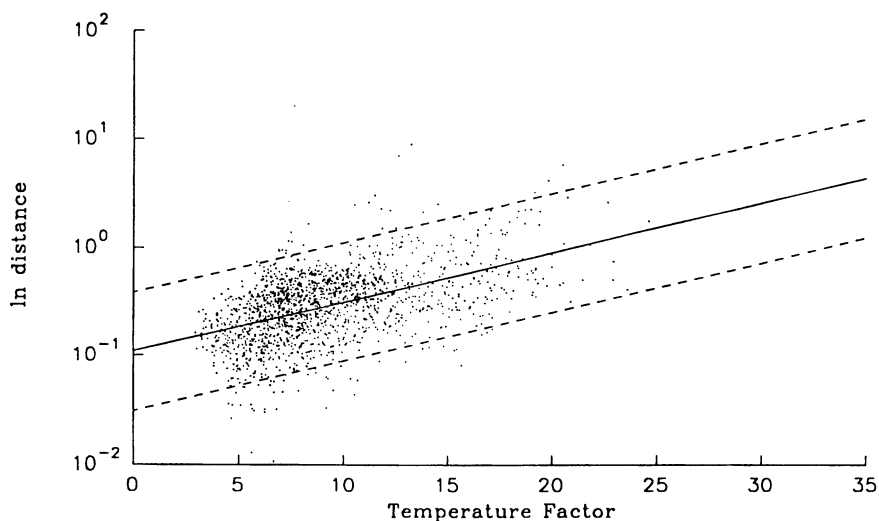


Figure 5. Semi-logarithmic plot of distances between equivalent atoms in native subtilisin BPN' from crystals grown at pH 6.0 from ammonium sulfate and a variant subtilisin BPN'(M50F/ N76D/ G169A/ Q209C/ Y217K and N218S) grown at pH 9.0 from acetone.

Table I. Values of the estimated error and variance for selected temperature factors (in Å)

	BPN' (pH 6) v BPN' M222S (pH 6)	BPN' (pH 6) v BPN' hextuple* (pH 9)
B = 5		
mean error	0.09	0.19
+ RMSE	0.17	0.35
B = 10		
mean error	0.13	0.31
+ RMSE	0.23	0.59
B = 15		
mean error	0.17	0.53
+ RMSE	0.32	1.00
B = 20		
mean error	0.24	0.90
+ RMSE	0.44	1.68

* hextuple variant (M50F/ N76D/ G169A/ Q209C/ Y217K and N218S)

flexible regions, having higher temperature factors, the threshold for the confidence level can be expected to be considerably higher. The possibility of this range of variation points to the importance of considering the crystallographic temperature factors in any modeling experiment.

Conclusions

The methods of crystallographic analysis are providing higher resolution structures at an accelerating rate as a consequence of improved methodologies and equipment for the acquisition of crystallographic data, faster computers which have stimulated development of refinement software and faster molecular graphics to analyze the data. The same technological advances have placed the tools to manipulate these structures in the hands of every biochemist.

For this reason it is equally necessary to disseminate information concerning the intrinsic sources of error and the methods for estimating this error. In addition to this overall error, it should be remembered that there are important clues regarding the reliability of specific segments of the structure in the crystallographic temperature factors.

Overall the estimates for the mean error of structure determined at high resolution (2.0 Å or better) range from 0.13 to 0.31 Å which are in relatively good agreement with estimates derived by other methods. R-factors with their associated estimates of the mean error can be useful guides for selecting model structures if more than one is available. However the mean error in the example of subtilisin would not be representative of the criterion for significance or for confidence bounds to be used for evaluating feasibility in modeling. The empirical estimate of the confidence levels outlined in this paper sets 0.6 Å the threshold for significant difference between relatively well ordered segments in closely related structures. This threshold varies as a function of the crystallographic temperature factors and can exceed 1.5 Å for segments with high crystallographic temperature factors. Current protein engineering experience suggests that the location and relative flexibility of a segment is of great importance in modeling and relating structure to function.

While this paper has focused on the deficiencies of the X-ray crystallographic models, it would be remiss to fail to note that overall these X-ray crystallographic structures from quite different crystal environments and conditions do share very high similarity. Any of the crystallographic models would constitute an excellent model for subtilisin under the differing conditions and thus would also serve as a reliable model for the molecule in solution.

Literature Cited

1. Eisenberg, D. In *The Enzymes* Boyer, P. D. Ed.; Academic Press: New York, NY, 1970 Vol. 1; pp 1-89.
2. Glusker, J. P. and Trueblood, K. N. *Crystal Structure Analysis*; Oxford University Press, London 1972.
3. Blundell, T. L. and Johnson, L. N. *Protein Crystallography*, Academic Press New York NY, 1976.
4. Smith, S. O., Farr-Jones, S., Griffin, R. G. and Bachovchin, W. W. *Science* **1989**, *244*, 961-964.
5. Cruikshank, D. W. J. *Acta Crystallogr.* **1949**, *2*, 65-82.
6. Luzzati, V. *Acta Crystallogr.* **1952**, *5*, 802-810.
7. Pantoliano, M. W., Whitlow, M., Wood, J. F., Dodd, S. W., Hardman, K. D., Rollence and Bryan, P. N. *Biochemistry* **1989**, *28*, 7205-7213.
8. Bott, R. and Ultsch, M. In *Fifth International Symposium on the Genetic of Industrial Microorganisms*; M Alacevic, D. Hranueli and Z. Toman Eds. Pliva Press Zagreb: 1987, pp 375-385.
9. Bott, R. and Frane, J. *Protein Engineering* **1990**, *3*, 649-657.
10. Bott, R., Dauberman, J., Caldwell, R., Mitchinson, C., Wilson, L., Schmidt, B., Simpson, C., Power, S., Lad, P., Sagar, H., Garycar, T. and Estell, D. In *Annals of the New York Academy of Sciences*, 1992, Vol 672, pp 10-19.

RECEIVED August 26, 1994

Chapter 3

Determination of Solution Conformation of Receptor-Bound Ligands by NMR Spectroscopy

A Transferred Nuclear-Overhauser-Effect Study of Cyclophilin and a Model Substrate

L. T. Kakalis¹ and I. M. Armitage^{1,2}

Departments of ¹Pharmacology and ²Diagnostic Radiology, Yale University School of Medicine, 333 Cedar Street, P.O. Box 208066, New Haven, CT 06520-8066

The receptor-bound conformation of weakly binding ligands may be determined from Transferred NOE (TRNOE) measurements. Cyclophilin (CyP) is the receptor of the immunosuppressant cyclosporin A (CsA) and a peptidyl prolyl isomerase (PPIase) that catalyzes the *cis-trans* isomerism of X-Pro peptide bonds via an unspecified mechanism. The conformation of substrates in the CyP binding site would provide insights into the enzymatic catalytic mechanism. Transferred NOE measurements indicate that the predominantly *trans* unbound model substrate suc-AAPF-pNA adopts a *cis* conformation when bound to CyP.

Interactions between ligands and macromolecules are fundamental processes for recognition, catalysis and regulation in biology. The structure of the ligand complexed macromolecules can sometimes be determined from either x-ray crystallography in the solid state (1, 2) or NMR spectroscopy in solution (3, 4). NMR approaches to the study of large macromolecular complexes require the use of methods for the simplification of the congested proton spectra in order to facilitate spectral analysis (4-6). In the case of strongly binding ligands, isotope-editing techniques permit the observation of only those protons that are scalar or dipolar coupled to the isotopically labeled nuclei of the ligand (7). Another approach involves difference spectroscopy of two receptor-ligand complexes prepared with either protonated or deuterated ligand (5, 6). In the case of recombinant proteins, a third approach relies on protein perdeuteration and the structure determination of the deuterated receptor-bound protonated ligand by conventional ¹H NMR (8). Central to any structure determination by NMR is the nuclear Overhauser effect (NOE) and its quantification. The Transferred NOE (TRNOE) is an extension of NOE measurements to

0097-6156/94/0576-0029\$08.00/0

© 1994 American Chemical Society

molecules undergoing facile chemical exchange between a free and a bound state and is ideally suited to studies of weakly binding ligands. The representative, rather than comprehensive, list of TRNOE applications in Table I is indicative of the usefulness of the method. It should be noted that in many of these studies, the large molecular size of the macromolecule-ligand complex precludes its straightforward structural investigation with state-of-the-art multidimensional NMR methods.

The TRNOE Principle.

The magnitude of the NOE, defined as the fractional change in the intensity of an NMR resonance upon the rf irradiation of another, is proportional to the cross-relaxation rate σ between the corresponding nuclei as represented by equation 1:

$$\sigma = \frac{\text{const}}{r^6} \left(\frac{6\tau_c}{1 + 4\omega^2\tau_c^2} - \tau_c \right) \quad (1)$$

The dependence of σ on the internuclear distance r is the basis of the use of NOE in structure determination whereas the dependence of σ on the correlation time τ_c is reflected in the sensitivity of NOE to molecular size and motion.

The TRNOE is an extension of NOE to systems in chemical exchange and, following its initial observation (31), has been systematically treated by Clore and Gronenborn (76, 77). It relies on the transfer of cross-relaxation information between two nuclei of the bound ligand to the free ligand resonances via chemical exchange. In the free state, the ligand is characterized by short correlation times ($\tau_c \sim 100$ ps) being either at the extreme narrowing limit ($\omega^2\tau_c^2 \ll 1$) where the NOEs are small and positive or at its boundary ($\omega^2\tau_c^2 \sim 1$) where the NOEs approach zero. When bound, the ligand's correlation time becomes that of the macromolecule/receptor ($\tau_c \sim 10$ ns or larger) and it is thus at the spin diffusion limit ($\omega^2\tau_c^2 \gg 1$) where the NOEs are large and negative. In the event of fast chemical exchange, the observed ligand cross-relaxation is the population-weighted average of cross-relaxation values in the free (F) and the bound (B) states

$$\sigma_{\text{obs}} = P_F \sigma_F + P_B \sigma_B \quad (2)$$

and negative bound-state NOEs are transferred to the free or exchange-averaged ligand resonances where they can be easily measured. These negative NOEs, by virtue of their large magnitude, dominate the observed NOEs thus identifying protons spatially close ($< 5 \text{ \AA}$) in the bound ligand conformation.

Table I. TRNOE Studies of Macromolecule-Ligand Complexes

<i>Macromolecule (kDa)</i>	<i>Ligand</i>	<i>Refs.</i>
CMP-KDO Synthetase (97.5)	Inhibitors	6
Hemoglobin (67)	ATP, GTP	10
Aspartate Transcarbamylase (300)	ATP, CTP, ITP	11
EPSP Synthetase (46.5)	Substrate/Product	12
Dihydrofolate Reductase (18)	NADP ⁺ , thio-NADP ⁺ , Inhibitors	13, 14
cAMP Receptor Protein (45)	cAMP, cGMP, Analogues	15
Dehydrogenases (80-336)	NAD ⁺ , NADP ⁺ , NADPH	16-21
Peroxidase (42)	Substrate	22
Kinases (80-230)	ADP, ATP	23-26
Methionyl-tRNA Synthetase (66)	ATP, ATP Analogue	26, 27
Isoleucyl-tRNA Synthetase	Ile, Val	28
Elastase (26)	Oligopeptide Substrates	29, 30
Neurophysin (22)	Peptide (oxytocin)	31-34
Thrombin (36)	Fibrinogen, Hirudin and Platelet Receptor Peptides	35-40
Ribonucleotide Reductase (171)	Peptide Inhibitor	41
Troponin C (18.5)	Troponin I Peptides	42, 43
Molecular Chaperones (60, 70)	Peptides	44, 45
Ricin B (34)	Disaccharides	46, 47
BSA (66)	Prostaglandins	21, 48
Antibody Fragment (50)	Opiates	49, 50
Antibody Fragments (25, 50)	Peptide Antigens	51-56
Antibody Fragments	Lipid Antigens	57
Monoclonal Antibody (150)	Carbohydrate Antigen	58
Catalytic Antibody (50)	Substrate	59
HIV Reverse Transcriptase (117)	AZT, T Triphosphates	60
DNA Binding Protein (75)	DNA Undecamer	61
tRNA ^{Asp} (20)	Codon, Wobble Codon	62
DNA Polymerase I Fragment (68)	Substrates, Templates	63, 64
Phospholipase A2 (30)	Substrate Analogues	65
Acetylcholine Receptor	Acetylcholine	66, 67
Rhodopsin	G Protein Peptide	68
G Protein (40)	Peptide (mastoparan)	69
Phospholipid Vesicles	Peptide (mastoparan)	70-72
Phospholipid Vesicles	Mating Factor Peptide	71, 73
Phospholipid Vesicles	Drug (chlorpromazine)	74
Phospholipid Bilayers	Enkephalin Analogues	75

The observation of TRNOEs is subject to the condition:

$$|P_{\text{B}\sigma_{\text{B}}}| \gg |P_{\text{F}\sigma_{\text{F}}}| \quad (3)$$

so that $\sigma_{\text{Obs}} \sim P_{\text{B}\sigma_{\text{B}}}$ (see eq. 2) and to the existence of fast exchange. For the following macromolecule (M) - ligand (L) interaction scheme



fast exchange requires

$$k_{\text{off}} \geq 10 R_{1\text{F}} \quad (5)$$

which normally translates into $k_{\text{off}} > 10 \text{ s}^{-1}$, $R_{1\text{F}}$ being the free ligand longitudinal relaxation rate. Assuming a diffusion-controlled $k_{\text{on}} \sim 10^8 \text{ M}^{-1} \text{ s}^{-1}$, one would expect substantial TRNOEs for ML dissociation constants between 0.1 μM to 1 mM. For weaker binding, the bound population fraction P_{B} is too low and the bound state contributes insignificantly to the overall relaxation (condition 3 breaks down). For stronger binding, the exchange is no longer fast (condition 5 breaks down) and information regarding the bound state conformation is not transferred to the free ligand resonances.

Experimental Aspects of TRNOE Measurements. The experiments for TRNOE measurements are no different from those routinely used for regular NOE measurements, i.e. the 1D steady state/truncated driven NOE experiment and the transient NOE experiment in its 1D or 2D (NOESY) version (9). The former is generally more sensitive but ill suited for quantitative distance determinations whereas the latter is less susceptible to spin diffusion (48), i.e. the spread of magnetization throughout the molecular complex with concomitant loss in structural information. A significant difference from regular NOEs is the slower growth of TRNOEs, whose time-development is proportional to $P_{\text{B}\sigma_{\text{B}}}$ rather than simply σ_{B} , thus necessitating longer irradiation, development and mixing times.

Specific experimental and methodological aspects of TRNOE measurements have been previously discussed (78, 79). Ratios of ligand to macromolecule typically range between 10 and 30, depending on the dissociation constant and also the size of the protein since the TRNOE magnitude increases with the protein size. A useful variable is the population ratio $P_{\text{F}}/P_{\text{B}}$ which is related to the association constant and total concentrations of M and L (eq. 4). Provided condition 3 is satisfied, one strives to use as high a $P_{\text{F}}/P_{\text{B}}$ ratio as possible for maximal sensitivity and

minimal spin diffusion. Temperature is another important experimental factor. For certain ligand-macromolecule systems, data acquisition at the highest temperature allowed by the sample's thermal stability may be required in order to satisfy the fast exchange condition (58). Alternatively, the dissociation rate k_{off} may be increased by a suitable modification of the ligand (53). A complete treatise of variables such as mixing time, bound correlation time/protein size and fraction of bound ligand P_B in NOESY measurements (80) indicates that TRNOE-derived distances are most accurate for low M to L ratios (with P_B being 5% or less) that minimize spin diffusion and reasonable mixing times that provide good signal-to-noise. Other points that affect the development of TRNOEs include the presence of internal mobility of the bound ligand (33, 81), finite bound-free exchange rates (32, 82) and protein indirect relaxation effects (87).

Rigorous analysis of TRNOE data requires a relaxation matrix treatment that takes into account the chemical exchange rates. Fast chemical exchange does not present any complications (79, 83, 84). However, in the event of intermediate chemical exchange, the independently determined exchange rates must be included in the calculations (79, 82, 84). These may be obtained from relaxation and saturation transfer measurements (22, 58, 85, 86).

Technical developments have include improvements in solvent suppression and baseline correction (88) to facilitate both the detection and quantitation of the TRNOE signals and 3D TRNOE measurements (89) to resolve spectral overlap. Other advances include recent reports of heteronuclear (90, 91) and proton X-filtered (92) TRNOE NMR investigations that may further expand the usefulness of the approach. Additional structural information concerning the receptor binding site may be obtained from concentrated protein samples (ca. 1 mM) where weaker protein-ligand intermolecular TRNOEs may be observed (22, 23, 34, 51-53, 64, 65, 67). Identification of intermolecular TRNOEs can be facilitated by $T_{1\rho}$ (56) or T_2 (58) filtering to remove protein NOE crosspeaks which are of comparable intensity and linewidth.

The Peptidyl Proline Isomerase Cyclophilin.

Cyclophilin (CyP) is the 17.8 kDa cytosolic receptor of the immunosuppressive drug cyclosporin A (CsA) and a peptidyl-prolyl *cis-trans* isomerase (PPIase) that catalyzes the *cis-trans* isomerization of X-Pro imide bonds, a catalysis strongly inhibited by CsA (93-95). While a possible relationship between immunosuppression and PPIase inhibition is unlikely (96, 97), X-Pro isomerization and its catalysis by CyP is, nevertheless, important *per se* for its role in protein folding (98).

Proposed mechanisms for the PPIase catalysis include nucleophilic attack on the carbonyl carbon of the X-Pro peptide bond (99) or protonation of the X-Pro peptide bond nitrogen (100), both leading to the formation of tetrahedral intermediates with decreased double bond

character for the X-Pro peptide bond and lowered activation energy barrier for the *cis-trans* interconversion. However, steady-state kinetic investigations of the CyP PPIase activity and specificity (101-103) and site-directed mutagenesis studies (104) argue against either nucleophilic or acid/base catalysis. According to a third "catalysis by distortion" mechanism (101, 104, 105), non-covalent enzyme-substrate interactions stabilize a substrate transition state with a non-planar X-Pro peptide bond, thereby destroying its resonance stabilization. Thus, the elucidation of the CyP-bound substrate conformation would be particularly revealing with regard to the PPIase mechanism.

The x-ray structure of CyP complexed with the model substrate N-acetyl-AlaAlaProAla-amidomethylcoumarin (ac-AAPA-amc) at 2.8 Å resolution showed the AP imide bond in the *trans* conformation with the substrate binding site being identical to that of CsA (106). Subsequent work at 2.3 Å resolution identified the CyP residues in the active site and revealed a structure consisting of a dimer of CyP-substrate complexes, each CyP molecule accommodating a *cis* ac-AAPA-amc molecule in its active site while also being associated with a partially disordered *trans* tetrapeptide (107). This dimer arrangement, however, could be the result of the observed stacking of six aromatic groups (four tetrapeptide coumarins and two CyP Trp indoles) and may not be biologically relevant. A third x-ray study at 1.64 Å resolution of CyP complexed with the dipeptide AP, in all likelihood a poor CyP substrate as it is for the PPIase FKBP (103), identified only the *cis* conformer as protein-bound and provided a detailed description of the protein active site (108). A solvent-assisted catalysis mechanism was suggested that involved the desolvation of the substrate upon entering the hydrophobic CyP active site and its stabilization by a protein-bound water molecule (108). In view of the conflicting reports regarding the CyP-bound substrate conformation, an experimental approach free of artifacts from crystal packing forces was highly desirable and this led to the selection of TRNOE methods for the determination of the CyP-bound substrate conformation in solution.

The Selection of a Suitable Substrate. Several Pro-containing oligopeptides are PPIase substrates as evidenced by the appearance of exchange crosspeaks in their NOESY spectra recorded in the presence of PPIase (109). However, not all of them appear to be equally effective, as indicated from competition assays against the standard PPIase substrate (103, 110). In addition to a dependence on the type of amino acid preceding Pro (102), catalysis appears to be affected by the presence of the C-terminal charge and its unfavorable accommodation in the hydrophobic PPIase active site (110).

CyP substrates can be quantitatively evaluated by NMR, using saturation transfer methods to measure rate constants (111). The measured CyP-catalyzed *cis-trans* interconversion rate of AAPA was found to be substantially slower than that of the standard CyP substrate N-succinyl-

AlaAlaProPhe-paranitroanilide (suc-AAPF-pNA) which may explain the absence of AAPA TRNOEs (Fig. 1). The best estimates of the steady-state kinetic constants for the CyP catalysis of the suc-AAPF-pNA *cis-trans* isomerism are $K_M \sim 1$ mM and $k_{cat} \sim 1.3 \times 10^4$ s⁻¹ at 0°C (100). The latter value indicates an efficient enzyme and sets a lower limit to all unimolecular steps in the catalytic mechanism, including the dissociation rate of the product from the Michaelis-Menten complex (112), thus satisfying the fast exchange condition for TRNOEs.

Materials and Methods. Recombinant human CyP was overexpressed in *E. coli*, harvested, and purified as previously detailed (104). Suc-AAPF-pNA was selected as a suitable substrate for the TRNOE studies of the CyP mechanism. Its ¹H NMR spectrum was assigned by standard 2D NMR methods (3). Attention was focused on the A2αH to P3δH or P3αH resonance NOEs which can be used to distinguish between the *trans* and *cis* AP conformers (Fig. 2). In the presence of CyP at 25°C, the averaged A2αH resonance was broadened beyond detection, thus precluding the observation of any NOEs. Experiments were, therefore, conducted at 5°C where the fast exchange condition should still be satisfied. As a result of limited substrate solubility, the more sensitive steady state/truncated driven NOE experiment was used for an assessment of the CyP-bound substrate conformation.

NOE difference spectra were acquired with the standard pulse sequence. The residual water resonance was presaturated with a 14 Hz field strength for 2 s and preirradiation at the selected frequency was applied for 400 ms with a field strength of 6 Hz, the strongest one that maintained selectivity of irradiation. Spectra were acquired by subtracting a 16-scan off-resonance control from a 16-scan on-resonance data set and adding the FID differences until a good signal-to-noise ratio in the Fourier transformed difference spectra was achieved after 60 cycles. Each difference spectrum was phased using the parameters of an identically acquired control spectrum phased all positive and referenced to the HDO resonance at 5.02 ppm vs TSP at 5°C. These difference spectra are displayed in Figs. 3-5 and, being identically acquired and processed, can be directly intercompared.

The Conformation of the Cyclophilin Substrate. An important distinction with regard to the conformation of a Pro-containing peptide is whether the X-Pro peptide bond is *cis* or *trans* (3). In the *trans* form, the XαH and PδH₂ protons are in close proximity whereas in the *cis* form, this is so for the XαH and PαH protons, a difference resulting in distinct, characteristic NOEs for the two conformers (Fig. 2). In aqueous solution, the conformation of the uncomplexed standard CyP model substrate suc-AAPF-pNA is ca. 90% *trans* (100). The *cis-trans* isomerization is sufficiently slow on the NMR timescale to give rise to two sets of

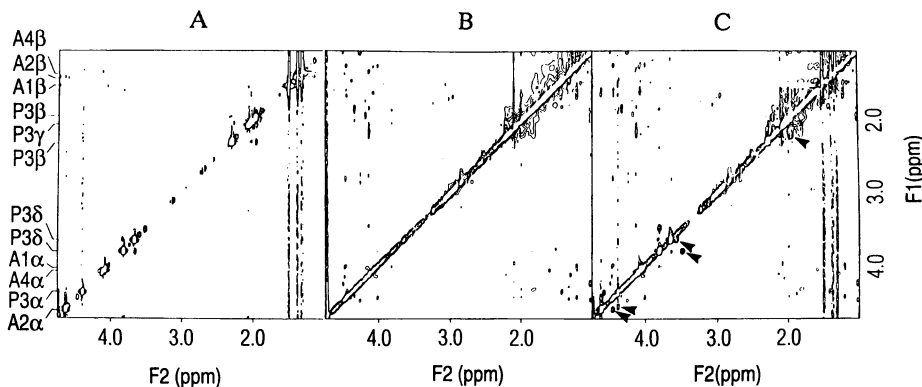


Figure 1. The aliphatic region of the pure phase absorption 500 MHz NOESY spectra (200 ms mixing time) of 10 mM AAPA (A), 1 mM CyP (B), and 10 mM AAPA plus 1 mM CyP (C) in buffered D₂O solutions, pH 6.8 at 25°C. The assignments of the dominant *trans* conformer are provided in A. Exchange crosspeaks due to the CyP-catalyzed, *cis-trans* interconversion are marked with an arrow in C. However, the substrate exchange between the bound and the free states is not sufficiently fast to produce TRNOEs.

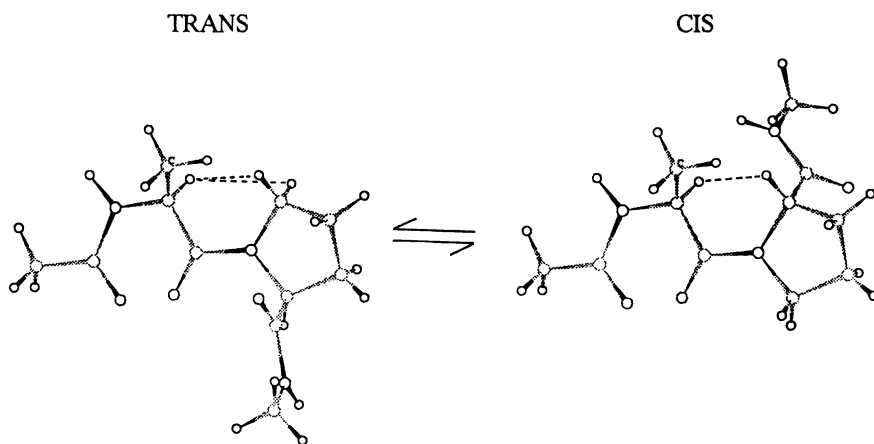


Figure 2. The *trans* and *cis* forms of the A-P peptide bond in CH₃CO-Ala-Pro-NHCH₃. In each form, the broken lines indicate the short distances A α H-P δ H₂ (*trans*) and A α H-P α H (*cis*) which give rise to the diagnostic NOEs.

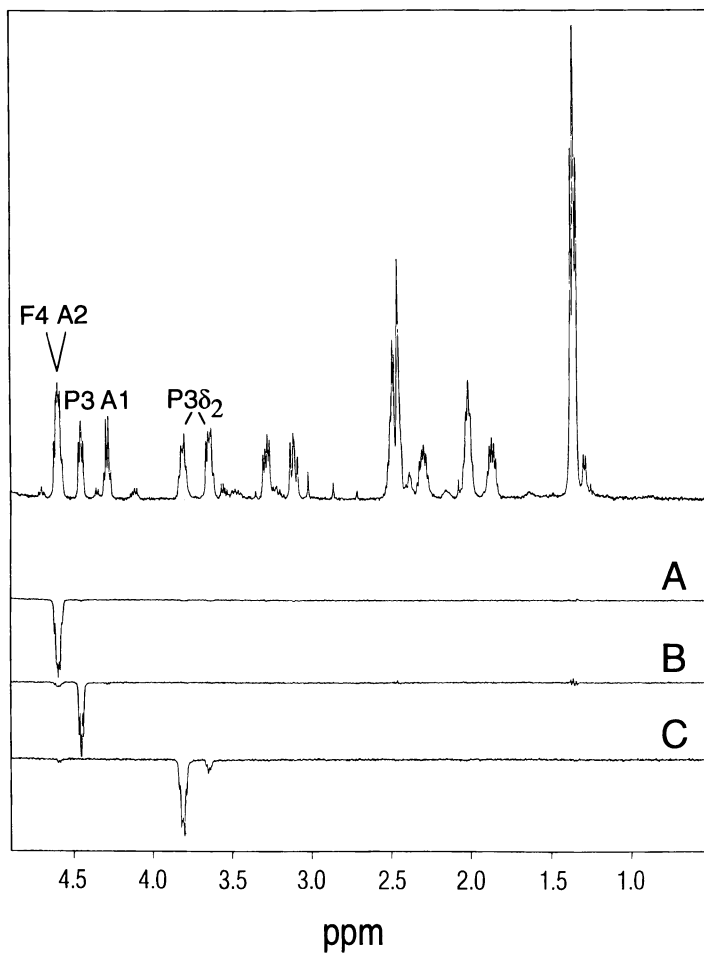


Figure 3. NOE difference spectra (500 MHz) for 0.86 mM suc-AAPF-pNA in buffered D₂O solutions at pH 6.8 and 5°C. On resonance irradiation frequencies: *trans* A2 α H (A), *trans* P3 α H (B) and *trans* P3 δ H (C).

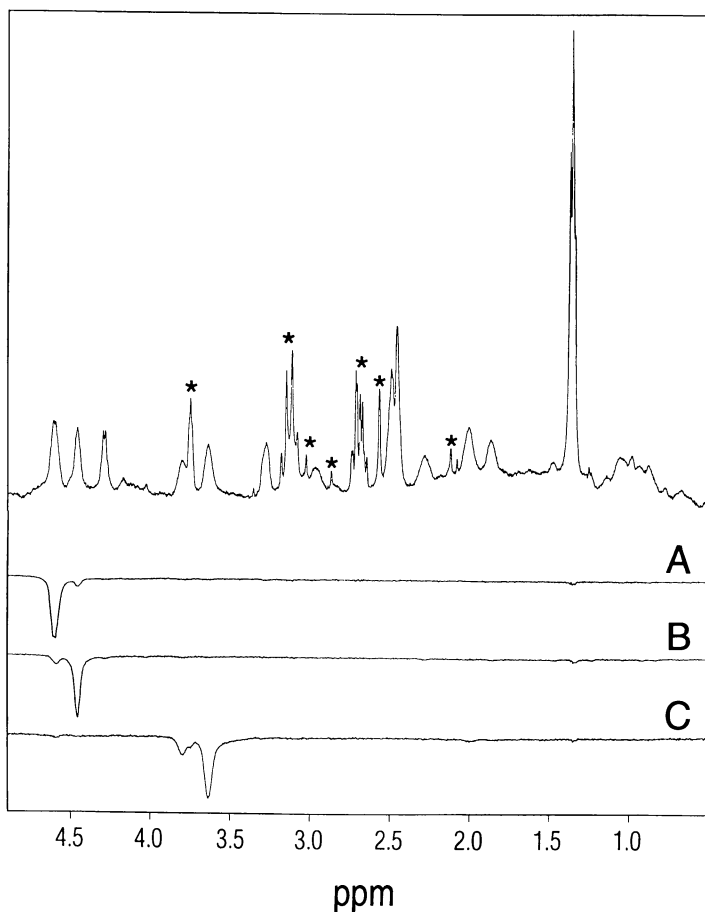


Figure 4. NOE difference spectra (500 MHz) for 0.86 mM suc-AAPF-pNA plus 40 μ M CyP in buffered D_2O solutions at pH 6.8 and 5°C. Asterisk-marked peaks (top) are from buffer components (DTT and EDTA). The on-resonance irradiation frequencies are those for the exchange-averaged A2 α H (A), P3 α H (B), and P3 δ H (C) peaks.

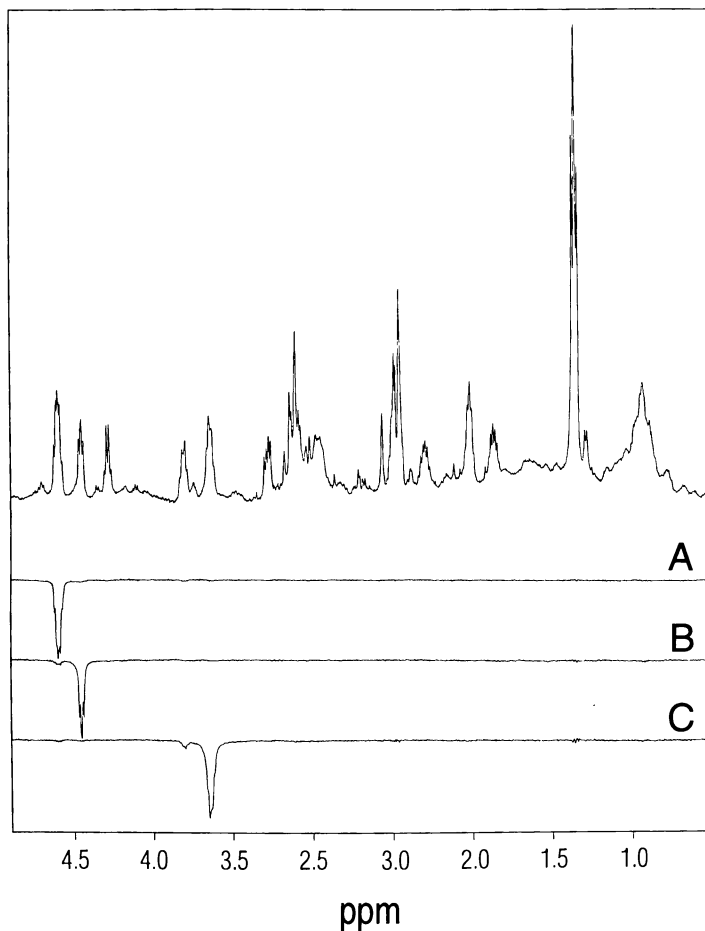


Figure 5. NOE difference spectra (500 MHz) for suc-AAPF-pNA (0.86 mM) plus CyP (40 μ M) plus CsA (80 μ M) in buffered D₂O solutions at pH 6.8 and 5°C. The on-resonance irradiation frequencies are those of the *trans* A2 α H (A), *trans* P3 α H (B), and *trans* P3 δ H (C) peaks.

resonances that correspond to a major (*trans*) and a minor (*cis*) population in a 10:1 ratio (Fig. 3, top).

In the 1D NOE difference spectra of free suc-AAPF-pNA, the observed NOEs are small with the exception of that between the geminal Pro δ H₂ (Fig. 3C) and, surprisingly for a molecule of this size, negative. This may result from slower molecular tumbling at 5°C, particularly in the ca. 40% more viscous D₂O solution (Chem. Abstr. 1941, 6169) and perhaps also from molecular association. In the presence of CyP, there is a marked broadening of all peptide resonances (Fig. 4 top) and the *cis* and *trans* resonances are averaged as a result of the enhanced CyP-catalyzed *cis-trans* interconversion. No NOE between A2 α H and P3 δ H₂ is observed (Figs. 4A and 4C) but there is a significant NOE between A2 α H and P3 α H (Figs. 4A and 4B) that must originate from the CyP-bound *cis* conformer. Upon addition of the PPIase inhibitor CsA, the NOE pattern and intensity (Fig. 5) reverts to that for the unbound substrate (Fig. 3) indicating that the A2 α H-P3 α H TRNOE (Fig. 4) originates from a specific interaction of the substrate with the CyP active site and not from nonspecific substrate binding.

The aim of this steady-state NOE study was to determine whether the CyP-bound substrate adopts a *cis* or a *trans* conformation. The subsequent quantitative evaluation of additional transient NOE measurements showed that the CyP-bound substrate adopts a *cis*-like conformation with the A-P peptide bond twisted no more than 40° out of planarity (113).

Acknowledgments. This research was supported by the NIH grants GM40660 and GM49858.

Literature Cited

1. Blundell, T.L.; Johnson, L.M. *Protein Crystallography*; Academic Press: New York, 1976.
2. Stezowski, J.J.; Chandrasekhar, K. *Annu. Rep. Med. Chem.* **1986**, *21*, 293-302.
3. Wüthrich, K. *NMR of Proteins and Nucleic Acids*; John Wiley & Sons, Inc.: New York, 1986.
4. Otting, G. *Cur. Opin. Struct. Biol.* **1993**, *3*, 760-768.
5. Fesik, S.W.; Zuideweg, E.R.P.; Olejniczak, E.T.; Gampe, Jr., R.T. *Biochem. Pharmacol.* **1990**, *40*, 161-167.
6. Fesik, S.W. *J. Med. Chem.* **1991**, *34*, 2937-2945.
7. Otting, G.; Wüthrich, K. *Quart. Rev. Biophys.* **1990**, *23*, 39-96.
8. Hsu, V.L.; Armitage, I.M. *Biochemistry* **1992**, *31*, 12778-12784.
9. Neuhaus, D.; Williamson, M.P. *The Nuclear Overhauser Effect in Structural and Conformational Analysis*; VCH Publishers: New York, 1989.
10. Gronenborn, A.M.; Clore, G.M.; Brunori, M.; Giardina, B.; Falcioni, G.; Perutz, M.F. *J. Mol. Biol.* **1984**, *178*, 731-742.

11. Banerjee, A.; Levy, H.R.; Levy, G.C.; Chan, W.W.C. *Biochemistry* **1985**, *24*, 1593-1598.
12. Leo, G.C.; Castellino, S.; Sammons, R.D.; Sikorski, J.A. *Biorg. Med. Chem. Lett.* **1992**, *2*, 151-154.
13. Albrand, J.P.; Birdsall, B.; Feeney, J.; Roberts, G.C.K.; Burgen, A.S.V. *Int. J. Biol. Macromol.* **1979**, *1*, 37-41.
14. Feeney, J.; Birdsall, B.; Roberts, G.C.K.; Burgen, A.S.V. *Biochemistry* **1983**, *22*, 628-633.
15. Gronenborn, A.M.; Clore, G.M. *Biochemistry* **1982**, *21*, 4040-4048.
16. Gronenborn, A.M.; Clore, G.M. *J. Mol. Biol.* **1982**, *157*, 155-160.
17. Gronenborn, A.M.; Clore, G.M.; Jeffery, J. *J. Mol. Biol.* **1984**, *172*, 559-572.
18. Gronenborn, A.M.; Clore, G.M.; Hobbs, L.; Jeffery, J.; *Eur. J. Biochem.* **1984**, *145*, 365-371.
19. Ehrlich, R.S.; Colman, R.F. *Biochemistry* **1985**, *24*, 5378-5387.
20. Banerjee, A.; Levy, H.R.; Levy, G.C.; LiMuti, C.; Goldstein, B.M.; Bell, J.E. *Biochemistry* **1987**, *26*, 8443-8450.
21. Andersen, N.H.; Eaton, H.L.; Nguyen, K.T. *Magn. Reson. Chem.* **1987**, *25*, 1025-1034.
22. La Mar, G.; Hernández, G.; de Ropp, J.S. *Biochemistry* **1992**, *31*, 9158-9168.
23. Fry, D.C.; Kuby, S.A.; Mildvan, A.S. *Biochemistry* **1985**, *24*, 4680-4694.
24. Rosevear, P.R.; Fox, T.L.; Mildvan, A.S. *Biochemistry* **1987**, *26*, 3487-3493.
25. Rosevear, P.R.; Powers, V.M.; Dowhan, D.; Mildvan, A.S.; Kenyon, G.L. *Biochemistry* **1987**, *26*, 5338-5344.
26. Landy, S.B.; Ray, B.D.; Plateau, P.; Lipkowitz, K.B.; Rao, B.D.N. *Eur. J. Biochem.* **1992**, *205*, 59-69.
27. Williams, J.S.; Rosevear, P.R. *J. Biol. Chem.* **1991**, *266*, 2089-2098.
28. Kohda, D.; Kawai, G.; Yokoyama, S.; Kawakami, M.; Mizushima, S.; Miyazawa, T. *Biochemistry* **1987**, *26*, 6531-6538.
29. Clore, G.M.; Gronenborn, A.M.; Carlson, G.; Meyer, E.F. *J. Mol. Biol.* **1986**, *190*, 259-267.
30. Meyer, E.F., Jr.; Clore, G.M.; Gronenborn, A.M.; Hansen, H.A.S. *Biochemistry* **1988**, *27*, 725-730.
31. Balaram, P.; Bothner-By, A.A.; Breslow, E. *J. Amer. Chem. Soc.* **1972**, *94*, 4017-4018.
32. Lippens, G.M.; Cerf, C.; Hallenga, K. *J. Magn. Reson.* **1992**, *99*, 268-281.
33. Nirmala, N.R.; Lippens, G.M.; Hallenga, K. *J. Magn. Reson.* **1992**, *100*, 25-42.
34. Lippens, G.; Hallenga, K.; Van Belle, D.; Wodak, S.J.; Nirmala, N.R.; Hill, P.; Russell, K.C.; Smith, D.D.; Hruby, V.J. *Biochemistry* **1993**, *32*, 9423-9434.

35. Ni, F.; Konishi, Y.; Frazier, R.B.; Scheraga, H.A.; Lord, S.T. *Biochemistry* **1989**, *28*, 3082-3094.
36. Ni, F.; Meinwald, Y.C.; Vásquez, M.; Scheraga, H.A. *Biochemistry* **1989**, *28*, 3094-3105.
37. Ni, F.; Konishi, Y.; Bullock, L.D.; Rivetna, M.N.; Scheraga, H.A. *Biochemistry* **1989**, *28*, 3106-3119.
38. Ni, F.; Konishi, Y.; Scheraga, H.A. *Biochemistry* **1990**, *29*, 4479-4489.
39. Zheng, Z.; Ashton, R.W.; Ni, F.; Scheraga, H.A. *Biochemistry* **1992**, *31*, 4426-4431.
40. Ni, F.; Ripoll, D.R.; Martin, P.D.; Edwards, B.F.P. *Biochemistry* **1992**, *31*, 11551-11557.
41. Bushweller, J.H.; Bartlett, P.A. *Biochemistry* **1991**, *30*, 8144-8151.
42. Campbell, A.P.; Sykes, B.D. *J. Mol. Biol.* **1991**, *222*, 405-421.
43. Campbell, A.P.; Van Eyk, J.E.; Hodges, R.S.; Sykes, B.D. *Biochim. Biophys. Acta* **1992**, *1160*, 35-54.
44. Landry, S.J.; Gierasch, L.M. *Biochemistry* **1991**, *30*, 7359-7362.
45. Landry, S.J.; Jordan, R.; McMacken, R.; Gierasch, L.M. *Nature* **1992**, *355*, 455-457.
46. Bevilacqua, V.L.; Thomson, D.S.; Prestegard, J.H. *Biochemistry* **1990**, *29*, 5529-5537.
47. Bevilacqua, V.L.; Kim, Y.; Prestegard, J.H. *Biochemistry* **1992**, *31*, 9339-9349.
48. Andersen, N.H.; Nguyen, K.T.; Eaton, H.L. *J. Magn. Reson.* **1985**, *63*, 365-375.
49. Glasel, J.A.; Borer, P.N. *Biochem. Biophys. Res. Commun.* **1986**, *30*, 1267-1273.
50. Glasel, J.A. *J. Mol. Biol.* **1989**, *209*, 747-761.
51. Anglister, J.; Levy, R.; Scherf, T. *Biochemistry* **1989**, *28*, 3360-3365.
52. Levy, R.; Assulin, O.; Scherf, T.; Levitt, M.; Anglister, J. *Biochemistry* **1989**, *28*, 7168-7175.
53. Anglister, J.; Zilber, B. *Biochemistry* **1990**, *29*, 921-928.
54. Zilber, B.; Scherf, T.; Levitt, M.; Anglister, J. *Biochemistry* **1990**, *29*, 10032-10041.
55. Scherf, T.; Hiller, R.; Naider, F.; Levitt, M.; Anglister, J. *Biochemistry* **1992**, *31*, 6884-6897.
56. Scherf, T.; Anglister, J. *Biophys. J.* **1993**, *64*, 754-761.
57. Bruderer, U.; Peyton, D.H.; Barbar, E.; Fellman, J.H.; Rittenberg, M.B. *Biochemistry* **1992**, *31*, 584-589.
58. Glaudemans, C.P.J.; Lerner, L.; Daves, Jr., G.D. Kovác, P.; Venable, R.; Bax, A. *Biochemistry* **1990**, *29*, 10906-10911.
59. Campbell, A.P.; Tarasow, T.M.; Masefski, W.; Wright, P.E.; Hilvert, D. *Proc. Natl. Acad. Sci. USA* **1993**, *90*, 8663-8667.
60. Painter, G.R.; Aulabaugh, A.E.; Andrews, C.W. *Biochem. Biophys. Res. Commun.* **1993**, *191*, 1166-1171.

61. Clore, G.M.; Gronenborn, A.M.; Greipel, J.; Maass, G. *J. Mol. Biol.* **1986**, *187*, 119-124.
62. Gronenborn, A.M.; Clore, G.M.; McLaughlin, L.W.; Graeser, E.; Lorber, B.; Giegié, R. *Eur. J. Biochem.* **1984**, *145*, 359-364.
63. Ferrin, L.J.; Mildvan, A.S. *Biochemistry* **1985**, *24*, 6904-6913.
64. Ferrin, L.J.; Mildvan, A.S. *Biochemistry* **1986**, *25*, 5131-5145.
65. Plensniak, L.A.; Boegeman, S.C.; Segelke, B.W.; Dennis, E.A. *Biochemistry* **1993**, *32*, 5009-5016.
66. Behling, R.W.; Yamane, T.; Navon, G.; Jelinski, L.W. *Proc. Natl. Acad. Sci. USA* **1988**, *85*, 6721-6725.
67. Fraenkel, Y.; Gershoni, J.M.; Navon, G. *FEBS Let.* **1991**, *291*, 225-228.
68. Dratz, E.A.; Furstenau, J.E.; Lambert, C.G.; Thireault, D.L.; Rarick, H.; Schepers, T.; Pakhlevanians, S.; Hamm, H.E. *Nature* **1993**, *363*, 276-281.
69. Sukumar, M.; Higashijima, T. *J. Biol. Chem.* **1992**, *267*, 21421-21424.
70. Wakamatsu, K.; Higashijima, T.; Fujino, M.; Nakajima, T.; Miyazawa, T. *FEBS Let.* **1983**, *162*, 123-126.
71. Wakamatsu, K.; Okada, A.; Higashijima, T.; Miyazawa, T. *Biopolymers* **1986**, *25*, S193-S200.
72. Wakamatsu, K.; Okada, A.; Miyazawa, T.; Ohya, M.; Higashijima, T. *Biochemistry* **1992**, *31*, 5654-5660.
73. Wakamatsu, K.; Okada, A.; Miyazawa, T.; Masui, Y.; Sakakibara, S.; Higashijima, T. *Eur. J. Biochem.* **1987**, *163*, 331-338.
74. Kuroda, Y.; Kitamura, K. *J. Amer. Chem. Soc.* **1984**, *106*, 1-6.
75. Milon, A.; Miyazawa, T.; Higashijima, T. *Biochemistry* **1990**, *29*, 65-75.
76. Clore, G.M.; Gronenborn, A.M. *J. Magn. Reson.* **1982**, *48*, 402-417.
77. Clore, G.M.; Gronenborn, A.M. *J. Magn. Reson.* **1983**, *53*, 423-442.
78. Rosevear, P.R.; Mildvan, A.S. *Methods Enzymol.* **1989**, *177*, 333-358.
79. Campbell, A.P.; Sykes, B.D. *Annu. Rev. Biophys. Biomol. Struct.* **1993**, *22*, 99-122.
80. Campbell, A.P.; Sykes, B.D. *J. Magn. Reson.* **1991**, *93*, 77-92.
81. Campbell, A.P.; Sykes, B.D. *J. Biomol. NMR* **1991**, *1*, 391-402.
82. London, R.E.; Perlman, R.E.; Davis, D.G. *J. Magn. Reson.* **1992**, *97*, 79-98.
83. Landy, S.B.; Rao, B.D.N. *J. Magn. Reson.* **1989**, *81*, 371-377.
84. Ni, F. *J. Magn. Reson.* **1992**, *96*, 651-656.
85. Akasaka, K. *J. Magn. Reson.* **1979**, *36*, 135-140.
86. Dubois, B.W.; Evers, A.S. *Biochemistry* **1992**, *31*, 7069-7076.
87. Zheng, J.; Post, C.B. *J. Magn. Reson. B* **1993**, *101*, 262-270.
88. Ni, F. *J. Magn. Reson.* **1992**, *99*, 391-397.
89. Ni, F. *J. Magn. Reson.* **1992**, *100*, 391-400.

90. Batta, G.; Kövér, K.E.; Székely, Z.; Sztaricskai, F. *J. Amer. Chem. Soc.* **1992**, *114*, 2757-2758.
91. Wang, S.X.; Caines, G.H.; Schleich, T. *J. Magn. Reson. B* **1993**, *102*, 47-53.
92. Oschkinat, H.; Schott, K.; Bacher, A. *J. Biomol. NMR* **1992**, *2*, 19-32.
93. Handschumacher, R.E.; Harding, M.W.; Rice, J.; Drugge, R.J. *Science* **1984**, *226*, 544-547.
94. Takahashi, N.; Hayano, T.; Suzuki, M. *Nature* **1989**, *337*, 473-475.
95. Fischer, G.; Wittmann-Liebold, B.; Lang, K.; Kiefhaber, T.; Schmid, F.X. *Nature* **1989**, *337*, 476-478.
96. Walsh, C.T.; Zydowsky, L.D.; McKeon, F.D. *J. Biol. Chem.* **1992**, *267*, 13115-13118.
97. Zydowsky, L.D.; Etzkorn, F.A.; Chang, H.Y.; Ferguson, S.B.; Stolz, L.A.; Ho, S.I.; Walsh, C.T. *Protein Sci.* **1992**, *1*, 1092-1099.
98. Schmid, F.X.; Mayr, L.M.; Mücke, M.; Schönbrunner, E.R. *Adv. Protein Chem.* **1993**, *44*, 25-66.
99. Fischer, G.; Berger, E.; Bang, H. *FEBS Let.* **1989**, *250*, 267-270.
100. Kofron, J.L.; Kuzmic, P.; Kishore, V.; Colón-Bonilla, E.; Rich, D.H. *Biochemistry* **1991**, *30*, 6127-6134.
101. Harrison, R.K.; Stein, R.L. *Biochemistry* **1990**, *29*, 1684-1689.
102. Harrison, R.K.; Stein, R.L. *Biochemistry* **1990**, *29*, 3813-3816.
103. Harrison, R.K.; Stein, R.L. *J. Amer. Chem. Soc.* **1992**, *114*, 3464-3471.
104. Liu, J.; Albers, M.W.; Chen, C.-M.; Schreiber, S.L.; Walsh, C.T. *Proc. Natl. Acad. Sci. U.S.A.* **1990**, *87*, 2304-2308.
105. Stein, R.L. *Adv. Protein Chem.* **1993**, *44*, 1-24.
106. Kallen, J.; Spitzfaden, C.; Zurini, M.G.M.; Wider, G.; Widmer, H.; Wüthrich, K.; Walkinshaw, M.D. *Nature* **1991**, *353*, 276-279.
107. Kallen, J.; Walkinshaw, M.D. *FEBS Let.* **1992**, *300*, 286-290.
108. Ke, M.; Mayrose, D.; Cao, W. *Proc. Natl. Acad. Sci. U.S.A.* **1993**, *90*, 3324-3328.
109. Justice, Jr., R.M.; Kline, A.D.; Sluka, J.P.; Roeder, W.D.; Rodgers, G.H.; Roehm, N.; Mynderse, J.S. *Biochem. Biophys. Res. Commun.* **1990**, *171*, 445-450.
110. Park, S.T.; Aldape, R.A.; Futer, O.; DeCenzo, M.T.; Livingston, D.J. *J. Biol. Chem.* **1992**, *267*, 3316-3324.
111. Hsu, V.L.; Handschumacher, R.E.; Armitage, I.M. *J. Amer. Chem. Soc.* **1990**, *112*, 6745-6747.
112. Hammes, G.G. *Enzyme Catalysis and Regulation*; Academic Press: New York, **1982**; pp. 38-42.
113. Kakalis, L.T.; Armitage, I.M. *Biochemistry* **1993**, in press.

RECEIVED January 13, 1994

Chapter 4

Modeling Biologically Relevant Peptides Using Circular Dichroism with Synchrotron Radiation and High-Temperature Molecular Dynamics

L. L. France¹, P. G. Piatti¹, I. Toth², J. F. E. Newman¹, and F. Brown¹

¹Plum Island Animal Disease Center, Agricultural Research Service,
U.S. Department of Agriculture, P.O. Box 848,
Greenport, NY 11944-0848

²Department of Pharmaceutical Chemistry, The School of Pharmacy,
29-39 Brunswick Square, London, United Kingdom WC1N 1AX

Five antigenic variants of foot-and-mouth disease virus, serotype A12, differ only at positions 148 and 153 in the viral capsid protein VP1. The structural properties of immunogenic synthetic peptides corresponding to residues 141-160 of VP1 have been investigated using UV circular dichroism (CD) and high temperature molecular dynamics. Results indicate that the structures of these peptides are relatively insensitive to substitutions at residue 153, except for Pro, but highly sensitive to the difference between Phe and Leu at residue 148. Molecular models suggest that Pro-153 induces an inverse γ -turn at residues 152-154, and that Leu-148 induces a type II' β -turn at residues 148-151. These results correlate well with serological data, and suggest that α -helix formation plays a dominant role in antigen-antibody interactions for this virus.

The icosahedral capsid of foot-and-mouth disease virus (FMDV) contains 60 copies of each of four structural proteins (VP1-4). Peptides corresponding to residues 141-160 of the capsid protein VP1 are highly immunogenic. This region, situated within the immunodominant loop of the virus (*I*), contains both the cell attachment site and the major antigen of the virus (*2-8*). Five antigenic variants of FMDV, serotype A12, obtained from one field isolate of the virus, differ only at residues 148 and 153 within the immunodominant loop (*5,9,10*). Synthetic peptides corresponding to residues 141-160 of VP1 in each of these viruses (Table I) show distinct serological differences (discussed in the chapter by F. Brown *et al.* of this volume) and suggest that the FL, FS, and FQ peptides comprise a structurally similar group, whereas the FP and LP peptides appear to be distinct from this group and from each other. However, the latter two appear to be more similar to each other than to the former group.

We have investigated the structural basis of the serological results cited above using UV circular dichroism (CD) and molecular modeling. CD spectra of

0097-6156/94/0576-0045\$08.00/0

© 1994 American Chemical Society

Table I. Substituted amino acids at positions 148 and 153 of VP1 in FMDV, serotype A12. The sequence of the immunodominant loop (residues 133-160) is: Gly-Thr-Asn-Lys-Tyr-Ser-Ala-Ser-Gly-Ser-Gly-Val-Arg-Gly-Asp-148^{*}-Gly-Ser-Leu-Ala-153^{*}-Arg-Val-Ala-Arg-Gln-Leu-Pro. The immunogenic synthetic peptides include residues 141-160 (underlined).

Residue	FP	LP	FL	FS	FQ
148	Phe	Leu	Phe	Phe	Phe
153	Pro	Pro	Leu	Ser	Gln

the five synthetic peptides (residues 141-160, Table I) indicate that in aqueous solution the peptides are structurally almost identical, but in 100% trifluoroethanol (TFE) they display a range of α -helix-forming properties which correlate well with the serological data. These properties appear to be insensitive to substitutions at position 153, except for Pro, but sensitive to the difference between Phe and Leu at position 148. Furthermore, the CD data are consistent with the existence of a conserved β -turn at the cell attachment site.

High temperature molecular dynamics were used to search conformational space for accessible low energy configurations of each peptide. These molecular models indicate that Pro-153 induces an inverse γ -turn at residues 152-154, and that one or more β -turns form at the cell attachment site in each of the peptides. The configurational similarity of the models in the region 148-153 correlates reasonably well with the serological data cited above.

Materials and Methods

Peptides were synthesized as described in (11). Peptide samples for CD measurements were prepared in 10 mM sodium phosphate buffer (pH 7.0) or in 100% trifluoroethanol (TFE). CD spectra were measured with the UV CD spectrometer at Port U9B of the National Synchrotron Light Source (NSLS) at Brookhaven National Laboratory (12). CD and absorption spectra were measured simultaneously (13) from 330 nm to 178 nm at increments of 0.5 nm. Peptide concentrations were determined using the amide extinction coefficient at 205 nm, after correction for phenylalanine content (14). The optical path length was adjusted so that the total absorption of sample plus buffer was ≤ 1.0 . Optical path lengths were varied from 4 μm to 1 mm, using a "Gray" cell (15) to obtain path lengths $< 100 \mu\text{m}$. The spectrometer and optical path lengths were calibrated as described in (16). CD spectra are reported in units of $\Delta\epsilon$ per amide.

CD spectra were analyzed for secondary structure content using the program Varselec (17,18), supplied by Dr. W. C. Johnson, Jr. (Oregon State University, Corvallis, OR), which analyzes a CD spectrum for five classes of secondary structure: α -helix (α), antiparallel (β_A) and parallel (β_P) β -sheet, β -turns (T), and "Other" (O). The T category includes all types of β -turns. The O category includes all structures not covered by the first four categories, such as aperiodic structures and aromatic contributions. For each experimental CD spectrum, the

values given by all successful fits were averaged to yield the final set of values, and the standard deviation for the mean structural content in each category was $\leq 2\%$ of the total structural content in each case. The root mean square error for each successful fit was typically $\leq 0.200 \text{ M}^{-1}\text{cm}^{-1}$ per amide. The minimum structural content was constrained to be $\geq -5.0\%$, and the total structural content was allowed to float between 95% and 105%.

Primary structure analysis was conducted using programs written in Basic (Version 7.0). The indices published by Kyte and Doolittle (19), Hopp and Woods (20), and Karplus and Schultz (21) were used to calculate hydropathic, antigenicity and flexibility profiles, respectively. Predictions of α -helix, β -strand and β -turn sites were made using the method of Chou and Fasman (22). The technique described by Eisenberg et al. (23) was used to calculate helical hydrophobic moments. Molecular models were constructed on a Silicon Graphics Iris Indigo using software programs from Biosym Technologies of San Diego (graphics displays were obtained using Insight II and molecular dynamics were conducted using Discover).

Results and Discussion

CD Spectra. In aqueous solution the five peptides display CD spectra typical of "random coil" peptides (Figure 1). The spectra are dominated by a large negative band at 198 nm, and show a small negative band at $\approx 220 \text{ nm}$. Although the amplitudes of the minima ($\Delta\epsilon_{198 \text{ nm}}$) varied from -3.2 to -4.3 $\text{M}^{-1}\text{cm}^{-1}\text{amide}^{-1}$, analysis of the CD spectra for secondary structure content showed no significant differences among the peptides (Table II). Moreover, in aqueous conditions, the peptide structures were independent of concentration over the range $\approx 0.05 - 10 \text{ mg/ml}$.

Table II. Secondary structure content of peptides in aqueous solution predicted by analysis of CD spectra

Peptide	Concentration mg/ml	Percent (%) secondary structure content					
		α	β_A	β_P	Turns	Other	Total
FP	6.87	7	28	2	23	38	98
LP	7.25	4	32	0	23	40	99
FL	5.58	8	24	2	25	39	98
FS	6.58	4	30	2	21	41	98
FQ	5.80	7	27	0	24	41	99

The biologically active conformation of a short, immunogenic peptide (i.e., when it is bound to its B-cell receptor or to an antibody) is likely to be quite different from its (aqueous) solution structure, due to the exclusion of water

molecules from the antigen-antibody interface (24). Although there are notable exceptions, short peptides (less than ≈ 20 residues) are not expected to form periodic secondary structures in aqueous solution, since water is a strong hydrogen bonding agent and competes with the peptide backbone atoms for the hydrogen bonds which define secondary structural elements (25-29). Trifluoroethanol (TFE) is a weak hydrogen bonding agent with a moderately low dielectric constant, 26 at 25°C, which is intermediate between that of water, 79 at 25°C, and that expected for the interior of a protein, ≈ 1 to 4 (30-32). In TFE, intramolecular electrostatic interactions, particularly hydrogen bonding, are enhanced, and periodic secondary structural elements, particularly α -helices, can be induced.

If a peptide is capable of forming an amphiphilic α -helix, the fractional α -helix content is expected to be concentration-dependent, since at high peptide concentrations the α -helices will be stabilized by intermolecular interactions (29). A helical wheel projection (Figure 2) shows that the entire region including residues 146-159 is able to form a highly amphiphilic α -helix (33). The mean hydrophobicity and mean hydrophobic moment for this putative α -helix in each of the five peptides range from -0.02 to -0.09 and from 0.63 to 0.67, respectively. These values are typical of highly amphiphilic α -helices found on protein surfaces (23). Residue 148 is located in the middle of the hydrophobic arc, where its side-chain can interact only with those of hydrophobic residues. A hydrophobic residue at this position is therefore expected to stabilize the helix, whereas a hydrophilic residue here should destabilize it. In contrast, residue 153 is located near the interface between the hydrophobic and hydrophilic arcs. The Gly residues at positions 146 and 149 provide no steric interference for the side-chain of residue 153, which therefore may interact with the hydrophilic Arg-157 and Ser-150, or the hydrophobic Ala-156 side-chains. The hydrophobic character of residue 153 is thus not expected to be critical to the stability of the helix.

CD spectra were measured for each of the peptides in 100% TFE at 5 to 8 peptide concentrations, ranging from ≈ 0.1 mg/ml to > 10 mg/ml. These spectra indicate significant amounts of α -helix, and display maxima at 190 nm, minima at 208 nm, and smaller negative bands at 220 nm (Figures 1 and 3). Moreover, for each peptide, the secondary structure content in TFE is strongly concentration-dependent (Table III). An isodichroic dichroic point at 198 nm was observed in these spectra for each peptide (Figure 3). An isodichroic point at 198 nm indicates a two-state, helix-to-sheet sheet transition (34). The values listed in Table III show that for every peptide, the decrease in α -helix upon going from high to low peptide concentration is equal to the increase in β -sheet ($\beta_A + \beta_P$) to within $\pm 4\%$ of the total structural content.

The percentage of α -helix content (α) at each peptide concentration (c) was plotted, and the equation,

$$\frac{\alpha - \alpha_{MAX}}{\alpha_{MAX} - \alpha_{MIN}} = 1 - \exp\{-Hc\} , \quad (1)$$

was fit to each data set using three floating parameters (α_{MAX} , α_{MIN} , H) (Figure 4). In equation 1, α_{MIN} is the minimum amount of α -helix predicted at infinitely dilute peptide concentrations, α_{MAX} is the maximum amount of α -helix observed at highest

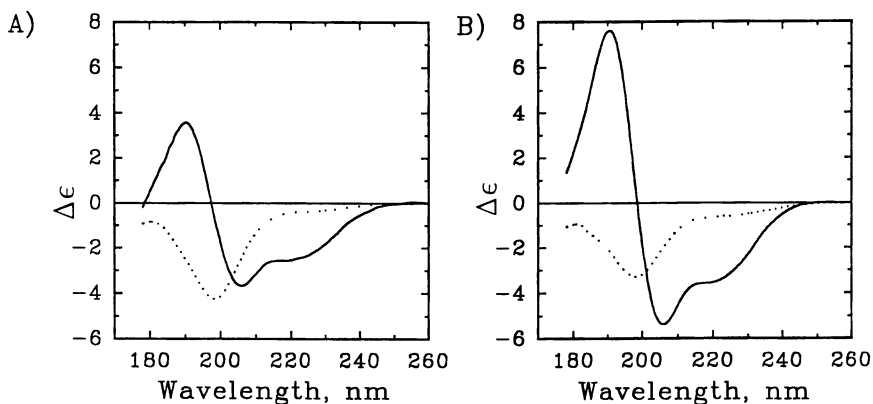


Figure 1. CD spectra of FP (A) and FL (B) in aqueous solution (···) and in 100% TFE (—). Peptide concentrations: aqueous FP, 6.8 mg/ml; aqueous FL, 5.6 mg/ml; FP in TFE, 13.7 mg/ml; FL in TFE, 14.4 mg/ml.

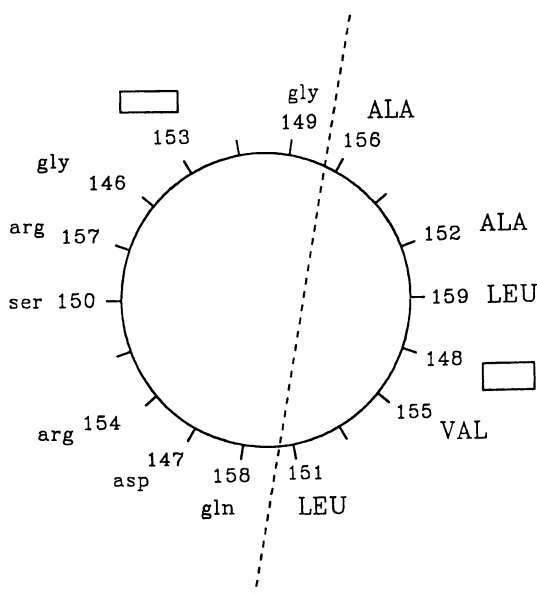


Figure 2. Helical wheel projection of residues 146-159. Uppercase, hydrophobic residues; lowercase, hydrophilic or neutral (Gly) residues. □, substituted sites. Dashed line divides the helix into a hydrophobic arc (right) and hydrophilic arc (left).

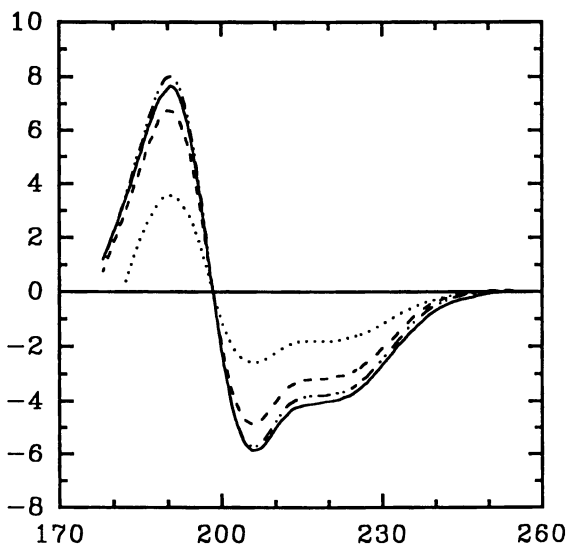


Figure 3. Four CD spectra of FS in 100% TFE at different peptide concentrations: 13.6 mg/ml (—), 5.64 mg/ml (—•—), 0.596 mg/ml (---), 0.106 mg/ml (···). An isodichroic point occurs at 198 nm.

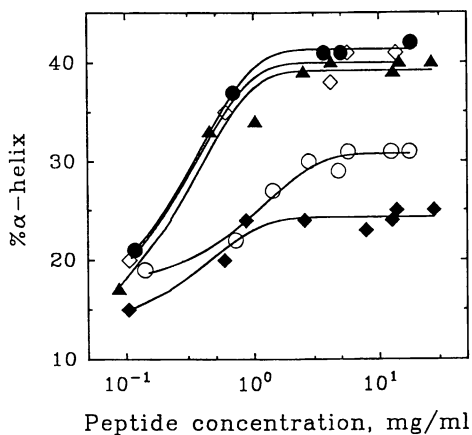


Figure 4. Fraction of α -helix content at different peptide concentrations. FQ (●), FS (◇), FL (▲), LP (○), FP (◆).

concentrations, and H is the reciprocal of the concentration at which 63% of the change from α_{MIN} to α_{MAX} is achieved. The parameter α_{MIN} is a measure of the importance of intramolecular interactions in α -helix formation. The parameter H is a measure of the importance of intermolecular interactions in α -helix formation. The parameter α_{MAX} depends on both intra- and intermolecular interactions.

Table III: Secondary structure content predicted from CD spectra of peptides in 100% TFE at selected low and high (bold) peptide concentrations

Peptide	Concentration, mg/ml	Percent (%) secondary structure content					
		α	β_A	β_P	Turns	Other	Total
FP	0.105	15	23	7	23	33	101
	13.7	25	18	2	26	30	101
LP	0.140	19	21	6	21	34	101
	12.4	31	18	0	28	25	102
FL	0.0881	17	30	5	16	32	100
	14.4	40	13	1	28	17	99
FS	0.106	20	26	7	17	28	99
	13.6	41	13	0	32	15	101
FQ	0.117	21	23	9	20	27	100
	17.8	42	13	2	33	9	99

Values for the α_{MIN} , H and α_{MAX} (Table IV) derived from the data in Figure 4 show that: (1) Pro-153 limits the maximum α -helix content for these peptides to $\leq 31\%$; (2) except for Pro, residue 153 appears to have no significant effect on the helix-forming properties of the peptides; (3) these parameters appear to be quite sensitive to the difference between Leu and Phe at residue 148.

Pro has a cyclic side-chain which restrains its dihedral angle ϕ to values near -75° (35) and introduces steric constraints on the residue immediately preceding it (36). Furthermore, since Pro has no imide hydrogen, it cannot act as a hydrogen bond donor. Consequently, although Pro frequently occurs at the N-cap position of an α -helix, it is rarely found at any other position within α -helices (22,37). However, Pro frequently occurs at the first external position on the C-terminal side of an α -helix (36), where it acts as a helix "terminator". In the peptides discussed here, it appears that the latter is the case, since the sensitivity of the helix-forming properties to residue 148 suggests that Leu-148 and Phe-148 are involved in the α -helices at high peptide concentrations.

Although the relative hydrophobicity of Leu with respect to Phe varies with the particular scale (i.e., with the technique used to measure hydrophobicity), the

Table IV: Values for the parameters α_{\max} , α_{\min} , and H (equation 1) for five peptides, derived from data in Figure 4

Peptide	α_{\max}	st. dev.	α_{\min}	st. dev.	H	st. dev.
FP	24.4	0.5	12.3	1.8	2.28	0.65
LP	30.8	0.5	17.2	1.2	0.80	0.16
FL	39.2	0.8	11.5	3.3	2.75	0.67
FS	40.0	1.0	13.0	3.5	2.83	0.79
FQ	41.3	0.3	13.5	1.2	2.68	0.26

helix-forming or helix-stabilizing ability of Leu is generally believed to be stronger than that of Phe. First, the Chou-Fasman helix-forming parameter P_{α} is greater for Leu than for Phe (22). Second, the thermodynamic scale of O'Neill and Degradó gives Leu a higher index for helix-formation than Phe (38). Third, Baldwin found that Leu enhanced α -helix formation in short peptides to a greater extent than Phe, and suggested that the bulky aromatic ring of Phe introduces restraints on the side-chain rotomers available to Phe (25). Our data also indicate that Leu enhances α -helix formation to a greater extent than Phe. A comparison of the values for LP and FP shows that the presence of Leu-148 increases both α_{\max} and α_{\min} by 5-6% (\approx one amide) and decreases the value of H by 65%, relative to Phe-148. The higher value of α_{\min} for LP relative to FP suggests that the intramolecular interactions between Leu-148, Leu-151 and Ala-152 (see Figure 2) are stronger than the analogous interactions involving Phe-148. Moreover, the lower value of H for LP relative to FP indicates that intermolecular interactions are less effective in stabilizing the α -helix in LP than in FP, a further indication that Leu-148 is more buried within the folded peptide.

It is interesting that under all conditions investigated here (i.e., at all peptide concentrations in aqueous solution and in TFE), each peptide CD spectrum showed a β -turn content $\geq 15\%$ (three amides). The latter value is equivalent to one β -turn, and suggests the possibility that a conserved β -turn may exist.

Primary structure analysis. Primary structure analysis predicts that the more hydrophilic N-terminal region of the peptides (residues 141-150) is relatively exposed and flexible, whereas the more hydrophobic C-terminal region (residues 151-160) is likely to be less exposed and more structured (19-21). Sites predicted for α -helices, β -strands and β -turns, using the method of Chou and Fasman (22), are indicated for each peptide in Figure 5. Although in no case does a β -strand prediction exceed that for an α -helix, each predicted α -helix does contain a segment with a relatively strong β -strand prediction, consistent with the idea that a helix-to-sheet transition could occur, as suggested by the isodichroic point observed at 198 nm in the CD spectra of each peptide in TFE. Three of the four predicted β -turns involve at least one residue from the cell attachment site (residues 145-147). In each of the peptide sequences, the latter site is invariably included in the region predicted to be the most probable site for a linear epitope (20,21).

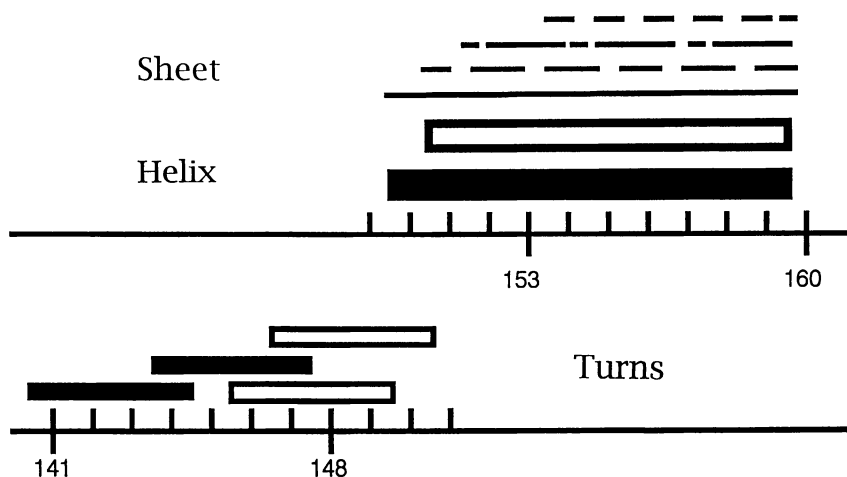


Figure 5. Secondary structure predictions using Chou-Fasman technique. Upper graph, α -helices: solid rectangle, FL, FS, FQ (residues 150-159); open rectangle, LP, FP (residues 150-159). Strong β -sheet propensity within the predicted helical regions: solid line, FL (residues 150-159); medium dashes FQ (residues 151-159); long and short dashes, FS (residues 142-159); short dashes LP, FP (residues 154-159). Lower graph, β -turns: solid rectangles, all five peptides (141-144, 144-147); open rectangles, FP, FL, FS, FQ (146-149, 147-150).

Molecular modeling. Molecular dynamics and energy minimization were conducted to search conformational space for accessible low energy conformations of each peptide. Polypeptides corresponding to the entire immunodominant loop region of each variant (residues 133-159) were constructed in order to take advantage of the coordinates solved for the residues at the base of the loop in the x-ray diffraction structure of the serotype O1 virus (*I*). In the latter structure residues 132-134 and 157-159 protrude from the capsid surface, forming two extended anti-parallel "arms", separated by $\approx 8 \text{ \AA}$. The coordinates for the residues between 134 and 157 could not be solved, presumably due to the mobility and flexibility of the loop (*I*). A sequence alignment of the serotype O1 and A12 viruses shows that residues 132-159 in the former correspond to residues 133-159 in the latter, if the position of the highly conserved Arg-Gly-Asp sequence at the cell attachment site is retained as residues 145-147 in both serotypes, since there is a deletion in the A12 sequence.

Eight initial conformations were constructed for each variant, using standard secondary structural elements (e.g., right-handed α -helix, β -strand, etc.). These eight initial structures were identical for each variant, except at residues 148 and 153. Since the biologically active conformation of the loop (i.e., when it is bound to its B-cell receptor or to an antibody) is the structure of interest, water molecules were omitted, and the dielectric constant was set to one. All side chains and terminals were defined to be in their neutral states, retaining only the partial charges on the atoms, as given by the force field. The latter practice prevents unscreened charges from playing an exaggerated role in the search (*39,40*). Each structure was energy-minimized using distance restraints (with force constants set to $1 \text{ kcal M}^{-1} \text{ \AA}^{-1}$) to bring the three alpha carbons at each end (in residues 133-135 and 157-159) to the relative distances specified by the O1 structure. Molecular dynamics were then conducted at 900 K, saving a structure every 2.5 ps until 8 structures were generated. Molecular dynamics were continued on each of the latter structures for 5.0 ps while the temperature was reduced to 300 K. This was followed by energy minimization. Distance restraints remained active during the entire procedure. This process generated 64 different configurations of each variant loop, each of which represents a local minimum on the potential energy surfaces of the folded polypeptides. The relative energies of each of the 64 conformations for a particular variant were compared in order to select the lowest energy structure. In cases where there was more than one conformation with a relative energy within 2 kcal M^{-1} ($\approx 3 \text{ kT}$) of the lowest energy, the structure which contained the most frequently occurring conformation in the region of antigenic interest was selected.

The five representative structures are shown in Color Plate 1. The first four N-terminal residues (133-136) have been truncated to show clearly that each structure appears to consist of two domains. In each model, the N-terminal domain appears to be both larger and more well-folded, whereas the C-terminal domain appears to be both smaller and more extended. The degree of folding in a protein or protein domain can be measured by its compactness. The Connolly algorithm (available in Insight II) was used to measure the solvent accessible surface areas of interest (*41*). This method involves rolling a probe with a radius of 1.4 \AA over the van der Waals surface and calculating the surface area of the volume from

NOTE: The color plates can be found in a color section in the center of this volume.

which the probe is excluded. Molecular volumes were calculated, using the volume included in the van der Waals envelope. The compactness of a region was then calculated as the ratio of the solvent accessible surface area to the surface area of a sphere of equal volume. This is equivalent to the roughness index (R) described by Richards (42), who found that globular proteins generally yield a value of $R \approx 2.0$. Although Richards noted no size-dependency for R , Zehfus and Rose have shown that a similar ratio, the coefficient of compactness, does increase with increasing size of the segment, up to ≈ 40 residues (43). We have therefore limited comparisons of R values to segments with equal numbers of residues.

Values of R for these structures range from 2.00 to 2.18 (Table V), indicating that each model is fairly well-folded and nearly as compact as globular proteins in general. To locate compact subunits (i.e., well-folded domains) within the structures, R was calculated for each 7-residue segment in each structure, and those segments with minimum R values were defined as the nucleus of a compact subunit. These segments were then lengthened on both sides until the average R value for the segment increased by $\geq 5\%$. The resulting regions were defined as domains (Table V). While all five structures contain domains in the N-terminal region, only two (FL and FS) showed regions with comparable low R values in the C-terminal region, indicating the presence of a second compact subunit.

Table V: Values for the roughness index, R , for the entire loop (133-159), and for the residues included in the N-terminal (R_N) and C-terminal (R_C) domains

	R 133-159	N-terminal domain residues	R_N	C-terminal domain residues	R_C
FP	2.15	137-144	0.96		
LP	2.18	141-151	1.1		
FL	2.18	137-143	1.2	152-159	1.2
FS	2.10	137-144	0.93	152-158	0.93
FQ	2.00	138-144	0.80		

The site of the N-terminal compact subunit appears to be related to residue 148. For those structures containing Phe-148, the domain is located on the N-terminal side of the cell attachment site (residues 137-144). For the LP structure, the N-terminal domain is shifted toward the C-terminal, to residues 141-151. In the latter structure, the alpha carbons of Leu-148 and Leu-151 are 5 Å apart, so that residues 148-151 form a β -turn (type II'). This turn appears to be stabilized by van der Waals contacts between the side-chains of Leu-148 and Leu-151. The net effect of this interaction is to bring the β -turn into the N-terminal domain, and shift the position of the domain toward the C-terminal.

Pro-153 appears to induce an inverse γ -turn from residues 152-154. The average values measured for the dihedral angles of the central residue in inverse

γ -turns in proteins are $\phi = -79^\circ$ and $\psi = +69^\circ$ (44). Pro is thus particularly suited to form an inverse γ -turn because its cyclic side-chain restrains its dihedral angle ϕ to values near -75° (44). The LP structure contains an H-bond from the amide nitrogen of Arg-154 to the oxygen of Ala-152, with the Pro-153 dihedral angles of $\phi = -81^\circ$ and $\psi = 70^\circ$, meeting all criteria for the turn. While the FP structure appears to lack the H-bond due to a somewhat rotated Pro dihedral angle ($\phi = -77^\circ$ and $\psi = 114^\circ$), a superposition of all the heavy atoms from the alpha carbon of Leu-151 to the amide nitrogen of Arg-154, inclusive, yields an RMSD of 0.36 Å, indicating that these two structures are nearly identical in this region (Color plate 2). Moreover, NMR data indicates that the FP peptide forms an inverse γ -turn around Pro-153 (P. Mascagni, personal communication).

In each of the representative structures, two or more β -turns involve the cell attachment site (the sequence Arg-Gly-Asp at residues 145-147). Although no one type of β -turn is conserved in all five structures, and no one set of four residues forms a β -turn in all five structures, turns at 146-149 and 147-150 occur in three and four of the structures, respectively. The propensity of this region to form β -turns is indicated by primary structure analysis (Figure 5), and arises from the density of Gly, Ser, Arg and Asp residues at or near the cell attachment site (45). Murata *et al.* (46) have shown that analysis of the CD spectrum of poly(RGD) for secondary structure content indicates 41% β -turns. Furthermore, Siligardi *et al.* (11) have predicted type II β -turns at residues 144-147 of the FP and LP peptides, based upon CD spectra, using solvent titrations. Although there appears to be a general consensus that a β -turn(s) occurs at the cell attachment site, the number and/or type(s) of turns is not yet clear. The unequivocal identification of the structure at this site is important since the Arg-Gly-Asp sequence at 145-147 is highly conserved in FMDVs, and there is evidence that not only is it involved in cell attachment, but it also comprises at least part of a neutralizing epitope (2).

Superposition of the alpha carbons in the region most likely to include the major antigen of serotype A12 (\approx residues 146-157) were conducted to investigate whether a correlation with serological data could be found (described in the chapter by F. Brown *et al.*, this volume). That is, does there exist a region in these representative structures which is structurally similar for FL, FS and FQ, only? A modestly good correlation was found for residues 148-153 (Table VI). This correlation decreased as the segment length was increased on either side. The mean value for the 10 possible superpositions listed in Table VI is 2.2 Å. There are five RMSD values ≤ 2.2 Å, indicating structural similarity. Of these five pairs, three coincide with the serological data (the pairs FS/FQ, FL/FS and FL/FQ, are structurally similar in this region), and two do not (the pairs LP/FQ and FP/FL, show RMSD values ≤ 2.2 Å, but the serological data do not indicate high cross-reactivity). The probability of guessing three correct matches, given five guesses and 10 possible selections is 8.3%. This correlation suggests that the region 148-153 comprises at least part of the dominant immunogen.

Conclusions

The CD spectra show that in aqueous solution the five immunogenic peptides are highly similar in structure and typical of "random coil" peptides which

Table VI: RMSD values (Å) for the superposition of the alpha carbons of residues 148-153 of the structures shown in Color Plate 1

	FP	LP	FL	FS	FQ
FP	0.0	2.9	1.1	2.3	2.6
LP		0.0	2.8	2.7	1.6
FL			0.0	1.8	2.2
FS				0.0	1.7
FQ					0.0

are believed to exist in a dynamic conformation comprised of, to a large extent, dihedral angles characteristic of β -strands and β -turns (47). However, in 100% TFE, the peptides show a range of α -helix-forming properties which correlate well with serological data indicating similarity between FL, FS and FQ, and differences between FP and LP and the former group. These α -helix-forming properties appear to be sensitive to substitution of Leu for Phe at residue 148, but insensitive to substitutions at residue 153, except for Pro. It seems likely that the sensitivity at position 148 arises from the position of this residue in the middle of the hydrophobic arc of a highly amphiphilic α -helix. In contrast, residue 153 is situated on this helix where it can interact with either hydrophobic or hydrophilic residues, so that only Pro, a known α -helix "terminator", affects the helix-forming properties at this position.

Searching conformational space using high temperature molecular dynamics, simulated annealing and energy minimization provides a different technique by which to investigate the structural basis of the antigen-antibody specificities revealed in the serological data. The models structures selected represent the global minimum for the polypeptides, within the limits of this conformational search. These structures indicate that the immunodominant loops of all five variants contain a relatively compact domain in the N-terminal region, and that the position of this domain depends upon residue 148. When Leu-148 is present, van der Waals contacts between the side-chains of Leu-148 and Leu-151 induce a type II' β -turn (residues 148-151) which is folded into the N-terminal domain.

The serological data and the helix-forming parameters obtained from CD indicate that although FP and LP are more similar to each other than to the remaining variants, they show significant structural differences from each other. The CD data suggest that the similarity between FP and LP arises from the helix-terminating effect of Pro, and the molecular models indicate (not inconsistently) that Pro-153 induces an inverse γ -turn at residues 152-154 in both LP and FP. The molecular models also suggest that the serological differences between LP and FP may arise from significant structural differences in the region on the N-terminal side of Pro-153, caused by attractive van der Waals forces between the side chains of Leu-148 and Leu-151 which do not occur between Phe-148 and Leu-151 in the remaining structures. These interactions could also explain the differences in α_{MIN} , α_{MAX} and H, observed for LP and FP. Since residues 148-153 are likely to be involved in the dominant immunogen, it appears that the two major determinants

of immunological specificity in these five antigenic variants are steric constraints imposed by the bulky aromatic ring of Phe-148 which may reduce the rotomers available to the side chain and thus prevent close interactions with the Leu-151 and Ala-152 side-chains, and steric constraints imposed by the cyclic side-chain of Pro-153.

Acknowledgements

We thank Dr. J. C. Sutherland (Biology Department, Brookhaven National Laboratory) for his assistance and the use of his CD spectrometer at Port U9B of the National Synchrotron Light Source (NSLS), and A. Emrick, D. C. Monteleone and J. Trunk (Biology Department, Brookhaven National Laboratory) and P. M. Brashich (Plum Island Animal Disease Center, USDA/ARS) for their technical assistance. We also thank Dr. W. C. Johnson, Jr. (Oregon State University) for providing a copy of the program Varselec. The NSLS is supported by the Office of Chemical Research and The Office of Materials Research, USDOE. The CD spectrometer at port U9B of the NSLS is supported by the Office of Health and Environmental Research, USDOE.

Literature Cited

1. Acharya, R.; Fry, E.; Stuart, D.; Fox, G.; Rowlands, D.; Brown, F. *Nature* **1989**, *337*, 709-716.
2. Fox, G.; Parry, N. R.; Barnett, P. V.; McGenn, B.; Rowlands, D. J.; Brown, F. *J. Gen. Virol.* **1989**, *70*, 625-637.
3. Baxt, B.; Becker, Y. *Virus Genes* **1990**, *4*, 73-83.
4. Liebermann, H.; Dolling, R.; Schmidt, D.; Thalmann, G. *Acta virol.* **1991**, *35*, 90-93.
5. Rowlands, D. J.; Clarke, B. E.; Carroll, A. R.; Brown, F.; Nicholson, B. H.; Bittle, J. L.; Houghten, R. A.; Lerner, R. A. *Nature* **1983**, *306*, 694-697.
6. Geysen, H. M.; Meloen, R. H.; Barteling, S. J. *Proc. Natl. Acad. Sci. USA* **1984**, *81*, 3998-4002.
7. Francis, M. J.; Fry, C. M.; Rowlands, D. J.; Bittle, J. L.; Houghten, R. A.; Lerner, R. A.; Brown, F. *Immunology* **1990**, *61*, 171-176.
8. Francis, M. J.; Hastings, G. Z.; Clarke, B. E.; Brown, A. L.; Beddell, C. R.; Rowlands, D. J.; Brown, F. *Immunology* **1990**, *69*, 171-176.
9. Moore, D. M.; Vakharia, V.; Morgan, D. O. *Virus Res.* **1989**, *14*, 281-296.
10. Baxt, B.; Vakharia, V.; Moore, D. M.; Franke, A. J.; Morgan, D. O. *J. Virol.* **1989**, *63*, 2143-2141.
11. Siligardi, G.; Drake, A. F.; Mascagni, P. M.; Rowlands, D.; Brown, F.; Gibbons, W.A. *Eur. J. Biochem.* **1991**, *199*, 545-551.
12. Sutherland, J. C.; Desmond, E. J.; Takacs, P. Z. *Nucl. Instr. & Meth.* **1982**, *172*, 195-199.
13. Sutherland, J. C.; Keck, P. C.; Griffin, K. P.; Takacs, P. Z. *Nucl. Instr. & Meth.* **1982**, *195*, 375-379.

14. Scopes, R. K. *Anal. Biochem.* **1974**, *59*, 277-282.
15. Gray, D. M.; Lang, D.; Kuner, E.; Vaughan, M.; Sutherland, J. *Anal. Biochem.* **1984**, *136*, 247-250.
16. France, L. L.; Kieleczawa, J.; Dunn, J. J.; Hind, G.; Sutherland, J. C. *Biochim. Biophys. Acta* **1992**, *1120*, 59-68.
17. Manavalen, P.; Johnson, W. C., Jr. (1987) *Anal. Biochem.* **1987**, *167*, 76-85.
18. Johnson, W. C., Jr. *Pro. Struct. Funct. Gen.* **1990**, *7*, 4108-4116.
19. Kyte, J.; Doolittle, R. F. *J. Mol. Biol.* **1982**, *157*, 104-132.
20. Hopp, T. P.; Woods, K. R. *Proc. Natl. Acad. Sci. USA* **1981**, *78*, 3824-3828.
21. Karplus, P. A.; Schulz, G. E. *Naturwissenschaften* **1985**, *72*, 212-213.
22. Chou, P. Y.; Fasman, G. D. *Ann. Rev. Biochem.* **1978**, *47*, 251-276.
23. Eisenberg, D.; Wesson, M.; Wilcox, W. In *Prediction of Protein Structure and the Principles of Protein Conformation*; Ed. Fasman, G.D.; Plenum Press: New York, NY, 1989; pp. 635-646.
24. Roitt, I. *Essential Immunology*; Blackwell Scientific Publications, Oxford, UK, 1991; p. 71.
25. Marqusee, S.; Robbins, V. H.; Baldwin, R. L. *Proc. Natl. Acad. Sci. USA* **1991**, *86*, 5286-5290.
26. Chakrabarty, A.; Schellman, J. A.; Baldwin, R. L. *Nature* **1991**, *351*, 586-588.
27. Lyu, P. C.; Liff, M. I.; Marky, L. A.; Marky, L. A.; Kallenbach, N. R. *Science* **1990**, *250*, 669-673.
28. Merutka, G.; Lipton, W.; Shalongo, W.; Park, S. -H.; Stellwagen, E. *Biochemistry* **1990**, *29*, 7511-7515.
29. O'Neil, K. T.; DeGrado, W. F. *Science* **1990**, *250*, 646-651.
30. *Handbook of Chemistry and Physics, 46th edition*; Weast, R. C.; Selby, S.M.; Hodgman, C. D., Eds.; The Chemical Rubber Co.: Cleveland, OH, 1965-66.
31. Barcelo, J. R.; Otero, C. *Spectrochim. Acta* **1962**, *18*, 1231.
32. Gilson, M. K.; Rashin, A.; Fine, R.; Honig, B. *J. Mol. Biol.* **1985**, *183*, 503-516.
33. Schiffer, M.; Edmundson, A. B. *Biophys. J.* **1968**, *8*, 29-39.
34. Greenfield, N.; Fasman, G. D. *Biochemistry* **1969**, *8*, 4008-4116.
35. Altmann, K. -H.; Wojcik, J.; Vasquez, M.; Scheraga, H. A. *Biopolymers* **1990**, *30*, 107-120.
36. Schimmel, P. R.; Flory, P. J. *J. Mol. Biol.* **1968**, *201*, 105-120.
37. Richardson, J. S.; Richardson, D. C. *Science* **1988**, *240*, 1648-1652.
38. O'Neil, K. T.; DeGrado, W. F. *Science* **1990**, *250*, 646-651.
39. Mackay, D. H. J.; Cross, A. J.; Hagler, A. T. In *Prediction of Protein Structure and the Principles of Protein Conformation*; Fasman, G., Ed.; Plenum Press: New York, NY, 1989; pp. 317-358.
40. Hagler, A. T., Osguthorpe, D. J., Dauber-Osguthorpe, P., Hemple, J. C. *Science* **1985**, *227*, 1309-1315.
41. Connolly, M. L. *Science* **1983**, *221*, 709-713.
42. Richards, F. M. *Ann. Rev. Biophys. Bioeng.* **1977**, *6*, 151-176.

43. Zehfus, M. H.; Rose, G. D. *Biochemistry* **1986**, *25*, 5759-5765.
44. Milner-White, E. J.; Ross, B. M.; Ismail, R.; Belhadj-Mostefa, K.; Poet, R.; *J. Mol. Biol.* **1988**, *204*, 777-782.
45. Wilmot, C. M.; Thornton, J. M. *J. Mol. Bio.* **1988**, *203*, 221-232.
46. Murata, J.; Saiki, I.; Ogawa, R.; Nishi, N.; Tokura, S.; Azuma, I. *Int. J. Peptide Protein Res.* **1991**, *38*, 212-217.
47. Loret, E. P.; Georgel, P.; Johnson, W. C., Jr.; Ho, P. S. *Proc. Natl. Acad. Sci. USA* **1992**, *89*, 9734-9738.

RECEIVED April 14, 1994

Chapter 5

Determination of Secondary Structures of Proteins Using Vibrational Circular Dichroism

Timothy A. Keiderling¹, Petr Pancoska^{1,2}, Vladimir Baumruk¹,
Marie Urbanova^{1,2}, Vijai P. Gupta¹, Rina K. Dukor^{1,3},
and Dongfang Huo¹

¹Department of Chemistry, University of Illinois at Chicago,
845 W. Taylor Street, Chicago, IL 60607–7061

²Faculty of Mathematics and Physics, Department of Chemical Physics,
Charles University, Ke Karlovu 3, Prague 2, Czech Republic

Vibrational Circular Dichroism (VCD) has been shown to be both qualitatively and quantitatively more sensitive to protein secondary structure than are its two “parent” techniques, IR and CD spectroscopy. In this lecture, recent developments and applications of VCD to protein structure have been reviewed.

Electronic circular dichroism in the ultraviolet (ECD) has become an indispensable tool for qualitative characterization of proteins in solution. However accessible amide transitions are limited in number, are broad and overlapping, and the resulting UV spectral bands are often sensitive to environmental or local perturbations. Vibrational spectroscopies, such as infrared (IR) and Raman, also have an established role for characterization of secondary structures of proteins and peptides. While exhibiting many resolved transitions, these are generally limited to measurement of relatively small frequency shifts characteristic of the effects of conformation on bond strengths or to perturbations due to hydrogen bonding. Vibrational CD (VCD) [and Raman optical activity, not covered in this talk] has developed as a hybrid of these techniques and has recently found application in the biomolecular structural studies (1). VCD can be used to correlate data for several different spectrally resolved transitions that involve different localized vibrations of the molecule; and each of these features will have a distinct bandshape dependence on molecular stereochemistry.

³Current address: Bioinformatics, Amoco Technologies, Naperville, IL 60566

0097–6156/94/0576–0061\$08.00/0
© 1994 American Chemical Society

Empirical correlation of spectral features with secondary structure has historically been the most profitable route for stereochemical utilization of both electronic CD and vibrational (IR and Raman) spectroscopies. The difference in the origins of CD measured in the two spectral regions suggests that they would bear a complementary relationship that could enhance the quality and quantity of structural information derivable from either one alone and compensate for shortcomings of each. A series of studies on peptides of varying sequence and length have borne this out. VCD has a distinctively shorter length dependence that does ECD which leads to its having more sensitivity to the variety of secondary structure types seen in proteins (1). Quantitative approaches to a uniform systematic analysis of VCD, FTIR and ECD data, with the eventual goal of carrying out a coupled analysis, is a major topic of our ongoing studies. The characteristic patterns observable and the insight gained into several protein systems from the simple, bandsape variation point of view are the focus of this report.

Experimentally, VCD spectra are routinely measured on either dispersive or FTIR-based instruments, both of which have been described in detail in the literature (2). To date, protein dispersive VCD obtained at $\sim 10\text{ cm}^{-1}$ resolution by averaging several repetitive scans over the band of interest have a signal-to-noise ratio (S/N) advantage over FTIR-VCD spectra (3). Our spectra are usually obtained on very concentrated (up to 50 mg/ml in D_2O and 200 mg/ml in H_2O) solutions in short path length (25 μm and down to 6 μm for H_2O) sample cells with CaF_2 windows (4). In non-aqueous environments, for model peptides or solvent perturbation tests of protein structure, lower concentrations and longer path lengths are possible. For purposes of comparison and further spectral analyses, higher resolution and better S/N FTIR absorption spectra are obtained on the same or more dilute samples, and ECD spectra ($>180\text{ nm}$) are obtained with much more dilute samples. All spectra are systematically treated using the SpectraCalc package of programs for data manipulation.

VCD Studies of Peptide Models. A number of studies on polypeptides and oligopeptides have established the regularities of VCD spectra for amide vibrational modes (5-8). The amide I band (amide I' in D_2O), mainly C=O stretch at $\sim 1650\text{ cm}^{-1}$, is the most characteristic and easiest to study with VCD. The amide II (N-H deformation and C-N stretch) VCD seems less sensitive to variation in secondary structure (9), and the amide III (same) is quite weak and mixed with non-amide modes (10). Typical model peptide results are shown in Figure 1 for N-H protonated α -helix and random coil model systems in H_2O . Suitable model β -sheets are difficult to measure under these conditions due to solubility and aggregation problems. The α -helical result is maintained over a range of solvents and peptide lengths and compositions. The primary variation in the α -helical band shape is due to N-H deuteration, which causes the

amide I' to have three features (-+-) rather than just the positive couplet pattern seen in Figure 1 and shifts the amide II down to $\sim 1450\text{ cm}^{-1}$ with loss of intensity. These studies have demonstrated that VCD of these amide modes exhibits a remarkable independence of the type of side chains on the peptide and a resolution of aromatic contributions in contrast to ECD (5). The β -sheet VCD is less universal, being clearest (two weak negative bands at ~ 1615 and $\sim 1685\text{ cm}^{-1}$) for the amide I' transitions of aggregated antiparallel structures in D_2O solution (1,6). The 'random-coil' form in many polypeptides and proteins gives rise to a large *negative* couplet (5-7). Such a pattern implies that, in these 'random-coils', substantial local ordering exists that is similar in nature to that of left-handed helical poly-L-proline II, as has been supported by oligomer studies (7). Additionally, less common structures such as the 3_{10} helix give distinct band shapes, particularly when several transitions are studied (8). This library of experience with VCD has enabled us to develop a qualitative understanding of the VCD of proteins which are the focus of this report (1).

We have recently demonstrated that the characteristic VCD band shapes arise in large part from just the near-neighbor interactions in an oligopeptide by carrying out *ab initio* level magnetic field perturbation calculations of peptide VCD spectra (11). Calculations for just a dipeptide constrained to ϕ, ψ angles characteristic of the secondary structures of interest yielded band shapes and relative intensities in very good agreement with the experimental results for polymers. These support our empirical observations, based on oligomer studies, that VCD for peptides has a relatively shorter range length dependence than does ECD.

Results and Discussion

Qualitative VCD patterns for proteins. In extending the application of VCD to proteins, we have obtained reproducible VCD spectra and profitably compared it to ECD and FTIR data for a range of proteins (1,3,4,9,12). As compared to IR spectra, the sign variation inherent in VCD gives it an effective resolution advantage in differentiating between proteins. Furthermore, the short range effects lead to different types of secondary structure contributing on a comparable basis to VCD, leading to more variation in band shape than in ECD, which is singularly dominated by the α -helical contribution. Our first experiments were focussed on the amide I' band, C=O stretch in D_2O solution, due to its ease of detection and the relatively simpler sample preparations (more dilute, longer path). For example, hemoglobin is in the class of proteins whose secondary structure is dominated by the α -helix, while concanavalin A is dominated by its β -sheet components. Each of these has similarly shaped ECD with the main difference being intensity (1,3). In VCD, they have different amide I' band shapes with oppositely signed overall patterns. Furthermore these VCD bands are significantly shifted in frequency from

each other so that the negative components arising in each are completely resolved. It is this shape reversal combined with shifts corresponding to the band width that leads to VCD evidencing more sensitivity to the structural variation than ECD (3,12). Thus globular proteins with a mix of **a** and **b** components have VCD spectra resembling a simple linear combination of these forms. As a consequence, one can state that the qualitative aspects of the protein secondary structure are more simply represented in the protein VCD than in the protein ECD. The same qualitative advantage is evident in the amide I VCD for proteins in H₂O, as is illustrated in Figure 2 where the VCD of chymotrypsin, myoglobin, and ribonuclease A are compared.

The same sort of advantage, only stronger, can be said to obtain for the protein VCD spectra in comparison to IR absorbance spectra, since the IR only evidences minor shifts in the overall bandshape with conformational change as compared to the overall band shape variations seen in VCD. Of course, with FTIR it is possible to resolution enhance the IR absorbance and improve the discriminatory capability of that technique. However the IR analysis is typically dependent on frequency assignment of features which we have shown to be non-unique (13). Since VCD and FTIR sample the same transitions, it was possible to use VCD to demonstrate that bands occurring at specific FTIR frequencies can have opposite VCD signs in different proteins. For example transitions above 1670 cm⁻¹ can occur with negative signs in proteins of predominantly helical character and with positive signs in proteins of little helix and substantial β -sheet and coil character. There is no way that these transitions that all occur in the same frequency interval and are often assigned to turns can arise from the same structural unit and have different VCD sign patterns. Since the only way that the same conformational unit in two peptides can give rise to opposite VCD signs is for them to have the opposite handedness, this work clearly demonstrates that FTIR frequencies have an ambiguous, at best, correlation to structure. The strength of VCD (or ECD for that matter) is its unique dependence on the local chirality of the chromophores being studied. In this case, for the amide I band, the chromophores are highly localized C=O stretches whose chirality arises almost entirely from coupling to their near neighbors.

For proteins in H₂O solution (4), the amide I VCD is sometimes more intense than is the corresponding amide I' VCD in D₂O solution, but the patterns simply correlate back to the basic peptide studies, with the most difference between the amide I' and I VCD evident in mixed α and β systems. The α -helical dominated proteins, such as myoglobin, have two signed VCD features in H₂O, with the lower negative band that is seen in D₂O disappearing and the positive becoming stronger and broader as is clear in Figure 2. On the other hand, the β -sheet dominant proteins, such as chymotrypsin, have a dominant negative band to lower frequency (~1635 cm⁻¹) much as seen for the same samples in D₂O.

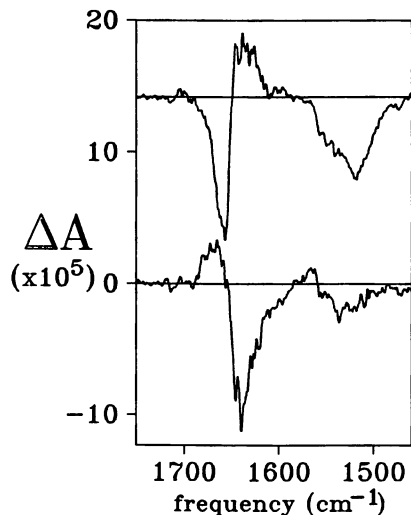


Figure 1. VCD spectra of $(PKELLEKL)_N$, a model α -helical peptide (top), and poly-L-lysine at pH7, a random coil polypeptide (bottom), in H_2O for the amide I and II bands.

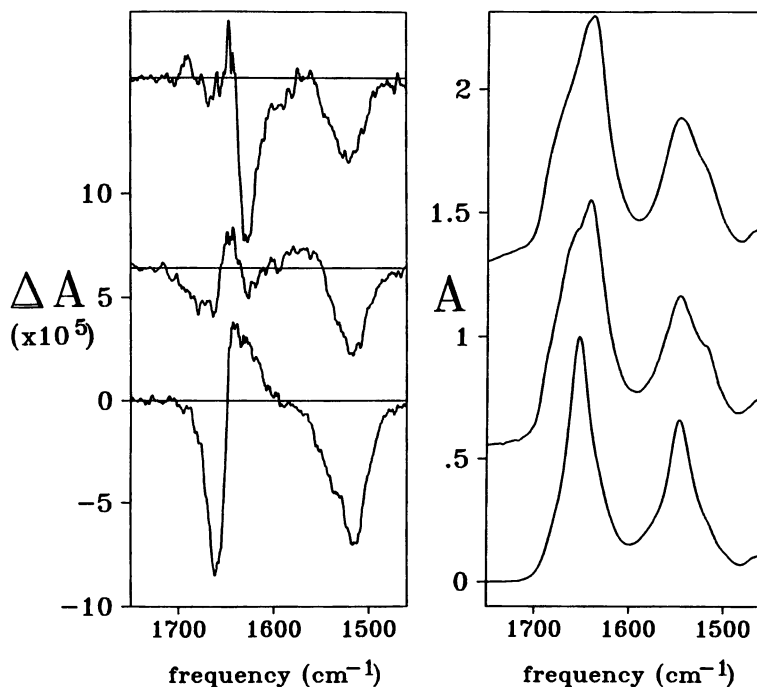


Figure 2. VCD (left) and FTIR (right) spectra of chymotrypsin, ribonuclease A, and myoglobin (top to bottom) in H_2O for the amide I and II bands.

The H₂O based measurements illustrated here (4) allow simultaneous study of the amide II VCD for the same sample in the same cell at the same concentration. For subsequent quantitative studies this has an advantage in stabilizing the spectral data set. The amide II evidences less shape variation and little shift with secondary structure change. The α -helical amide II VCD is predominantly negative and falls to lower frequency than the absorbance maximum while the β -sheet amide II VCD is a somewhat weaker negative couplet centered on the absorbance. Again the mixed structure proteins have a VCD that looks roughly like a linear combination of the two more extreme forms. These amide II patterns have a variance with secondary structure roughly comparable to that found with ECD.

Quantitative Secondary Structure Studies with VCD. A statistical analysis of the amide I' VCD data for 13 proteins to fit their fractional secondary structures as derived from x-ray crystal structures resulted in a method of secondary structure analysis that gives more precise fits, particularly for the β -sheet, under defined conditions, than with the same statistical method using ECD data (Table I) (12). These analyses were done with a limited set of spectral parameters, keeping only those found to have a statistically reliable dependence on secondary structure parameters. This is in contrast to many current methods which use multiple spectral parameters and achieve apparently very precise fits. Expansion of the data base to include more proteins and inclusion of the VCD of the amide II band for the same expanded set of proteins in H₂O leads to a stable analysis and the capability of fitting more secondary structural elements (14). Our preliminary results indicate that using the VCD spectra measured for the amide I and II bands on the same sample in H₂O, and thereby avoiding the complications of deuterium exchange, can lead to relatively more precise fits (15). But in both approaches, increasing the number of proteins beyond the original 13 (to 23) leads to some decrease in the precision of statistical fit when the analysis is done under the same conditions. Of course with more parameters used, the precision improves. In all cases, better fits are obtained with parameter sets that have contributions from both spectral transitions.

However, the main goal of spectral analyses of secondary structure is not fitting but prediction of secondary structure for proteins whose x-ray structures are unknown. In our initial study we tested prediction by leaving one protein out of the analysis and optimizing the regression on just 12 protein spectra. Then the equations were used to predict the structure of the protein left out. This was repeated for each protein in the set and the average prediction error computed and compared to the error of the fit. As expected the predicted values were worse than the fit values, but in the initial attempt the difference was not too large (Table I). Unfortunately, we have found that the improvement of the fit as described above with more parameters and/or data from more than one transition is *not* matched by an improvement of the quantitative

predictive capability of the method, tested as described above for one protein left out of the set, even with the larger set of proteins (14). [In an alternate approach, predictive capabilities of the method are also being tested on structures for various species and under different environmental conditions.]

Table I. Standard Deviations of Fits and Predictions of Protein Secondary Structure Using VCD (Amide I' Data Only) and ECD Data for 13 "Known" Proteins

(%)	α	β	bend	turn	"other"
$\sigma_{\text{fit}}^{\text{VCD}}$	6	6	4	-	5
σ_{predict}	9	7	5	-	6
$\sigma_{\text{fit}}^{\text{ECD}}$	8	10	4	-	3

SOURCE: Adapted from ref. 12.

Applications. Despite the seemingly large errors found in the quantitative approaches, VCD, as well as both ECD and resolution enhanced FTIR, evidence high sensitivity to small conformational changes in proteins that can be induced by environmental perturbation. Consequently a number of applications for VCD have arisen that can be interpreted usefully on even a qualitative basis. For example, the α -lactalbumin crystal structure shows the same folding pattern as found for lysozyme but the VCD spectra of the two in D₂O are significantly different. If the solvent for α -lactalbumin is altered to contain 33% propanol (16) or the pH is lowered (Figure 3), a better match with the lysozyme VCD is found. Our data thus indicate that the crystal and aqueous solvent environment structures for α -lactalbumin are not the same. Variations of the spectra with species (and consequently, sequence) indicate that the low pH structures are much more uniform than are the neutral pH ones (17). The higher the homology in sequence with the human α -lactalbumin sequence (the closest available to the baboon one studied crystallographically) the less is the neutral pH spectral deviation from the human α -lactalbumin VCD, ECD and FTIR spectra. At low pH these spectra tend to converge on the low pH human α -lactalbumin result. Quantitative analyses of these spectra indicate that the proteins in the lower pH environment, normally ascribed to the partially denatured, molten globule or A-state, have higher α -helical fractions than at neutral

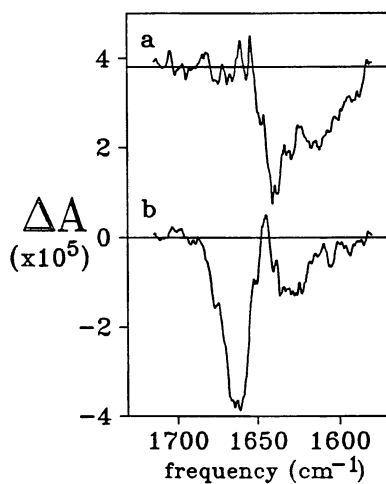


Figure 3. VCD of human α -lactalbumin in D_2O at (a) pH7.2 and (b) pH2.5 with added Ca^{2+} .

pH. These interpretations are consistent for the VCD, ECD and FTIR results, but the change is amplified in the VCD spectra.

Similarly, we studied two growth factors proteins that had been analyzed, on the basis of frequency shift, by FTIR to have a significant α -helix content. However no other technique (ECD, NMR, x-ray) gives data in agreement with such proposed structures. VCD spectra confirmed that the proteins had little helix and indicated that the frequency positions of the misattributed FTIR bands was probably a consequence of the high level of solvation of the small peptide fragment environment and not of secondary structure (18). This study provided an excellent example of the ambiguity that can arise using just IR frequencies to assign secondary structure.

From the other point of view, ECD studies of glycoamylase yield unusual values for the α -helix content due to interference from the glycosylated residues. The sugar groups contribute to the mass, making concentration measurements difficult, and contribute to the CD in the far uv just where one needs precise data for a reliable ECD-based structural determination. VCD avoids this problem due to the inherent resolution of the vibrational spectrum and to the normalization to the peptide absorption which can be used to circumvent the need for precise concentration determinations. We were thus able to use VCD to study the denaturation from high helix (room temperature) to substantial sheet forms due to aggregation of the denatured protein at high temperatures (19).

Acknowledgments. We thank the National Institutes of Health for continued support of this research (GM 30147 to TAK) and the National Science Foundation (INT91-07588 to TAK with PP) for help maintaining our international collaboration.

Literature Cited

1. Keiderling, T. A. and Pancoska, P. In *Biomolecular Spectroscopy Part B*, Ed. R. E. Hester and R. J. H. Clark, *Advances in Spectroscopy*, John Wiley and Sons, Chichester, 1993, Vol. 21, p.267-315. Keiderling, T. A. *Nature* **1986** 322, 851-852. Keiderling, T. A. In *New Techniques and Applications of Physical Chemistry to Food Systems (Physical Chemistry of Food Processes, Volume II)* Ed. I. Bainau, H. Pessen and T. F. Kumosinski, Van Nostrand Reinhold, New York, 1993, pp. 307-337.
2. Keiderling, T. A., in 'Practical Fourier Transform Infrared Spectroscopy' (Ferraro, J. R., and Krishnan, K., Eds.) p. 203-284, Academic, San Diego, 1990; Keiderling, T. A., *Appl. Spectr. Rev.*, **1981**, 17, 189-226.
3. P. Pancoska, S. C. Yasui, and T. A. Keiderling, *Biochemistry* **1989** 28, 5917-5923.

4. V. Baumruk and T. A. Keiderling *J. Amer. Chem. Soc.* **1993** 115 6939-6942.
5. Lal, B. B. and Nafie, L. A. *Biopolymers* **1982** 21, 2161-83. Sen, A. C. and Keiderling, T. A. *Biopolymers* **1984** 23, 1519-32. Yasui, S. C., Keiderling, T. A. and Katakai, R. *Biopolymers* **1987** 26, 1407-1412. Yasui, S. C. and Keiderling, T. A., *Biopolymers* **1986** 25 5. Yasui, S. C., Keiderling, T. A. and Sisido, M., *Macromolec.* **1987** 20 2403. Dukor, R. K. and Keiderling, T. A. in *Peptides 1988, Proceedings of the 20th European Peptide Symposium* (Bayer, E., and Jung, G., eds) deGruyter, Berlin, 1989, p. 519-521.
6. Yasui, S. C. and Keiderling, T. A., *J. Am. Chem. Soc.* **1986** 108, 5576-5581. Paterlini, M. G., Freedman, T. B. and Nafie, L. A. *Biopolymers* **1986** 25, 1751-1765.
7. Dukor, R. K., Keiderling, T. A. and Gut, V. *Int. J. Pept. Prot. Res.* **1991** 38 198-203. Dukor, R. K. and Keiderling, T. A. (1991) *Biopolymers* **1991** 31 1747-1761.
8. Yasui, S. C., Keiderling, T. A., Formaggio, F., Bonora, G. M., and Toniolo, C. *J. Am. Chem. Soc.* **1986** 108, 4988-93.
9. Gupta, V. P. and Keiderling, T. A. *Biopolymers* **1992** 32 239-248
10. Diem, M., Oboodi, M. R., Alva, C. *Biopolymers* **1984** 23 1917-1930. Roberts, G. L., Lee, O., Calienni, J. and Diem M. *J. Amer. Chem. Soc.* **1988** 110 1749-52. Roberts, G. L., Lee, O. and Diem M. *J. Phys. Chem.* **1992** 96 548-554. Malon, P., Kobrinskaya, R. & Keiderling, T. A. *Biopolymers* **1988** 27 733-746.
11. Bour, P. and Keiderling, T. A. *J. Amer. Chem. Soc.* **1993** 115 9602-9607.
12. Pancoska, P., Yasui S. C., and Keiderling, T. A. *Biochemistry* 30 **1991** 5089-5103. Pancoska, P., and Keiderling, T. A., *Biochemistry* **1991** 30 6885-6895.
13. Pancoska, P., Wang, L. and Keiderling, T. A. *Prot. Sci.* **1993** 2 411-419.
14. Pancoska, P., Urbanova M., Gupta, V. P., and Keiderling, T. A. *Anal. Bioch.* (to be submitted)
15. V. Baumruk, Pancoska, P. and T. A. Keiderling (to be submitted)
16. Urbanova M., Dukor, R. K., Pancoska, P., Gupta, V. P., Keiderling, T. A. *Biochemistry* **1991** 30 10479-10485
17. Urbanova, M., Pancoska, P and Keiderling T. A. (to be submitted)
18. Dukor, R. K. Keiderling, T. A., Prestrelski, S. A., and Arakawa, T. *Arch. Biochem. Biophys.* **1992** 298 678
19. Urbanova, M., Pancoska, P and Keiderling T. A. *Biochim. Biophys. Acta* **1993** (in press)

RECEIVED June 1, 1994

Chapter 6

Global-Secondary-Structure Analysis of Proteins in Solution

Resolution-Enhanced Deconvolution Fourier Transform Infrared Spectroscopy in Water

Thomas F. Kumosinski and Joseph J. Unruh

Eastern Regional Research Center, Agricultural Research Service,
U.S. Department of Agriculture, 600 East Mermaid Lane,
Philadelphia, PA 19118

Previous studies comparing the global secondary (2°) structure of proteins by Fourier deconvolution FTIR were performed in D_2O . D_2O however increases hydrophobic interactions leading to spurious 2° structural changes. We have now performed FTIR experiments in H_2O on globular proteins with varying amounts and types of 2° structure. A method has been developed to increase the sensitivity of the analysis of these FTIR spectra. Calculation of the component 2° structural vibrational bands was accomplished by fitting both amide I and amide II envelopes by nonlinear regression analysis. The method entails fitting of: Fourier deconvoluted spectra, second derivative spectra, and refits of the component bands to the original spectra. Criteria for acceptance of the analysis was that the fractional areas from all three methods were in agreement. Results show good agreement with known X-ray crystallographic structures, and allow prediction of 2° structures for non-crystallizable proteins.

During the past several decades, controversy has existed within the literature concerning the methodology appropriate for analyzing experimental results to obtain the global secondary structure of proteins in solution; whether the experimental method used was circular dichroism (CD), Fourier transform infrared (FTIR) or vibrational circular dichroism (VCD). Traditional CD experiments were analyzed using a sum of Gaussian bands (I). In that publication, the criterion for the correct number of bands was established when the theoretical bands not only fit the CD data, but also when they were mathematically transformed (via a Krönig-Cramers transformation) to theoretical

0097-6156/94/0576-0071\$09.44/0
© 1994 American Chemical Society

optical rotatory dispersion (ORD) curves, with these curves agreeing with experimentally determined ORD results. However, this type of analysis necessitated performing ORD as well as CD experiments, which became cumbersome and costly for most investigators. The next method for analyzing CD results utilized factor analysis, and hypothesized that the CD spectra for an unknown protein was equal to a linear combination at each wavelength of pure secondary structural elements — such as α -helix, random, β -sheet, etc. To develop a basis set for analyses, model polypeptides known to adapt to almost pure structure (α -helix or β -sheet, etc.) were analyzed. However, this basis set was short lived because the fits to unknown protein data were poor.

With the increase in the number of protein structures determined from X-ray crystallography, many scientists developed new basis sets using the calculated secondary structure from the X-ray crystal structure of proteins. At first, factor analysis seemed appropriate due to the extremely low signal-to-noise and high error of the CD experiment. However, the questions of which class of proteins should be used and how the secondary structure should be calculated from the X-ray crystallographic structure became a never ending stumbling block for proteins with low α -helix and high β -sheet or turn conformations. It should be noted that many investigators who used basis set methodology have reported only the results of their calculations. They usually did not show a plot of the theoretical versus the experimental spectrum. If a good fit between the experimental and theoretical curve is not achieved, the basis set used is inappropriate for analysis and a new basis set should be sought. The question is whether such a basis set does, in fact, really exist.

When FTIR instruments with extremely good precision, accuracy and signal-to-noise were developed, current investigations of the IR spectra of proteins were possible. Unfortunately, factor analysis of the spectra rather than deconvoluting the spectra into its component bands was practiced.

Examining theoretical principles, other research groups (such as Mantsch (2) and Susi (3,4)) noted that the amide I band was a sum of badly overlapped Gaussian or Lorentzian bands. They adapted a methodology using calculated second derivative spectra, Fourier deconvolution algorithms, and nonlinear regression analysis for deconvoluting the amide I envelope into individual component bands. However, controversy exists to this day concerning the number of bands, the fraction of Lorentzian character, and the choice of parameter values for Fourier deconvolution.

In this paper we now present FTIR experiments in H₂O on thirteen globular proteins with varying types and amounts of 2° structures. Analysis of the spectral data using Fourier deconvolution, second derivative, and band curve-fitting techniques allows the individual 2° structural components to be distinguished and compared (3,5) with the known X-ray crystallographic data.

Methods

Infrared Measurement. The individual proteins were prepared as 4% solutions in 20 mM, pH = 6.7 imidazole buffer. All samples were introduced into a demountable cell with CaF₂ windows separated by a 12 μ m Teflon spacer. Spectra were obtained using a Nicolet 740 FTIR spectrometer equipped with the Nicolet 660 data system. Following nitrogen purge of the sample chamber to reduce water vapor to a minimum, data collection was carried out. Each spectrum consisted of 4096 double-sided interferograms, co-added, phase-corrected, apodized (Happ-Genzel function), and fast-Fourier transformed. Nominal instrument resolution was 2 cm⁻¹, with one data point every 1 cm⁻¹. Water vapor absorption was routinely subtracted from all spectra (6-7).

Data Analysis. Difference spectra obtained by subtraction of buffer absorption from the respective protein solution absorptions were used to calculate second-derivative spectra by a simple analytical procedure that used every data point (3). Second-derivative spectra served as sensitive indicators for identifying individual peak positions used in subsequent processing. The unresolved spectra were subjected to Fourier deconvolution (FD) using an algorithm developed from the one described by Kauppinen et al. (2). The deconvolution was undertaken with a number of resolution enhancement factors. Qualitatively, under-FD was judged by the absence of peak position indications in the spectra and over-FD by the appearance of side lobes and deconvolved noise in the flat portions of the spectra (5,8). The methodology used will be illustrated for lysozyme.

All spectra were deconvolved (decomposed into their component structural elements) using a Gauss-Newton nonlinear iterative curve-fitting program developed at this laboratory, which assumes Gaussian band envelopes for the resolved components. In practice, the three parameters of each band (height, peak frequency, and half-width at half-height) were allowed to float during the iterations, as was the baseline. Integrated areas were calculated for those peaks that correspond to conformational elements, such as helices, sheets, turns, and loops (9). The areas serve to estimate the fraction of the various secondary elements in the protein molecule. Note the terms deconvolution, deconvolving and deconvolved will refer to the nonlinear regression fitting procedure.

Sample Calculation: FTIR Analysis of Lysozyme. A typical FTIR spectrum of hen's egg white lysozyme showing just the amide I and amide II regions is in Figure 1 (outer envelope). The spectrum is considered to be a sum of the variety of individual absorption bands arising from the specific structural components of the protein — such as α -helix, β -sheets and turns. Fitting it directly with an undefined number of Gaussian bands by nonlinear regression would be a daunting task. To alleviate this dilemma, we first examine the

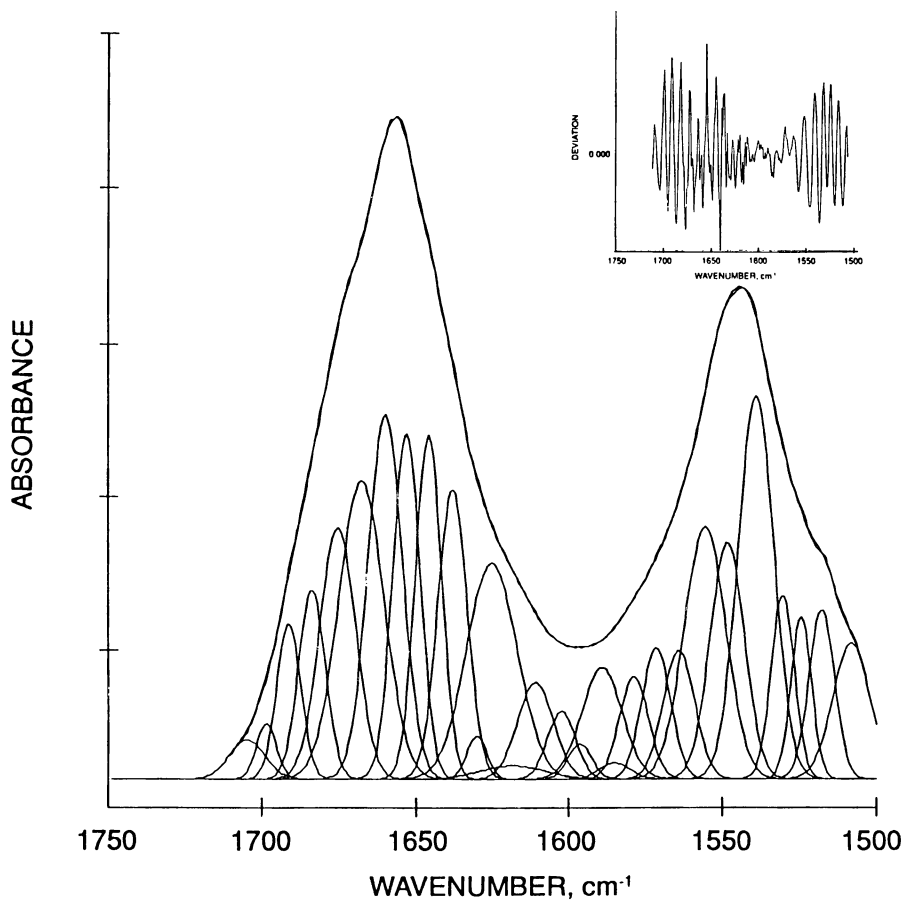


Figure 1. FTIR spectrum showing amide I and amide II bands of lysozyme in aqueous solution. Outer envelope double line is connected original spectrum. Line on outer envelope and individual component peaks underneath are the results of nonlinear regression analysis as described in text. Inset shows plot of residuals (connected by line) of the differences between the calculated and experimental absorbances vs. frequency.

second derivative of the spectrum (inset, Figure 2) to determine the number of component bands and the approximate positions of these bands.

The next step in the analysis is to enhance the resolution of the original spectrum via the FD algorithm developed by Kauppinen et al. (2). Care must be taken to choose the proper values for the band width and resolution enhancement factor used in this algorithm, so that the FTIR spectrum is not over- or under-deconvoluted. As the deconvolution procedure progresses, analysis of the FD spectrum by nonlinear regression analysis is used in an iterative fashion to determine the proper FD parameters.

Quantitative criteria to insure correct deconvolution are: (1) correlation of all band assignments with the second derivative peaks; (2) agreement of calculated and experimental baselines; (3) a root-mean-square value of the fit \leq instrumental noise; (4) a successful fit to the original spectrum of the model using fixed frequencies found by fitting the FD spectrum. In practice, attainment of these criteria may require several cycles of FD and regression, until an optimal fit is achieved. Criterion 4 involves using the results of the regression analysis of the FD spectrum (Figure 2) to provide the number of bands and their frequencies, which are then fixed in a model to perform a nonlinear regression analysis of the original spectrum.

The final fit to the lysozyme FTIR spectrum is shown in Figure 1 with its 28 component peaks. The inset (Figure 1) shows the residuals of the regression are reasonably random, indicating the model is a reliable fit to the data. Calculated relative areas under the component bands of the original spectrum are in good agreement with those calculated from results of the regression analysis of the FD spectrum.

Further validation of the calculated components of the amide I and II bands can be obtained by mathematical comparison of the second derivative FTIR spectrum obtained from the original with the calculated second derivative obtained from the model fit. The results of such a comparison for lysozyme are shown in Figure 3, where the jagged line represents the fitted model and the smooth line the experimental data. The inset of this figure shows a reasonable pseudo-random residual plot which further establishes the reliability of this methodology for quantitatively resolving FTIR results of proteins into their individual component absorption bands.

Results

Rational for Deconvoluting into Component Gaussian Peaks. We want to note some problems in reported protein 2° structure results. First, some of the groups still using factor analysis have reported protein structures within their database which do not add up to 100% structure (10-11); others, however, do (12). Second, the same calculated 2° structure from X-ray crystallography was used whether the experimental method was CD, FTIR or VCD. This assumption may have serious problems since the theoretical parameters

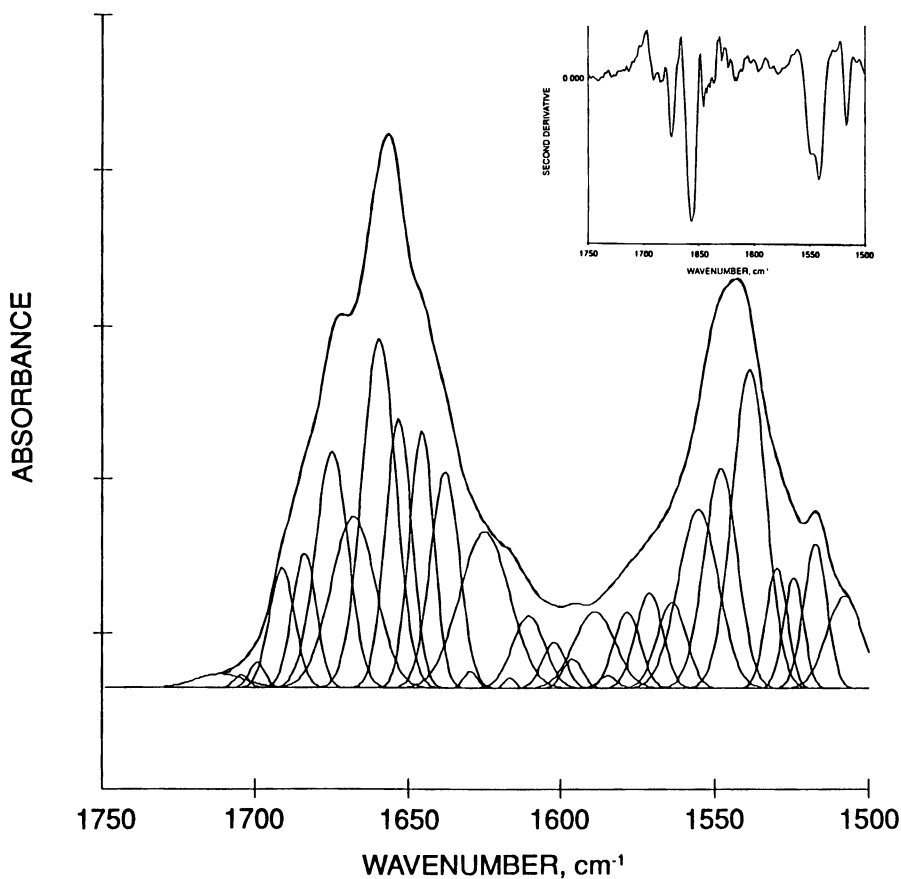


Figure 2. Fourier deconvolution of FTIR spectrum of lysozyme in Fig. 1. Lines on outer envelope and individual component peaks underneath were found by regression analysis as described in text. Double line is connected experimental data. Insert shows second derivative of original spectrum.

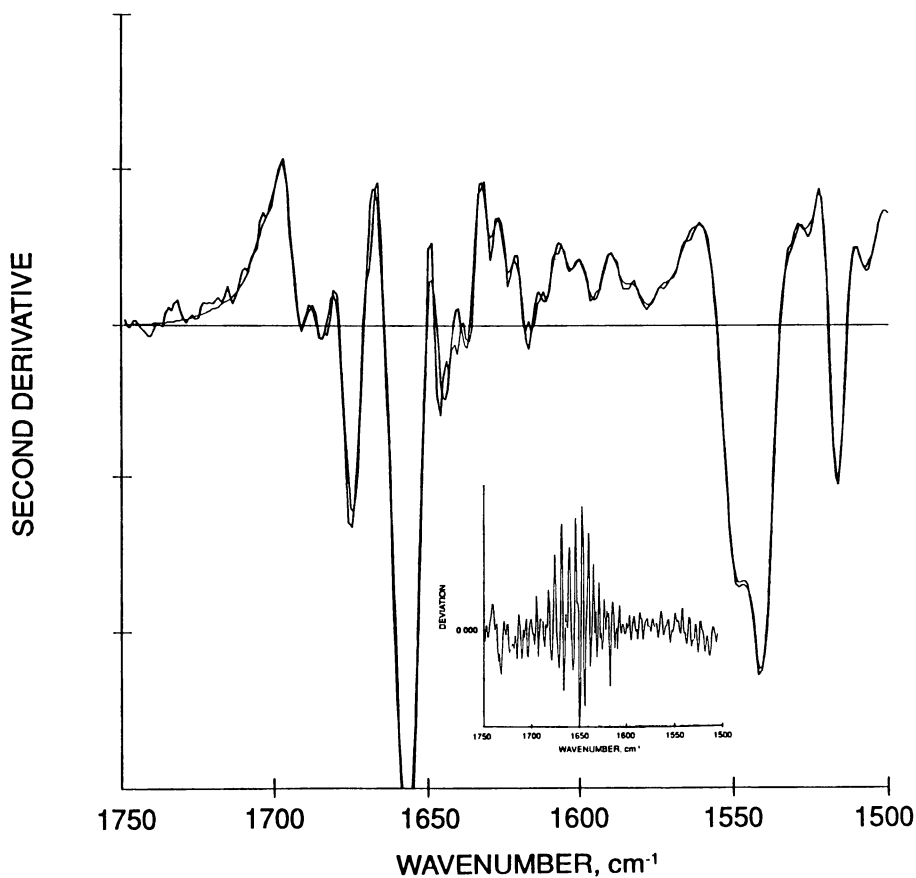


Figure 3. Second derivative FTIR spectrum of amide I and II bands of lysozyme in aqueous solution. Smooth line is connected experimental data. Jagged line on outer envelope is the results of final regression analysis as described in text. Insert shows plot of connected residuals between calculated and experimental second derivative results vs. frequency.

measured by these methods all differ: CD measures $\text{Im} \langle r\tau \times P \rangle$, for electronic transitions; VCD measures $\text{Im} \tau$ too, but for vibrational transitions; and FTIR measures the transition dipole $\langle r \rangle$, for vibrational transitions. Third, changes in the shape and position of the spectra, for which factor analysis relies, may not reflect any structural changes in a protein. To illustrate these problems we shall use the results obtained from the analysis of the lysozyme spectrum.

Using the component Gaussian bands calculated in the analysis of the original FTIR spectrum (Figure 1) of lysozyme, we have calculated a theoretical FTIR spectrum (see Figure 4A, dashed line). We also calculated an altered spectra by changing the shape of the 1676, 1659 and 1638 cm^{-1} bands (solid line in Figure 5A). The heights of the respective bands were divided by a factor of 1.2 while the half-width at half-height was multiplied by 1.2. This results in a different overall shape of the calculated amide I curve, while the resulting fractional areas of these bands remain constant. In this study, we shall refer to this modified theoretical spectrum as envelope 1.2 and the exact theoretical spectrum of lysozyme as the exact envelope.

In addition, an envelope of 1.5 was also constructed and is presented in Figure 4A (double solid line). The dashed line and the solid lines in Figure 4A represent the exact and the 1.2 envelope, respectively. It can be seen in Figure 4A that as the factor is increased from 1.0 (for the exact), to 1.2 and 1.5, dramatic changes occur in the amide I envelope. Here, with a factor of 1.0 only one band appears in the amide I region where with factors of 1.2 and 1.5 four bands appear, three of which are near the affected 1676, 1659 and 1638 cm^{-1} bands. The resolution of these three bands increases with an increasing factor. This overall shape change however, does not affect the calculated areas of the peaks. The effect on the area was corroborated by fitting the theoretical curves to a sum of 29 peaks via a nonlinear regression analysis. The calculated areas of the absorption bands at 1676, 1659, and 1638 cm^{-1} remained constant. Thus factor analysis would reveal changes in conformation while regression analysis would not.

Since FTIR has an extremely large signal-to-noise ratio and yields very precise spectra, it is capable of seeing very small conformational changes in proteins as a function of varying environmental conditions. It has been suggested (5) that the difference spectra between FTIR second derivative spectra can at certain frequencies reflect small 2° structural changes due to changes in environmental conditions. However, care must be taken using this methodology for several reasons. The first reason being that the areas of the two amide I envelopes must be exactly the same and it is never clear where the amide I envelope ends at the low end of the frequency range. The other reason will be discussed below using the results presented in Figure 4.

Figure 4B shows the calculated second derivative spectrum of the exact envelope of lysozyme as a dashed line. The double and single line in Figure 4B represent the difference between the calculated second derivative spectra of the exact envelope and the 1.2 as well as the 1.5 envelope, respectively. It can

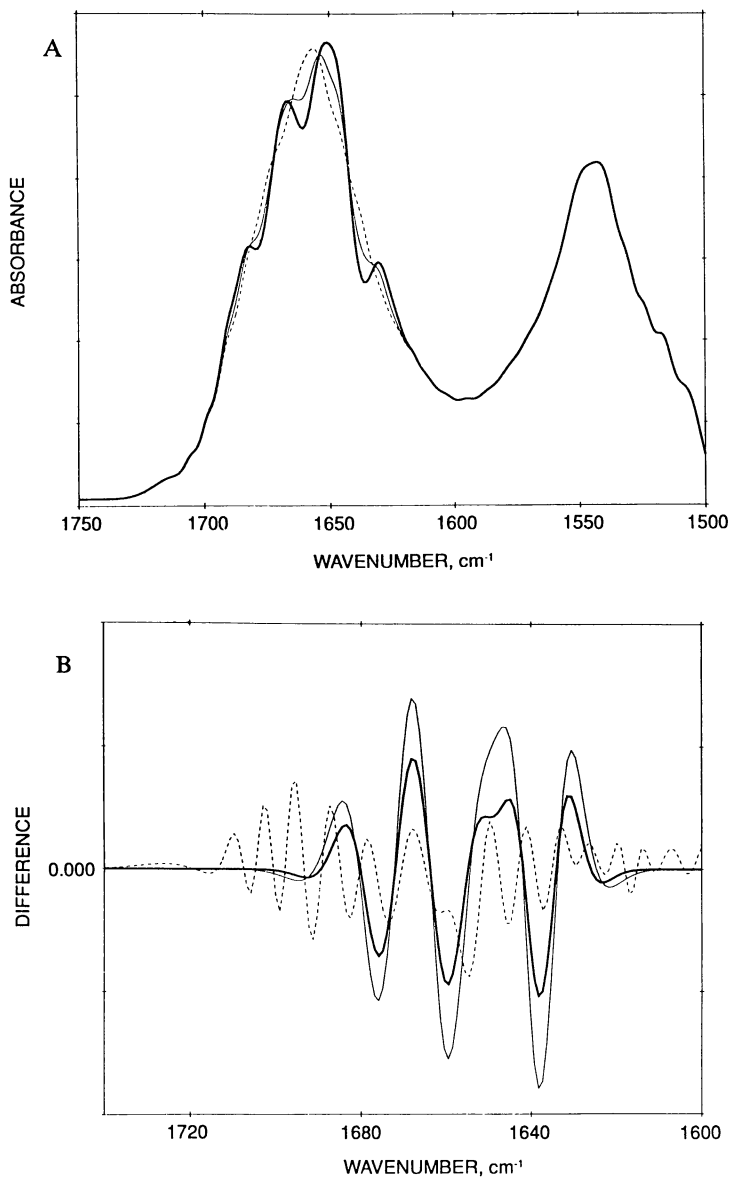


Figure 4. Theoretical spectrum from analysis of lysozyme results in Figure 1A. dashed line, exact curve from component bands of Figure 1; solid line, height decreased and width-at-half-height increased by a factor of 1.2 for 1676, 1959 and 1638 cm^{-1} bands; double line, same as single line but with a factor of 1.5. B. dashed line, calculated second derivative of exact curve in Figure 4A dashed line; solid line, difference between calculated second derivative of 4A double line and 4B dashed line; double line, difference between calculated second derivative of 4A single line and 4B dashed line.

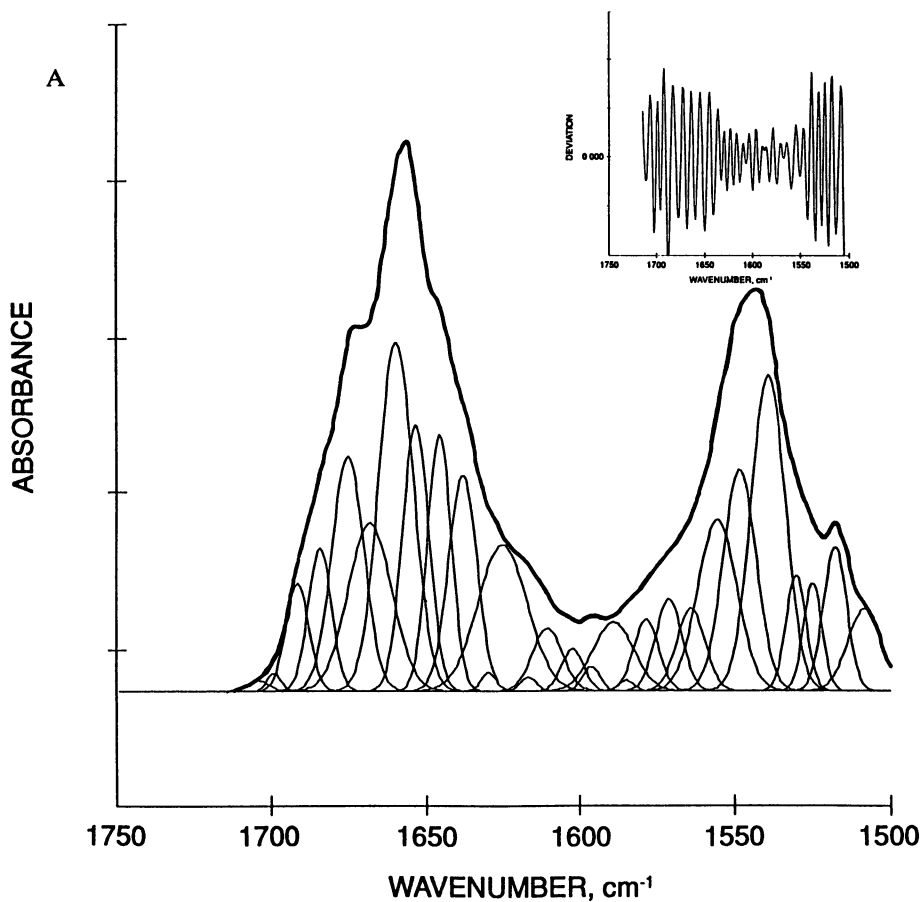


Figure 5. Best fit for Fourier deconvoluted lysozyme FTIR spectrum using non-linear regression analysis, half width at half height, W , of 9 cm^{-1} and a resolution enhancement factor of 2.5: single lines are experimental points, double lines are theoretical sum of component peaks shown in single line. A. for 29 peaks and B. for 28 peaks. Insets are plots of residuals between individual fits and experimental values.

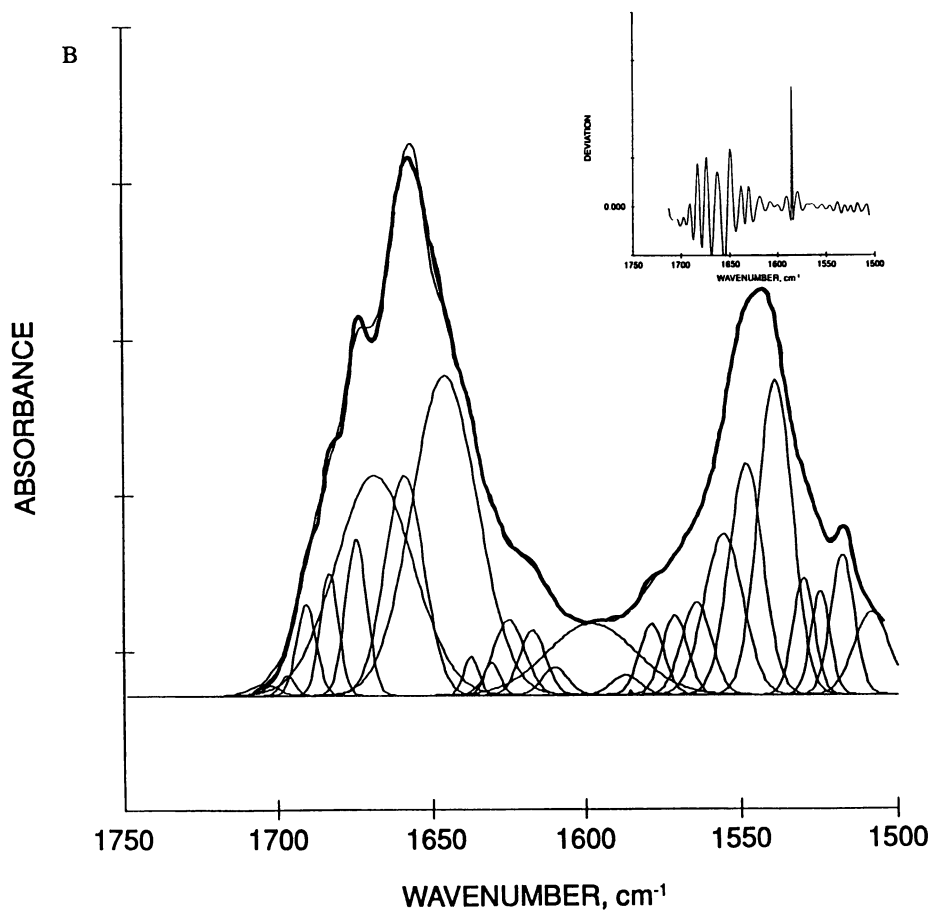


Figure 5. Continued.

easily be seen that the difference spectra in the amide I region (double and single line and in Figure 4B) are larger in magnitude than the exact calculated second derivative envelope of lysozyme. Such large differences may be easily interpreted as substantial conformation changes but, in reality, no change in area of the 1676, 1659, and 1638 cm^{-1} bands has occurred — only the distribution of their Gaussian envelopes has changed. Therefore, second derivative difference spectra are not valid methods to study protein conformational changes.

The only sure methodology is to deconvolve the FTIR spectrum into its component Gaussian bands. But to insure accurate deconvolving, two important aspects of the methodology must be considered: 1, using the proper number of bands in the calculations; and 2, using both the amide I and II envelopes in the calculations. Using too few bands and not considering the amide II envelope could lead to serious errors and misinterpretations.

Rational for the Parameters of the Non Linear Regression. Controversy exists among researchers concerning the deconvolving of the FTIR amide I and II into their component bands. The determination of the number of bands and the character of the bands (whether they are of pure Gaussian or a combination of Gaussian and Lorentzian character) requires careful consideration. And, when using FD to ascertain the number of bands, the magnitude of applied FD variables must be justified.

Traditional FTIR experiments were performed in D_2O where no amide II envelope exists, and only the amide I envelope was deconvolved into its component bands. So, when experiments were eventually performed in H_2O , the amide II band while present was not always deconvolved. Analysts then also routinely applied conservatively low values for the resolution enhancement factor (REF) with a large half-width at half-height value of 13 cm^{-1} , in order to avoid overdeconvolution of the spectrum. And then in the next step, used only the smallest number of component peaks for the nonlinear regression analysis.

Their rational was to avoid the possibility of distorting the experimental spectrum. However, no studies analyzing the same spectrum using nonlinear regression analysis with varying FD parameters have as yet been performed, thus the parameter limits before causing distortion are not known. Using higher REFs and narrower half-widths during FD increases the number of component bands. Properly choosing to use this increased number of bands (with equal half-widths at half height) in the nonlinear regression fit of the experimental spectrum results in much lower root-mean-square (RMS) values.

While nonlinear regression analysis using an increased number of component bands is more difficult and more time consuming (and cannot be easily performed on microcomputers), the correct number of peaks must be obtained to insure the correct assignments of the 2° structure of the protein spectra. If one band is used when two really exist, then a larger amount of disordered, helical or extended structure could be calculated due to an incorrect assignment. With our methodology we use the maximum number of component bands to fit the theoretical curve to the experimental data, that yields the

lowest root-mean-square value. We also fit the amide II along with the amide I envelope, and we allow a zero slope baseline to float to a calculated value. However, we still will maintain a low REF for FD of the spectra of 2.5 with a half-width at half-height of 9 cm^{-1} for all spectra. Furthermore, we only use pure Gaussian component bands for fits to FD, original and calculated second derivative amide I and II spectra. Attempts to use a constant fraction of Lorentzian character and optimize the value of the fraction of Lorentzian character were unsuccessful, resulting in non convergence of the Gauss-Newton nonlinear regression program. It should also be noted that during our calculations, the use of too many bands resulted in the heights of the excess bands approaching a zero or negative value. Thus, the excuse that excess bands yield better fits to the data is not valid. This statement only applies to the use of polynomial curve fitting algorithms and not to nonlinear regression analysis.

The effects of choosing too few component bands are discussed as follows. Figure 5A shows the fit of the Fourier deconvoluted amide I and II envelopes of lysozyme with 29 component Gaussian bands. The fit is excellent with no discernable difference between the resulted theoretical and experimental curves. However, when the band at 1659 cm^{-1} is removed and only 28 peaks are utilized, non linear regression analysis yields a poor fit with some component bands becoming much broader than others (see Figure 5B). Also, it can be seen that the fit to the amide II bands is affected even though no band was removed from that range of frequencies. In addition, as seen in Table II, the RMS for 28 peaks is six times larger than for 29 peaks, i.e. 0.00157 verses 0.000251, respectively.

This behavior is also seen in the fits of 25 peaks verses 24 peaks for trypsin, 24 peaks verses 23 peaks for elastase and 17 peaks verses 16 peaks for myoglobin. In all cases, the fits using one less peak in the range of 1659 cm^{-1} are extremely poor, with some extremely broad peaks in the amide I region which in turn influences the fit of the amide II region. Also the RMS of the poor fits are on the order of a factor of 3 to 6 times higher than the best fits (see Table I). It should be noted, that FD causes the component bands to have almost equal half-width at half-heights. As described by Byler and Susi (3), the appearance of broad component bands in the results of nonlinear regression analysis was considered unacceptable. Despite the improved fits, the number of peaks which exist under the amide I or II envelopes should be based upon more theoretical concepts to prove the above hypothesis.

Table I. Influence of number of Gaussian peaks on RMS

Protein	N	RMS	N ₂	RMS ₂
Lysozyme	29	0.000251	28	0.00157
Trypsin	25	0.000362	24	0.00140
Elastase	24	0.000624	23	0.00174
Myoglobin	17	0.000306	16	0.00154

Table II. % Extended content of globular proteins

	FTIR	Ps	R	BS	XBS
Hemoglobin	7.7	24.5	8.2	25	0
Myoglobin	10.4	10.5	5.9	24	0
Cytochrome C	10.5	19.4	11.6	34	10
Lysozyme	29.6	25.6	30.5	21	19,16
Ribonuclease	41.3	23.4	37.5	50	46,40
Papain	22.2	39.6	19.8	32	29
PTI	31.8	25.9	29.5	52	50
α -Chymotrypsin	38.8	34.2	37.9	51	49,34
Trypsin	40.6	38.1	39.6	55	56
Elastase	36.6	43.8	33.6	45	47,52
Carbonic Anhydrase	41.7	26.7	36.0	49	45,40
β -Lactoglobulin	44.5	18.5	44.2	50	-
CON A	44.2	32.9	44.5	60	60,51
Oxytocin	0	0	0	-	-

FTIR = Average error $\pm 1.6 \text{ cm}^{-1}$; **Ps** = Modeling predicted % sheet structure (26); **R** = Ramachandran; **BS**, **XBS** = Byler, D.M. and Susi, H. (4), Data and X-ray respectively.

In recent articles by Torii and Tasumi (13-15), the theoretical FTIR amide I spectrum was calculated using the three dimensional structure of lysozyme from X-ray crystallography and a Gaussian envelope of each peptide oscillator with a half-width at half-height of 3.0 cm^{-1} . Their calculations were used to qualitatively compare theoretical spectra of several proteins with their experimental counterparts in D_2O , so the force constants used were optimized to agree with D_2O and not H_2O results. For this reason we cannot directly compare our experimental results for lysozyme with their theoretical spectrum. We can, however, deconvolve their spectrum for lysozyme into the component Gaussian peaks using nonlinear regression analysis and compare the number of peaks with our experimental FD spectrum. Here, it must be emphasized that the theoretical spectrum must contain at least but not less than the same number of bands in our experimentally analyzed spectrum.

Figure 6A shows deconvolved theoretical FTIR spectra from Torii and Tasumi (13) with the best fit of the sum of 14 Gaussian bands. Attempts to use less resulted in poor fits while the addition of more peaks caused the height of the extra bands to approach a zero or negative value. The experimental FD FTIR spectrum of lysozyme using an REF of 3.8 (a value considered much too large by most investigators) is shown in Figure 6B. Here, the amide I region is fit to the sum of 14 Gaussian peaks with success. A low RMS value (to within 0.1%) and a pseudo random deviation pattern was obtained with the 14 band fit. No additional bands could be successfully added to the 14 band fit. In addition, none of these 14 bands had extremely broad half-widths at half-height just as in the calculated bands of Figure 1 which used only 10 bands for the amide I region (comparing same region, 1690-1620 cm^{-1} only 10 component bands).

Attempts to fit the 14 bands to experimental data using a universally more accepted REF value of 2.5 ended with some bands, especially those with the highest and lowest frequency, to become unacceptably broad (see Figure 6C).

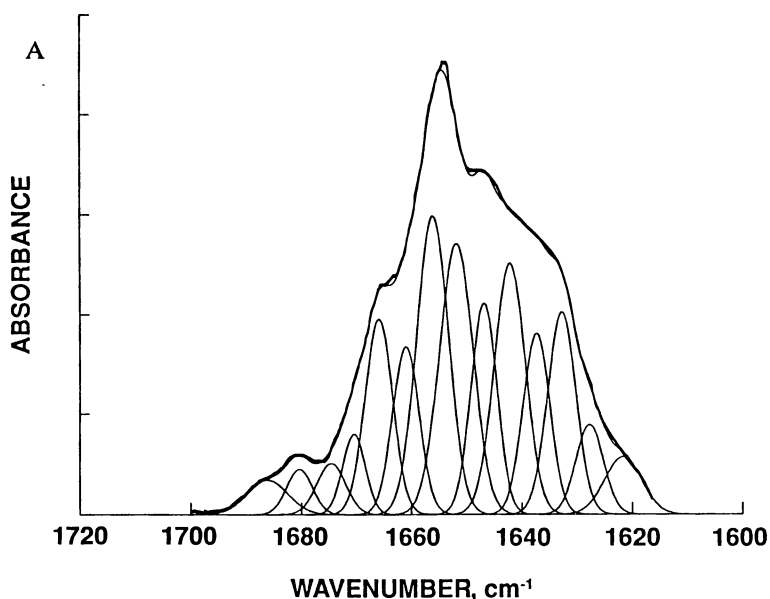


Figure 6. Best fits by non-linear regression analysis using amide I band of lysozyme. A: from theoretical calculations of Torii and Tasumi (13-15). B: Fourier deconvoluted lysozyme spectra using a resolution enhancement factor of 3.8. C: same as 6B with a factor of 2.5. *Continued on next page.*

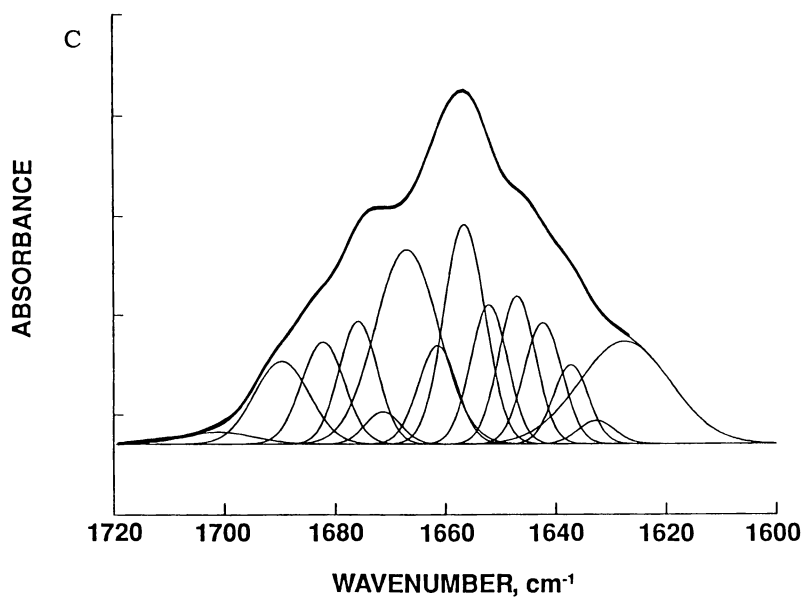
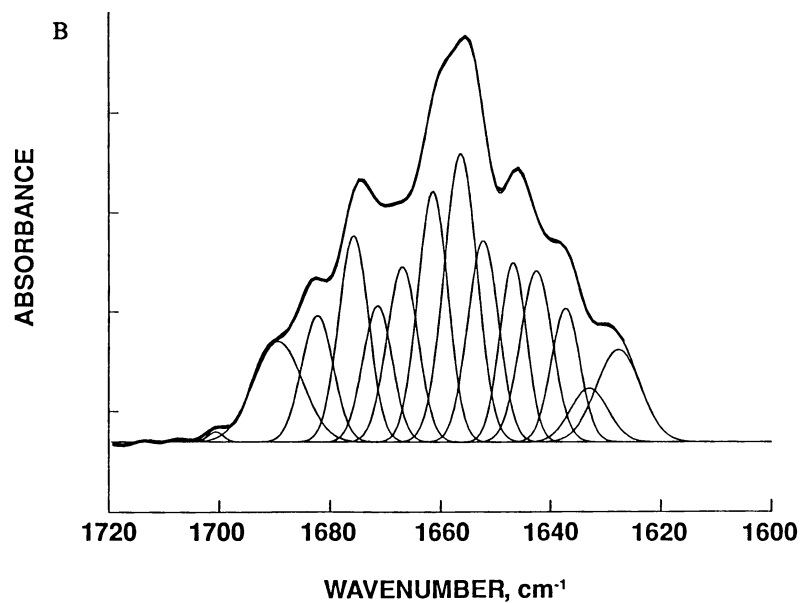


Figure 6. Continued.

However, if the amide II band is fit simultaneously with the amide I (using REF of 2.5 and fitting 29 bands), inordinately large values of the half-width at half-height are not present (see Figure 2). Hence, it is extremely important to simultaneously deconvolve the amide I and II envelopes if the FD spectra is generated using lower REF values (such as 2-3). Calculations excluding the amide II envelope may lead to large errors in the estimated secondary structures assigned to a protein.

Even though the calculated spectra of Torii and Tasumi are based on D₂O force constants, we will attempt to compare the 2° structure calculated from the theoretical with the experimental spectra for all the reported proteins in a separate paper. It also, should be noted that the experimental spectra analyzed with 29 peaks (Figure 2), only has 10 peaks in the amide I region. The four extra peaks determined from the theoretical spectra are well within the strand and turn region and could easily be summed to obtain the total turn and strand structure.

Discussion

Contribution of ASN and GLN Side Chain. Recently, Venyaminov and Kalnin (16) performed FTIR experiments on amino acids in water and deconvolved the spectra into component bands to ascertain the influence of side chains on the overall amide I envelope. They found that large bands due to the side chains of asparagine (ASN) and glutamine (GLN) exist within the amide I envelope. In subsequent articles (12,17), they attempted to eliminate all side chain contributions to the amide I and II by subtracting, on a molar basis, the amino acids from the amide I and II envelopes of proteins. They assumed the absorptivity of amino acids and proteins were equivalent, and they found a band at 1668 cm⁻¹ which is invariant to changes in environmental conditions for both ASN and GLN. They have assigned this band as the C=O stretch of the GLN and ASN side chains. In addition, they show in their figures a small broad band at 1650 cm⁻¹ in all of their GLN and ASN studies, which also maintained the same frequency as the solvent or pH were varied. Since this band also exists in their side chain analysis of arginine, we expect that this band may be a C-N deformation of GLN and ASN. Now we will attempt to ascertain if any of the amide I component Gaussian bands calculated for the FTIR of the 14 proteins correlate with the amount of GLN and ASN present in these proteins.

Figure 7 is a plot of the area % of the 1651 and 1667 cm⁻¹ bands verses the % GLN and ASN for the proteins listed in Table II. The % areas of the 1651 cm⁻¹ for the first three predominately helical proteins, i.e. hemoglobin, myoglobin and cytochrome C, are eliminated from this analysis. The best fit line and the 95% confidence lines (dashed lines) were calculated and are also shown in Figure 7. The results of the linear regression analysis yields an intercept value of 2.83 (S.E. = 2.06, σ = 0.184) and a slope of 0.948 (S.E. = 0.292, σ = 0.00388). Thus, it appears the intercept value could statistically

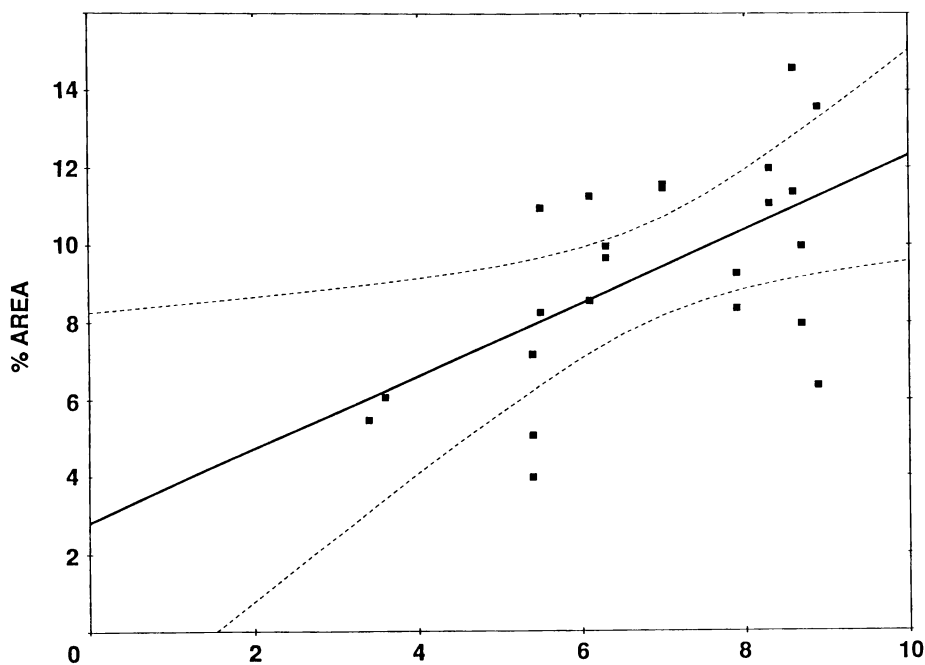


Figure 7. Linear regression analysis for area percent of 1667 and 1651 cm^{-1} bands versus present GLN and ASN in Protein for the 14 proteins listed in Tables II-IV. Filled squares: data points; dashed lines: confidence curves at level 0.95; double line: best straight line using linear regression analysis with no weighted points.

have a zero value, and the slope a value of near unity. While the analysis suggests a correlation, it should be viewed as an assumption. Only when the analysis of at least 50 proteins yields similar results would this assumption be considered proven.

For this report, we will assign the 1667 and 1651 cm^{-1} bands to GLN and ASN side chain modes. Most likely 1651 cm^{-1} is a C-N deformation and 1667 cm^{-1} is a C=O stretch. The true fraction of GLN and ASN should be subtracted from the fractional areas of these bands and any excess area arising should be assigned to the appropriate global secondary structure. If the percentages of areas are less than the percent of GLN and ASN residues present, then the experimental areas should be subtracted from the amide I envelope and all the remaining bands should be renormalized to unity.

Ramachandan Analysis. Of paramount importance to obtaining the global secondary structure of proteins (by correctly interpreting the assignments of the FTIR amide I bands) is the correct calculation of the secondary structure from the results of X-ray crystallography. Not only the amount of α -helix, turn and extended conformation is important, but the length of the helix and extended conformation as well as whether internal backbone hydrogen bonding exists may be relevant descriptors for correlation with the % areas of the component bands in the amide I region. Until recently, researchers in this field used the values provided in the Brookhaven Protein Data Bank. The values depended on the definitions of conformation adopted by each crystallographer. The definitions can, also, change over a period of time. What is needed is an algorithm consistent with FTIR results to be used on all X-ray crystallographic structures. To date, no consensus in the scientific community for the appropriate algorithm has been found.

Kalnin et al. (12) has recently subdivided both the helices and sheets into hydrogen bonded and non hydrogen bonded conformations, which along with the turn and all other conformations, forms a basis of six instead of four conformations. Their calculations, however, are correlated with FTIR results using factor analysis instead of nonlinear regression analysis. In their analysis normalized structures (i.e. the total conformation of an individual protein adds up to 100%) show reasonable correlation with FTIR results. The same good correlations were not obtained by the use of factorial analysis (11,18) in which structure normalization was ignored and only four conformations were considered.

An algorithm developed recently in Liebman's laboratory (19,20) shows good agreement with FTIR results when deconvolving the amide I into component bands (19,20). However, these experiments were performed only for serine proteases in D_2O . While this method appears to be promising it is still in its infancy, more correlations between experimental (in H_2O) and calculated conformations must be reported.

In this study, we use the traditional Ramachandran plot calculated from the X-ray crystallographic structure of proteins in conjunction with the secondary conformations reported in the Protein Data Banks. We strongly stress that the choice of Ramachandran analysis does not in any way imply that the transitional dipolar coupling mechanism of Krimm et al. (21,22) and Torii et al. (13-15) may not apply to the vibrational spectroscopy of proteins. We believe that this mechanism is correct.

Here, we use the Ramachandran plot only as a method for correlating reported fractional values. If discrepancies exist we shall use other molecular modeling techniques available in the Sybyl Molecular Modeling software (such as inspection by ribbon, hydrogen bonding) in conjunction with the Ramachandran analysis. The latter analysis, of course, does not take into account the required minimum number of sequential residues to sustain a periodic conformation. In Figures 8A and 8B the Ramachandran plots are given for myoglobin and lysozyme, respectively. Figure 8A shows a template Ramachandran with theoretical curves for predominately helical proteins. The dashed lines in the upper left are for a β sheet structure. The solid circle next to it is defined as the second position for a β turn type II. Directly under this circle is the theoretical envelope for a one residue inverse "a" (γ) turn. Directly under the inverse "a" turn is the area which defines the 3/10 helix, the double line area represents the α -helix region. At the lower right hand region is an area representative of a type II β turn, and above it is the area common for an "a" (γ) turn.

Figure 8B uses a template plot which is more consistent with proteins containing a large amount of turn conformation. Here the dashed lines and double lines represent the sheet and α -helix region, respectively, as in Figure 8A. Adjacent to the sheet region in the second position for the type II B turn and directly under that region is the envelope for the one center inverse "a" turn. The third position for a type I and II β turn region is below the inverse "a" turn. Directly above the double lined acceptable region for an α -helix is the envelope for the second position of the type III β turn. In the lower right hand quadrant and proceeding upward are the acceptable regions for the conformations of the second positive of "a" type II β turn, an "a" turn, the third position of type I and II β turn, and the second position of type III β turn, respectively. Shown in Figure 8B (as "+") are the calculated phi, phi results from the Sybyl modeling program for the X-ray crystallographic structure of lysozyme (6LYZ).

A three dimensional Ramachandran plot, where the residue number is plotted on the third axis, can also be calculated using the Sybyl molecular modeling software. With this plot, adjacent residues within a periodic structure as well as those residues which are part of a turn conformation can easily be determined. With all the above analyses the global secondary structure calculated from the crystallographic data for each of the proteins studied in the FTIR were calculated, the results along with the corresponding experimental values are presented in Tables II, III, and IV for strand, helix, and turn plus irregular, respectively. The results for each conformation will be discussed in separate sections for the remainder of this report.

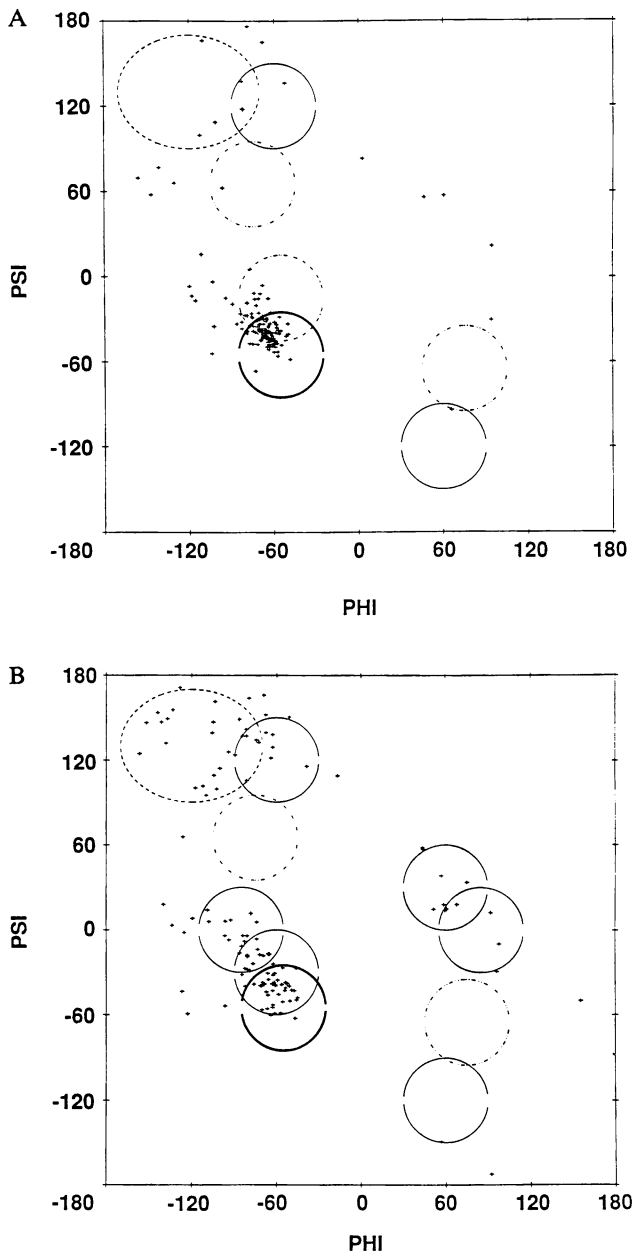


Figure 8. Ramachandran plot from X-ray crystallographic structure of A: myoglobin and B: lysozyme.

Table III. % Helix content of globular proteins

	α				3/10, Bent Strand	
	FTIR	P _H	R	BS	FTIR	R
Hemoglobin	76.7	49.7	81.6	74	1.2	-
Myoglobin	76.4	73.9	78.1	76	1.9	-
Cytochrome C	46.2	36.9	44.6	51	2.2	4.0 BE
Lysozyme	25.6	20.9	27.1	41	4.2	5.5 BE
Ribonuclease	10.0	29.8	18.0	21	10.3	8.8
Papain	15.2	11.3	18.8	27	17.4	15.6
PTI	4.2	8.6	12.0	10	0	6.9 BE
α - Chymotrypsin	11.9	16.7	12.4	12	3.5	6.8
Trypsin	17.4	12.1	10.4	16	0	8.3
Elastase	14.1	10.0	9.7		0	8.2
Carbonic Anhydrase	14.2	22.4	14.2	13	1.7	7.1
β -Lactoglobulin	15.8	63.0	5.0	10	3.1	9.2 BE
CON A	13.6	16.9	1.7	-	2.2	8.8 BE
Oxytocin	0	0	0	-	-	-

FTIR = Average error α helix ± 0.8 cm⁻¹; P_H = Modeling predicted % α -helix (26); R = Ramachandran; BE = Bent strand; BS = Byler, D.M. and Susi, H. (4).

Table IV. % Non-periodic content of globular proteins

	Turn or Twisted Strand			Irregular		
	FTIR	P _T	R	FTIR	P _I	R
Hemoglobin	4.3	7.5	7.5	10.1	18.4	10.8
Myoglobin	2.3	6.5	2.1	8.8	9.2	14.6
Cytochrome C	37.9	20.4	35.8	3.1	23.3	3.1
Lysozyme	28.1	42.6	26.6	12.5	10.9	9.6
Ribonuclease	24.2	40.3	26.0	14.2	6.5	9.7
Papain	43.3	35.4	39.6	2.0	13.7	6.2
PTI	60.5	56.9	49.6	3.5	8.6	8.9
α -Chymotrypsin	25.2	31.1	22.8	20.6	18.0	20.1
Trypsin	26.3	34.5	28.8	15.7	15.2	16.7
Elastase	31.8	26.7	33.6	17.4	19.6	16.0
Carbonic Anhydrase	22.3	22.4	26.0	20.1	28.6	16.3
β -Lactoglobulin	17.3	18.7	21.2	19.3	14.8	20.4
CON A	22.2	17.3	22.7	17.7	32.9	22.3
Oxytocin	100.0	100.0	100.0	0	0	0

FTIR = Average error: turn or twisted strand $\pm 1.4 \text{ cm}^{-1}$, loop $\pm 0.8 \text{ cm}^{-1}$;
P_T = Modeling predicted % turn or twisted strand (26); **R** = Ramachandran;
P_I = Modeling predicted % irregular.

Extend (Strand) Content. Table II shows a comparison between the experimentally determined global 2° structure of the extended conformation (column 2, heading FTIR) along with values calculated by the composite Ramachandran analysis (column 4, heading R). The amount of extended conformation was determined from FTIR amide I results by summing the component bands from 1638 to 1623 cm^{-1} . Bands at frequencies lower than 1623 cm^{-1} were closer to 1615 cm^{-1} and were presumed to be caused by unprotected side chain carboxyl groups (21). Note that we assumed the bands (1638-1623 cm^{-1}) quantitate not only the amount of sheet conformation but also extended or strand non hydrogen bonded structures as did Prestrelski et al. (18-19). For proper comparison, we summed all points in the upper left hand quadrant of the Ramachandran plot for phi values above 130°, as well as those

within the sheet region. Connectivity for any extended structure was determined using three dimensional Ramachandran (not shown).

For comparison, the fifth and six columns of Table II, labeled BS and XBS are the experimentally determined (FTIR) and calculated (X-ray crystallographic) conformation results of Byler and Susi (3) where the FTIR results were determined in D₂O.

As seen in Table II, comparison between the experimental (FTIR) and calculated (R) extended conformation is excellent. Only in myoglobin, a high helical protein, is the deviation greater than 4%. In fact, the average deviation between the experimental (FTIR) and calculated (R) strand structure for these 14 proteins is 1.8%. This value is lower than we had hoped, since Mantsch (23) had recently reported that FTIR cannot determine the absolute value of a conformation to within 2% and the average deviation for the 12 proteins in the Byler and Susi report (BS vs R) is 9.2%. Inspection of their results in Table II shows that in almost all instances their values are much higher than those in this study. The discrepancy could either be due to the use of D₂O which will not exchange 100% of all the protein protons and may cause an increase in hydrophobic interactions (24), or to the use of two few bands which leads to serious mis-assignments.

Next, we compared our FTIR results for % extended in trypsin, α -chymotrypsin and elastase with those reported by Liebman's group (19-20). We obtained FTIR values of 40.6, 38.8 and 36.6% for trypsin, α -chymotrypsin and elastase, respectively. Liebman reported FTIR values (determined in D₂O) of 39, 45 and 46, respectively, and calculated values (using their own algorithms) of 42, 42 and 47, respectively. While all experimental values for trypsin and α -chymotrypsin agree equally with their calculated values, our elastase value of 36.6% agrees far better with their calculated value of 37 than their experimental (in D₂O) value of 46%. This adds further support for our methodology (spectra in H₂O) and the supposition that the amount of GLN and ASN side chain residues must be subtracted for the 1667 and 1651 cm⁻¹ component amide I bands.

Shown in column 3 of Table II with a heading of P_S is the predicted amount of sheet structure using a secondary structure sequence based prediction algorithm by Garnier et al. (25). Other algorithms were attempted but this method yielded the most comparable results. In Tables III and IV, the P_H, P_T, and P_I for the amount of α -helix, turn and irregular conformation calculated by this algorithm are also presented.

Helical Content. Table III lists the calculated and experimental results for the α -helix (column 1-3), 3/10 helix (column 5, 6), and those reported by Byler and Susi (3) — BS (column 4). Here the calculated average deviation between our experimental (from the 1659 and 1651 cm⁻¹ bands) and calculated % helix is twice as high (3.6%) as for the extended structure (1.8%), but is still

acceptable. Close inspection of these differences may provide a rational for the change.

Not counting the high helical proteins (hemoglobin and myoglobin) we observe that the largest differences occur for ribonuclease, the serine proteases (i.e. Pancreatic Trypsin Inhibitor (PTI), trypsin, and α -chymotrypsin), concanavalin A (CONA) and β -lactoglobulin. The last two proteins in this series contain significant amounts of β -barrel structures. Such structures appear as antiparallel β -sheets which are highly bent. The Ramachandran plots also yield points in the lower left quadrant which is normally considered a forbidden region. The Ramachandran plots for the serine proteases all contain some phi, psi angles in the region. Hence, the discrepancy in the experimental and helical constant for concanavalin A, β -lactoglobulin and perhaps the serine proteases may be caused by β -barrel which could have bands at $1658 \pm 2 \text{ cm}^{-1}$. More experimental and theoretical studies must be performed before this hypothesis can be concluded.

But ribonuclease contains no β -barrel and inspection of the work by Kalnin, et al. (12) may provide an answer. In this study, the α -helix was subdivided into an ordered and unordered class as was the sheet structure. They calculate values of 13% and 10% for their ordered and unordered α -helix conformation and find experimental values of 11% and 8% respectively. Upon inspection of the ribbon structure, a major distortion of the helical region of ribonuclease is found. In addition, a value of 10.3% has been obtained from the excess area of the 1667 cm^{-1} which we have assigned as a 3/10 helix, in accordance with the results of Krimm and Bandekar (21). If the α -helix and 3/10 helix values are summed they add up to more acceptable values. However, the Ramachandran plot for ribonuclease shows 11 residues within the type III or 3/10 helix region which calculates to a theoretical value of 8.8% for these possible conformations. It should also be stressed that Kalnin et al. (12) calculates an ordered α -helix conformation of 27% for lysozyme which agrees well with our experimental and theoretical values of 25.6% and 27.1%. Therefore, we believe that the discrepancy for the ribonuclease helical structure is a result of its structure which our Ramachandran analysis could not adequately calculate.

It should be stated at this time that the 1676 cm^{-1} band was also summed along with the excess area for the 1651 cm^{-1} as well as the 1658 cm^{-1} for obtaining the total α -helix content of hemoglobin and myoglobin. The 1676 cm^{-1} band represented approximately 17% of the total helical structure. The areas of the 1658 cm^{-1} and 1651 cm^{-1} bands summed to 63%. A value of 63% was also calculated as the amount of unordered helix content in myoglobin by Kalnin et al. (12). Moreover close inspection of the ribboned structures of hemoglobin and myoglobin reveal highly distorted helical segments which could not be observed using Ramachandran analysis. The 1676 cm^{-1} band, assigned by Krimm and Bandekar (21) as a turn may be reflective of a type III β -turn. Such a turn would have phi, psi values overlapping the α -helical region of a Ramachandran plot (see Figure 8B). Nevertheless, for high helical proteins

(i.e. above 55%) it may be more prudent for investigators to utilize UV circular dichroism analysis. Because as Torii and Tasumi (13-15) have recently reported, a serious overlap of E and A bands for α -helices with varying lengths occurs, thus resulting in theoretical amide I envelopes which contain bands well below the 1650 cm^{-1} region. But with lower helical proteins, FTIR correlates much better than circular dichroism since the turn conformation can be more easily determined.

Finally, the excess areas in the 1667 cm^{-1} band above the GLN and ASN side chain contribution is shown in Table III as a 3/10 helix or bent strand i.e. a possible "a" (γ) turn. These small values cannot be easily correlated with Ramachandran analysis and no firm assignments are made. However, we do not observe any large amounts of 3/10 helix in lysozyme especially in the 1638 cm^{-1} region where Prestrelski et al. (20,21) concluded that such 3/10 helical bands exist. While we have not performed any experiments on α -lactalbumin, we still concur with the assignment of Krimm and Bandekar (21) that the 3/10 helix is in the range of 1665 cm^{-1} rather than in the low range of $1638 - 1640\text{ cm}^{-1}$ as reported by Prestrelski et al. (19-20). Perhaps this discrepancy can be explained by the fact that their experiments were performed in D_2O rather than H_2O .

Turn and Irregular Content. Table IV shows the turn and irregular content determined experimentally from analysis of the FTIR amide I band and calculated from the three dimensional Ramachandran analysis of X-crystallographic structure of the 14 listed proteins. The turn content was determined from the sum of all amide I bands from 1670 cm^{-1} to 1694 cm^{-1} . The irregular content was calculated from the normalized area of the $1646 \pm 2\text{ cm}^{-1}$ band. The irregular theoretical structure was calculated as all other structure not defined by this analysis. Good agreements between the experimental and theoretical values were observed with average calculated deviations of 2.9% and 2.6% for the turn and irregular content, respectively. These values are well within the deviations observed in the strand and α -helix content i.e. 1.8% and 3.6%. However, it should be noted that in the case of cytochrome C and β -lactoglobulin, phi and psi values exist in the upper right hand region of the Ramachandran plot. Although this region has been considered forbidden, closer inspection shows that these phi and psi values are the result of twisted sheets. Since no other proteins in this database exhibited phi and psi values in this region, we have concluded that the 1676 cm^{-1} may also be assigned to a twisted strand. However, more studies must be performed before a definite assignment can be made.

Conclusions

Calculation of the component 2° structural elements of the vibrational bands, i.e., approximately 25 Gaussian bands, was accomplished by fitting both the

Table V. Secondary structure assignments [this study]

1681 - 1695 cm ⁻¹	Turn I, I'
1673 - 1679 cm ⁻¹	Turn II, II', III, III', twisted sheet
1667 - 1669 cm ⁻¹	GLN (C=O) & ASN (C-O) side chain, 3/10 helix, bent strand
1657 - 1661 cm ⁻¹	α -Helix (A Band)
1651 - 1653 cm ⁻¹	GLN (N-H) & ASN (C-N) side chain, α -helix (E Band)
1643 - 1648 cm ⁻¹	Disordered, irregular, gentle loop
1622 - 1638 cm ⁻¹	Extended strand, rippled and pleated sheets

amide I and II bands using nonlinear regression analysis of: the Fourier deconvoluted spectra, the second derivative spectra, and the original spectrum. Fixed frequencies used in the original spectra analysis were obtained from both the FD spectra and 2nd derivative analyses. The criterion for acceptance of any analysis was that the fractional areas calculated from all three methods were in agreement. Results clearly show that 2° structural conformations determined in water were in better agreement with global 2° structure analysis of X-ray structures than the previously reported values determined in D₂O. Also with H₂O the types of turns can be correlated with the X-ray structure, and 2° structure elements can be calculated from the amide II band to be used for validation of amide I assignments. In addition, resolution of amide I spectra in H₂O is greater than that in D₂O. The deterioration of resolution of FTIR spectra in D₂O results primarily from incomplete exchange of protein protons to deuterons. The results lay the foundation for the study of conformational changes in proteins induced by ligands, cosolutes or perhaps structural changes from site directed mutagenesis (9).

In this study, we have presented a method for analyzing the FTIR of proteins in water and determining their global 2° structure. Analysis of 14 proteins whose X-ray crystallographic structures are known, showed agreement between experimental and theoretical 2° structure content to within 4%. The bands which are assigned to these structures are shown in Table 5 along with their tentative structural assignments. While the results are excellent, it must be stressed that only after a database of at least 50 proteins is obtained, can any definite conclusions be reached. It is hoped that this study will inspire others investigators to adopt this methodology and add more information to increase this database above 14 proteins. In this laboratory, we too will continue to add to this database.

Acknowledgments

Reference to a brand or firm name does not constitute an endorsement by the U.S. Department of Agriculture over others of a similar nature not mentioned.

Literature Cited

1. Townend, R.; Kumosinski, T. F.; Timasheff, S. N. *J. Biol. Chem.*, **1967**, *242*, 4538-4545.
2. Kauppinen, J. K.; Moffatt, D. J.; Mantsch, H. H., Cameron, D. G. *Appl. Spec.*, **1981**, *35*, 271-276.
3. Byler, D. M.; Susi, H. *Biopolymers*, **1986**, *25*, 469-487.
4. Susi, H.; Byler, D. M. *Methods. Enzymol.*, **1986**, *130*, 290-311.
5. Byler, D. M.; Farrell, Jr., H. M. *J. Dairy Science*, **1989**, *72*, 1719-1723.
6. Birke, S. S.; Dien, M. *Biochemistry*, **1992**, *31*, 450-455.
7. Cantor, C. R.; Schimmel, P. R. In *Biophysical Chemistry Part III: Techniques for the Study of Biological Structure and Function*, Freeman, W. H., Ed., **1980**, pp. 687-791.
8. Dong, A.; Huang, P.; Caughey, W. S. *Biochemistry*, **1990**, *29*, 3303-3308.
9. Dousseau, F.; Pezolet, M. *Biochemistry*, **1990**, *29*, 8771.-8779.
10. Pancoska, P.; Yasui, S. C.; Keiderling, T. A. *Biochemistry*, **1991**, *30*, 5089-5103.
11. Pancoska, P.; Keiderling, T. A. *Biochemistry*, **1991**, *30*, 6885-6895.
12. Kalnin, N. N.; Baikalov, I. A.; Venyaminov, S. Y. *Biopolymers* **1990**, *30*, 1273-1280.
13. Torii, H.; Tasumi, M. *J. Chem. Phys.*, **1992**, *96*, 3379-3387.
14. Torii, H.; Tasumi, M. *J. Chem. Phys.*, **1992**, *97*, 86-91.
15. Torii, H.; Tasumi, M. *J. Chem. Phys.*, **1992**, *97*, 92-98.
16. Venyaminov, S. Y.; Kalnin, N. N. *Biopolymers*, **1990**, *30*, 1243-1258.
17. Venyaminov, S. Y.; Kalnin, N. N. *Biopolymers*, **1990**, *30*, 1259-1271.
18. Pancoska, P.; Wang, L.; Keiderling, T. A. *Protein Science*, **1993**, *2*, (in press).
19. Prestrelski, S. J.; Williams, A. L.; Liebman, M. N. *Proteins, Structure, Function and Genetics*, **1992**, *14*, 430-439.
20. Prestrelski, S. J.; Byler, D. M.; Liebman, M. N. *Proteins, Structure, Function and Genetics*, **1992**, *14*, 440-450.
21. Krimm, S.; Bandekar, J. *Adv. Protein Chem.*, **1986**, *38*, 181-364.
22. Krimm, S.; Bandekar, J. *Biopolymers*, **1980**, *19*, 1-29.
23. Surewicz, W. K., and Mantasch, H. H., and Chapman, D. *Biochemistry*, **1993**, *32*, 389-394.
24. Timasheff, S. N. In *Protides of the Biological Fluids, 20th Colloquium*, Peters, H. ed., **1973**, p. 511-519.
25. Garnier, J.; Osguthorpe, D.; Robson, B. *J. Mol. Biol.*, **1978**, *120*, 97-120.

RECEIVED June 8, 1994

Chapter 7

Molecular Modeling of Apolipoprotein A-I Using Template Derived from Crystal Structure of Apolipophorin III

Eleanor M. Brown, Thomas F. Kumosinski, and Harold M. Farrell, Jr.

Eastern Regional Research Center, Agricultural Research Service,
U.S. Department of Agriculture, 600 East Mermaid Lane,
Philadelphia, PA 19118

Plasma lipoprotein particles are the major vehicles for lipid transport in the circulatory systems of animals. Apolipoprotein A-I (apo A-I), the primary high density lipoprotein (HDL), serves as a cofactor in the esterification of cholesterol and mediates the transport of cholesterol esters to the liver for utilization. Structure-function studies of the apolipoproteins are hindered by the limited crystallographic data available. Apolipophorin III (apo Lp-III) of flying insects is responsible for delivery of lipids for utilization by flight muscles. The 2.5-Å resolution structure for apo Lp-III from the African migratory locust, *Locusta migratoria*, is used in the development of a template for the construction of partial models of canine, human and avian apo A-I. Residues 7 to 156 of apo Lp-III were aligned with residues 72 to 236 of apo A-I using alanine as a spacer residue. Four of the five helices of apo Lp-III were preserved in the apo A-I model. Helix 4, the longest of the original helices and the one with the most inserted residues, was separated into two shorter helices connected by a nonhelical strand. Amphipathic character was similar in all models. Electrostatic interactions were more important in the apo A-I models, resulting in increased stability of about 6 kcal/mol/residue.

Lipids serve as structural components of membranes and as energy depots in the cells of all living organisms. Thus, the transport of these sparingly water soluble molecules through the aqueous environment of an organism is of major biochemical importance. Discrete complexes of lipids and proteins, the lipoproteins of mammals and birds or their companion complexes, lipophorins of insects, have evolved as primary transport vehicles for lipids. Lipoprotein particles consist of a core of neutral lipids, stabilized by a surface monolayer of polar lipids complexed with one or more proteins. The protein components of lipoproteins, called apolipoproteins, are amphipathic in nature, having both hydrophobic and hydrophilic regions. Although the number of discrete apolipoproteins and their specific functions vary, comparable proteins are often found in different species. Apolipoproteins A-I (apo A-I) and apo B are respectively, the major protein components of high density (HDL) and very low density (VLDL) lipoproteins of mammals and birds. Most mammalian lipoprotein classes also contain

0097-6156/94/0576-0100\$08.00/0
© 1994 American Chemical Society

apolipoprotein E (apo E), a protein that functions with apo B in directing the delivery and redistribution of lipids to cells that express receptors (1,2).

Research over the past 30 years has resulted in complete amino acid sequences and a wealth of detail concerning the biosynthesis, physical characteristics and metabolism of apolipoproteins. Acquisition of structural data at the molecular level has, however, been hindered by the noncrystalline character of most apolipoproteins. This lack of crystallographic data inspired the development of predictive techniques for approximating suitable structures. Correlation of global secondary structures determined spectrophotometrically with local secondary structures predicted from amino acid sequences have been used to suggest structural motifs. Spectrophotometric studies, mainly circular dichroism, show high levels of helical structure for apolipoproteins in aqueous solution and still higher levels when lipids are included in the system (3-5). Through analysis of apolipoprotein primary structure, Boguski and coworkers (6) established a pattern of repeating sequences punctuated by highly conserved proline residues for this family of highly homologous proteins (6-10). Algorithms for the prediction of secondary structure in globular proteins (11,12) and for the analysis of hydrophathy patterns (13,14) were applied to the development of models for the apolipoproteins (15,16). A high potential for the formation of amphipathic helices, characterized by opposing hydrophobic and hydrophilic faces, has been established as a common motif (6,17-19).

Mammalian apo A-I is synthesized in the liver and the intestine (20). Among its functions are the activation of lecithin:cholesterol acyltransferase (LCAT) and the transport of cholesterol from peripheral tissues to the liver for metabolism (21). Synthesis of avian apo A-I occurs in peripheral tissues as well as in the liver, particularly at the time of hatching (22). In addition to LCAT activation it is thought to function in the mobilization of yolk lipids (22). The biochemistry and physiology of insects resemble in a general fashion the corresponding metabolic pathways of vertebrates (23). Two exceptions are that the insect has an open circulatory system in which the hemolymph is enclosed by membranes and a specialized tissue, the fat body, that combines many of the functions of the vertebrate liver and adipose tissue. The majority of hemolymph lipids are found in a single lipoprotein particle, called lipophorin, that is synthesized in the fat body and circulates in the hemolymph. All lipophorins have been observed to have at least two apolipophorins (apo Lp-I and apo Lp-II). A third apolipophorin (apo Lp-III) that functions in the transport of lipid through the hemolymph from storage depots to flight muscles during prolonged flight is found mainly in insects that use lipids to fuel flight (23).

Crystal structures at 2.5-Å resolution have now been published for the 18-kDa apo Lp-III from the African migratory locust, *Locusta migratoria* (24) and the 22-kDa N-terminal, receptor-binding domain of human apo E (25). Despite the functional differences of apo E and LpIII, the crystal structures are remarkably similar. Nolte and Atkinson (26) used the data from these crystal structures in an integrated approach that included statistical analyses, information theory and sequence homology to predict secondary structure and possible tertiary folding patterns for apo A-I and apo E-3 in lipid environments. Delivery of lipids from storage depots to the site of metabolism, the common functional characteristic of apo Lp-III and apo A-I, would argue for a degree of structural similarity. In this study, computer-based molecular modeling has been applied to the search for common structural elements in these functionally related proteins.

Methods

Sequence Comparison. Amino acid sequences for apo Lp-III, human apo A-I (HuA-I), canine apo A-I (DgA-I) and chicken apo A-I (ChA-I) are available in the Protein Identification Resource (PIR) (27). The overall similarity of apo Lp-III to apo

A-I and the amphipathic helical potential of both was established previously (28). The positions of residues 7 to 156 of apo Lp-III were resolved in the crystal structure of this protein (24). IALIGN, an interactive alignment program distributed with the PIR (27) was used to align residues (7 - 156) of apo Lp-III with each of the apo A-I proteins. The best alignment used one to four residue gaps to spread this 150-residue segment of apo Lp-III over a 165-residue segment of apo A-I.

Segments of apo Lp-III that were identified as helical from the crystal structure (24) were evaluated with the "strip of helix" template of Vazquez and coworkers (29) to determine amphipathic potential. The best fit for each helix was then applied to the corresponding segment of DgA-I, HuA-I and ChA-I using the convention of Brasseur (30) that classifies Pro as unique, Gly as equivalent to any other residue, Asn, Asp, Arg, Gln, Glu, His, Lys, Ser, Thr as hydrophilic residues and Ala, Ile, Leu, Met, Phe, Trp, Tyr, Val as hydrophobic residues.

Molecular Models. A three dimensional representation of apo Lp-III (residues 7 - 156) derived from a residue by residue reading of the positions of α -carbons in the published crystal structure (24) was constructed using molecular modeling software (SYBYL, v. 5.5) (31) on a SGI 4D/35 workstation (Silicon Graphics, Mountain View, CA). (Reference to a brand or firm name does not constitute endorsement by the U.S. Department of Agriculture over others of a similar nature not mentioned.) This model was refined by SYBYL-directed energy minimization using the Kollman United Atom force field (32,33). As a template for apo A-I, the model called apo Lp-IIIa was derived from the apo Lp-III structure by the insertion of alanine residues at each of the gap positions identified by IALIGN.

Because the DgA-I sequence most closely matched that of apo Lp-IIIa this model was constructed first. The SYBYL command "mutate residue" was used to substitute residues 72 - 236 of DgA-I into the apo Lp-IIIa sequence. The DgA-I model was then subjected to energy minimization by the SIMPLEX method (31), a rapid nonderivative calculation that moves one atom at a time to reduce any overlapping of atoms that may be present in the initial structure. Refinement by the SIMPLEX method was continued until no atom in the model had a calculated force on it greater than 1000 kcal/mol \AA . The conjugate gradient method was then used with the Kollman United Atom force field (32,33) to further refine the model. A distance dependent dielectric function was used in the force field to implicitly account for the dielectric screening due to solvent molecules (which are not explicitly included in the models). When the total energy for the model was reduced to -1000 kcal/mole, the partially refined DgA-I model was used as a template for mutation to the HuA-I and ChA-I models.

Refinement of models was by SYBYL-directed energy minimization carried out *in vacuo* using a root-mean square (rms) force of 0.01 kcal/mole \AA as the cutoff value for ending the minimization. At the completion of this minimization step, models were subjected to a 300K molecular dynamics simulation for 1psec followed by an additional minimization step using the same limits as the initial minimization.

Conformational details of the energy minimized models were obtained by analysis of the dihedral angles ϕ and ψ . An α -helical segment was identified when both ϕ and ψ were between -30° and -75° for six or more consecutive residues. Single residues with dihedral angles outside these limits could be accommodated as distortions in longer helical segments. This definition of α -helical conformation was applied to the refined conformation of apo Lp-III (24) so that in this study the amino acids of apo Lp-III in helices 1 through 5 (H1 - H5) are residues 7 - 25 (H1), residues 35 - 59 (H2), residues 70 - 91 (H3), residues 95 - 121 (H4) and residues 136 - 155 (H5). Schiffer-Edmunson (34) amphipathic helical wheel projections (not shown) were used in the evaluation of the amphipathic character of the refined helical segments. The angle between residues along an α -helical backbone was assumed to be 100° (34). Circular plots from an end-on perspective of the helix were constructed, the magnitude

in degrees of each hydrophobic sector was measured, and the average hydrophobicity of this hydrophobic face was calculated using the Eisenberg normalized hydrophathy scale (13).

Results

Sequence Comparisons. Amino acid residues 7 - 156 of apo Lp-III are contained in five long α -helices connected by short loops (24). Figure 1 shows the alignment of apo Lp-IIIa with DgA-I, HuA-I and ChA-I. The best alignment produced by the program IALIGN for apo Lp-III with DgA-I, HuA-I and ChA-I used one to four-residue gaps to align residues 7 - 156 of apo Lp-III with residues 72 - 236 of DgA-I or ChA-I or residues 73 - 237 of HuA-I. A Pro-Pro repeat near the N-terminus of HuA-I causes this sequence to be shifted by one position relative to DgA-I or ChA-I (26). For simplicity, residues in the modeled apo A-I fragments will be referred to as though all were numbered 72 - 236. Although gaps are an accepted feature of sequence alignments, the model builder must find another way to space out the residues. Alanine was chosen as the spacer residue in building apo Lp-IIIa because it has a small, nonionic side chain lacking notable functionality except for a high probability of being found in helical structures. At 16 positions, the residues in Lp-IIIa, DgA-I, HuA-I and ChA-I are identical, 17 additional identities between DgA-I and apo Lp-IIIa as well as numerous functional identities are observed.

The amphipathic potentials of sequence fragments proposed for helical conformation may be found in Table I. Values of 1 or 0 were assigned for residues that did or did not fit, respectively, the "strip of helix" template (29). Average values for the 5 helices of apo Lp-III were optimized by varying the starting position for the "strip of helix" formalism (29). This optimum analysis was then applied to each of the other sequences. The highest hydrophobic potentials are calculated for sequences of residues having hydrophobic side chains at alternating third or fourth positions. The five helices of apo Lp-III with 19, 25, 22, 27, and 20 residues each are notably longer than those commonly found in globular proteins, and thus provide a useful test of the "strip of helix" template. The amphipathic potentials for the five apo Lp-III helices fall in a narrow range between 0.67 and 0.80. The insertion of alanine residues in an arbitrary fashion to fill the gap positions lengthened helices 3, 4, and 5 to 24, 33 and 21 residues, respectively, still within the range of helical lengths (19 to 35 residues) found in apo E (25). The additional alanine residues had relatively little effect on the amphipathic potentials (0.61 to 0.80). The range of amphipathic potentials was greater for the apo A-I sequences (0.45 to 0.88). However, all values for helices 1, 2, 3 and 5 were greater than 0.5, suggesting that each of these helices if stable would be suitable for a lipid-aqueous interface.

Table I. Amphipathic potentials of predicted helical segments in apolipoprotein models

Model	H1		H2		H3		H4		H5	
	#res	AP	#res	AP	#res	AP	#res	AP	#res	AP
Lp-III	19	0.79	25	0.80	22	0.77	27	0.70	21	0.67
Lp-IIIa	19	0.79	25	0.80	24	0.67	33	0.61	22	0.64
DgA-I	19	0.58	25	0.84	24	0.75	33	0.48	22	0.68
HuA-I	19	0.53	25	0.88	24	0.71	33	0.45	22	0.73
ChA-I	19	0.53	25	0.88	24	0.58	33	0.52	22	0.82

Amphipathic potentials (AP) are the best average value for a helical segment of (#res) residues measured with the "strip of helix" (29).

Molecular Modeling. Energy minimized models of apo Lp-III, apo Lp-IIIa and the apo A-I proteins are shown in Figure 2a and Color Plate 3. Apo Lp-III and Lp-IIIa are compared in Figure 2 to show the effects of alanine residues on the complete model (2a) and on individual helical segments H3, H4, and H5 (2b, 2c, and 2d). The template model apo Lp-IIIa and the apo A-I protein fragments were compared in terms of stability, amphipathic character and general appearance with the crystallographically determined structure of apo Lp-III (24). All models including apo Lp-III were subjected to the same refinement procedures. Values of the potential energy per residue for each helical segment and for each entire sequence fragment are listed in Table II. The apo Lp-III structure was slightly more stable with an energy of -22 kcal/mol/residue than apo Lp-IIIa (-20 kcal/mol/residue). Stabilization energies of the apo A-I models are essentially identical (-24 kcal/mol/residue) and greater than either apo Lp-IIIa or apo Lp-III. In Color Plate 3, lateral views of the energy minimized Lp-III, DgA-I, HuA-I and ChA-I emphasize the potential for electrostatic interactions, while end-on views emphasize the arrangement of hydrophobic sidechains.

Table II. Energetic evaluation of the refined models

	Lp-III	Lp-IIIa	DgA-I	HuA-I	ChA-I
Number of residues	150	165	165	165	165
Energy, kcal					
bond stretching	20.7	23.1	28.9	29.2	31.8
angle bending	123.4	154.0	189.6	208.5	212.2
torsional	198.1	235.7	308.3	324.6	340.9
out of plane	26.1	35.8	37.7	41.5	44.5
1-4 van der Waals	218.7	235.6	247.2	260.2	270.3
van der Waals	-993.0	-1094.3	-1074.8	-1131.8	-1155.6
1-4 electrostatics	1684.0	1813.3	1497.1	1566.7	1524.1
electrostatic	-4497.8	-4567.3	-5214.7	-5182.6	-5208.4
H-bond	-69.3	-70.2	-64.5	-69.8	-76.3
Total Energy, kcal	-3289.0	-3234.3	-4045.2	-3953.6	-4016.4
kcal/mol/residue	-21.9	-19.6	-24.5	-24.0	-24.3
H1 (residues)	(7-25)	—	(72-90)	(73-88)	(72-90)
kcal/mol/residue	-16.1		-18.5	-16.6	-17.4
H2 (residues)	(35-59)	—	(100-124)	(101-125)	(100-124)
kcal/mol/residue	-16.1		-20.1	-19.4	-19.2
H3 (residues)	(70-91)	(70-91)	(135-158)	(136-159)	(135-158)
kcal/mol/residue	-17.6	-20.7	-18.4	-16.2	-17.6
H4 (residues)	(96-121)	(96-121)	(166-196)	(167-197)	(166-196)
kcal/mol/residue	-15.5	-13.8	-16.7	-16.9	-17.6
H5 (residues)	(136-155)	(136-155)	(207-235)	(208-236)	(207-235)
kcal/mol/residue	-15.8	-15.4	-14.0	-13.4	-14.9

These energy calculations, based on the sequence segments initially assigned to H1 - H5 (see Figure 1), are given to illustrate the stability of the structures. Sequences for helices 3 - 5 (H3, H4, H5) of apo Lp-IIIa contain inserted alanine residues, see Methods.

NOTE: The color plates can be found in a color section in the center of this volume.

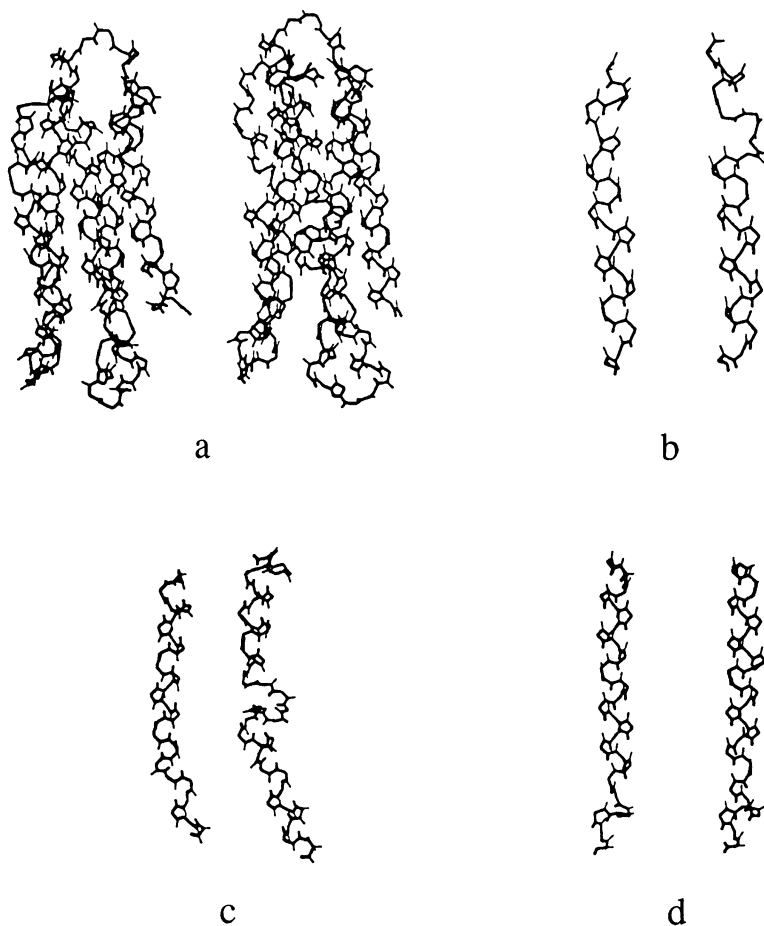


Figure 2. Energy refined models of apo Lp-III and the template apo Lp-IIIa constructed by inserting alanine residues into gap positions identified when the apo Lp-III sequence was aligned with canine, human and chicken apo A-I sequences. Backbone structures for the complete models are in panel (a). The effects of inserted alanine residues on H3, H4 and H5 are displayed in panels (b), (c) and (d) respectively. In each display the apo Lp-III is to the left of the apo Lp-IIIa.

Helix 1. With 19 residues, H1 was the shortest helical segment. The substitution of apo A-I sequences into this helix had less effect than might have been expected. The C-terminal residue of H1 was Glu in all of the models. A second Glu found in position four of apo Lp-III, DgA-I and HuA-I was replaced by Leu in ChA-I (Figure 1). Nevertheless, the helical character of H1 (residues 72 - 90) in both DgA-I and ChA-I withstood energy refinement to a stabilizing energy of -17 to -18 kcal/mol/residue (Table II). The amphipathic character of H1 was diminished as the hydrophobic sector was narrowed from 140° to 100° and the average hydrophobicity of this sector was reduced from 1.0 to 0.8 (Table III). Although the corresponding HuA-I sequence differed from DgA-I only in the replacement of Val by Gly at position nine, upon refinement this helix was shortened by three residues from the C-terminal end, the hydrophobic sector reduced to 60°, and the average hydrophobicity to 0.7.

Table III. Amphipathic analysis of energy refined helices

Model	Lp-III	Lp-IIIa	DgA-I	HuA-I	ChA-I
H1	N7-E25	N7-E25	D72-E90	D73-K88	E72-E90
Hb sector	140°	140°	100°	60°	100°
Av. Hb	0.98	0.98	0.78	0.70	0.81
H2	P35-S59	P35-S59	L100-E124	L101-E125	L100-E124
Hb sector	100°	100°	100°	100°	100°
Av. Hb	0.80	0.80	0.70	0.89	0.95
H3	S70-T91	S70-S85	Q137-R150	E136-R149	L135-L151
Hb sector	140°	120°	100°	100°	120°
Av. Hb	0.93	0.79	0.95	0.88	0.82
H4					
H4A	A96-S121	Q98-T107	D168-K181	Q172-E179	D167-R181
Hb sector	100°	140°	80°	100°	80°
Av. Hb	0.56	0.91	1.10	0.94	1.18
H4B	—	Q117-a121d	S187-S196	A190-A194	P186-V195
Hb sector	—	60°	100°	—	60°
Av. Hb	—	0.45	0.45	—	0.66
H5	E136-V156	L134-A155	L213-I232	L214-T237	E212-L235
Hb sector	120°	80°	100°	100°	100°
Av. Hb	0.74	0.53	1.05	0.96	0.86

The size of the hydrophobic (Hb) sector of a helix was determined from helical wheel projections (34). The average hydrophobicity (Av. Hb), of this sector was calculated by the method of Eisenberg (13). Single letter designations are used for amino acid residues, (a121d) designates the fourth alanine residue inserted between residues 121 and 122 of Lp-III.

Helix 2. The 25-residue H2 was marked by Asp in position two, Glu in position ten and Leu in the first position past the C-terminal residue in all models. The length of this helix and the size of the hydrophobic sector (100°) were conserved in all of the apo A-I models. The average hydrophobicity of the hydrophobic sector showed minor variation among the sequences but with values of 0.7 to 0.95, the comparison with a hydrophobicity of 0.8 for apo Lp-IIIa was generally favorable. As in H1, the apo A-I models of H2 were marginally more stable than apo Lp-III (Table II).

Helix 3. The first gap position in the sequence alignment required the insertion of two alanine residues between residues sixteen and seventeen of H3 in apo Lp-III. Although alanine residues were inserted into the sequence with α -helical conformational angles ($\phi, \psi = -58^\circ, -47^\circ$) (31), the result after refinement was a kink in the helix that essentially halted the progression of that helix (Figure 2b). Thus, the sequence contributing to H3 was increased in length from 22 to 24 residues by the addition of two alanine residues, while the helical structure was reduced in length to 16 residues. In the alignment of the 24-residue H3 sequence of apo Lp-IIIa with apo A-I two identities were observed, Leu in position six and His in position twenty. In the refined apo Lp-IIIa model this histidine residue was no longer in the helical structure. The shortened H3, with its associated nonhelical strand, was energetically slightly more stable than H3 of Lp-III (Table II). In refined apo A-I models H3 was further shortened to 14 or 15 residues by the removal of one or two residues from either end. At $100^\circ - 120^\circ$ the hydrophobic sectors of the helical wheel projections for apo Lp-IIIa and the apo A-I proteins were less than the 140° for apo Lp-III, but still within the range of apo Lp-III helices as were the average hydrophobicity values (0.8 - 0.9) (Table III). H3 in the apo A-I models was energetically more similar to apo Lp-III than to apo Lp-IIIa (Table II).

Helix 4. The most dramatic effects on the model structure were in the region of H4. With 27 residues, H4 was the longest of the native helices and coincidentally the only one for which our criteria and that of Breiter and coworkers (24) agreed exactly as to position and length. The alignment of apo Lp-III with the apo A-I sequences dictated the insertion of two alanine residues in the loop region between H3 and H4, six residues into the H4 sequence, and four more to the C-terminus of H4 (Figure 1). The aggregate effect of these insertions on H4 of the apo Lp-IIIa model was to unwind the N-terminal turn of this helix and break H4 into two helical segments, H4A and H4B, separated by a nonhelical strand (Figure 2c). The four alanine residues following the C-terminal serine residue of H4 were incorporated into H4B. The energy of this final conformation in Lp-IIIa was -21 kcal/mol/residue suggesting stability similar to H4 of apo Lp-III (Table II). The sequence alignment (Figure 1) shows several identical residues among the proteins in the region of H4 -- Pro in the N-terminal position, Asp in position three and Leu in positions five, eight and twelve. After energy minimization, the N-terminal Pro and at least two adjacent residues had lost their helical character. Both DgA-I and ChA-I preserved longer helical segments, H4A (14 or 15 residues) and H4B (10 residues), with a shorter nonhelical strand (4 or 5 residues) than did Lp-IIIa. In contrast to the 13-residue nonhelical strand in Lp-IIIa, composed mostly of residues that generally favor helix formation, the five-residue strand in DgA-I contains three Gly residues. In the HuA-I model, the helical strand equivalent to H4A was only eight residues long being shortened from both ends relative to DgA-I and ChA-I. This shortened H4A structure was followed by a ten-residue nonhelical strand, containing two glycine residues, and then by five residues with α -helical dihedral angles, a sequence too short to evaluate by the helical wheel method. In the apo A-I models H4A had a relatively small ($80^\circ - 100^\circ$) hydrophobic sector with an unusually high average hydrophobicity (0.9 - 1.2). H4B, on the other

hand, was much more hydrophilic with an average hydrophobicity for the hydrophobic sector of 0.5 - 0.6.

Helix 5. The 20-residue H5 of apo Lp-III required a single alanine insertion in an already alanine rich region for alignment with the apo A-I proteins. In all refined models, the start of H5 was shifted two or three residues in the N-terminal direction. H5 in the refined models became a 20 to 23-residue helix with a two or three-residue kink just past the site of the alanine insertion in apo Lp-IIIa (Figure 2d). H5 in the DgA-I model was shortened by four residues from the C-terminal end. In other apo A-I models the helical conformation was maintained to within one residue of the C-terminal end.

Loops. Loops that separate helical segments play an essential role in determining intrahelical interactions that stabilize the entire structure. The insertion of an additional three residues in the region of the very short loop (3 residues) separating H3 from H4 caused the first turn at the start of H4 to unwind, although it did not change the orientation of H4. Residues inserted at the beginning of the loop separating H4 from H5 in apo Lp-III were incorporated into the H4B of apo Lp-IIIa, again having little effect on the orientation of the following helical segment (Figure 2a). Substitution of apo A-I sequences into the apo Lp-IIIa model had very little effect on the conformation of the loop regions despite the relatively few identical residues in these regions. The average hydrophobicity of amino acid residues in the loop regions of these models were in a narrow range between 0.3 and -0.6, that is neither particularly hydrophilic nor hydrophobic. As may be seen in Color Plate 3, there are many more ionized side chains in apo A-I than in apo Lp-III. The loop separating H2 from H3 contains 4 or 5 charged side chains in each of the proteins. Other loops of apo Lp-III contain one or no ionizable side chains, while those loops in apo A-I contain up to 7 ionizable side chains.

Role of Proline. Proline residues tend to have a greater effect on local secondary structure in a molecule than any other single residue. Of six proline residues in the crystalline fragment of apo Lp-III (24) Pro33 and Pro129 are in loops, Pro35 and Pro95 are located at the N-termini of helices while Pro118 and Pro120 are located within H4. The ϕ and ψ angles (-52° to -60° and -23° to -60°) for these six proline residues are more similar to those generally observed for residues in α -helical structures than to angles typically associated with proline residues. In the apo A-I models Pro98 and Pro208 are aligned with Pro33 and Pro129 of apo Lp-III in loop regions. Pro35 is replaced by a leucine residue in apo A-I with no effect on initiation of H2. The Gln94-Pro95-Ala96-Asp97 sequence that initiates H4 in apo Lp-III is replaced by ala-Pro-Ala-ala-Asp in apo Lp-IIIa and Ala-Pro-Tyr-Ser-Asp in apo A-I (Figure 1) with the introduction of a γ -turn centered on the proline residue similar to the proline based turns in the modeling of milk proteins (35). This turn shifts the start of H4 to at least Asp167 (Figure 1). The helical Pro-X-Pro sequence (residues 118 - 120) of apo Lp-III is not found in apo A-I, however, Pro120 and Pro208 in H2 and H3 respectively of apo A-I have α -helical ϕ , ψ angles (-52° , -35°).

Discussion

Pinker and coworkers (36) found that the effects on conformation and conformational stability of alanine substitutions in the helices of myoglobin could not be reliably predicted. In a study using mutant myoglobins, destabilization could be predicted when bulky internal residues were replaced, but results were variable when substitution was in the capping region of a helix. Surface substitutions in mid-helix were less detrimental than at the ends, but because of the disruption of interactions

between neighboring residues, the effects on conformation and stability were not completely predictable. Insertion of alanine residues into a helical segment, had the additional effect of displacing formerly neighboring residues and disrupting the stabilizing interactions of those residues. Insertion of alanine residues with helical conformational angles into a nonhelical loop region might be expected to cause a major disruption of conformation. In fact, the effect in these less structured loop regions was less than in the already helical segments. Had it been feasible to add an entire turn of helical structure at each insertion point in an existing helix, the result in terms of maintaining the helix might have been better. Aside from the effects on individual helical and loop segments detailed above, the stabilization energy (Table II) for this rather arbitrary model differed little from that of the apo Lp-III model.

The extent to which a mutant sequence can maintain the conformation, stability and properties of the template used for its construction gives some clue as to the requirements for a given structure. The signature characteristic of apolipoproteins is a series of helical segments connected by short loops (6,19). Similarities between apo A-I sequences are much greater than those between apo Lp-III and apo A-I, and one would expect excellent results in building a model for ChA-I or DgA-I based on an experimentally determined model of HuA-I. However, no such model is available at this time. An obvious difference (Color Plate 3) between apo Lp-III and apo A-I is in the number of ionizable side chains available for electrostatic interactions. The increased stabilization of apo A-I models relative to apo Lp-III is almost entirely due to improvements in the electrostatic energy (Table II).

In comparative studies of human and chicken apo A-I, circular dichroism spectra show a 10 to 15% greater amount of helical structure for the chicken protein relative to human apo A-I (3-5). In the present study, the loss of H4B in the HuA-I model combined with a greater tendency to unwind the ends of other helices resulted in 10 - 12% fewer residues in helical conformations for HuA-I when compared with DgA-I and ChA-I models.

The use of alanine, a helix stabilizing residue, as the spacer residue was undertaken with the expectation that helical structure would be less affected than if residues from an apo A-I sequence were inserted directly into the Lp-III structure. In fact, the insertion of alanine residues in the helical segments of Lp-III had a disrupting effect on helical structure that, to some extent, was ameliorated when apo A-I sequences were substituted into the Lp-IIIa structure. The insertion of any residue into a stable helical structure shifts the relative positions of other residues in at least two turns of the helix resulting in disruption stabilizing electrostatic and hydrophobic interactions. If the inserted residue can participate in new helix stabilizing interactions the disruption may be minimal. Alanine does not participate in the electrostatic interactions that contribute most to favorable energies for the apo A-I models.

The N-terminal residues 1 - 71 of apo A-I were ignored in these models because that portion of the sequences showed little homology with Lp-III (24) and had little predicted helical structure (26,30,34). The importance of this less helical N-terminal region would appear to be in the activation of LCAT, as several of the epitopes critical for the activation of LCAT are located in the N-terminal third of HuA-I (37,38). As the library of experimentally determined apolipoprotein structures is expanded, other models, better suited for apo A-I, that include the nonhelical regions will undoubtedly become apparent.

Both Nolte and Atkinson (26) and Brasseur and coworkers (30) have recently used molecular modeling techniques to develop models for apo A-I based largely on analysis of the amino acid sequence with respect to repeating segments, hydrophobic potential and predicted secondary structures modified by experimentally estimated conformations. Their models each predict six or seven helical regions beginning with residue 68 of the human apo A-I amino acid sequence. In our models of canine and chicken apo A-I the positions of H1, H2, H4A and H4B are in reasonable agreement

with the locations of helices predicted by others (26,30). H3 in our model is shifted several residues toward the C-terminus of the protein but does overlap the other predictions. Initially much of the region between residues 145 and 162 predicted (26,30) to be weakly helical, was modeled as helical, but the structure did not withstand energy refinement.

Total energy for each of the apo A-I models is -24 kcal/mole/residue, significantly better than the -20 kcal/mole/residue calculated for the energy minimized apo Lp-IIIa model. The much larger number of electrostatic interactions in apo A-I (Color Plate 3, Table II) is primarily responsible for the greater stability of these models. These electrostatic interactions may also be critical to the differences in function between these molecules. Color Plate 3 shows the Asp72 to Thr236 fragment of DgA-I as fitted to the apo Lp-IIIa template. Although apo Lp-III and the apo A-I proteins all have significant potential for the formation of amphipathic helices, the actual number of identities across these sequences is only 10%. Thus the substitution of apo A-I sequences into the apo Lp-IIIa template was reasonably successful, and may represent another path in the search for structural correlations among distantly related proteins.

Acknowledgments

The authors wish to thank Drs. James Chen, Edyth Malin and Gregory King for technical assistance and helpful suggestions.

Literature Cited

1. Chapman, M.J. *Meth. Enzymol.* **1986**, *128*, 70-143.
2. Weisgraber, K.H. *Adv. Protein Chem.* **1994**, *45*, 249-302.
3. Lux, S.E.; Hirz, R.; Shragar, R.I.; Gotto, A.M. *J. Biol. Chem.* **1972**, *247*, 2598-2606.
4. Osborne, J.C., Jr.; Brewer, H.B., Jr. *Ann. N.Y. Acad. Sci.* **1980**, *348*, 104-121.
5. Brown, E.M. *Poultry Science* **1989**, *68*, 399-407.
6. Boguski, M.S.; Freeman, M.; Elshourbagy, N.A.; Taylor, J.M.; Gordon, J.I. *J. Lipid Res.* **1986**, *27*, 1011-1034.
7. McLachlan, A.D. *Nature* **1977**, *267*, 465-466.
8. Lou, C.-C.; Li, W.-H.; Moore, M.N.; Chan, L. *J. Mol. Biol.* **1986**, *187*, 325-340.
9. De Loof, H. *Ann. Biol. Clin.* **1988**, *46*, 10-15.
10. Brasseur, R. *J. Biol. Chem.* **1991**, *266*, 16120-16127.
11. Chou, P.Y.; Fasman, G. D. *Adv. Enzymol.* **1978**, *47*, 45-148.
12. Garnier, J.; Osguthorpe, D.J.; Robson, B. *J. Mol. Biol.* **1978**, *120*, 97-120.
13. Eisenberg, D. *Annu. Rev. Biochem.* **1984**, *53*, 595-623.
14. Kyte, J.; Doolittle, R. F. *J. Mol. Biol.* **1982**, *157*, 105-132.
15. Andrews, A.L.; Atkinson, D.; Barratt, M.D.; Finer, E.G.; Hauser, H.; Henry, R.; Leslie, R.B.; Owens, N.L.; Phillips, M.C.; Robertson, R.N. *Eur. J. Biochem.* **1976**, *64*, 549-563.
16. Mahley, R.W.; Innerarity, T.L.; Rall, S.C. Jr.; Weisgraber, K.W. *J. Lipid Res.*, **1984**, *25*, 1277-1294.
17. Segrest, J.P.; Jackson, R.L.; Morrisett, J.D.; Gotto A.M., Jr. *FEBS Lett.* **1974**, *38*, 247-253.
18. Li, W.H.; Tanimura, M.; Luo, C.C.; Datta, S.; Chan, L. *J. Lipid Res.* **1988**, *29*, 245-271.
19. Segrest, J.P.; Garber, D.W.; Brouillette, C.G.; Harvey, S.C.; Anantharamaiah, G.M. *Adv. Protein Chem.* **1994**, *45*, 303-369.

20. Smith, L.C.; Pownall, H.J.; Gotto, A.M. Jr. *Annu. Rev. Biochem.* **1978**, *47*, 751-777.
21. Glomset, J.A. *J. Lipid Res.* **1968**, *9*, 155-167.
22. Blue, M-L.; Ostapchuk, P.; Gordon, J.S.; Williams, D.L. *J. Biol. Chem.* **1982**, *257*, 11151-11159.
23. Soulages, J.L.; Wells, M.A. *Adv. Protein Chem.* **1994**, *45*, 371-415.
24. Breiter, D.R.; Kanost, M.R.; Benning, M.M.; Wesenberg, G.; Law, J.H.; Wells, M.A.; Rayment, I.; Holden, H.M. *Biochemistry* **1991**, *30*, 603-608.
25. Wilson, C.; Wardell, M.R.; Weisgraber, K.H.; Mahley, R.W.; Agard, D.A. *Science* **1991**, *252*, 1817-1822.
26. Nolte, R.T.; Atkinson, D. *Biophys. J.* **1992**, *63*, 1221-1239.
27. National Biomedical Research Foundation, Georgetown University Medical Center, Washington DC 20007. Release 28.0 March 31, **1991**.
28. Cole K.D.; Fernando-Warnakulasuriya, G.J.P.; Boguski, M.S.; Freeman, M.; Gordon, J.I.; Clark, W.A.; Law, J.H.; Wells, M.A. *J. Biol. Chem.* **1987**, *262*, 11794-11800.
29. Vazquez, S.R.; Kuo, D.Z.; Botsitis, C.M.; Hardy, L.W.; Lew, R.A.; Humphreys, R.E. *J. Biol. Chem.* **1992**, *267*, 7406-7410.
30. Brasseur, R.; Lins, L.; Vanloo, B.; Ruysschaert, J-M.; Rosseneu, M. *Proteins* **1992**, *13*, 246-257.
31. SYBYL, Macromolecular modeling software. version 5.5. TRIPOS Associates, Inc. St. Louis, MO.
32. Weiner, S.J.; Kollman, P.A.; Case, D.A.; Singh, U.C.; Ghio, C.; Alagona, G.; Profeta, S.; Weiner, P.K. *J. Am. Chem. Soc.* **1984**, *106*, 765-784.
33. Weiner, S.J.; Kollman, P.A.; Nguyen, D.T.; Case, D.A. *J. Comput. Chem.* **1986**, *7*, 230-252.
34. Schiffer M.; Edmunson A.B. *Biophys. J.* **1967**, *7*, 121-125.
35. Kumosinski, T.F.; Brown, E.M.; Farrell, H.M., Jr. *J. Dairy Sci.* **1991**, *74*, 2879-2888.
36. Pinker, R.J.; Lin, L.; Rose, G.D.; Kallenbach, N.R. *Protein Science* **1993**, *2*, 1099.
37. Marcel, Y.L.; Provost, P.R.; Koa, H.; Raffai, E.; Dac, N.V., Fruchart, J-C.; Rassart, E. *J. Biol. Chem.* **1991**, *266*, 3644-3653.
38. Banka, C.L.; Bonnet, D.J.; Black, A.S.; Smith, R.S.; Curtiss, L.K. *J. Biol. Chem.* **1991**, *266*, 23886-23892.

RECEIVED July 8, 1994

Chapter 8

Testing a Model of the Extracellular Domain of Human Tissue Consistent with Fourier Transform Infrared Spectroscopy

J. B. A. Ross¹, C. A. Hasselbacher¹, Thomas F. Kumosinski³, Gregory King³, T. M. Laue⁴, A. Guha², Y. Nemerson^{1,2}, W. H. Konigsberg⁵, E. Rusinova¹, and E. Waxman¹

Departments of ¹Biochemistry and ²Medicine, Mount Sinai School of Medicine, New York, NY 10029

³Eastern Regional Research Center, Agricultural Research Service, U.S. Department of Agriculture, 600 East Mermaid Lane, Philadelphia, PA 19118

⁴Department of Biochemistry, University of New Hampshire, Durham, NH 03824

⁵Department of Biochemistry, Yale University, New Haven, CT 06510

Tissue Factor (TF) is a membrane-anchored cell-surface protein that in complex with the serine protease Factor VIIa initiates blood coagulation upon tissue damage. We have cloned and expressed the soluble, cytoplasmic domain of TF (residues 1-218) (sTF) for analysis of structure and function. Global secondary structural elements were determined using FTIR spectroscopy. The amide I band assignments indicated *ca.* 15% α -helix, 23% extended strands, the remainder being turns, loops, β -sheet, and 'other' structure. Secondary structure prediction algorithms using a knowledge-based approach that was constrained to the FTIR-determined structural elements were used to generate a working model of sTF, which was energy minimized and equilibrated at 300 K using a Kollman force field. The predictions of this model were tested by analytical ultracentrifugation, proteolytic cleavage, and absorption and fluorescence spectra of Trp \rightarrow Tyr and Trp \rightarrow Phe mutants of sTF.

Following tissue damage, blood coagulation is initiated by the complexation of the transmembrane protein tissue factor (TF) with the circulating serine protease, factor VIIa (VIIa). TF is a glycoprotein consisting of an extracellular domain (residues 1-219), a single transmembrane domain (residues 220-242), and a cytoplasmic domain (residues 243-263) with a sequence that contains a half-cysteine residue thioesterified to palmitate or stearate (for review see (1)). After complexation with TF, the proteolytic activity of VIIa towards its natural substrate, factor X (X), increases by many orders of magnitude (2).

0097-6156/94/0576-0113\$08.00/0

© 1994 American Chemical Society

Study of the three-dimensional structure of TF is complicated by the requirement that full-length TF be reconstituted in phospholipid vesicles or solubilized in detergent. To facilitate studies on the three-dimensional structure of TF, we have prepared a soluble TF construct (TF₁₋₂₁₈; sTF) which lacks the transmembrane and cytoplasmic domains (1).

Here we describe a model for the three-dimensional structure of sTF, developed using a knowledge-based approach for protein structure prediction. The prediction algorithms were constrained to the secondary structural elements of sTF, determined separately by FTIR spectroscopic analysis of the protein amide I absorption bands. The model for the sTF structure was tested by evaluating which of its features are consistent with experimental results obtained from proteolytic cleavage of sTF (3), absorption and fluorescence spectra of sTF and Trp→Tyr and Trp→Phe mutants (4), and analytical ultracentrifugation (5).

Experimental Procedures

sTF. Recombinant sTF was cloned, expressed, and purified as described by Waxman et al. (5). The authenticity of the protein was established by amino acid composition, NH₂-terminal sequence analysis (10 residues), and carboxyl terminal analysis by carboxypeptidase P digestion. The amino terminal sequence was that expected for human TF. Carboxyl terminal digestion, however, yielded exclusively Arg (residue 218) rather than Glu (residue 219) as predicted by the sequence of the cDNA. We presume that residue 219 was removed by *E. coli* proteases.

Proteolytic Cleavage. Limited proteolytic cleavage was by subtilisin digestion in 20 mM Tris buffer, pH 8, using an sTF:subtilisin ratio of 650:1 (w/w), as described (3). Two fragments recovered after purification by ion-exchange on a mono-Q column followed by HPLC corresponded to residues 1-84 and 87-218 of the 1-218 residue polypeptide chain.

sTF tryptophan mutants. sTF has four tryptophan residues. These are located at positions 14, 25, 45, and 158 of the polypeptide chain. Four functionally active mutants were prepared as described (4). In each, a different tryptophan residue was mutated, using site-directed mutagenesis, either to tyrosine or phenylalanine: W14F, W25Y, W45Y, and W158F.

Ultraviolet Absorption and Fluorescence Spectroscopy. Concentrations of sTF and sTF mutants were determined by ultraviolet absorption spectroscopy (6). Since the number of tyrosine, phenylalanine, and tryptophan residues varied in the different mutants, it was necessary to determine the relative absorption spectra of the wild-type and mutant proteins. This was accomplished by diluting and denaturing an aliquot of each protein stock in 6 M guanidinium chloride (Gdm•Cl). The concentration of denatured protein was then calculated from knowledge of the protein tryptophan and tyrosine content and the extinction coefficients of tryptophan and tyrosine of a protein denatured in 6 M Gdm•Cl (6). Finally, the concentrations and properly scaled absorbance spectra of the native proteins were calculated from the dilution factors. By computer subtraction of each nondenatured mutant spectrum (three tryptophans)

from that of native sTF (four tryptophans), the individual absorbance spectra were obtained for each of the four tryptophan residues in their native protein environment.

Fluorescence emission spectra (7 nm bandpass) were obtained at 24° C using 295 nm excitation (7 nm bandpass) to avoid contribution from tyrosine residues. To avoid emission intensity artifacts resulting from molecular rotation and segmental motions, all fluorescence measurements were made using 'magic angle' polarization conditions (7).

Ultracentrifugation Experiments. Sedimentation equilibrium studies were performed, as previously described (8), at 23° C in a Beckman Model E analytical ultracentrifuge equipped with an electronic speed control, RTIC temperature controller, Rayleigh interference optics, and a pulsed laser diode light source (670 nm). Data was acquired using a television-camera-based, on-line data acquisition and analysis system. Samples were loaded into short column cells (0.7 mm) at several concentrations up to 1 mg/mL total protein. The data from experiments at all concentrations were analyzed globally using the nonlinear least squares analysis package NONLIN (9).

Sedimentation velocity experiments were performed on a Beckman XLA analytical centrifuge equipped with absorption optics that allow scanning at multiple wavelengths and interfaced to a microcomputer for data acquisition. The sedimentation coefficient was determined from the movement of the boundary (10) and corrected to standard conditions (20.0°C) and zero protein concentration using conventional methods (11,12).

FTIR Spectroscopy. The spectrum of sTF in 50 mM Tris and 100 mM NaCl, pH 7.5, was obtained using a Nicolet 740 spectrometer. The water-subtracted spectrum (13), was obtained for the frequency region from 1800 to 1350 cm^{-1} , and enhanced by Fourier deconvolution (14,15). Amide I band-structure assignments for the frequency region from 1710 to 1585 cm^{-1} were based on the results of Surewicz & Mantsch (16), Casal *et al.*, (17), and Krimm & Bandekar (18). The precision of the assignments reflects the enhancement factor used in the Fourier deconvolution; the bandwidth of the actual measurement is 2 cm^{-1} . The results of the deconvolution procedure were checked by reconstructing the observed spectrum from the resolved bands. We assume that the fraction of residues composing each secondary structural element is proportional to the relative percent area of the associated vibrational band.

Structure Modeling Procedures

To generate the working model for the three-dimensional structure of sTF, knowledge-based secondary and tertiary structure prediction algorithms were used. Since TF is not homologous to any known proteins, initial test structures were generated with the program SYBYL/Biopolymer (TRIPOS), which uses Bayes Statistics (19,20), information theory (21), and neural networks (22).

The model was built in three parts, each consisting of about 80-residue fragments, maintaining disulfide pairs within fragments. An additional constraint, maintained throughout the model building (fragment and final structures), was that the global secondary structure be consistent with that determined from deconvolution FTIR analysis, described above.

We searched for periodic structures. In that way, each piece could be assembled without hypothetical loops or turns at the junctions. All calculations were performed in vacuum (no waters were added) in order to maintain a reasonable calculation time. Each folded fragment was energy minimized and equilibrated at 300 K using molecular dynamics. This procedure employed a Kollman force field (Amber) using a united-atom approach with essential hydrogens, a non-bonded cutoff of 8 Å, and a dielectric constant that varied with distance. We assumed here that equilibrium was reached when the radius of gyration and root-mean square fluctuations of all atoms obtained nearly constant values. Bonded distances of the peptide backbone were constrained so as to maintain periodic structures (23). When all constraints were met, the pieces were joined. The final working model of sTF was then energy minimized and equilibrated at 300 K until equilibrium was reached as described above.

Results and Discussion

At present, neither x-ray crystal structure nor NMR models exist for sTF. Since no such models are currently available, we developed a theoretical model for sTF, constrained by secondary structural information obtained from FTIR spectra.

We recognize, of course, that no one has yet succeeded in predicting the correct three-dimensional structure of a protein *de novo* from sequence information alone. Therefore, the theoretical model we describe here should not be construed as a prediction of the "correct" structure of sTF. Rather, it should be considered as a low-resolution working model for generating ideas and testing hypotheses. In addition, it will be of interest to compare this model with an x-ray or NMR-based model when either one or both become available.

Our approach to developing a theoretical model for sTF was to constrain the structure-prediction algorithms, described above, by only accepting those structures that contained the elements of secondary structure which we previously obtained through analysis of the protein's FTIR spectrum.

We then evaluated the resulting working model by comparing its key structural features with experimental data that provides independent information about these features. The symmetry of the global structure was assessed from the hydrodynamic behavior of the protein, measured by sedimentation velocity experiments. Susceptibility to proteolytic cleavage was used to locate a surface-accessible region of the polypeptide chain. Additional surface regions were identified by the location of glycosylation sites. Buried versus exposed regions were identified from the absorption and fluorescence spectra of each of the four tryptophan residues.

FTIR Spectrum of sTF. The water-subtracted FTIR spectrum (13) and its second derivative from 1800 to 1350 cm^{-1} are compared in Figure 1. The frequency region of immediate interest from 1710 to 1585 cm^{-1} , containing the Amide I resonances and enhanced by Fourier deconvolution (14,15) is shown in Figure 2.

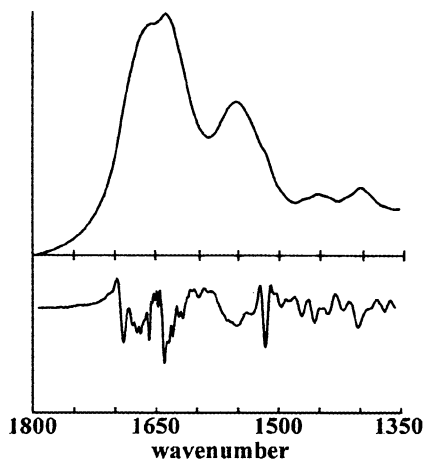


Figure 1. The FTIR spectrum (top) and its second derivative (bottom).

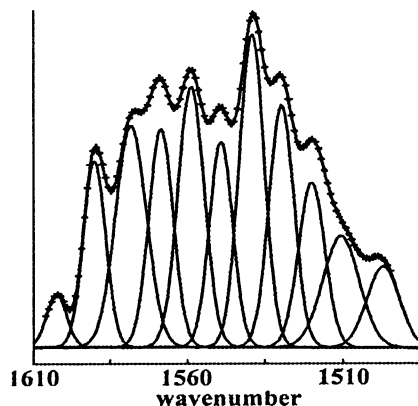


Figure 2. The deconvolved amide I region of the FTIR spectrum.

The characteristic frequencies for various secondary structural elements obtained after Fourier self-deconvolution (16-18) are listed in Table I. The salient result of this analysis is that sTF appears to contain relatively little α -helix; the Amide I assignments indicate that the secondary structure largely consists of extended strands, loops, and turns.

Table I. Amide I Band and Structure Assignments for sTF

Peak (cm ⁻¹)	% Area	Structural Element
1620	9.4 ± 0.4	extended strand
1630	13.1 ± 0.6	extended strand
1639	17.0 ± 0.7	extended strand
1649	10.3 ± 0.4	irregular, loop, or other
1659	14.9 ± 0.7	α -helix
1668	11.5 ± 0.5	turns
1678	14.8 ± 0.6	turns
1690	9.0 ± 0.4	turns

FTIR-Consistent model of sTF. Two aspects of the final, energy-minimized, FTIR-consistent model of sTF are shown in color plates 4 and 5. The first aspect (plate 4) is a ribbon tracing of the polypeptide backbone, including the four tryptophan residues at positions 14, 25, 45, and 158. The model has been edited to show the subtilisin cleavage between residues 86 and 87. The second aspect (plate 5) is a ribbon with the amino acid side chains.

NOTE: The color plates can be found in a color section in the center of this volume.

Testing the FTIR-Consistent, Theoretical Model by Comparison with Experiment.

The first feature of the model that we examined was the global structure; two black-and-white space-filling views are shown in Figure 3. The overall three-dimensional structure of the model predicts that sTF is an asymmetric protein, which is consistent with the results from analytical ultracentrifugation experiments. Using a monomer molecular weight of 24,700 daltons, obtained by sedimentation equilibrium, we calculate a value of 1.16 for the friction ratio. On this basis, if modeled as a prolate ellipsoid, sTF has molecular axes of 59 X 32 X 32 Å. The theoretical model, besides being asymmetric, contains a deep cleft, which would raise the frictional coefficient compared with that of a smoothly shaped, spherical protein of equal volume; modeling

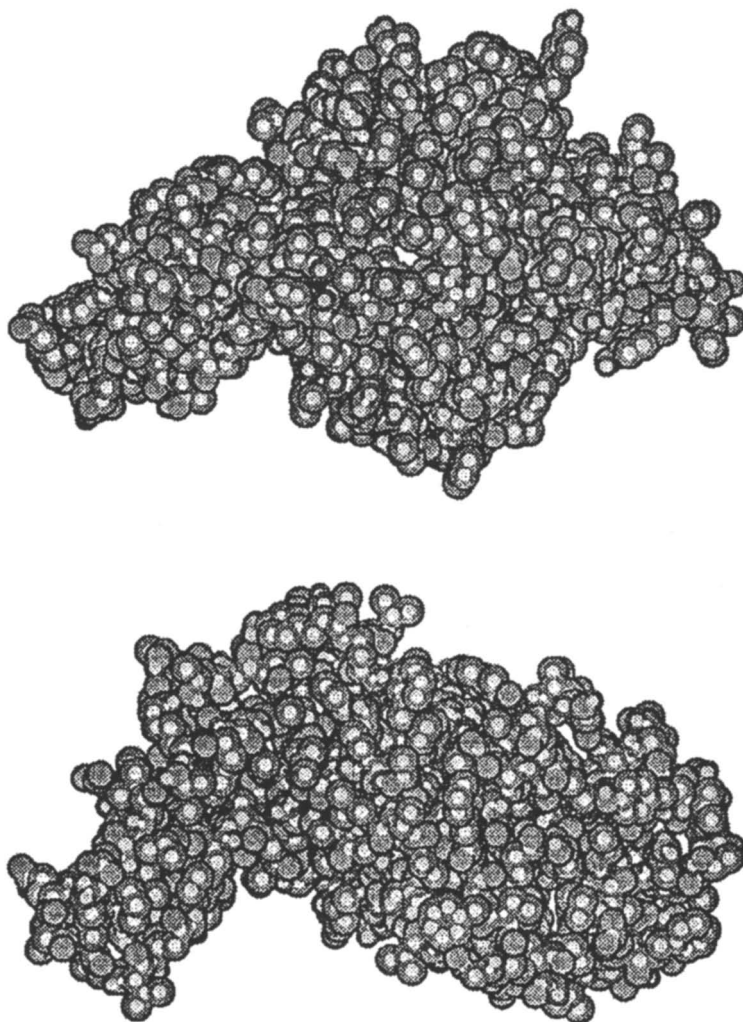


Figure 3. Two views of the predicted structure of human sTF.

the analytical ultracentrifuge results in terms of a prolate ellipsoid could thus be in error. It is interesting to note, however, that VIIa, which is activated by binding to membrane-bound TF, is also highly asymmetric (5). Interpretation of the friction ratios of VIIa and the sTF:VIIa complex in terms of prolate ellipsoids suggests that both VIIa and the complex with sTF might have diameters similar to that of sTF. If so, one might conclude that the protease and its cofactor stack end-to-end. This is of interest since these proteins function in a flowing environment that can subject them to considerable shear-force, and our laboratory has shown that the kinetics of production of Xa from X, the substrate for the TF:VIIa complex, are strongly shear-dependent (24).

The second feature of the model to be examined was the location of known surface regions. These are the protease digestion site and glycosylation sites. The theoretical model for sTF appears to be in accord with the results obtained from subtilisin cleavage experiments. The region of the polypeptide chain that includes residues 85 and 86, where the cleavage site is localized, is surface-exposed in the model (plate 4). The prediction of the locations of glycosylation sites also appears to be successful. The residues subject to N-linked glycosylation are Asn-11, Asn-124, and Asn-137. In the model (see Figure 4, below), these residues all appear to be at the surface of the folded polypeptide chain.

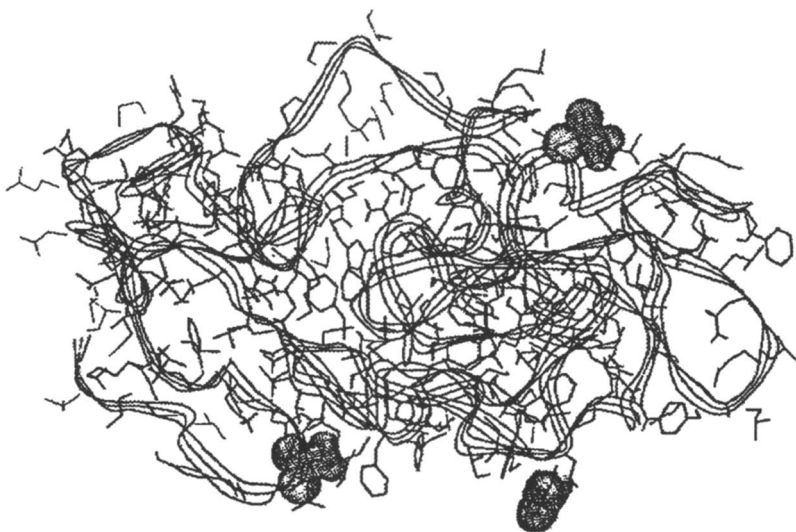


Figure 4. Glycosylation sites are highlighted by a Van der Waals representation of the atoms of the side chains of Asn 11, Asn 124, and Asn 137.

The third feature of the model to be examined was the local environment of each tryptophan residue within a radius of 10 Å. Figure 5 shows Trp-45 as an example. Since the solvent exposure of a tryptophan indole side chain is reflected in the energy distribution of its absorption and fluorescence spectra -- 'buried' tryptophans have 'red-shifted' absorption and 'blue-shifted' fluorescence emission spectra compared with those of residues that are 'solvent-exposed'. In addition, the fluorescence quantum yield is affected by nearest neighbor interactions. For example, quenching can result from interactions with disulfide bridges, histidine side chains, positively charged amino groups, carbonyl groups, or peptide bonds (25). According to a calculation of favorable free energy due to Van der Waals contacts in the model,

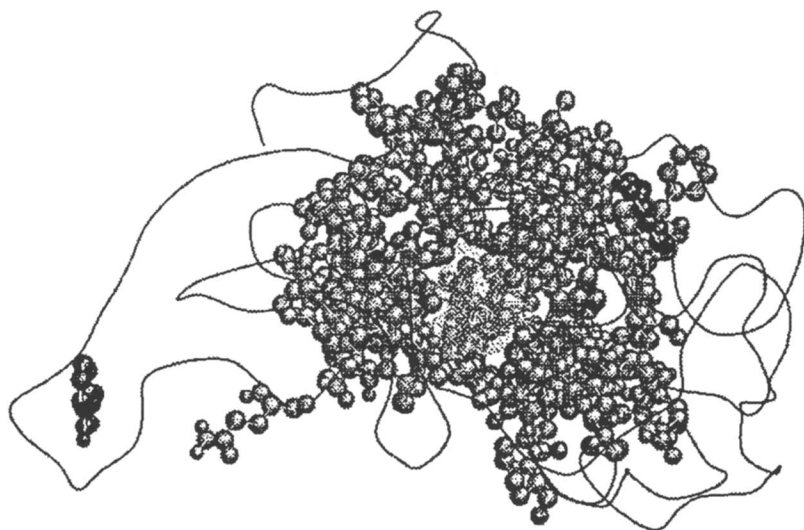


Figure 5. Side chains of residues having atoms within 10 Å of Trp-45 (Van der Waals surface). The indole rings identify Trp-25 and Trp-158.

the indole rings of Trp-14 and Trp-45 have the greatest amount of contact with other atoms in the protein (-21.5 and -23.3 kcal/mol, respectively). Trp-25 is intermediate and Trp-158 has the least Van der Waals interaction of the four residues (-17.5 and -14.4 kcal/mol, respectively). Thus, the model predicts that the indole rings of Trp-14 and Trp-45 will be the most shielded from bulk solvent. Moreover, the greatest number of positively charged side chains is in the vicinity of Trp-158, suggesting that the fluorescence of this residue has the greatest probability of being quenched. Using the absorption spectrum of N-acetyltryptophanamide (NATA) in dioxane as a reference model for a 'buried' tryptophan, one would conclude from the difference absorption spectrum of Trp-14, derived from the STF mutant W14F, that

Trp-14 is not accessible to aqueous solvent; the absorption spectrum of Trp-14 is identical with that of NATA in dioxane. On the same basis, Trp-25 would be considered 'buried'. While the absorbance spectrum of Trp-158 suggests it also is 'buried', the vibrational bands that make up its absorption band are broader than those of NATA in dioxane, suggesting that the indole ring experiences a less restricted, more heterogeneous environment. Similar data are observed for Trp-45, suggesting that the latter two tryptophans occupy regions in sTF that are more flexible, yielding a greater distribution of local interactions for the indole sides. Approximately 75% of the fluorescence of sTF is associated with Trp-45. The remaining 25% is associated with Trp-14; Trp-25 and Trp-158 have negligible fluorescence yields. While the low fluorescence yield of Trp-158 correlates with the model prediction of a high density of positively charged residues in local environment of this residue, no such elements stand out in the local environment of Trp-25. On the other hand, it is interesting to note that the two emissive residues, Trp-14 and Trp-45, are the residues predicted to have the largest Van der Waals interaction with the protein matrix. In terms of testing the model, however, the most important observation is that the relative energies of the emission spectra of Trp-14 and Trp-45 are consistent with their absorption spectra, confirming the conclusion that Trp-14 is less solvent accessible than Trp-45.

In conclusion, we find that experimental data from analytical centrifugation, protease digestion, location of glycosylation sites, and spectra of the four tryptophan residues are in accord with the FTIR-consistent theoretical model for sTF. In spite of the apparent 'success' of the model, according to these experimental criteria, this particular model by no means should be regarded as a final representative of the three-dimensional structure of sTF. There are important reasons for caution. Perhaps the most important is the fact that water plays a critical role in protein folding, and for ease of computation the contribution of water was not taken into account in our structure prediction. The motivation for our approach to structure prediction was to find out whether the constraints set by the FTIR-determined secondary structural elements would yield a model for a protein of unknown structure with experimentally verifiable features. We look forward to comparison of this low-resolution working model with a model determined by NMR or x-ray crystal diffraction studies. For these reasons, at present we prefer to view the low-resolution working model as a guide for generating experiments to understand the structure and function of TF in the initiation of blood coagulation.

Acknowledgments

Supported in part by grants HL-29019 and GM-39750 from the National Institutes of Health, and in part by grant DIR-9002027 from the National Science Foundation.

Literature Cited

1. Bach, R. Initiation of coagulation by Tissue Factor *CRC Crit. Rev. Biochem.* 1988, 23, 339-368.
2. Silverberg, S. A., Nemerson, Y., Zur, M. *J. Biol. Chem.* 1977, 252, 8481-8488.
3. Konigsberg, W.H., Guha, A., Singanallore, T.V., Lin, T.C., Ross, J.B.A., Nemerson, Y., in preparation.
4. Hasselbacher, C.A., Waxman, E., Rusinova, E., Guha, A., Konigsberg, W.H., Nemerson, Y., Ross, J.B.A., in preparation.
5. Waxman, E., Laws, W.R., Laue, T. M., Nemerson, Y., Ross, J. B. A. *Biochemistry*, 1993, 32, 3005-3012.
6. Waxman, E., Rusinova, E., Hasselbacher, C.A., Schwartz, G.P., Laws, W.R., Ross, J.B.A., *Anal. Biochem.* 1993, 210, 425-428.
7. Badae, M. G., & Brand, L. *Methods Enzymol.* 1979, 61, 378-425.
8. Laue, T. M., Shah, B. D., Ridgeway, T. M., and Pelletier, S. M. *Analytical Ultracentrifugation in Biochemistry and Polymer Science*, S. Harding, A. Rowe, Eds., Royal Society of Chemistry, London, 1992, pp. 90-125.
9. Johnson, M. L., Correia, J. C., Yphantis, D. A., Halvorson, H. R. *Biophys. J.* 1981, 36, 575-588.
10. Stafford, W. F. III *Anal. Biochem.* 1992, 203, 295-301.
11. Teller, D. C. *Methods Enzymol.* 1973, 27, 346-441.
12. Van Holde, K. E. *Physical Biochemistry*, 2nd. Edition. Prentice-Hall, Englewood Cliffs, NJ, 1985, p. 117.
13. Dousseau, F., Pezolet, M. *Biochemistry* 1990, 29, 8771-8779.
14. Susi, H., Byler, D.M. *Methods Enzymol.* 1986, 130, 290-311.
15. Mantsch, H.H., Casal, H.L., Jones, R.N. Resolution enhancement of infrared spectra of biological systems *Spectroscopy of Biological Systems*, Clark, R.J.H., Hester, R.E., Eds., Wiley & Sons, NY, 1986.
16. Surewicz, W.K., Mantsch, H.H. *Biochem. Biophys. Res. Commun.* 1988, 150, 245-251.
17. Casal, H.L., Kohler, U., Mantsch, H.H., *Biochim. Biophys. Acta* 1988, 957, 11-20.
18. Krimm, S., Bandekar, J. *Adv. Prot. Chem.* 1986, 38, 181-364.
19. Maxfield, F.R., Scheraga, H.A. *Biochemistry* 1976, 15, 5138-5153.
20. Kabsch, W., Sander, C. *Biopolymers* 1983, 22, 2577-2637.
21. Garnier, J., Osguthorpe, D., Robson, B. *J. Mol. Bio.* 1978, 120, 97-120.
22. Qian, N., Sejnowski, T. *J. Mol. Biol.* 1988, 202, 865-884.
23. Kumosinski, T.F., Smuda, E., Farrell, Jr. H. Molecular Dynamics and annealing of peptides and proteins: solvent and salt effects, Biophysical Society Meeting abst., 1991.
24. Gemmell, C. H., Broze, G. J., Jr., Turitto, V.T., Nemerson, Y. *Blood* 1990, 76, 2266-2271.
25. Longworth, J.W. Luminescence of Polypeptides and Proteins, *Excited States of Proteins and Nucleic Acids*, Steiner, R.F., Weinryb, I., eds., Plenum Press, NY, 1971, pp. 319-484.

RECEIVED September 8, 1994

Chapter 9

Computer-Generated Working Models of α -Crystallin Subunits and Their Complex

Patricia N. Farnsworth^{1,2}, Thomas F. Kumosinski³, Gregory King³,
and Barbara Groth-Vasselli²

Departments of ¹Physiology and ²Ophthalmology, University of Medicine and Dentistry—New Jersey Medical School, Newark, NJ 07103
³Macromolecular and Cell Structure, Eastern Regional Research Center, Agricultural Research Service, U.S. Department of Agriculture, 600 East Mermaid Lane, Philadelphia, PA 19118

The 3D structure of α -crystallin is the missing link for defining, at the molecular level, its functions as both a structural and chaperone protein involved in maintaining lens transparency. α -Crystallin has not been crystallized and its large aggregate size precludes 2D NMR. Therefore, computer assisted molecular modeling was used to construct energy minimized 3D working models of α -crystallin subunits, α A and α B, and their complex. This provides a basis for our speculation concerning the stoichiometry and orientation of these subunits within the quaternary structure of its oligomers. A comparison of these working models with existing experimental data provides a high level of confidence in the accuracy of our 3D models.

α -Crystallin is one of a number of globular proteins recruited for maintaining lens transparency and establishing its refractive properties (1,2). The most prevalent crystallins are designated as α -, β - and γ -crystallins. Originally, they were considered highly stable, relatively inert structural proteins unique to the lens; however, more recent studies suggest the crystallins are derived from multifunctional pre-existing proteins (3-5) that have been recruited to participate in lens supramolecular order. α -Crystallin is a water soluble protein composed of two gene products designated as α A2 (acidic pKi, 5.6-5.9) and α B2 (basic pKi, 7.1-7.4) (6). In most animals (7) each gene for α -crystallin subunits consists of three exons and two introns which suggests a molecular organization similar to the β - and γ -crystallins, namely, N- and C-terminal domains with an interconnecting segment. The α -crystallin subunits share approximately 60 % homology with each other. Homology exceeds 83 % in the most highly conserved residues between His 101 and Arg 123 of

0097-6156/94/0576-0123\$08.00/0
© 1994 American Chemical Society

α B₂ and His 97 and Arg 119 of α A₂. These sequences are within the region of homology with the small heat shock protein (hsp) (8).

In vivo, the primary gene products are subject to serine phosphorylation by specific cAMP-dependent kinases (9,10). The phosphorylated subunits are designated α A₁ and α B₁, while the unphosphorylated forms are α A₂ and α B₂. In general, the effect of protein phosphorylation is to significantly increase the net negative fixed charges within a cell which produces an influx of cations with a concomitant volume increase (11). In addition, the protein surface charge pattern is altered which may modify the interactive behavior of α -crystallin subunits to form an oligomer and/or the interactive behavior of an α -crystallin oligomer with other proteins.

The presence of α -crystallin in all vertebrate lenses and its abundance (20-30 % of crystallins) are evidence of its importance in lens functional integrity. Our earlier studies established that the crystallins in the outer cortical fiber cells are essentially in a solution phase (short-range order) while in the inner nuclear region, they coexist in both solution and solid-like phases (12). This transition constitutes a continuing maturation of lens fiber cells as they are displaced inward during lens growth which continues throughout life. An exquisite control of supramolecular order during maturation is implied by the consistency of chemical composition, morphology, and growth of age-matched lenses of a given species. The progressive shift from cortex to nucleus of α -crystallin from the soluble to the insoluble fraction of lens crystallins suggests that alterations of its aggregative properties are important in normal fiber cell maturation and the stabilization of the essentially inert nuclear region. Conversely, its presence as the major constituent of large light scattering elements during lens opacification suggests that modification of its normal aggregative properties is also central to cataractogenic processes (13,14).

As a polydisperse protein with a dynamic quaternary structure (15), α -crystallin can readily respond to the changing cell environment associated with normal fiber cell maturation and cataractogenic stress. A significant sequence homology with small heat shock proteins (8) and the identification of α B as a hsp (16) in the lens led to the investigation of the functional role of α -crystallin as a hsp. Since the subunits of α -crystallin have been identified in other tissues under both normal and pathologic conditions (17-21), the relevance of our investigation of α -crystallin reaches beyond the confines of the lens and cataractogenesis.

The 3-D structure of α -crystallin is basic to our understanding of the molecular mechanisms involved in fiber cell maturation, stabilization of the lens nuclear region and maintenance of lens transparency; therefore, we have used computer-assisted molecular modeling to construct minimized working models of α -crystallin subunits and their complex. The very slow

rate of sequence modifications during the evolution of α -crystallin suggests that there is a highly conserved structure/function relationship (1). On a more global level, it is basic to the function of small hsp in many cell types of various species. The fact that the subunits share homology with small hsp (8) and α B is known to interact with other heat shock proteins *in vivo* (22) provides an opportunity to probe the structure-function relationship of these very important polypeptides. Our "working models" of the subunits, their complex and the construction of a quaternary structure of the aggregate will be used for predictive purposes in designing experiments and for confirming or modifying our working models. There are several 3 layered models suggested for the quaternary structure of the oligomer; however, there is mounting evidence that the most reasonable quaternary structure is a micelle (23).

Modeling of α -Crystallin Subunits and Complex

Hardware. The HP9000/845 main frame running the UNIX operating system is interfaced with a Silicon Graphics personal Iris work station. SYBYL, Del Phi and Insight II softwares are loaded on Iris and files are stored in the main frame. A Tektronix 4693DX printer is directly attached to the Iris work station to obtain a hard copy of the figures on the screen.

Software. The SYBYL molecular modeling package (version 6.0) obtained from Tripos Associates, Inc. was used. The Kollman (AMBER) force field was utilized for all energy minimization and molecular dynamics simulations (24, 25). A united atom approach, using only essential hydrogen bonding protons, was used in order to increase the speed of the calculation. Hence, van der Waals radii of carbon atoms were increased to account for the lack of hydrogens. This united atom approach is widely used when dealing with proteins in excess of thirty residues. Electrostatic interactions were added to the calculation by using united atom partial charges according to Weiner et al. (25). A non-bonded cutoff of 8 Å was employed for van der Waals and electrostatic non-bonded interactions. All calculations were performed in vacuum using a dielectric constant that varies with distance.

Modeling of α -Crystallin. The sequences for bovine α -crystallin subunits, α A₂ and α B₂ were used for modeling their 3D structures. The sequence of these subunits are highly conserved. For example, α A₂-crystallin polypeptide is changing at a rate of only 3 % every hundred million years (1). Therefore, the models of these subunits are applicable to most mammals. Molecular modeling techniques along with secondary structure sequence based prediction algorithms (26) as well as experimental global secondary structure were utilized to construct energy minimized 3D working

models of the subunits. The local minimization problem for these energy minimized structures was solved by using molecular dynamics with a Kollman (AMBER) force field and constraint bond distances for backbone atoms until equilibrium structures were achieved. All reported structures were the final structure achieved during molecular dynamics. This eliminates bad contacts and achieves the more favorable hydrogen bond and van der Waals interactions. The total energy obtained includes bond stretching, angle bending, torsional, out of plane bending, 1-4 van der Waals and van der Waals energy. Energies on the order of $-10 \text{ kcal M}^{-1} \text{ residue}^{-1}$ or lower relative to the unfolded protein were used as a criterion for acceptable structures. Thus the 3D structures described herein should not be construed as high resolution structures but should be regarded as low resolution "working" models for generating ideas and testing hypotheses.

3D Structure of Working Models

αA_2 Subunit. The backbone configuration of the αA subunit consisting of 173 amino acid residues is presented in Figure 1. There are well defined N-(left) and C-terminal (right) domains with an interconnecting α -helical peptide chain. An analysis of the Ramachandran plot of the working model of αA (Figure 2) reveals that the secondary structure contains approximately 37 % α -helix, 9 % extended structure and extensive turns and loops. In the model, the C-terminal extension consists of 10 amino acids. This is in agreement with a 2D NMR analysis of the spectra of αA crystalline aggregates that the last 8 residues of the subunits in the C-terminal domain exhibit conformational flexibility (27). Since the N-terminal domain is hydrophobic and the C-terminal domain is highly charged, the αA_2 subunit is amphipathic and capable of aggregating to form micelle-like structures of varied size and shape (see Survey of the Literature, subheading, *iv*).

αB_2 Subunit. The αB_2 subunit which contains 175 residues shares similarities with αA_2 but differs in conformation. Similar to αA_2 , the αB_2 backbone configuration (Figure 3) has two domains and an α -helical connecting peptide. However, the tertiary structure of αB_2 differs from αA_2 . The long axis of the hydrophobic N-terminal domain of αB_2 is parallel rather than perpendicular to the C-terminal (right) domain. This difference creates a cleft between the two domains and provides a putative binding site for αA_2 . The folding pattern is most likely due to the presence of 9 prolines in the N-terminal domain in contrast to the 5 found in αA_2 . The additional prolines in αB_2 may influence the bend in the backbone which changes the positioning of the N-terminal domain. Similar to αA_2 , the N-terminal domain is very hydrophobic while the C-terminal domain is highly charged resulting in an

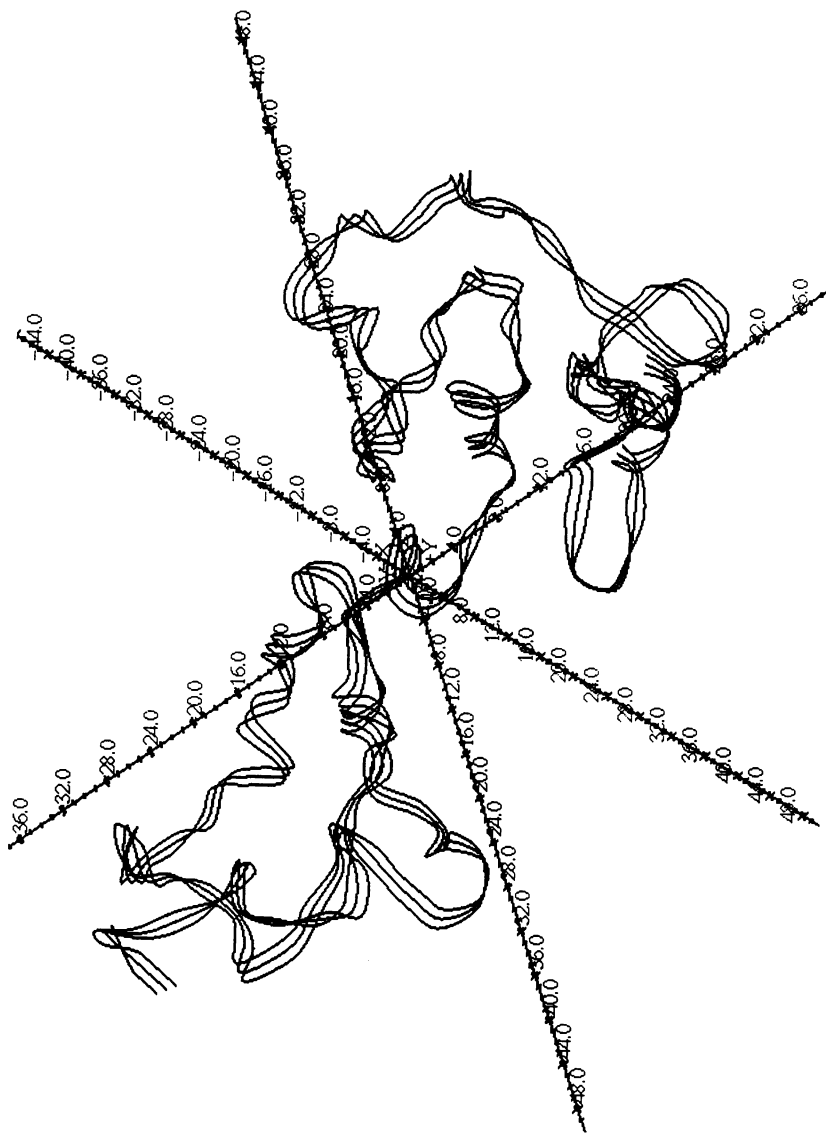


Figure 1. Backbone of the "working" model of the αA_2 subunit of bovine α -crystallin.

amphipathic molecule. In Color Plate 6, the green spheres represent hydrophobic amino acid side chains while the van der Waals surfaces of the negative and positive charged residues are depicted in red and purple, respectively. The well documented stability of the molecule (28-30) appears to be related to the combination of numerous ion pairs and hydrophobic interactions.

$\alpha A_2/\alpha B_2$ Complex. The complex of $\alpha A_2/\alpha B_2$ is seen in Figure 4 and Color Plate 7. In Figure 4, the backbone structure of αA_2 is defined as a 3-stranded ribbon, while the backbone of αB_2 is a ribbon. A portion of the C-terminal domain of αA_2 is bound by both hydrogen bonds and hydrophobic interactions in the cleft region of αB_2 (Color Plate 7). The hydrophobic N-terminus of αA_2 projects out from the cleft at an angle which is adjacent and parallel to the hydrophobic N-terminal domain of αB_2 . This results in an amphipathic complex that can readily participate in a micelle-like quaternary structure with variable molecular weights ranging from 6×10^5 to 5×10^7 (30-32). The C-termini of both subunits remain free from the complex as defined by 2D NMR (27). The ratio of $\alpha A/\alpha B$ reported in normal bovine lenses is three to one. Although the configuration of a 3:1 complex of $\alpha A_2/\alpha B_2$ remains controversial, our model suggests that this ratio may be appropriate. The available surface contacts remaining on αB in the complex will probably accommodate only 2 additional αA subunits (see Survey of the Literature, *iv*). The Ramachandran plot (Figure 5) of our "working" model indicates that the secondary structure of the complex contains approximately 24 % α -helix, 7 % extended structure and a substantial amount of loops and turns.

Secondary Structure Validation of "Working" Models

FTIR Spectroscopic Studies. Analysis of Fourier transform infrared (FTIR) spectroscopic studies (33) of the secondary structure of bovine α -crystallin in H_2O revealed 27 % α -helix, 18 % extended structure, 41 % turns and 14 % loops. These data agree favorably with Ramachandran plots generated from our working 3D models of αA_2 and the $\alpha A_2/\alpha B_2$ complex which contained 37 and 24 % α -helix, respectively. However, Lamba et al. (34) also analyzed α -crystallin by FTIR but concluded that the predominant secondary structure was extended β -sheet with very little α -helix. The disparity in FTIR results may, in part, be related to the use of D_2O by Lamba et al. in place of H_2O . Although D_2O was used routinely, the use of H_2O offers significant advantages. The results of an FTIR spectroscopic study comparing H_2O and D_2O showed that the secondary structure of 15

NOTE: The color plates can be found in a color section in the center of this volume.

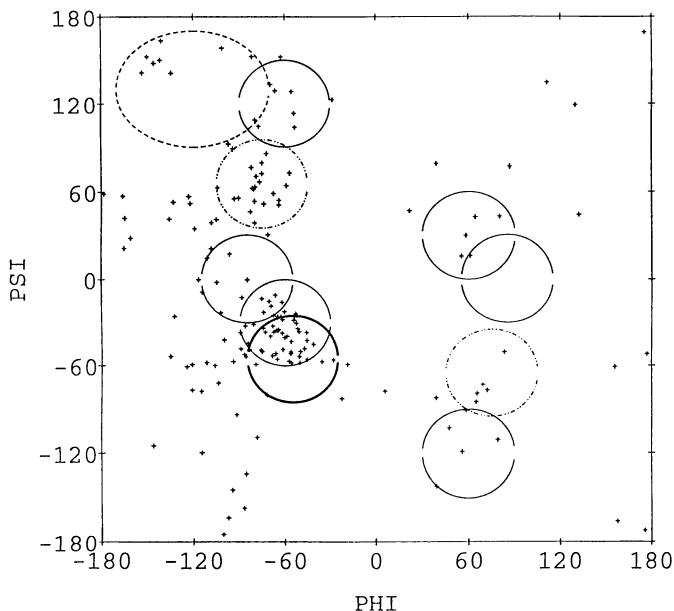


Figure 2. A Ramachandran plot is the phi psi backbone dihedral angle calculated from Sybyl on the predicted structure of the working model of α A2. In this and Figure 5 the dashed lines are allowable envelopes for β -sheet, double lines indicate α -helix and all others are various turns. This plot of the α A₂ model provides evidence for approximately 37 % α -helix.

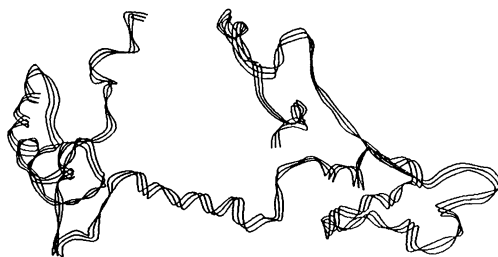


Figure 3. Backbone of the "working" model of the α B₂ subunit.



Figure 4. Backbone of the "working" model of the α A₂ and α B₂ complex.

proteins analyzed in H₂O were in better agreement with x-ray crystallographic data (33). Also, more information is obtained in H₂O since the amide I and amide II bands may be analyzed concurrently. A significant disadvantage of D₂O for α -crystallin analysis is its enhancement of hydrophobic interactions which serve as a driving force for subunit aggregation and therefore, have the potential for altering results (35). The high signal of the water peak in D₂O saturates the detector, and a slope as well as the baseline correction must be utilized for proper subtraction of the solvent spectrum from the solution spectrum (36). These problems are eliminated with the use of H₂O. Finally, the methods of Fourier deconvolution of the spectrum and the peak assignments of α -crystallin are fundamental to the quantitation of secondary structure. The analysis of FTIR in H₂O agrees with the "working" models and meets the required quantitative criteria described recently by Kumosinski and Farrell (36).

Circular Dichroism (CD) Spectroscopic Studies. CD spectroscopy is a widely used technique for estimation of protein secondary structure. However, light scattering created by the inherent tendency of α -crystallin to aggregate presents a source of error in the analysis. Early attempts to predict α -crystallin secondary structure based on far UV CD measurements showed one minimum (216 nm) for α -crystallin (37-38) and reaggregated α A (37) while reaggregated α B had a somewhat different spectrum with a minimum at 206-209 nm (37). These results suggested that α -crystallin contained approximately 55-63 % β -sheet. However, recent CD spectra of α -crystallin showed two minima at 220 and 209 with a positive peak at 190 nm which are indicative of α -helix (39). In a recent paper by Merck et al. (40), the CD spectra were recorded for native calf α -crystallin, recombinant α A and HSP 25. The spectrum for recombinant α A (0.35 mg/ml) displayed two minima typical of α -helix; however, at 0.70 mg/ml, α A and HSP 25 displayed a single minimum suggesting β -sheet. We suggest that the loss of the 206-209 minimum is related to a concentration dependent increased aggregation resulting in light scattering. Our experimental confirmation for such aggregation is presented in Figure 6 where absorbance is plotted against α -crystallin concentration. The plot is non-linear (curve fitted to the data represents a third order least square polynomial, $R^2 = 0.997$). The non-linearity of the curve denotes a deviation from Beer's law and indicates increased aggregation resulting in light scattering. This occurs at α -crystallin concentrations greater than approximately 0.2 mg/ml. This is in accord with Bloemendal et al. (41) who also observed light scattering at low α -crystallin concentrations.

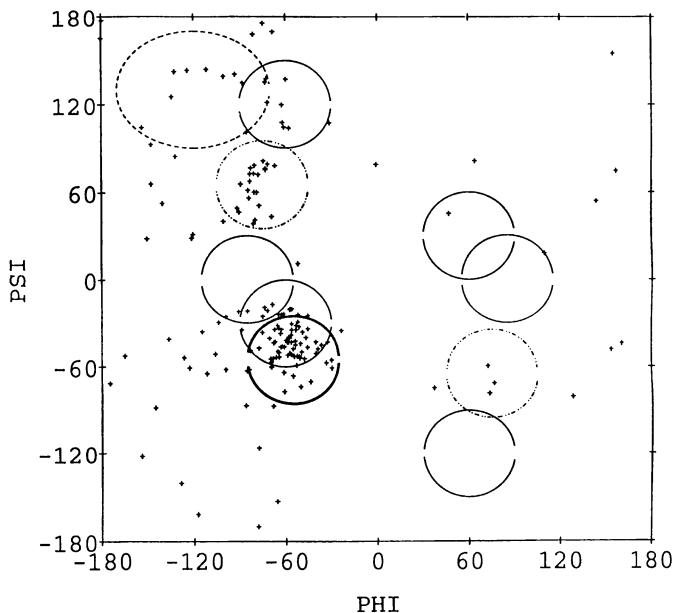


Figure 5. Ramachandran plot of the α A/ α B complex provides evidence for 24% α -helix.

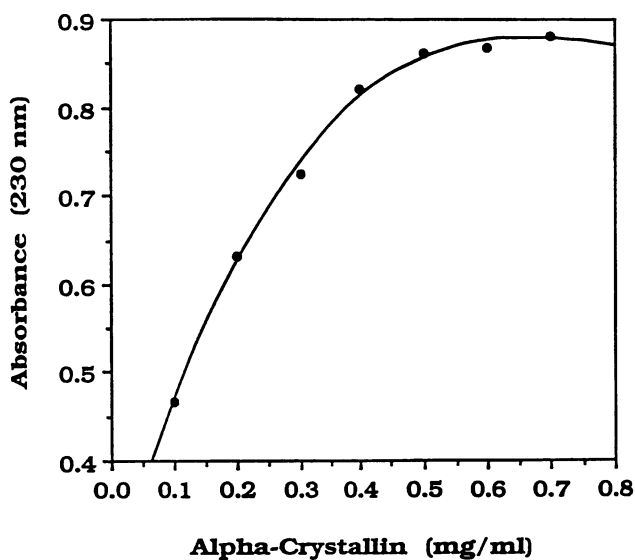


Figure 6. A plot of absorbance vs. calf α -crystallin concentration. Deviations from linearity at concentrations greater than ca. 0.2 mg/ml indicates the presence of light scattering.

Predicted Secondary Structure of the Subunits. A comparison of the predicted secondary structure for bovine γ -II-crystallin and α A and α B and four small heat shock proteins that have significant homology with both subunits is presented in Table I (26). With respect to γ -II crystallin, analysis of the Ramachandron plot based on the x-ray crystal structure yields approximately 55 % β -sheet, 35 % turns and 10 % other. This is in agreement with the predicted secondary structure of γ -II based on the algorithm by Garnier, Osguthorpe and Robson presented in Table I. Thus, this provides validity for predicting a substantial amount of α -helix (38-51 %) for the α A and α B subunits and the hsp's using this algorithm. The comparable values of the α -crystallin subunits and the various hsp is not surprising since the region of homology between the hsp and the subunits includes the helical connecting segment between the two domains of both α A and α B. The percentage of α -helix in the α A and α B subunits are in good agreement with the "working" models and FTIR and CD spectroscopic measurements at low concentrations. The absence of α -helix in γ -II crystallin suggests that the structural homology with α -crystallin predicted by other investigators is unfounded.

Table I. Comparative Study of Predicted Secondary Structure

<i>Species</i>	<i>Protein</i>	<i>α-Helix</i>	<i>β-Sheet</i>	<i>Reverse Turn</i>	<i>Random Coil</i>
Spiny Dogfish	α A	37.9	4.5	31.6	26.0
Human		46.5	2.9	26.7	23.8
Bovine		43.9	4.6	24.3	27.2
Bovine	α B	51.4	3.4	20.0	25.1
Drosophila	hsp22	59.2	4.0	16.7	20.1
	23	37.1	41.9	7.0	14.0
	26	34.9	39.7	13.9	11.5
	27	32.4	39.9	14.6	13.1
Bovine	γ -crystallin	0.0	52.9	37.9	9.2

Tertiary Structure of the Subunits

It is generally accepted that the molecular structure of α -crystallin subunits has a 2 domain symmetrical β -sheet configuration similar to the β - and γ -crystallins. Each domain has a Greek-key motif containing 4

antiparallel β -strands (42). However, a comparison of the Eisenberg hydrophobic moment of α B and γ -crystallins reveals significant differences. α B is an asymmetric molecule with a hydrophobic N-terminal domain while the γ -crystallin x-ray diffraction derived 3D structure is essentially symmetrical in charge distribution and 3D structure. In addition, residues critical to the folding pattern of either β - or γ -crystallin and the internal sequence symmetry required for the Greek-key configuration are absent in both α A₂ and α B₂ (43, 44).

To further investigate possible tertiary structural homology between α - and γ -crystallin, computer assisted molecular modeling was used to superimpose the primary structure of α A on the β -pleated sheet, Greek-key structure of γ -crystallin. The resultant α A₂ structure was presented to molecular dynamics until equilibrium was again established. Following these maneuvers, the final energy level of this α A₂ structure was 120 kcal M⁻¹ higher than our working model presented in Figure 1. A comparison of the backbone and side chains of γ -crystallin and α A₂ are seen in Color Plate 8. From these figures, it is obvious that the folding pattern of α A₂ is not compatible with that of γ -crystallin. The numerous hydrophobic amino acids on the surface, the presence of buried or isolated charged residues (red, negative and purple, positive) within the interior and the unbalanced distribution of charges throughout the structure clearly demonstrate that a β -pleated sheet structure is highly unlikely for this molecule. In addition, the Ramachandran plot (not illustrated) does not resemble that of either γ -crystallin or our working model of α A₂.

Agreement of Working Models of α A, α B and their Complex with the Literature

Our "working" 3-D models are based on predictive, experimental and molecular modeling techniques. All three are required. These models are in agreement with a number of biochemical, physical chemical and solution structural data in the literature. These resulting dynamic structures will be used for predictive purposes in designing experiments to confirm and/or modify these models. Speculation is in progress concerning the stoichiometry and orientation of these subunits within a micellar type oligomer which occurs in solution.

Survey of the Literature. A survey of the literature on α -crystallin over the past several decades revealed significant agreement between the experimental data and our 3D models.

- The working model for each subunit has two distinct N- and C-terminal

domains with an interconnecting α -helical peptide. Each gene responsible for the expression of α A and α B subunits has three exons and two introns (45). There is a striking correspondence between the exons on a genetic level and the expression of three well defined structural units evident in our models.

- The recognized *in vivo* serine phosphorylation sites, Ser 122 in α A and Ser 43 and/or 45 and 59 in α B (9,10) have been localized within the models and, as required, are found on the surface of both the subunits and the complex. From our observations of the 3D structures of the phosphorylation sites for Ser 122 and Ser 59 we found that the sites are essentially identical; an arginine extends into the solvent and the characteristic turn dictated by proline extends the serine to the surface. This is not surprising since the primary sequences (RLPSN) and (RAPSW) for Ser 122 (α A) and SER 59 (α B), respectively, are essentially identical. The other proven phosphorylation sites for either Ser 43 or 45 in α B have significantly different primary sequences (9) and conformation. Therefore the two different serine kinases, predicted in laboratory studies (9), correlate with our computer analysis.

- Two dimensional ^1H NMR spectroscopy of either homo- or heterogeneous α -crystallin aggregates revealed that they have short and flexible C-terminal extensions on both subunits (27). Each subunit and their complex in our models also have comparable C-terminal extensions. In addition, the susceptibility to proteolysis of these extensions (46,47) and the peptide following residue 101 in α A (48) further support our configuration of the subunits and complex.

- There is considerable evidence to support a micellar quaternary structure for α -crystallin (23, 49-51). Our "working" models of α -crystallin subunits and complex are amphipathic and, therefore, eminently suited for formation of "protein micelles". Using TEM, the micelles (800 kd aggregates on average) are composed of 40-42 subunits with a 3/1 ratio of α A₂ to α B₂ and range in diameter between 15-17 nm (52). We were able to construct an aggregate of similar size and number of subunits based on the dimensions of our computer-generated subunits and complex dimensions.

- Recent studies of Boyle et al. (53) provide support for the localization of denatured proteins within the "central cavity" of α -crystallin aggregates. A "central cavity" is compatible with the hydrophobic N-terminal domains of the subunits facing the interior of the aggregate. The interior of the micelle thus provides binding sites for exposed hydrophobic regions of denatured protein and may explain the role of α -crystallin as a molecular chaperone (54).

- To probe the high order structure (β -sheet or α -helix) of α -crystallin, a new analytical method is being developed based on detection of the rate of

peptide amide hydrogen exchange with D₂O (55). This method uses mass spectroscopy with continuous flow HPLC. Although it is not possible to discriminate between these two secondary structures at this time, one can determine segments in the protein with a slow rate of hydrogen exchange, an indication of high order. Preliminary data from David Smith (56) revealed that the connecting segment between the α A domains has a low hydrogen exchange rate. This observation is compatible with the high order α -helical structure in our model of α A.

• The C-terminal lysine of α B crystallin is an amine donor substrate for tissue transglutaminase (57) and, therefore, is on the surface as observed in our model of α B and the complex.

Significance

The 3D structure of α -crystallin is basic to our understanding of the molecular mechanisms involved in fiber cell maturation, stabilization of the lens nuclear region and the maintenance of lens transparency. On a more global level, it is basic to the function of small hsp in many cell types of various species. The fact that the subunits share homology with small heat shock proteins (8) and α B is known to interact with other heat shock proteins *in vivo* (22) provides an opportunity to probe the structure-function relationship of these very important polypeptides. Our working models of the subunits, their complex and the future construction of a more precise quaternary structure will provide the framework for a myriad of studies in our own as well as other laboratories.

Acknowledgment

This research was supported by an NEI Shannon Award. Additional support was from Research to Prevent Blindness, Inc. and the Lions Sight Foundation of N.J. We are indebted to Dr. M.C. Reddy and David Palmisano, graduate student, for their valuable participation.

Literature Cited

1. Bloemendal, H. *Molecular and Cellular Biology of the Eye Lens*; John Wiley and Sons: New York, 1981; pp 221-278.
2. Wistow, G.J.; Piatigorsky, J. *Annu. Rev. Biochem.* **1988**, *57*, 479-504.
3. Bloemendal, H.; de Jong, W.W. In *Progress in Nucleic Acid Research and Molecular Biology*; Cohn, W.E. and Moldave, K., Eds.; Academic Press: San Diego, CA, 1991, Vol. 41; pp 259-281.
4. Piatigorsky, J. *J. Biol. Chem.* **1992**, *267*, 4277-4280.
5. Wistow, G.J.; Piatigorsky, J. *Science* **1987**, *236*, 1554-1556.

6. Schoenmakers, J.G.C.; Bloemendal, H. *Nature (London)* **1968**, *220*, 790-791.
7. Wistow, G. *FEBS. Lett.* **1985**, *181*, 1-6.
8. Ingolia, T.D.; Craig, E.A. *Proc. Natl. Acad. Sci. USA. Genetics* **1982**, *79*, 2360-2364.
9. Chiesa, R.; Gawinowicz-Kolks, M.A.; and Spector, A. *J. Biol. Chem.* **1987**, *262(4)*, 1438-1441.
10. Chiesa, R.; Gawinowicz-Kolks, M.A.; Kleiman, N.J.; Spector, A. *Biochem. Biophys. Res. Com.* **1987**, *144*, 1340-1347.
11. Farnsworth, P.N.; Groth-Vasselli, B.; Van Inwegen, J.; Macdonald, J. Christopher; Mathur, R.; Reddy, M.C. In *Eye Lens Membranes and Aging*; G.F.J.M.Vrensen and J. Clauwaert, Eds.; EURAGE: Leiden, Netherlands, 1991, Vol. 15; pp 303-317.
12. Morgan, C.F.; Schleich, T.W.; Caines, G.H.; Farnsworth, P.N. *Biochem.* **1989**, *28*, 5065-5074.
13. Spector, A.; Garner, M.H.; Garner, W.H.; Roy, D.; Farnsworth, P.N.; Shyne, S. *Science* **1979**, *204*, 1323-1326.
14. Farnsworth, P.N.; Burke, P.; Wagner, B.J.; Fu, S.C.-J.; Regan, J. *TIB.* **1980**, *6*, 133-136.
15. van den Oetelaar, P.J.M.; van Someren, P.F.H.M.; Thomson, J.A.; Siezen, R.J.; Hoenders, H.J. *Biochem.* **1990**, *29*, 3488-3493.
16. Klemenz, R.; Frohli, E.; Steiger, R.H.; Schafer, R.; Aoyama, A. *Proc.Natl. Acad. Sci. USA.* **1991**, *88*, 3652-3656.
17. Iwaki, T.; Wisniewslid, T.; Iwaki, A.; Corbin, E.; Tomokane, N.; Tateishi, J.; Goldman, J.E. *Amer. J. Path.* **1992**, *140*, 345-356.
18. Iwaki, T.; Iwaki, A.; Miyazono, M.; Goldman, J.E. *Cancer* **1991**, *68*, 2230-2240.
19. Klemenz, R.; Andres, A.C.; Frohli, E.; Schafer, R.; Aoyama, A. *J. Cell Biol.* **1993**, *120(3)*, 639-645.
20. Srinivasan, A.N.; Nagineni, C.N.; Bhat, S.P. *J. Biol. Chem.* **1992**, *267(32)*, 23337-23341.
21. Bhat, S.P.; Horwitz, J.; Srinivasan, A.; Ding, L. *Eur. J. Biochem.* **1991**, *102*, 775-781.
22. Merck, K.B.; Groenen, P.J.T.A.; Voorter, C.E.M.; Horwitz, J.; Bloemendal, H.; de Jong, W.W. *J. Biol. Chem.* **1993**, *268(2)*, 1046-1052.
23. Augusteyn, R.C.; Koretz, J.F. *FEBS Lett.* **1987**, *222*, 1-5.
24. Weiner, S.J.; Kollman, P.A.; Nguyen, D.T.; Case, D.A. *J. Comput. Chem.* **1986**, *1*, 230.
25. Nemerson, Y.; Kumosinski, T.F.; Curley, D.; Liebman, M.N.; Konigsberg, W.H.; Guha, A.; Ross, J.B.A. *Biophys. Jr.* **1992**, *61*, A65.
26. Garnier, J.; Osguthorpe, D.J.; Robson, B. *J. Mol. Biol.* **1978**, *120*, 97-120.

27. Carver, J.A.; Aquilina, A.; Truscott, R.J.W.; Ralston, G.B. *FEBS. Lett.* 1992, 311(2), 143-149.
28. Maiti, M.; Kono, M.; Chakrabarti, B. *FEBS Lett.* 1988, 236, 109-114.
29. Castoro, J.A.; Bettelheim, F.A. *Lens Eye Toxic. Res.* 1989, 6, 781-793.
30. Chiou S.-H.; Azar, P. *J. Prot. Chem.* 1989, 8, 1-7.
31. Spector, A.; Li, L.-K.; Augusteyn, R.G.; Schneider, A.; Freund, T. *Biochem. J.* 1971, 124, 337-343.
32. Kramps, H.A.; Stols, A.L.H.; Hoenders, H.J.; de Groot, K. *Eur. J. Biochem.* 1975, 50, 503-509.
33. Kumosinski, T.F.; Farrell, H.M., Jr. *Biophys. Jr.* 1993, 64, A171.
34. Lamba, O.P.; Borchman, D.; Sinha, S.K.; Shah, J.; Renugopalakrishnan, V.; Yappert, M.C. *Biochem. Biophys. Acta.* 1993, BXP34482, 1-11.
35. Timasheff, S.M.; In *Protides of the Biological Fluids*; 20th Colloquium; Ed. H. Peeters; 1973, p. 511-519.
36. Kumosinski, T.; Farrell, H.M. *Trends Food Sci. & Tech.* 1993, 4, 169-175.
37. Li, L.-K.; Spector, A. *Exp. Eye. Res.* 1974, 19, 49-57.
38. Horwitz, J. *Exp. Eye. Res.* 1976, 23, 471-481.
39. Reddy, M.C.; Palmisano, D.; Groth-Vasselli, B.; Farnsworth, P.N. New Jersey Medical School, unpublished data, 1993.
40. Merck, K.B.; Groenen, P.J.T.A.; Voorter, C.E.M.; de Haard-Hoekman, W.A.; Horwitz, J.; Bloemendal, H.; de Jong, W.W. *J. Biol. Chem.* 1993, 268(2), 1046-1052.
41. Bloemendal, M., van Amerongen, H., Bloemendal, H. and van Grondelle, R. *Eur. J. Biochem.* 1989, 184, 427-432.
42. Summers, L.; Slingsby, C.; White, H.; Narebor, M.; Moss, D.; Miller, L.; Mahadevan, D.; Lindley, P.; Driessen, H.; Blundell, T.; et al.; Human Cataract Formation (Ciba Foundation Symposium); Pitman; London, 1984, Vol. 106; pp. 219-236.
43. Argos, P.; Siezen, R.J. *Eur. J. Biochem.* 1983, 131, 143-148.
44. Siezen, R.J.; Owen, E.A.; Kubota, Y.; and Ooi, T. *Biochim. Biophys. Acta* 1983, 748, 49-55.
45. Wistow, G.J.; Piatigorsky, J. *Ann. Rev. Biochem.* 1988, 57, 479-504.
46. Stauffer, J.; Rothschild, C.; Wandel, T.; Spector, A. *Invest. Ophthalmol.* 1974, 13, 135-146.
47. deJong, W.W.; van Kleef, F.S.M.; Bloemendal, H. *Eur. J. Biochem.* 1974, 48, 271-276.
48. van Kleef, F.S.M.; de Jong, W.W.; Hoenders, H.J. *Nature.* 1975, 258, 264-266.
49. Augusteyn, R.C.; Koretz, J.F.; and Schurtenberger, P. *Biochim. Biophys. Acta.* 1989, 99, 293-299.
50. Radlick, L.W.; Koretz, J.F. *Biochim. Biophys. Acta.* 1992, 1120, 193-200.

51. Augusteyn, R.C.; Ghiggino, K.P.; Putilina, T. *Biochim. Biophys. Acta* **1993**, *116*, 61-71.
52. Koretz, J.; Augusteyn, R.C. *Curr. Eye Res.* **1988**, *7*, 25-30.
53. Boyle, D.; Gopalakrishnan, S.; Takemoto L. *Biochem. Biophys. Res. Commun.* **1993**, *192(3)* 1147-1154.
54. Horwitz, J. *Invest. Ophthalmol. Vis. Sci.* **1993**, *34(1)*, 10-19.
55. Smith, J.B.; Thevenon-Emeric, G.; Smith, D.L.; Green, B. *Anal. Biochem.* **1991**, *193*, 118-129.
56. Smith, D., University of Purdue, personal communication, 1993.
57. Groenen, P.J.T.A.; Bloemendal, H.; de Jong, W.W. **1992**, *Eur. J. Biochem.* *205*, 671-674.

RECEIVED March 29, 1994

Chapter 10

Three-Dimensional Energy-Minimized Model of Human Type II "Smith" Collagen Microfibril

James M. Chen^{1,3} and Adrian Sheldon²

Departments of ¹Chemistry and ²Enzyme Biochemistry, OsteoArthritis Sciences, Inc., 1 Kendall Square, Building 200, Cambridge, MA 02139

The development of a molecular model of a type II collagen "Smith" microfibril is described. The model is a complex of five individual collagen triple-helical molecules, and is based on structural parameters known for collagen. Advantages of these three-dimensional models are that the stereochemistry of all the sidechain groups is accounted for and specific atomic contacts or interactions between atoms can now be studied. This model is useful for: development of therapeutics for collagen related diseases; development of synthetic collagen tissues; design of chemical reagents (i.e., tanning agents) to treat collagen-related products; and understanding the structure-function aspects of collagen folding, stability and interaction.

1. Background

Collagens, whose main functions are to provide an extracellular scaffold, are the most abundant mammalian proteins. The macromolecular proteins are the major structural components of skin, cartilage, tendons and ligaments, blood vessels, cornea and bone. 19 types of collagen molecules with varying amino acid sequences have been described. Types I, II and III are known as fibril-forming collagens, with type I primarily found in skin, type II found in cartilage, and type III found in blood vessels (1-4). These three collagen types form triple-helical molecules, unlike the other types which also contain regions that are non-triple-helical.

Each fibril-forming collagen polypeptide chain is about 1050 amino acid residues in length, including telopeptides, and assumes a left-handed helical secondary conformation, with 3.3 residues per turn. Type I collagen triple helices contain two $\alpha 1$ (typeI) chains and one, homologous but distinct, $\alpha 2$ (typeI) chain [described as $[\alpha 1(I)]_2\alpha 2(I)$] wound into a right-handed triple helix, whereas the type II collagen triple helix is comprised of three $\alpha 1$ (typeII) polypeptides. This triple helix structure is a semi-flexible rod-shaped complex, the length and diameter of which are approximately 300nm and 1.3nm, respectively. These triple helices pack together to form microfibrils, which in turn associate to form collagen fibers (approximately 3nm and 20-200nm diameter, respectively). The microfibril unit exists *in vitro*, and

³Current address: DuPont Agricultural Products, Stine-Haskell Research Center, P.O. Box 30, Building 300, Newark, DE 19714

0097-6156/94/0576-0139\$10.16/0

© 1994 American Chemical Society

is proposed to be an intermediate during the formation of fibers (5-7). The variability in fiber size is thought to result mainly from tissue-specific differences in intermolecular crosslinking and chemical structure.

Although the amino acid sequences of collagen polypeptides are complex, a repeated Gly-X-Y tripeptide motif is apparent (for example, in types I, II and III); approximately 33% of the residues are glycine and 25% are proline/hydroxyproline. Glycine is important to the structure since it is sufficiently compact to pack inside the central portion of each triple helix. Hydroxyproline, formed by post-translational modification, is also important in stabilizing both the triple helix and fiber via intrachain, interchain, or water-bridged hydrogen bond interactions (8). These interactions help to make the collagen molecule stable at mammalian body temperature.

Collagens exhibit a high degree of polymorphism (1, 9, 10). Since they are ubiquitous and multipurpose structural proteins, they can form a diverse range of fibrillar structures *in vivo*. Collagen *in vitro* has been observed to form various structures such as segment long spacing crystals, fibrillar long spacing aggregates, and obliquely banded and nonbanded fibrils. Many of these forms have been analyzed using the techniques of x-ray diffraction, freeze fracture and electron microscopy (9, 11-13). These and other studies have provided much information on the three-dimensional structure of collagen packing.

The longitudinal packing arrangement has been characterized, but the nature of the lateral packing of the collagen triple helices within the microfibrils has not yet been well defined. Although x-ray diffraction and electron microscopic analysis indicate that the order of packing may depend both on the type and function of the tissue (10, 12) and sample preparation (14), previous studies had shown that the lateral packing has crystalline properties (15-18). In light of this, models of the microfibril have been proposed, such as the Smith five-stranded helical microfibril (19) and that proposed by Veis and Yuan (four-stranded; 20). In these models, the collagen molecules are staggered by 1 D-unit. The length of a D-period is approximately 67nm, corresponding to about 234 amino acid residues (21, 22). Each triple helix, when arranged in the microfibril, has a stagger of slightly less than one-quarter relative to each other; that is, each triple helix is displaced by 0, 1, 2, 3 or 4 D-periods relative to laterally packed adjacent molecules, where $D=1/4.4$ of the molecular length (see 23, 24 for review and diagrams).

The Smith model is able to explain the negative staining banding patterns of transverse collagen fibril sections. The light bands correspond to regions of more dense lateral packing where adjacent collagen molecules overlap laterally. The length of this "overlap" is about 0.4D (19, 25). The dark bands correspond to "gap" regions, domains of low density molecular packing noted by Hodge and Petruska (25), where a separation exists between adjacent collagen molecules along the same longitudinal axis; this end-to-end separation is approximately 0.6D (19, 25). Thus, no end-to-end interactions occur between adjacent collagen molecules along the longitudinal axis. The above model emphasizes a rope-like structure for the microfibril with an overall left-handed supercoil of pitch 20D/11 (i.e. between 115-200nm in length; 26-28).

Other microscopy studies imply, however, that the lateral packing arrangement of collagens has properties which are less crystalline and more liquid-like (21, 29). The octafibril model is one such fluid model (21). It is proposed by this model that there is no intermediate substructure; this is supported by ^2H and ^{13}C NMR data showing that there is significant mobility in the intermolecular collagen interactions (30-32).

Models which emphasize crystalline properties of collagen packing are the quasihexagonal (33-35) and five-stranded "compressed" microfibril model (36). The quasihexagonal molecular crystal model was proposed based on the observed

periodicity from optical diffraction analyses of electron micrographs, which suggested concentrically oriented crystalline domains. Collagen molecules simply packed in a cylindrical, hexagonal lateral array do not produce the correct X-ray reflections, but modification of the lattice spacings and tilting of the collagen molecules by 4-5° to the fibril axis ("quasi-hexagonal") does give rise to the desired X-ray diffraction pattern. The compressed microfibril model mentioned above contains five collagen strands compressed laterally; this distortion results in the microfibril occupying a unit cell similar to that of the quasi-hexagonal arrangement.

In 1991, Chen et al. (37) developed an energy-minimized three-dimensional collagen microfibril model using molecular modeling techniques. Their goal was to develop a model, based on the earlier "Smith" microfibril model (19), in order to describe both intra- and inter-fibrillar interactions. This recent energetic model consisted of 5 triple helices symmetrically packed in a left-handed superhelical arrangement; the polypeptide sequence used was 15(Gly-Pro-Hyp)₁₂. Analysis demonstrated that van der Waals interactions are important in microfibril formation, and that electrostatic interactions are important in microfibril stability and the specificity by which collagen molecules pack within the structure. A preliminary fibrillar model of bovine type I collagen, incorporating the primary amino acid sequence for this collagen, was described by Chen et al. (23, 24).

To date, no analogous computer model has been developed for type II collagen. This collagen type is the primary form found in cartilage; other collagens (e.g., type IX) are minor components in this tissue. Modeling of type II collagen is attractive since the 3 helical chains are identical; i.e., the structure is described as $[\alpha 1(\text{II})]_3$. Determination of the structure of type II collagen is of intense interest since pathological conditions such as arthritis involve the degradation of this structure, leading to loss of integrity of this load-bearing tissue. A microfibril model for type II collagen (and other forms) can be created by substituting the native amino acid sequence into linked 15(Gly-Pro-Hyp)₁₂ templates mentioned above. Furthermore, since a single D-spacing is the repeating unit making up collagen fibers, a native microfibril model can be constructed simply by using this unit as a building block; it is an important concept that all of the possible intra- and inter-molecular interactions in collagen can be incorporated into this repeating unit. The construction of such a model would permit a three-dimensional analysis of amino acid sidechain interactions which contribute to triple helical and microfibril interaction and stability, crosslink sites of importance, and exterior surface contour and properties. For example, study of the exterior surface could depict regions or sites where proteolytic enzymes (e.g., collagenase) and other molecules or synthetic compounds can interact with collagen type II.

This chapter describes the construction of a microfibril model containing the repeating motif, a D-spacing unit, of native type II collagen. The approach used was to construct a 15(Gly-Pro-Hyp)₃₀₀ template model. The primary amino acid sequence of human type II collagen was then substituted into the 15(Gly-Pro-Hyp)₃₀₀ model, followed by structural refinement methods. Each of the polypeptide chains within the final fibrillar structure has a length of 300 amino acids, corresponding to about 1.3D; the reason for choosing this particular length was that it allowed for the construction of a unit that contained a single gap region flanked by overlap regions. The characteristics of this model are described, and its features are compared with experimental data.

2. Methods

2.1. Potential Energy Function.

Molecular modeling was performed on an SGI Crimson Elan workstation (Silicon Graphics, Inc.). Calculations were performed using the molecular modeling software SYBYL, version 5.5, developed by TRIPOS Associates, Inc. (1992). SYBYL v.5.5 contains a combination of computational tools for efficient and reliable modeling of both large proteins and small molecules. For this work, emphasis was placed upon utilizing the Biopolymer methods within SYBYL for modeling of protein structures. The molecular mechanics method along with the Kollman force-fields (38) were used to refine the protein structures constructed. The conjugate gradient method for energy minimization was used to minimize the potential energy of the proteins. For the collagen fibril structures, the United Atoms approach was utilized in order to improve the efficiency of all calculations since a single fibrillar model contains over thirty thousand atoms (38; SYBYL, v.5.5). Nonbonded interactions were not computed beyond a distance of 8Å from each atom. The 1-4 interaction terms were given a scaling factor of 0.5 in accordance with Weiner et al. (39). Water molecules were not explicitly included in order to account for solvation, but a distance-dependent dielectric function, $D = (R_{ij} + 1)$, where D is the dielectric function and R_{ij} is the distance between atoms i and j , was used to implicitly account for solvation effects, as all calculations were carried out *in vacuo*. The amino- and carboxyl-termini contained N-acetyl and NHCH_3 groups, respectively, to minimize any possible end effects. The convergence criterion for all energy minimizations was a root-mean-square (rms) derivative of 0.01kcal/mol-Å.

2.2. Microfibril Modeling Strategy.

The general modeling strategy involved the construction of a "Smith" 15(Gly-Pro-Hyp)₃₀₀ microfibril model starting from a shorter 15(Gly-Pro-Hyp)₁₂ microfibril model (37), followed by incorporation of the human type II collagen amino acid sequence into the 15(Gly-Pro-Hyp)₃₀₀ model. The final stages consisted of structural refinement using both interactive graphics manipulations and energy minimization methods. Details for these procedures are given below.

2.3. Backbone and Sidechains.

An intermediate step in constructing a model of the type II fiber was the incorporation of the complete collagen type II amino acid sequence into the 15(Gly-Pro-Hyp)₃₀₀ template model. Substitutions were then made for each amino acid type into all their respective positions along the three-dimensional fibril template of 15(Gly-Pro-Hyp)₃₀₀. Before energy minimizing the modified model, unfavorable steric contacts between sidechain atoms were removed by an algorithm which rotated all the sidechain torsional angles in an iterative manner until no further bad contacts were found (SYBYL, v.5.5). Energy minimization of the structure with its corresponding sidechains was carried out in a two step process, described in detail later. First, all the polypeptide backbone atoms were constrained to remain fixed in their original positions while only the sidechains were allowed to move during energy minimization. Finally, the polypeptide backbone constraints were removed and the potential energy of the backbone and sidechains were minimized.

2.4. Molecular Dynamics Simulations.

As a final step in the structural refinement procedure, molecular dynamics was performed at different stages of the microfibril construction in order to "repair" and structurally refine regions in the model which had undergone modifications. For example, when amino acid substitutions are made at specific positions in the models, both the backbone and sidechain parameters for the new residues had to be re-minimized. This was because the use of energy minimization alone may not necessarily be effective in allowing a system (i.e., the modified structure) to escape from a local energy minimum previously defined for the unmodified structure; hence, molecular dynamics was used to better and more thoroughly explore the possible conformations allowed for each new substitution and/or modification. Constrained molecular dynamics was also performed on the sidechain groups of the type II microfibril model for optimization of sidechain contacts or interactions. As described in the Potential Energy Function section, the Kollman force field with the United atoms parameter set was used (SYBYL, v.5.5). The list of non-bonded interactions was updated every 25 femtoseconds and the non-bonded cutoff distance was set at 8.0Å. Time steps used were 1 femtosecond and the temperature was changed according to the "annealing" procedure (SYBYL, v.5.5). Simulations were carried out until system equilibration was obtained. The final equilibrated model was re-minimized in order to derive the final minimized model. Due to the large size of these molecular models, no attempts were made to perform rigorous (i.e., greater than 100 picoseconds after system equilibration) molecular dynamics simulations as this was not the purpose of this initial study.

2.5. Construction of the 15(Gly-Pro-Hyp)₃₀₀ Microfibril Model.

The energy-minimized "Smith" model of the 15(Gly-Pro-Hyp)₁₂ microfibril was used as the starting template (23, 37) for building the final 15(Gly-Pro-Hyp)₃₀₀ microfibril template model. The 15(Gly-Pro-Hyp)₁₂ molecule was duplicated and aligned longitudinally. Following removal of the appropriate end groups (see Potential Energy Section) prior to coupling, the molecules were docked and positioned such that no unfavorable van der Waals contacts existed, and the length of each polypeptide chain was modified slightly (i.e., some C-terminal residues were "cropped" to place each chain's twist conformation into register) in order to maintain the specific left-handed helical conformation characteristic of collagen polypeptides. A covalent amide bond was then created joining the two molecules. In the molecular modeling process, this intermediate structure was a 15(Gly-Pro-Hyp)₂₂ extended microfibril structure. The structure then underwent structural refinement using both energy minimization and molecular dynamics methods (SYBYL, v.5.5).

The next stage of the microfibril construction involved modifying the 15(Gly-Pro-Hyp)₂₂ structure into a 15(Gly-Pro-Hyp)₄₂ structure, in the same manner as described above. This procedure was repeated until an energy-minimized microfibril of slightly over 300 residues per chain was constructed (Scheme 1a). As mentioned above, the removal of some unnecessary 15(Gly-Pro-Hyp)_n tripeptide segments was carried out; this is also shown in Scheme 1b and described in the Results and Discussion section. The final product exhibited symmetrical packing (C₅ rotational symmetry around the longitudinal axis) as the fiber was created based on the Smith model for collagen packing.

2.6. Incorporation of the Native Human Type II Sequence.

The model for the human type II collagen fiber was constructed by substituting the primary amino acid sequence in place of the Gly-Pro-Hyp model (Scheme 1b). The type II collagen sequence is known to contain most of the standard amino acids in addition to some post-translationally modified residues (40). The complete sequence containing 1014 amino acids per $\alpha 1$ chain is represented in a single D-space unit of

Scheme 1a

	300 Residues Length	
TH1	1	300
TH2	1	300
TH3	1	300
TH4	1	300
TH5	1	300

Scheme 1b

	300 Residues Length		
TH1	1		300
TH2	1	78	235
TH3	1		300
TH4	1		300
TH5	1		300

OVERLAP_n
GAP_n
OVERLAP_{n+1}

this model. It is important to note that because this model is a repeating unit of the collagen fiber, it is possible to study all of the relevant inter- and intra-molecular sidechain interactions. Since type II collagen contains 3 identical $\alpha 1$ chains, the microfibril contains both a C_3 (triple helices) and a C_5 (packing of triple helices) axis of symmetry. The fully constructed molecular model contains 2 overlap regions and 1 gap region as shown in Scheme 1b.

2.7. Structural and Energetic Refinements.

The type II Smith collagen microfibril model underwent a 3 step procedure for overall structural refinement. First, the orientation of sidechain atoms which made unfavorable van der Waals interactions with backbone or neighboring sidechains were modified by an algorithm that rotated all the sidechain torsional angles in an iterative manner until such interactions were removed. Second, a (subjective) visually-based procedure was performed interactively on the molecular modeling graphics workstation in order to further optimize sidechain interactions. During this step, the entire microfibril model was analyzed for unfavorable energetic interactions such as similarly-charged ligands positioned closely together; when such interactions were found, sidechain torsional angles were modified toward a more favorable orientation. The third structural refinement step involved the initial placement of force-constraints (SYBYL, v.5.5) upon the polypeptide backbone of the microfibril

followed by energy minimization. This "constrained-minimization" step was then followed by removal of the force-constraints from the polypeptide backbone which then allowed for the final and complete energy minimization of the whole structure.

3. Results and Discussion

3.1. Molecular Modeling of the 15(Gly-Pro-Hyp)₃₀₀ Microfibril.

The goal was to create a single three-dimensional model which represented a repeating motif in fiber-forming collagens. Negative staining patterns of transverse fiber sections (25) show that the length of fiber-forming collagens consists of a repeating unit referred to as a D-space. This D-space contains a single "overlap" and "gap" region (1, 9). The approximate length defined for a single D-space is 234 amino acid residues (21, 22). Although the repeating unit of a fiber is considered to be a single D-spacing, our microfibril structure corresponds to a length of 300 amino acid residues. This chain length represents a type II microfibril containing a single gap region surrounded on each end by overlap regions (Scheme 1b). Since the N- and C-terminus telopeptides interact at each end of the gap region, the surrounding overlap regions were modeled to allow for the study of both the above sets of telopeptide segments. Following the construction and energy refinement of the molecular model for the complete 15(Gly-Pro-Hyp)₃₀₀ microfibril template, the gap region within the above template model was created by removal of a segment corresponding to 156 amino acid residues in length from a single triple helical collagen molecule in the microfibril template model (Scheme 1b).

3.1.1. Construction of the 15(Gly-Pro-Hyp)₃₀₀ Microfibril from a 15(Gly-Pro-Hyp)₁₂. The energy-minimized "Smith" microfibril model of 15(Gly-Pro-Hyp)₁₂ (23, 24, 37) was used in the initial buildup of the full microfibril model. First, the 15(Gly-Pro-Hyp)₁₂ structure was duplicated so that two separate structures existed graphically. Each microfibril structure was docked and positioned longitudinally so that the carboxyl terminus of one microfibril was adjacent to the amino terminus of the second duplicate microfibril. In addition, the removal of each polypeptide's end groups was required to form an amide bond in order to connect the two structures. Second, since each polypeptide chain in the collagen microfibril model contains a specific left-handed helical conformation (37), corresponding peptide chains at the connecting regions were not necessarily positioned correctly to form a continuous polypeptide left-handed helical conformation for the newly "extended" microfibril. Hence, to place the helices in register, the length of each polypeptide chain at the connecting points were "cropped" to an appropriate length prior to amide bond formation. After repositioning each microfibril segment such that the chains to be joined were as close as possible, with no energetically unfavorable van der Waals contacts, the corresponding chains were connected in order to form a single extended microfibril structure of 15(Gly-Pro-Hyp)₂₂.

Before the next stage of microfibril extension, the 15(Gly-Pro-Hyp)₂₂ structure underwent structural refinement using energy minimization and molecular dynamics methods. With the inclusion of end groups at each polypeptide's terminus, the complete 15(Gly-Pro-Hyp)₂₂ microfibril structure was energy minimized. To fully optimize the modified microfibril model, especially around the region where the chains were modified and connected, molecular dynamics refinement was performed. In the 15(Gly-Pro-Hyp)₂₂ model, terminal regions of the microfibril (residues 1:6 and 17:22 of each polypeptide chain; see diagram below) were held fixed using force-constraints. These fixed regions were not allowed to move during the molecular dynamics routine; this is schematically shown below (TH = triple

helix). Another reason for constraining each chain terminus was to prevent deviations from its original symmetrical configuration as modeled according to the Smith model (37); deviations of the tethered ends would complicate the docking procedures during additional molecular modeling processes for extending the microfibrillar segments.

Placement of Force-Constraints During Structural Refinement Using Molecular Dynamics

	FIXED	Microfibril Template of 15(Gly-Pro-Hyp) ₂₂	FIXED ^a
TH1	1	6	17 22
TH2	1	6	17 22
TH3	1	6	17 22
TH4	1	6	17 22
TH5	1	6	17 22

^aFIXED indicates regions containing a Force-Constraining Potential applied to the polypeptide backbone regions. These constraints prevent the regions (1:6 and 17:22) from deviating from their original positions during molecular dynamics simulations.

Molecular dynamics (SYBYL, v.5.5) was carried out as described in the Methods Section. Dynamics relaxation was carried out where the system temperature was raised from 0 to 100 degrees Kelvin and returned to 25 degrees Kelvin using increments of 25 degrees. Five picoseconds of calculations at each interval was given to insure a proper and smooth transition between each temperature change. At the final dynamics temperature of 25 degrees Kelvin, an additional 10 picoseconds was given for system equilibration. Finally, the resulting structure was energy-minimized without using force constraints. This general procedure of minimization and dynamics was used in the preparation of a starting structure for the next stage of microfibril extension, i.e., 15(Gly-Pro-Hyp)₄₂. The extended microfibril was again duplicated on the graphics workstation and connected to form an extended structure close to twice its original size. This procedure was repeated until an energy-minimized microfibril of slightly over 300 residues per chain was constructed. The target model of the microfibril template was achieved after removal of the unnecessary 15(Gly-Pro-Hyp)_n tripeptide sequences (segment corresponding to 79:234 in the TH2 molecule as seen in Scheme 1b). It is seen in the model that the overlap segments are regions where the collagen molecules are packed more densely (i.e., containing all five collagen helices) and the gap segments are regions where the molecules are less densely packed (i.e., containing only four of five helices; see Scheme 1b), consistent with experimental observations.

3.2. Molecular Modeling of the Type II Microfibril Model.

The native type II collagen sequence contains most of the known naturally occurring amino acids in addition to specifically modified amino acids such as hydroxyprolines. It is reasonable to assume that the type of inter-helical collagen interactions in fibers will differ depending upon which regions of each molecule are packed together. That is, regions in the microfibril model containing bulky sidechains (e.g., Phe, Tyr, Leu, Ile) will pack differently from regions which contain amino acids having smaller sidechains (e.g., Ala, Ser). It has been shown

experimentally that the diameter of fiber-forming collagens is not uniform throughout the length of the fiber, but changes in a consistent and repeating fashion along its axial length (41). One can also assume that the mode of collagen packing, whether more similar to the "compressed" (36) or "symmetrical" (19) model, will differ depending upon the type (e.g., skin, tendon, cartilage or bone) and state (e.g., increased or decreased water content) of the collagen tissue.

The strategy used herein to construct an accurate type II collagen model was to initially prepare the full 15(Gly-Pro-Hyp)₃₀₀ microfibrillar template before incorporating the actual type II collagen sequence. As explained previously, the 15(Gly-Pro-Hyp)₃₀₀ model is a symmetrical packing (i.e., contains a C₅ rotational symmetry around the microfibrillar longitudinal axis) of five identical collagen triple helical structures in a configuration known as a "Smith" microfibril (19). Since the sequence of each collagen molecule is a repeat of the consensus tripeptide sequence (Gly-Pro-Hyp), the microfibrillar diameter is constant along the axial length (i.e., ~30.0Å). This Smith microfibril model will therefore allow for comparison and easy visual detection of structural changes that may occur both as the native sequence of type II collagen is incorporated into the 15(Gly-Pro-Hyp)₃₀₀ model, and the modified type II model is then structurally refined using energy minimization methods. Furthermore, minimization of the modified microfibrillar structure occurs more efficiently since the interactions of each polypeptide's backbone and the specific interactions of the proline and hydroxyproline sidechains had already been energetically optimized for the Smith microfibril packing. The process of structural modifications carried out here also prevents gross structural deviations from the original microfibrillar packing or peptide conformations; this may occur during the energy minimization step if caution is not taken.

3.2.1. Incorporation of the Native Type II Collagen Sequence into the "Smith" Microfibril Model of 15(Gly-Pro-Hyp)₃₀₀. In Scheme 2a, it is evident that within the Smith microfibril unit, containing five collagen molecules, the complete primary amino acid sequence for human type II collagen (1014 amino acids per $\alpha 1$ chain) is represented in a single D-space unit. The microfibril D-space model represents an actual repeating unit; a single collagen fiber is simply a stacking arrangement of the Smith microfibril model along the lateral (i.e., fiber width) and longitudinal (i.e., fiber length) directions. Furthermore, this type II model will represent all the possible inter- and intra-molecular sidechain interactions which can occur between type II collagen molecules packed into fibrillar units.

Since type II collagen contains 3 identical $\alpha 1$ polypeptide chains, both a C₃ and C₅ rotational symmetry exist along the triple helical and microfibrillar longitudinal axis, respectively, as mentioned previously. The type II microfibril model permits the analysis of how a specific amino acid sidechain interacts with adjacent collagen molecules, and due to this C₃ axis of symmetry in the triple helix, one can also study how the same amino acid sidechain interacts along the exterior surface of the microfibril (see Figures 1a, b). Figures 1a and 1b depict the cross-sections of the microfibril model in the overlap and gap regions, respectively. The "R" groups represent amino acid sidechains. As shown in these diagrams, the positions of each specific "R" group is found both in the interior and the exterior of the fibril models; thus, this model allows for the study of how each sidechain is positioned on the fibril surface and how each sidechain interacts with adjacent sidechains found on other collagen molecules.

As a final modification, shown in Scheme 2b, the actual Smith microfibril model used for sequence substitution contained a gap region in the second (TH2) collagen molecule. Residues 79:234 were deleted from each polypeptide chain (see Scheme

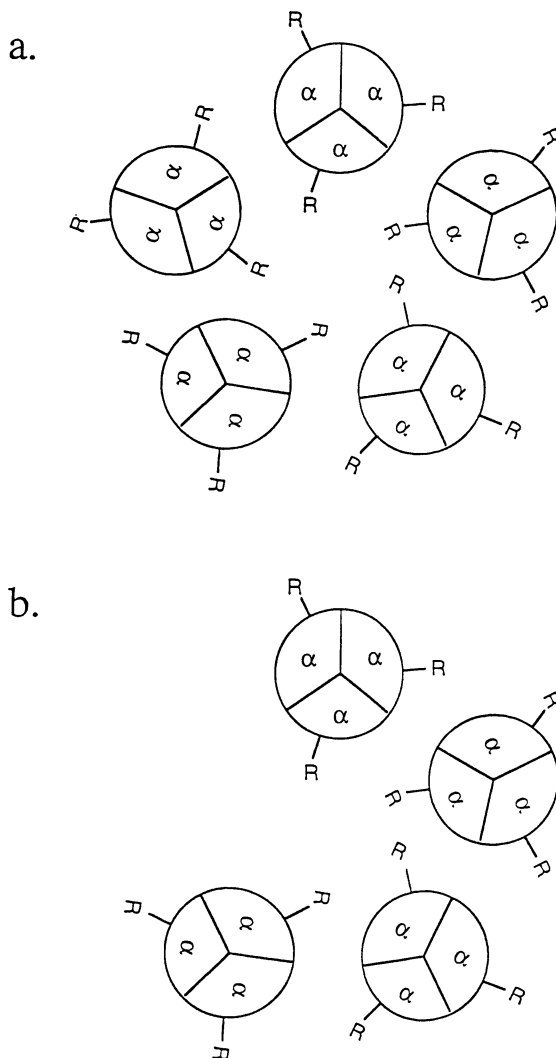


Figure 1. Cross-Sections showing the arrangement of collagen triple helices in the microfibril model. “ α ” represents a collagen polypeptide chain; “R” represents amino acid sidechains either exposed or buried within the microfibril center. a. Cross-section in overlap region containing five triple-helices; b. Cross-section in gap region containing four triple-helices.

Scheme 2a

<-----Single D-Space Unit----->

TH1	<u>1</u>	<u>234</u>
TH2	<u>937</u>	<u>1014</u>
TH3	<u>703</u>	<u>936</u>
TH4	<u>469</u>	<u>702</u>
TH5	<u>235</u>	<u>468</u>

OVERLAP_n GAP_n

Scheme 2b

<-----Modeled D-Space Unit----->

TH1	<u>1</u>	<u>234</u>	<u>300</u>
TH2	<u>937</u>	<u>1014</u>	<u>1</u> <u>66</u>
TH3	<u>703</u>	<u>936</u>	<u>1002</u>
TH4	<u>469</u>	<u>702</u>	<u>768</u>
TH5	<u>235</u>	<u>468</u>	<u>534</u>

OVERLAP_n GAP_n OVERLAP_{n+1}

1b) of the TH2 collagen molecule, and the modified structure was reminimized. Each polypeptide chain in the modified Smith microfibril was renumbered according to Scheme 2b. Amino acid substitutions into the microfibril structure were made for all the "X" and "Y" positions in the Gly-X-Y tripeptide repeats corresponding to the human type II collagen sequence. The orientations of all the glycines, prolines and hydroxyprolines found in the native type II sequence were left as is (that is, not substituted), since they had already been optimized using minimization methods during the original microfibril modeling. After the complete type II collagen sequence was incorporated into the microfibril template model, structural refinement was used to optimize both intra- and inter-helical interactions.

Telopectide chains are found within the gap region, carboxyl- and amino-terminal to amino acid residues 1014 and 1, respectively (Scheme 2b). The carboxyl-terminal telopeptides contain 19 amino acid residues and the N-terminal telopeptide contains 27 amino acid residues (40). Scheme 3 depicts the region in the type II microfibril where the set of telopeptides would reside (telopeptides are shown as xxxxx). Since it has been shown experimentally that these telopeptide segments are critical for fiber stability (42-44), the three-dimensional type II model provides important conformational constraints for future studies involving the prediction of telopeptide structures, which in turn will provide insight into how telopeptides stabilize fiber-forming collagens. The type II microfibril model described here does not contain these telopeptides but work is in progress to incorporate these structural domains using computational predictive methods.

3.2.2. Structural and Energetic Refinement of the Type II "Smith" Microfibril Model.

The type II Smith collagen microfibril model underwent a three step

Scheme 3

	←-----D-Space Unit----->		
TH1	<u>1</u>		<u>234</u> <u>300</u>
TH2	<u>937</u>	<u>1014</u> <i>XXXXXX</i>	<i>XXXXXX</i> ¹ <u>66</u>
TH3	<u>703</u>		<u>936</u> <u>1002</u>
TH4	<u>469</u>		<u>702</u> <u>768</u>
TH5	<u>235</u>		<u>468</u> <u>534</u>
	OVERLAP_n	GAP_n	OVERLAP_{n+1}

procedure for structural refinement similar to methods applied recently to other molecular models (45). First, sidechain atoms which had unfavorable van der Waals contacts with backbone or neighboring sidechains were reoriented by application of an algorithm that rotated all the sidechain torsional angles in an iterative manner until no further bad contacts were detected (SYBYL, v.5.5). The importance of this step prior to applying minimization methods is to prevent gross structural changes due to possible atomic overlaps produced during structural modifications and to increase the overall efficiency of the minimization routine. Second, also prior to applying the minimization methods, a subjective and interactive procedure was used on the molecular modeling workstation to further optimize sidechain interactions. Using three-dimensional stereoviewing, the entire type II microfibril model was scanned visually in order to determine if any unfavorable sidechain interactions such as interactions between similar charges existed. If such interactions were identified, sidechain torsional angles were adjusted before full structural refinement was applied using energy minimization methods. The rationale for this visual procedure was that, although the first step of the minimization routine uses an algorithm that attempts to remove bad contacts between overlapping atoms, this program may not place the substituted sidechains into the most favorable positions. Thus, this interactive work, utilizing the graphics capability of the program, is very important and increases the overall efficiency and performance of the energy minimizer, especially for a system containing over 37,000 "essential" atoms using the Kollman United Atoms force-field (39).

The third energy minimization step was performed in two separate stages. Initially, the polypeptide backbones of all chains within the microfibril model were constrained using force-constraints so that these regions were not allowed to deviate during energy minimization. This initial constraint step allowed for the optimization of only the sidechain conformations; the backbone conformation of the original 15(Gly-Pro-Hyp)₃₀₀ microfibril model remained unchanged. Since it was possible that some additional bad contacts may have been created during the interactive graphics work, the placement of these backbone force-constraints prevents gross structural changes within localized fibrillar regions during the energy minimization steps. After minimization of the sidechains, the backbone force-constraints were removed from the type II microfibril model and the whole structure was then reminimized.

3.3. The Energy-Minimized Type II Collagen Microfibril Model.

Consistent with the "Smith" microfibril model, Figure 2 contains the actual microfibril alignment of the three $\alpha 1$ chains of type II collagen. Collagen packing models are based on electron microscopy and X-ray diffraction analysis data. The two-dimensional linear arrangement of collagen molecules was determined by electron microscopy studies of both the negative and positive stained transverse sections of collagen fiber samples. Based on the negative staining pattern where tungstate or uranyl salts tend to deposit within regions where the collagen molecules are packed loosely, repeated intervals of light and dark bands are seen along the length of the stained fiber sections (1). A single D-interval is defined as a single set containing one light band and one dark band. Since the entire negatively-stained fiber has a pattern of repeating light and dark bands, the D-interval represents a repeating unit in fiber forming collagens. The present model explains this banding pattern as an array of adjacently aligned molecules staggered along their helical long axis by one D interval, where $4.4D$ intervals define the entire length of a single type I, II or III collagen triple helical molecule (1, 9).

Positively-stained fibers also demonstrate a pattern of distinct bands that are specific for the regions within the fiber packing where charged amino acid sidechains are clustered (1, 9). Within each D-interval, five sets of major bands are seen in transverse fiber sections prepared through positive staining methods. Within each set of the five major bands are also found specific sub-banding patterns. By comparing regions in the native collagen sequences where charged sidechains are localized with the experimental banding patterns of a positively-stained collagen sample, the length of a single D interval can be determined as 234 amino acid residues along a single polypeptide chain. Electron microscopy studies also indicate that the end-to-end alignment of collagen molecules contain a spacing referred to as a "gap" region, as explained previously. On the basis of the molecular model, this gap region corresponds to 156 amino acid residues in length ($0.67D$) and the overlap region corresponds to a length of 78 amino acids ($0.33D$).

In Figure 2, the two dimensional alignment of a full length triple-helical collagen molecule (labeled 1) and each corresponding region of four other adjacent molecules (labeled 2-5) are arranged as proposed by the Smith model, where five collagen triple helices are packed in a circular array while each adjacent molecule is staggered laterally by a $1D$ interval. The Smith model is consistent with experimental evidence where each collagen molecule (Figure 2), with respect to the adjacent molecules, are staggered laterally by 234 amino acid positions or a single D spacing unit (labeled D-Spacing). In the Smith microfibril model, the overlap region contains five collagen molecules whereas the gap region contains four collagen molecules per microfibril unit.

Table 1 contains the specific amino acid sequences aligned as shown in Figure 2. There are five sets of columns representing five collagen molecules. Each collagen molecule contains three individual polypeptide chains. As described for Figure 3, collagen molecule 1 is shown as the complete fiber forming sequence, 1014 amino acid residues in total, not including the chain terminal telopeptide segments. In this two dimensional alignment, one can determine which types of amino acid residues are within the vicinity of each other. Table 1 also allows one to infer which amino acids may form stabilizing interactions with each other, although one cannot identify specific atomic interactions without the actual three-dimensional models which provide both geometrical and stereochemical information.

The type II microfibril model represents a three-dimensional mapping of all the possible sidechain interactions and surface contours for each specific fiber-forming collagen. Structure-function studies are important for understanding how specific

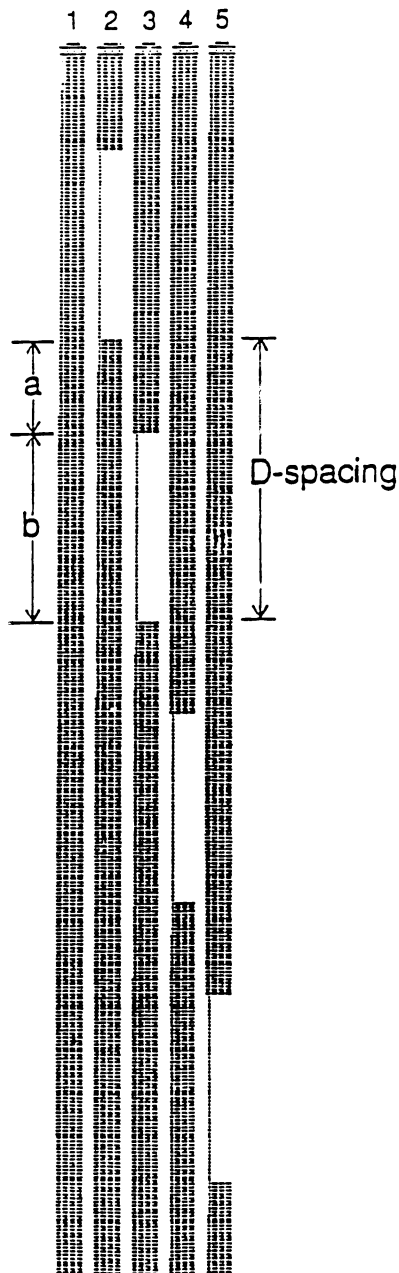


Figure 2. Microfibril alignment for $\alpha 1$ chains of type II collagen. Collagen triple helices are numbered 1 to 5. Overlap and gap regions are labelled "a" and "b," respectively.

amino acids contribute to fiber interactions and stability. As examples, studies of the inter-helical sidechain interactions may reveal possible reactive sites for synthetic modifying agents; cross-linking sites that occur naturally within fibers may also be studied (23, 24, 37). Analysis of the exterior surfaces of the fibril models may depict where certain proteins bind, i.e., proteolytic enzymes known to cleave collagens. Regions which interact with other collagen types or extracellular matrix components can also be studied using these models. Furthermore, telopeptides which are known to be essential for fiber stability can be modeled into the type II model since this structure may provide realistic spatial constraints for the conformational prediction of the telopeptides (23). The above are just some of the important properties of collagen which can be studied using these models. Although the recent synthetic microfibrillar models (23) proposed are sufficient for understanding the general structural aspects of collagen packing, the models containing the native sequences will yield structure-function relationships directed towards specific collagen types. Thus, the three-dimensional models of real collagen fibers are very important for studying realistic collagen systems.

3.4. Applications of the Three-Dimensional Microfibril Model for Type II Collagen.

Figure 3 depicts the same type II collagen microfibril alignment as shown in Figure 2 but where a single type IX collagen molecule is also shown relative to the type II microfibril. The chain direction of the type IX molecule is known to be anti-parallel to the chain direction of the type II collagens (46). Type IX collagen contains both regions which are able to form triple-helical structures, and regions that are non-triple-helical. Other points of interest are labelled such as: 1) sites where covalent cross-links exist between types II and IX collagens; 2) sites where matrix metalloenzymes cleave collagens; 3) regions where type IX collagens may make non-covalent contacts with type II collagens (i.e., overall alignment between both molecules in Figure 3); 4) regions where glycosaminoglycans interact with the collagens or make natural covalent cross-links; and 5) sites within the gap regions where telopeptides are known to interact for maintaining fibril stability [shown as curved lines at both the carboxyl- and amino-terminus regions of type II collagens]. Thus, Figure 3 depicts the general schematics of molecular alignments between two collagens of interest and indicates where specific sites occur, in a two-dimensional format. The diagram in Figure 3 also shows specific regions which may form inter- or intra-helical interactions. But an important disadvantage with a representation such as this is that one cannot study these interactions at the atomic level. However, since we have created a three-dimensional structural model representing two "a" units (overlap regions) and one "b" unit (gap region), we have the capability to study each site as identified in Figure 3 at its three-dimensional atomic level and detail.

3.4.1. Telopeptide Structure and Analysis. In the molecular model of type II collagen microfibrils, we can study the "pocket"-like cleft where telopeptides interact and pack. These three-dimensional models provide for the structural constraints which can then be used to determine the possible bioactive conformations of the three neighboring telopeptides. The interactions between telopeptides must be such that they not only stabilize each other, but they also form a tripeptide complex such that the complex fits within the "cleft" or pocket-like cavity which is formed within the gap region in the type II fiber models. Most of the recently published experimental studies using NMR (47-50), energetic model building (44) and conformational analysis (50) involve the study of a single telopeptide chain. The preferred conformation found for a single telopeptide chain

Table 1. Amino acid sequences of type II collagen aligned as in Figure 2. The five sets of columns represent the five collagen triple helices.

	COLLAGEN 1			COLLAGEN 2			COLLAGEN 3			COLLAGEN 4			COLLAGEN 5						
	A1	A1	A1	A1	A1	A1	A1	A1	A1	A1	A1	A1	A1	A1	A1				
1	GPM	GPM	GPM	313	GLQ	GLQ	GLQ	235	GRV	GRV	GRV	157	GPP	GPP	GPP	79	GPR	GPR	GPR
2	GPM	GPM	GPM	314	GLP	GLP	GLP	236	GPP	GPP	GPP	158	GPP	GPP	GPP	80	GPP	GPP	GPP
3	GPR	GPR	GPR	315	GPP	GPP	GPP	237	GSN	GSN	GSN	159	GEG	GEG	GEG	81	GPQ	GPQ	GPQ
4	GPP	GPP	GPP	316	GPS	GPS	GPS	238	GNP	GNP	GNP	160	GKP	GKP	GKP	82	GAT	GAT	GAT
5	GPA	GPA	GPA	317	GDQ	GDQ	GDQ	239	GPP	GPP	GPP	161	GDQ	GDQ	GDQ	83	GPL	GPL	GPL
6	GAP	GAP	GAP	318	GAS	GAS	GAS	240	GPP	GPP	GPP	162	GVP	GVP	GVP	84	GPK	GPK	GPK
7	GFQ	GFQ	GFQ	319	GPA	GPA	GPA	241	GPS	GPS	GPS	163	GEA	GEA	GEA	85	GQT	GQT	GQT
8	GFQ	GFQ	GFQ	320	GPS	GPS	GPS	242	GKD	GKD	GKD	164	GAP	GAP	GAP	86	GEP	GEP	GEP
9	GNP	GNP	GNP	321	GPR	GPR	GPR	243	GPK	GPK	GPK	165	GLV	GLV	GLV	87	GIA	GIA	GIA
10	GEP	GEP	GEP	322	GPP	GPP	GPP	244	GAR	GAR	GAR	166	GPR	GPR	GPR	88	GFK	GFK	GFK
11	GEP	GEP	GEP	323	GPV	GPV	GPV	245	GDS	GDS	GDS	167	GER	GER	GER	89	GEO	GEO	GEO
12	GVS	GVS	GVS	324	GPS	GPS	GPS	246	GPP	GPP	GPP	168	GFP	GFP	GFP	90	GPK	GPK	GPK
13	GPM	GPM	GPM	325	GKD	GKD	GKD	247	GRA	GRA	GRA	169	GER	GER	GER	91	GEP	GEP	GEP
14	GPR	GPR	GPR	326	GAN	GAN	GAN	248	GEP	GEP	GEP	170	GSP	GSP	GSP	92	GPA	GPA	GPA
15	GPP	GPP	GPP	327	GIP	GIP	GIP	249	GLQ	GLQ	GLQ	171	GAQ	GAQ	GAQ	93	GPQ	GPQ	GPQ
16	GPP	GPP	GPP	328	GPI	GPI	GPI	250	GPA	GPA	GPA	172	GLQ	GLQ	GLQ	94	GAP	GAP	GAP
17	GKP	GKP	GKP	329	GPP	GPP	GPP	251	GPP	GPP	GPP	173	GPR	GPR	GPR	95	GPA	GPA	GPA
18	GDD	GDD	GDD	330	GPR	GPR	GPR	252	GEK	GEK	GEK	174	GLP	GLP	GLP	96	GEE	GEE	GEE
19	GEA	GEA	GEA	331	GRS	GRS	GRS	253	GEP	GEP	GEP	175	GTP	GTP	GTP	97	GKR	GKR	GKR
20	GKP	GKP	GKP	332	GET	GET	GET	254	GDD	GDD	GDD	176	GTD	GTD	GTD	98	GAR	GAR	GAR
21	GKA	GKA	GKA	333	GPA	GPA	GPA	255	GPS	GPS	GPS	177	GPK	GPK	GPK	99	GEP	GEP	GEP
22	GER	GER	GER	334	GPP	GPP	GPP	256	GAE	GAE	GAE	178	GAS	GAS	GAS	100	GGV	GGV	GGV
23	GPP	GPP	GPP	335	GNP	GNP	GNP	257	GPP	GPP	GPP	179	GPA	GPA	GPA	101	GPI	GPI	GPI
24	GPQ	GPQ	GPQ	336	GPP	GPP	GPP	258	GPQ	GPQ	GPQ	180	GPP	GPP	GPP	102	GPP	GPP	GPP
25	GAR	GAR	GAR	337	GPP	GPP	GPP	259	GLA	GLA	GLA	181	GAQ	GAQ	GAQ	103	GER	GER	GER
26	GFP	GFP	GFP	338	GPP	GPP	GPP	260	GQR	GQR	GQR	182	GPP	GPP	GPP	104	GAP	GAP	GAP
27	GTP	GTP	GTP	0				261	GIV	GIV	GIV	183	GLQ	GLQ	GLQ	105	GNR	GNR	GNR
28	GLP	GLP	GLP	0				262	GLP	GLP	GLP	184	GMP	GMP	GMP	106	GFP	GFP	GFP
29	GVK	GVK	GVK	0				263	GQR	GQR	GQR	185	GER	GER	GER	107	GQD	GQD	GQD
30	GER	GER	GER	0				264	GER	GER	GER	186	GAA	GAA	GAA	108	GLA	GLA	GLA
31	GYP	GYP	GYP	0				265	GFP	GFP	GFP	187	GIA	GIA	GIA	109	GPK	GPK	GPK
32	GLD	GLD	GLD	0				266	GLP	GLP	GLP	188	GPK	GPK	GPK	110	GAP	GAP	GAP
33	GAK	GAK	GAK	0				267	GPS	GPS	GPS	189	GDR	GDR	GDR	111	GER	GER	GER
34	GEA	GEA	GEA	0				268	GEP	GEP	GEP	190	GDV	GDV	GDV	112	GPS	GPS	GPS
35	GAP	GAP	GAP	0				269	GKQ	GKQ	GKQ	191	GEX	GEX	GEX	113	GLA	GLA	GLA
36	GVK	GVK	GVK	0				270	GAP	GAP	GAP	192	GPE	GPE	GPE	114	GPK	GPK	GPK
37	GES	GES	GES	0				271	GAS	GAS	GAS	193	GAP	GAP	GAP	115	GAN	GAN	GAN
38	GSP	GSP	GSP	0				272	GDR	GDR	GDR	194	GKD	GKD	GKD	116	GDP	GDP	GDP
39	GEN	GEN	GEN	0				273	GPP	GPP	GPP	195	GGR	GGR	GGR	117	GRP	GRP	GRP
40	GSP	GSP	GSP	0				274	GPV	GPV	GPV	196	GLT	GLT	GLT	118	GEP	GEP	GEP
41	GPM	GPM	GPM	0				275	GPP	GPP	GPP	197	GPI	GPI	GPI	119	GLP	GLP	GLP
42	GPR	GPR	GPR	0				276	GLT	GLT	GLT	198	GPP	GPP	GPP	120	GAR	GAR	GAR
43	GLP	GLP	GLP	0				277	GPA	GPA	GPA	199	GPA	GPA	GPA	121	GLT	GLT	GLT
44	GER	GER	GER	0				278	GEP	GEP	GEP	200	GAN	GAN	GAN	122	GRP	GRP	GRP
45	GRT	GRT	GRT	0				279	GRQ	GRQ	GRQ	201	GEX	GEX	GEX	123	GDA	GDA	GDA

Table 1. Continued.

COLLAGEN 1				COLLAGEN 2			COLLAGEN 3			COLLAGEN 4			COLLAGEN 5						
A1	A1	A1		A1	A1	A1	A1	A1	A1	A1	A1	A1	A1	A1	A1				
46	GPA	GPA	GPA	0			280	GSP	GSP	GSP	202	GEV	GEV	GEV	124	GPQ	GPQ	GPQ	
47	GAA	GAA	GAA	0			281	GAD	GAD	GAD	203	GPP	GPP	GPP	125	GKV	GKV	GKV	
48	GAR	GAR	GAR	0			282	GPF	GPF	GPF	204	GPA	GPA	GPA	126	GPS	GPS	GPS	
49	GND	GND	GND	0			283	GRD	GRD	GRD	205	GSA	GSA	GSA	127	GAP	GAP	GAP	
50	GQP	GQP	GQP	0			284	GAA	GAA	GAA	206	GAR	GAR	GAR	128	GED	GED	GED	
51	GPA	GPA	GPA	0			285	GVK	GVK	GVK	207	GAP	GAP	GAP	129	GRP	GRP	GRP	
52	GPP	GPP	GPP	0			286	GDR	GDR	GDR	208	GER	GER	GER	130	GPP	GPP	GPP	
53	GPV	GPV	GPV	0			287	GET	GET	GET	209	GET	GET	GET	131	GPQ	GPQ	GPQ	
54	GPA	GPA	GPA	0			288	GAV	GAV	GAV	210	GPP	GPP	GPP	132	GAR	GAR	GAR	
55	GGP	GGP	GGP	0			289	GAP	GAP	GAP	211	GPA	GPA	GPA	133	GQP	GQP	GQP	
56	GFP	GFP	GFP	0			290	GTP	GTP	GTP	212	GFA	GFA	GFA	134	GVM	GVM	GVM	
57	GAP	GAP	GAP	0			291	GPP	GPP	GPP	213	GPP	GPP	GPP	135	GFP	GFP	GFP	
58	GAK	GAK	GAK	0			292	GSP	GSP	GSP	214	GAD	GAD	GAD	136	GPK	GPK	GPK	
59	GEA	GEA	GEA	0			293	GPA	GPA	GPA	215	GQP	GQP	GQP	137	GAN	GAN	GAN	
60	GPT	GPT	GPT	0			294	GPT	GPT	GPT	216	GAK	GAK	GAK	138	GEP	GEP	GEP	
61	GAR	GAR	GAR	0			295	GKQ	GKQ	GKQ	217	GEQ	GEQ	GEQ	139	GKA	GKA	GKA	
62	GPE	GPE	GPE	0			296	GDR	GDR	GDR	218	GEA	GEA	GEA	140	GEX	GEX	GEX	
63	GAQ	GAQ	GAQ	0			297	GEA	GEA	GEA	219	GQK	GQK	GQK	141	GLP	GLP	GLP	
64	GPR	GPR	GPR	0			298	GAQ	GAQ	GAQ	220	GDA	GDA	GDA	142	GAP	GAP	GAP	
65	GEP	GEP	GEP	0			299	GPM	GPM	GPM	221	GAP	GAP	GAP	143	GLR	GLR	GLR	
66	GTP	GTP	GTP	0			300	GPS	GPS	GPS	222	GPQ	GPQ	GPQ	144	GLP	GLP	GLP	
67	GSP	GSP	GSP	0			301	GPA	GPA	GPA	223	GPS	GPS	GPS	145	GKD	GKD	GKD	
68	GPA	GPA	GPA	0			302	GAR	GAR	GAR	224	GAP	GAP	GAP	146	GET	GET	GET	
69	GAS	GAS	GAS	0			303	GIQ	GIQ	GIQ	225	GPQ	GPQ	GPQ	147	GAA	GAA	GAA	
70	GNP	GNP	GNP	0			304	GPQ	GPQ	GPQ	226	GPT	GPT	GPT	148	GPP	GPP	GPP	
71	GTD	GTD	GTD	0			305	GPR	GPR	GPR	227	GVT	GVT	GVT	149	GPA	GPA	GPA	
72	GIP	GIP	GIP	0			306	GDK	GDK	GDK	228	GPK	GPK	GPK	150	GPA	GPA	GPA	
73	GAK	GAK	GAK	0			307	GEA	GEA	GEA	229	GAR	GAR	GAR	151	GER	GER	GER	
74	GSA	GSA	GSA	0			308	GEP	GEP	GEP	230	GAQ	GAQ	GAQ	152	GEQ	GEQ	GEQ	
75	GAP	GAP	GAP	0			309	GER	GER	GER	231	GPP	GPP	GPP	153	GAP	GAP	GAP	
76	GIA	GIA	GIA	0			310	GLK	GLK	GLK	232	GAT	GAT	GAT	154	GPS	GPS	GPS	
77	GAP	GAP	GAP	0			311	GER	GER	GER	233	GFP	GFP	GFP	155	GFQ	GFQ	GFQ	
78	GFP	GFP	GFP	0			312	GFT	GFT	GFT	234	GAA	GAA	GAA	156	GLP	GLP	GLP	
79	GPR	GPR	GPR	1	GPM	GPM	GPM	313	GLQ	GLQ	GLQ	235	GRV	GRV	GRV	157	GPP	GPP	GPP
80	GPP	GPP	GPP	2	GPM	GPM	GPM	314	GLP	GLP	GLP	236	GPP	GPP	GPP	158	GPP	GPP	GPP
81	GPQ	GPQ	GPQ	3	GPR	GPR	GPR	315	GPP	GPP	GPP	237	GSN	GSN	GSN	159	GEG	GEG	GEG
82	GAT	GAT	GAT	4	GPP	GPP	GPP	316	GPS	GPS	GPS	238	GNN	GNN	GNN	160	GKP	GKP	GKP
83	GPL	GPL	GPL	5	GPA	GPA	GPA	317	GDQ	GDQ	GDQ	239	GPP	GPP	GPP	161	GDQ	GDQ	GDQ
84	GPK	GPK	GPK	6	GAP	GAP	GAP	318	GAS	GAS	GAS	240	GPP	GPP	GPP	162	GVP	GVP	GVP
85	GQT	GQT	GQT	7	GPQ	GPQ	GPQ	319	GPA	GPA	GPA	241	GPS	GPS	GPS	163	GEA	GEA	GEA
86	GEP	GEP	GEP	8	GFQ	GFQ	GFQ	320	GPS	GPS	GPS	242	GKD	GKD	GKD	164	GAP	GAP	GAP
87	GIA	GIA	GIA	9	GNP	GNP	GNP	321	GPR	GPR	GPR	243	GPK	GPK	GPK	165	GLV	GLV	GLV
88	GFK	GFK	GFK	10	GEP	GEP	GEP	322	GPP	GPP	GPP	244	GAR	GAR	GAR	166	GPR	GPR	GPR
89	GEQ	GEQ	GEQ	11	GEP	GEP	GEP	323	GPV	GPV	GPV	245	GDS	GDS	GDS	167	GER	GER	GER
90	GPK	GPK	GPK	12	GVS	GVS	GVS	324	GPS	GPS	GPS	246	GPP	GPP	GPP	168	GFP	GFP	GFP
91	GEP	GEP	GEP	13	GPM	GPM	GPM	325	GKD	GKD	GKD	247	GRA	GRA	GRA	169	GER	GER	GER
92	GPA	GPA	GPA	14	GPR	GPR	GPR	326	GAN	GAN	GAN	248	GEP	GEP	GEP	170	GSP	GSP	GSP
93	GPQ	GPQ	GPQ	15	GPP	GPP	GPP	327	GIP	GIP	GIP	249	GLQ	GLQ	GLQ	171	GAQ	GAQ	GAQ
94	GAP	GAP	GAP	16	GPP	GPP	GPP	328	GPI	GPI	GPI	250	GPA	GPA	GPA	172	GLQ	GLQ	GLQ
95	GPA	GPA	GPA	17	GKP	GKP	GKP	329	GPP	GPP	GPP	251	GPP	GPP	GPP	173	GPR	GPR	GPR
96	GEE	GEE	GEE	18	GDD	GDD	GDD	330	GPR	GPR	GPR	252	GEX	GEX	GEX	174	GLP	GLP	GLP
97	GKR	GKR	GKR	19	GEA	GEA	GEA	331	GRS	GRS	GRS	253	GEP	GEP	GEP	175	GTP	GTP	GTP
98	GAR	GAR	GAR	20	GKP	GKP	GKP	332	GET	GET	GET	254	GDD	GDD	GDD	176	GTD	GTD	GTD
99	GEP	GEP	GEP	21	GKA	GKA	GKA	333	GPA	GPA	GPA	255	GPS	GPS	GPS	177	GPK	GPK	GPK

Continued on next page.

Table 1. Continued.

COLLAGEN 1			COLLAGEN 2			COLLAGEN 3			COLLAGEN 4			COLLAGEN 5							
A1	A1	A1	A1	A1	A1	A1	A1	A1	A1	A1	A1	A1	A1	A1					
100	GGV	GGV	GGV	22	GER	GER	GER	334	GPP	GPP	GPP	256	GAE	GAE	GAE	178	GAS	GAS	GAS
101	GPI	GPI	GPI	23	GPP	GPP	GPP	335	GNP	GNP	GNP	257	GFP	GFP	GFP	179	GPA	GPA	GPA
102	GPP	GPP	GPP	24	GPQ	GPQ	GPQ	336	GPP	GPP	GPP	258	GPQ	GPQ	GPQ	180	GPP	GPP	GPP
103	GER	GER	GER	25	GAR	GAR	GAR	337	GPP	GPP	GPP	259	GLA	GLA	GLA	181	GAQ	GAQ	GAQ
104	GAP	GAP	GAP	26	GFP	GFP	GFP	338	GPP	GPP	GPP	260	GQR	GQR	GQR	182	GPP	GPP	GPP
105	GNR	GNR	GNR	27	GTP	GTP	GTP	0				261	GIV	GIV	GIV	183	GLQ	GLQ	GLQ
106	GFP	GFP	GFP	28	GLP	GLP	GLP	0				262	GLP	GLP	GLP	184	GMP	GMP	GMP
107	GQD	GQD	GQD	29	GVK	GVK	GVK	0				263	GQR	GQR	GQR	185	GER	GER	GER
108	GLA	GLA	GLA	30	GHR	GHR	GHR	0				264	GER	GER	GER	186	GAA	GAA	GAA
109	GPK	GPK	GPK	31	GYP	GYP	GYP	0				265	GFP	GFP	GFP	187	GIA	GIA	GIA
110	GAP	GAP	GAP	32	GLD	GLD	GLD	0				266	GLP	GLP	GLP	188	GPK	GPK	GPK
111	GER	GER	GER	33	GAK	GAK	GAK	0				267	GPS	GPS	GPS	189	GDR	GDR	GDR
112	GPS	GPS	GPS	34	GEA	GEA	GEA	0				268	GEP	GEP	GEP	190	GDV	GDV	GDV
113	GLA	GLA	GLA	35	GAP	GAP	GAP	0				269	GKQ	GKQ	GKQ	191	GEK	GEK	GEK
114	GPK	GPK	GPK	36	GVK	GVK	GVK	0				270	GAP	GAP	GAP	192	GPE	GPE	GPE
115	GAN	GAN	GAN	37	GES	GES	GES	0				271	GAS	GAS	GAS	193	GAP	GAP	GAP
116	GDP	GDP	GDP	38	GSP	GSP	GSP	0				272	GDR	GDR	GDR	194	GKD	GKD	GKD
117	GRP	GRP	GRP	39	GEN	GEN	GEN	0				273	GFP	GFP	GFP	195	GGR	GGR	GGR
118	GEP	GEP	GEP	40	GSP	GSP	GSP	0				274	GPV	GPV	GPV	196	GLT	GLT	GLT
119	GLP	GLP	GLP	41	GPM	GPM	GPM	0				275	GPP	GPP	GPP	197	GPI	GPI	GPI
120	GAR	GAR	GAR	42	GPR	GPR	GPR	0				276	GLT	GLT	GLT	198	GPP	GPP	GPP
121	GLT	GLT	GLT	43	GLP	GLP	GLP	0				277	GPA	GPA	GPA	199	GPA	GPA	GPA
122	GRP	GRP	GRP	44	GER	GER	GER	0				278	GEP	GEP	GEP	200	GAN	GAN	GAN
123	GDA	GDA	GDA	45	GRT	GRT	GRT	0				279	GRQ	GRQ	GRQ	201	GEK	GEK	GEK
124	GPQ	GPQ	GPQ	46	GPA	GPA	GPA	0				280	GSP	GSP	GSP	202	GEV	GEV	GEV
125	GKV	GKV	GKV	47	GAA	GAA	GAA	0				281	GAD	GAD	GAD	203	GPP	GPP	GPP
126	GPS	GPS	GPS	48	GAR	GAR	GAR	0				282	GPP	GPP	GPP	204	GPA	GPA	GPA
127	GAP	GAP	GAP	49	GND	GND	GND	0				283	GRD	GRD	GRD	205	GSA	GSA	GSA
128	GED	GED	GED	50	GQP	GQP	GQP	0				284	GAA	GAA	GAA	206	GAR	GAR	GAR
129	GRP	GRP	GRP	51	GPA	GPA	GPA	0				285	GVK	GVK	GVK	207	GAP	GAP	GAP
130	GPP	GPP	GPP	52	GPP	GPP	GPP	0				286	GDR	GDR	GDR	208	GER	GER	GER
131	GPQ	GPQ	GPQ	53	GPV	GPV	GPV	0				287	GET	GET	GET	209	GET	GET	GET
132	GAR	GAR	GAR	54	GPA	GPA	GPA	0				288	GAV	GAV	GAV	210	GPP	GPP	GPP
133	GQP	GQP	GQP	55	GGP	GGP	GGP	0				289	GAP	GAP	GAP	211	GPA	GPA	GPA
134	GVM	GVM	GVM	56	GFP	GFP	GFP	0				290	GTP	GTP	GTP	212	GFA	GFA	GFA
135	GFP	GFP	GFP	57	GAP	GAP	GAP	0				291	GPP	GPP	GPP	213	GPP	GPP	GPP
136	GPK	GPK	GPK	58	GAK	GAK	GAK	0				292	GSP	GSP	GSP	214	GAD	GAD	GAD
137	GAN	GAN	GAN	59	GEA	GEA	GEA	0				293	GPA	GPA	GPA	215	GQP	GQP	GQP
138	GEP	GEP	GEP	60	GPT	GPT	GPT	0				294	GPT	GPT	GPT	216	GAK	GAK	GAK
139	GKA	GKA	GKA	61	GAR	GAR	GAR	0				295	GKQ	GKQ	GKQ	217	GEK	GEK	GEK
140	GEK	GEK	GEK	62	GPE	GPE	GPE	0				296	GDR	GDR	GDR	218	GEA	GEA	GEA
141	GLP	GLP	GLP	63	GAQ	GAQ	GAQ	0				297	GEA	GEA	GEA	219	GQK	GQK	GQK
142	GAP	GAP	GAP	64	GPR	GPR	GPR	0				298	GAQ	GAQ	GAQ	220	GDA	GDA	GDA
143	GLR	GLR	GLR	65	GEP	GEP	GEP	0				299	GPM	GPM	GPM	221	GAP	GAP	GAP
144	GLP	GLP	GLP	66	GTP	GTP	GTP	0				300	GPS	GPS	GPS	222	GPQ	GPQ	GPQ
145	GKD	GKD	GKD	67	GSP	GSP	GSP	0				301	GPA	GPA	GPA	223	GPS	GPS	GPS
146	GET	GET	GET	68	GPA	GPA	GPA	0				302	GAR	GAR	GAR	224	GAP	GAP	GAP
147	GAA	GAA	GAA	69	GAS	GAS	GAS	0				303	GIQ	GIQ	GIQ	225	GPQ	GPQ	GPQ
148	GPP	GPP	GPP	70	GMP	GMP	GMP	0				304	GPQ	GPQ	GPQ	226	GPT	GPT	GPT
149	GPA	GPA	GPA	71	GTD	GTD	GTD	0				305	GPR	GPR	GPR	227	GVT	GVT	GVT
150	GPA	GPA	GPA	72	GIP	GIP	GIP	0				306	GDK	GDK	GDK	228	GPK	GPK	GPK
151	GER	GER	GER	73	GAK	GAK	GAK	0				307	GEA	GEA	GEA	229	GAR	GAR	GAR
152	GEQ	GEQ	GEQ	74	GSA	GSA	GSA	0				308	GEP	GEP	GEP	230	GAQ	GAQ	GAQ
153	GAP	GAP	GAP	75	GAP	GAP	GAP	0				309	GER	GER	GER	231	GPP	GPP	GPP

Table 1. Continued.

COLLAGEN 1			COLLAGEN 2			COLLAGEN 3			COLLAGEN 4			COLLAGEN 5						
A1	A1	A1	A1	A1	A1	A1	A1	A1	A1	A1	A1	A1	A1	A1				
154	GPS	GPS	GPS	76	GIA	GIA	GIA	0		310	GLK	GLK	GLK	232	GAT	GAT	GAT	
155	GFQ	GFQ	GFQ	77	GAP	GAP	GAP	0		311	GHR	GHR	GHR	233	GFP	GFP	GFP	
156	GLP	GLP	GLP	78	GFP	GFP	GFP	0		312	GFT	GFT	GFT	234	GAA	GAA	GAA	
157	GPP	GPP	GPP	79	GPR	GPR	GPR	1	GPM	GPM	GPM	313	GLQ	GLQ	GLQ	235	GRV	GRV
158	GPP	GPP	GPP	80	GPP	GPP	GPP	2	GPM	GPM	GPM	314	GLP	GLP	GLP	236	GPP	GPP
159	GEG	GEG	GEG	81	GPQ	GPQ	GPQ	3	GPR	GPR	GPR	315	GPP	GPP	GPP	237	GSN	GSN
160	GKP	GKP	GKP	82	GAT	GAT	GAT	4	GPP	GPP	GPP	316	GPS	GPS	GPS	238	GNP	GNP
161	GDQ	GDQ	GDQ	83	GEL	GEL	GEL	5	GPA	GPA	GPA	317	GDQ	GDQ	GDQ	239	GPP	GPP
162	GVP	GVP	GVP	84	GPK	GPK	GPK	6	GAP	GAP	GAP	318	GAS	GAS	GAS	240	GPP	GPP
163	GEA	GEA	GEA	85	GQT	GQT	GQT	7	GPQ	GPQ	GPQ	319	GPA	GPA	GPA	241	GPS	GPS
164	GAP	GAP	GAP	86	GEP	GEP	GEP	8	GFQ	GFQ	GFQ	320	GPS	GPS	GPS	242	GKD	GKD
165	GLV	GLV	GLV	87	GIA	GIA	GIA	9	GNP	GNP	GNP	321	GPR	GPR	GPR	243	GPK	GPK
166	GPR	GPR	GPR	88	GFK	GFK	GFK	10	GEP	GEP	GEP	322	GPP	GPP	GPP	244	GAR	GAR
167	GER	GER	GER	89	GEQ	GEQ	GEQ	11	GEP	GEP	GEP	323	GPV	GPV	GPV	245	GDS	GDS
168	GFP	GFP	GFP	90	GPK	GPK	GPK	12	GVS	GVS	GVS	324	GPS	GPS	GPS	246	GPP	GPP
169	GER	GER	GER	91	GEP	GEP	GEP	13	GPM	GPM	GPM	325	GKD	GKD	GKD	247	GRA	GRA
170	GSP	GSP	GSP	92	GPA	GPA	GPA	14	GPR	GPR	GPR	326	GAN	GAN	GAN	248	GEP	GEP
171	GAQ	GAQ	GAQ	93	GPQ	GPQ	GPQ	15	GPP	GPP	GPP	327	GIP	GIP	GIP	249	GLQ	GLQ
172	GLQ	GLQ	GLQ	94	GAP	GAP	GAP	16	GPP	GPP	GPP	328	GPI	GPI	GPI	250	GPA	GPA
173	GPR	GPR	GPR	95	GPA	GPA	GPA	17	GKP	GKP	GKP	329	GPP	GPP	GPP	251	GPP	GPP
174	GLP	GLP	GLP	96	GEE	GEE	GEE	18	GDD	GDD	GDD	330	GPR	GPR	GPR	252	GEX	GEX
175	GTP	GTP	GTP	97	GKR	GKR	GKR	19	GEA	GEA	GEA	331	GRS	GRS	GRS	253	GEP	GEP
176	GTD	GTD	GTD	98	GAR	GAR	GAR	20	GKP	GKP	GKP	332	GET	GET	GET	254	GDD	GDD
177	GPK	GPK	GPK	99	GEP	GEP	GEP	21	GKA	GKA	GKA	333	GPA	GPA	GPA	255	GPS	GPS
178	GAS	GAS	GAS	100	GGV	GGV	GGV	22	GER	GER	GER	334	GPP	GPP	GPP	256	GAE	GAE
179	GPA	GPA	GPA	101	GPI	GPI	GPI	23	GPP	GPP	GPP	335	GNP	GNP	GNP	257	GPP	GPP
180	GPP	GPP	GPP	102	GPP	GPP	GPP	24	GPQ	GPQ	GPQ	336	GPP	GPP	GPP	258	GPQ	GPQ
181	GAQ	GAQ	GAQ	103	GER	GER	GER	25	GAR	GAR	GAR	337	GPP	GPP	GPP	259	GLA	GLA
182	GPP	GPP	GPP	104	GAP	GAP	GAP	26	GFP	GFP	GFP	338	GPP	GPP	GPP	260	GQR	GQR
183	GLQ	GLQ	GLQ	105	GNR	GNR	GNR	27	GTP	GTP	GTP	0				261	GIV	GIV
184	GMP	GMP	GMP	106	GFP	GFP	GFP	28	GLP	GLP	GLP	0				262	GLP	GLP
185	GER	GER	GER	107	GQD	GQD	GQD	29	GVK	GVK	GVK	0				263	GQR	GQR
186	GAA	GAA	GAA	108	GLA	GLA	GLA	30	GHR	GHR	GHR	0				264	GER	GER
187	GIA	GIA	GIA	109	GPK	GPK	GPK	31	GYP	GYP	GYP	0				265	GFP	GFP
188	GPK	GPK	GPK	110	GAP	GAP	GAP	32	GLD	GLD	GLD	0				266	GLP	GLP
189	GDR	GDR	GDR	111	GER	GER	GER	33	GAK	GAK	GAK	0				267	GPS	GPS
190	GDV	GDV	GDV	112	GPS	GPS	GPS	34	GEA	GEA	GEA	0				268	GEP	GEP
191	GZK	GZK	GZK	113	GLA	GLA	GLA	35	GAP	GAP	GAP	0				269	GKQ	GKQ
192	GPE	GPE	GPE	114	GPK	GPK	GPK	36	GVK	GVK	GVK	0				270	GAP	GAP
193	GAP	GAP	GAP	115	GAN	GAN	GAN	37	GES	GES	GES	0				271	GAS	GAS
194	GKD	GKD	GKD	116	GDP	GDP	GDP	38	GSP	GSP	GSP	0				272	GDR	GDR
195	GGR	GGR	GGR	117	GRP	GRP	GRP	39	GEN	GEN	GEN	0				273	GPP	GPP
196	GLT	GLT	GLT	118	GEP	GEP	GEP	40	GSP	GSP	GSP	0				274	GPV	GPV
197	GPI	GPI	GPI	119	GLP	GLP	GLP	41	GPM	GPM	GPM	0				275	GPP	GPP
198	GPP	GPP	GPP	120	GAR	GAR	GAR	42	GPR	GPR	GPR	0				276	GLT	GLT
199	GPA	GPA	GPA	121	GLT	GLT	GLT	43	GLP	GLP	GLP	0				277	GPA	GPA
200	GAN	GAN	GAN	122	GRP	GRP	GRP	44	GER	GER	GER	0				278	GEP	GEP
201	GKQ	GKQ	GKQ	123	GDA	GDA	GDA	45	GRT	GRT	GRT	0				279	GRQ	GRQ
202	GEV	GEV	GEV	124	GPQ	GPQ	GPQ	46	GPA	GPA	GPA	0				280	GSP	GSP
203	GPP	GPP	GPP	125	GKV	GKV	GKV	47	GAA	GAA	GAA	0				281	GAD	GAD
204	GPA	GPA	GPA	126	GPS	GPS	GPS	48	GAR	GAR	GAR	0				282	GPP	GPP
205	GSA	GSA	GSA	127	GAP	GAP	GAP	49	GND	GND	GND	0				283	GRD	GRD
206	GAR	GAR	GAR	128	GED	GED	GED	50	GQP	GQP	GQP	0				284	GAA	GAA
207	GAP	GAP	GAP	129	GRP	GRP	GRP	51	GPA	GPA	GPA	0				285	GVK	GVK

Continued on next page.

Table 1. Continued.

COLLAGEN 1			COLLAGEN 2			COLLAGEN 3			COLLAGEN 4			COLLAGEN 5							
A1	A1	A1	A1	A1	A1	A1	A1	A1	A1	A1	A1	A1	A1	A1					
208	GER	GER	GER	130	GPP	GPP	GPP	52	GPP	GPP	GPP	0	286	GDR	GDR	GDR			
209	GET	GET	GET	131	GPQ	GPQ	GPQ	53	GPV	GPV	GPV	0	287	GET	GET	GET			
210	GFP	GFP	GFP	132	GAR	GAR	GAR	54	GPA	GPA	GPA	0	288	GAV	GAV	GAV			
211	GPA	GPA	GPA	133	GQP	GQP	GQP	55	GGP	GGP	GGP	0	289	GAP	GAP	GAP			
212	GFA	GFA	GFA	134	GVM	GVM	GVM	56	GFP	GFP	GFP	0	290	GTP	GTP	GTP			
213	GPP	GPP	GPP	135	GFF	GFF	GFF	57	GAP	GAP	GAP	0	291	GPP	GPP	GPP			
214	GAD	GAD	GAD	136	GPK	GPK	GPK	58	GAK	GAK	GAK	0	292	GSP	GSP	GSP			
215	GQP	GQP	GQP	137	GAN	GAN	GAN	59	GEA	GEA	GEA	0	293	GPA	GPA	GPA			
216	GAK	GAK	GAK	138	GEP	GEP	GEP	60	GPT	GPT	GPT	0	294	GPT	GPT	GPT			
217	GEQ	GEQ	GEQ	139	GKA	GKA	GKA	61	GAR	GAR	GAR	0	295	GKQ	GKQ	GKQ			
218	GEA	GEA	GEA	140	GKE	GKE	GKE	62	GPE	GPE	GPE	0	296	GDR	GDR	GDR			
219	GQK	GQK	GQK	141	GLP	GLP	GLP	63	GAQ	GAQ	GAQ	0	297	GEA	GEA	GEA			
220	GDA	GDA	GDA	142	GAP	GAP	GAP	64	GPR	GPR	GPR	0	298	GAQ	GAQ	GAQ			
221	GAP	GAP	GAP	143	GLR	GLR	GLR	65	GEP	GEP	GEP	0	299	GPM	GPM	GPM			
222	GPQ	GPQ	GPQ	144	GLP	GLP	GLP	66	GTP	GTP	GTP	0	300	GPS	GPS	GPS			
223	GPS	GPS	GPS	145	GKD	GKD	GKD	67	GSP	GSP	GSP	0	301	GPA	GPA	GPA			
224	GAP	GAP	GAP	146	GET	GET	GET	68	GPA	GPA	GPA	0	302	GAR	GAR	GAR			
225	GPQ	GPQ	GPQ	147	GAA	GAA	GAA	69	GAS	GAS	GAS	0	303	GIQ	GIQ	GIQ			
226	GPT	GPT	GPT	148	GPP	GPP	GPP	70	GNP	GNP	GNP	0	304	GPQ	GPQ	GPQ			
227	GVT	GVT	GVT	149	GPA	GPA	GPA	71	GTD	GTD	GTD	0	305	GPR	GPR	GPR			
228	GPK	GPK	GPK	150	GPA	GPA	GPA	72	GIP	GIP	GIP	0	306	GDK	GDK	GDK			
229	GAR	GAR	GAR	151	GER	GER	GER	73	GAK	GAK	GAK	0	307	GEA	GEA	GEA			
230	GAQ	GAQ	GAQ	152	GEQ	GEQ	GEQ	74	GSA	GSA	GSA	0	308	GEP	GEP	GEP			
231	GPP	GPP	GPP	153	GAP	GAP	GAP	75	GAP	GAP	GAP	0	309	GER	GER	GER			
232	GAT	GAT	GAT	154	GPS	GPS	GPS	76	GIA	GIA	GIA	0	310	GLK	GLK	GLK			
233	GFP	GFP	GFP	155	GFQ	GFQ	GFQ	77	GAP	GAP	GAP	0	311	GHR	GHR	GHR			
234	GAA	GAA	GAA	156	GLP	GLP	GLP	78	GFP	GFP	GFP	0	312	GFT	GFT	GFT			
235	GRV	GRV	GRV	157	GPP	GPP	GPP	79	GPR	GPR	GPR	1	GPM	GPM	GPM	313	GLQ	GLQ	GLQ
236	GPP	GPP	GPP	158	GPP	GPP	GPP	80	GPP	GPP	GPP	2	GPM	GPM	GPM	314	GLP	GLP	GLP
237	GSN	GSN	GSN	159	GEG	GEG	GEG	81	GPQ	GPQ	GPQ	3	GPR	GPR	GPR	315	GPP	GPP	GPP
238	GNP	GNP	GNP	160	GKP	GKP	GKP	82	GAT	GAT	GAT	4	GPP	GPP	GPP	316	GPS	GPS	GPS
239	GPP	GPP	GPP	161	GDQ	GDQ	GDQ	83	GPL	GPL	GPL	5	GPA	GPA	GPA	317	GDQ	GDQ	GDQ
240	GPP	GPP	GPP	162	GVP	GVP	GVP	84	GPK	GPK	GPK	6	GAP	GAP	GAP	318	GAS	GAS	GAS
241	GPS	GPS	GPS	163	GEA	GEA	GEA	85	GQT	GQT	GQT	7	GFQ	GFQ	GFQ	319	GPA	GPA	GPA
242	GKD	GKD	GKD	164	GAP	GAP	GAP	86	GEP	GEP	GEP	8	GFQ	GFQ	GFQ	320	GPS	GPS	GPS
243	GPK	GPK	GPK	165	GLV	GLV	GLV	87	GIA	GIA	GIA	9	GNP	GNP	GNP	321	GPR	GPR	GPR
244	GAR	GAR	GAR	166	GPR	GPR	GPR	88	GFK	GFK	GFK	10	GEP	GEP	GEP	322	GPP	GPP	GPP
245	GDS	GDS	GDS	167	GER	GER	GER	89	GEQ	GEQ	GEQ	11	GEP	GEP	GEP	323	GPV	GPV	GPV
246	GPP	GPP	GPP	168	GFP	GFP	GFP	90	GPK	GPK	GPK	12	GVS	GVS	GVS	324	GPS	GPS	GPS
247	GRA	GRA	GRA	169	GER	GER	GER	91	GEP	GEP	GEP	13	GPM	GPM	GPM	325	GKD	GKD	GKD
248	GEP	GEP	GEP	170	GSP	GSP	GSP	92	GPA	GPA	GPA	14	GPR	GPR	GPR	326	GAN	GAN	GAN
249	GLQ	GLQ	GLQ	171	GAQ	GAQ	GAQ	93	GPQ	GPQ	GPQ	15	GPP	GPP	GPP	327	GTP	GTP	GTP
250	GPA	GPA	GPA	172	GLQ	GLQ	GLQ	94	GAP	GAP	GAP	16	GPP	GPP	GPP	328	GPI	GPI	GPI
251	GPP	GPP	GPP	173	GPR	GPR	GPR	95	GPA	GPA	GPA	17	GKP	GKP	GKP	329	GPP	GPP	GPP
252	GEK	GEK	GEK	174	GLP	GLP	GLP	96	GEE	GEE	GEE	18	GDD	GDD	GDD	330	GPR	GPR	GPR
253	GEP	GEP	GEP	175	GTP	GTP	GTP	97	GKR	GKR	GKR	19	GEA	GEA	GEA	331	GRS	GRS	GRS
254	GDD	GDD	GDD	176	GTD	GTD	GTD	98	GAR	GAR	GAR	20	GKP	GKP	GKP	332	GET	GET	GET
255	GPS	GPS	GPS	177	GPK	GPK	GPK	99	GEP	GEP	GEP	21	GKA	GKA	GKA	333	GPA	GPA	GPA
256	GAE	GAE	GAE	178	GAS	GAS	GAS	100	GGV	GGV	GGV	22	GER	GER	GER	334	GPP	GPP	GPP
257	GPP	GPP	GPP	179	GPA	GPA	GPA	101	GPI	GPI	GPI	23	GPP	GPP	GPP	335	GNP	GNP	GNP
258	GPQ	GPQ	GPQ	180	GPP	GPP	GPP	102	GPP	GPP	GPP	24	GPQ	GPQ	GPQ	336	GPP	GPP	GPP
259	GLA	GLA	GLA	181	GAQ	GAQ	GAQ	103	GER	GER	GER	25	GAR	GAR	GAR	337	GPP	GPP	GPP
260	GQR	GQR	GQR	182	GPP	GPP	GPP	104	GAP	GAP	GAP	26	GFP	GFP	GFP	338	GPP	GPP	GPP
261	GIV	GIV	GIV	183	GLQ	GLQ	GLQ	105	GNR	GNR	GNR	27	GTP	GTP	GTP	0			

Table 1. Continued.

COLLAGEN 1			COLLAGEN 2			COLLAGEN 3			COLLAGEN 4			COLLAGEN 5				
A1	A1	A1	A1	A1	A1	A1	A1	A1	A1	A1	A1	A1	A1	A1	A1	
262	GLP	GLP	GLP	184	GMP	GMP	GMP	106	GFP	GFP	GFP	28	GLP	GLP	GLP	0
263	GQR	GQR	GQR	185	GER	GER	GER	107	GQD	GQD	GQD	29	GVK	GVK	GVK	0
264	GGR	GER	GER	186	GAA	GAA	GAA	108	GLA	GLA	GLA	30	GHR	GHR	GHR	0
265	GFP	GFP	GFP	187	GIA	GIA	GIA	109	GPK	GPK	GPK	31	GYP	GYP	GYP	0
266	GLP	GLP	GLP	188	GPK	GPK	GPK	110	GAP	GAP	GAP	32	GLD	GLD	GLD	0
267	GPS	GPS	GPS	189	GDR	GDR	GDR	111	GER	GER	GER	33	GAK	GAK	GAK	0
268	GEP	GEP	GEP	190	GDV	GDV	GDV	112	GPS	GPS	GPS	34	GEA	GEA	GEA	0
269	GKQ	GKQ	GKQ	191	GKV	GKV	GKV	113	GLA	GLA	GLA	35	GAP	GAP	GAP	0
270	GAP	GAP	GAP	192	GPE	GPE	GPE	114	GPK	GPK	GPK	36	GVK	GVK	GVK	0
271	GAS	GAS	GAS	193	GAP	GAP	GAP	115	GAN	GAN	GAN	37	GES	GES	GES	0
272	GDR	GDR	GDR	194	GKD	GKD	GKD	116	GDP	GDP	GDP	38	GSP	GSP	GSP	0
273	GPP	GPP	GPP	195	GGR	GGR	GGR	117	GRP	GRP	GRP	39	GEN	GEN	GEN	0
274	GPV	GPV	GPV	196	GLT	GLT	GLT	118	GEP	GEP	GEP	40	GSP	GSP	GSP	0
275	GPP	GPP	GPP	197	GPI	GPI	GPI	119	GLP	GLP	GLP	41	GPM	GPM	GPM	0
276	GLT	GLT	GLT	198	GPP	GPP	GPP	120	GAR	GAR	GAR	42	GPR	GPR	GPR	0
277	GPA	GPA	GPA	199	GPA	GPA	GPA	121	GLT	GLT	GLT	43	GLP	GLP	GLP	0
278	GEP	GEP	GEP	200	GAN	GAN	GAN	122	GRP	GRP	GRP	44	GER	GER	GER	0
279	GRQ	GRQ	GRQ	201	GEK	GEK	GEK	123	GDA	GDA	GDA	45	GRT	GRT	GRT	0
280	GSP	GSP	GSP	202	GEV	GEV	GEV	124	GPQ	GPQ	GPQ	46	GPA	GPA	GPA	0
281	GAD	GAD	GAD	203	GPP	GPP	GPP	125	GKV	GKV	GKV	47	GAA	GAA	GAA	0
282	GPP	GPP	GPP	204	GPA	GPA	GPA	126	GPS	GPS	GPS	48	GAR	GAR	GAR	0
283	GRD	GRD	GRD	205	GSA	GSA	GSA	127	GAP	GAP	GAP	49	GND	GND	GND	0
284	GAA	GAA	GAA	206	GAR	GAR	GAR	128	GED	GED	GED	50	GQP	GQP	GQP	0
285	GVK	GVK	GVK	207	GAP	GAP	GAP	129	GRP	GRP	GRP	51	GPA	GPA	GPA	0
286	GDR	GDR	GDR	208	GER	GER	GER	130	GPP	GPP	GPP	52	GPP	GPP	GPP	0
287	GET	GET	GET	209	GET	GET	GET	131	GPQ	GPQ	GPQ	53	GPV	GPV	GPV	0
288	GAV	GAV	GAV	210	GPP	GPP	GPP	132	GAR	GAR	GAR	54	GPA	GPA	GPA	0
289	GAP	GAP	GAP	211	GPA	GPA	GPA	133	GQP	GQP	GQP	55	GGP	GGP	GGP	0
290	GTP	GTP	GTP	212	GFA	GFA	GFA	134	GVM	GVM	GVM	56	GFP	GFP	GFP	0
291	GPP	GPP	GPP	213	GPP	GPP	GPP	135	GFP	GFP	GFP	57	GAP	GAP	GAP	0
292	GSP	GSP	GSP	214	GAD	GAD	GAD	136	GPK	GPK	GPK	58	GAK	GAK	GAK	0
293	GPA	GPA	GPA	215	GQP	GQP	GQP	137	GAN	GAN	GAN	59	GEA	GEA	GEA	0
294	GPT	GPT	GPT	216	GAK	GAK	GAK	138	GEP	GEP	GEP	60	GPT	GPT	GPT	0
295	GKQ	GKQ	GKQ	217	GEQ	GEQ	GEQ	139	GKA	GKA	GKA	61	GAR	GAR	GAR	0
296	GDR	GDR	GDR	218	GEA	GEA	GEA	140	GEK	GEK	GEK	62	GPE	GPE	GPE	0
297	GEA	GEA	GEA	219	GQK	GQK	GQK	141	GLP	GLP	GLP	63	GAQ	GAQ	GAQ	0
298	GAQ	GAQ	GAQ	220	GDA	GDA	GDA	142	GAP	GAP	GAP	64	GPR	GPR	GPR	0
299	GPM	GPM	GPM	221	GAP	GAP	GAP	143	GPR	GPR	GPR	65	GEP	GEP	GEP	0
300	GPS	GPS	GPS	222	GPQ	GPQ	GPQ	144	GLP	GLP	GLP	66	GTP	GTP	GTP	0
301	GPA	GPA	GPA	223	GPS	GPS	GPS	145	GKD	GKD	GKD	67	GSP	GSP	GSP	0
302	GAR	GAR	GAR	224	GAP	GAP	GAP	146	GET	GET	GET	68	GPA	GPA	GPA	0
303	GIQ	GIQ	GIQ	225	GPQ	GPQ	GPQ	147	GAA	GAA	GAA	69	GAS	GAS	GAS	0
304	GPQ	GPQ	GPQ	226	GPT	GPT	GPT	148	GPP	GPP	GPP	70	GNP	GNP	GNP	0
305	GPR	GPR	GPR	227	GVT	GVT	GVT	149	GPA	GPA	GPA	71	GTD	GTD	GTD	0
306	GDK	GDK	GDK	228	GPK	GPK	GPK	150	GPA	GPA	GPA	72	GIP	GIP	GIP	0
307	GEA	GEA	GEA	229	GAR	GAR	GAR	151	GER	GER	GER	73	GAK	GAK	GAK	0
308	GEP	GEP	GEP	230	GAQ	GAQ	GAQ	152	GEQ	GEQ	GEQ	74	GSA	GSA	GSA	0
309	GER	GER	GER	231	GFP	GFP	GFP	153	GAP	GAP	GAP	75	GAP	GAP	GAP	0
310	GLK	GLK	GLK	232	GAT	GAT	GAT	154	GPS	GPS	GPS	76	GIA	GIA	GIA	0
311	GHR	GHR	GHR	233	GFP	GFP	GFP	155	GFQ	GFQ	GFQ	77	GAP	GAP	GAP	0
312	GFT	GFT	GFT	234	GAA	GAA	GAA	156	GLP	GLP	GLP	78	GFP	GFP	GFP	0
313	GLQ	GLQ	GLQ	235	GRV	GRV	GRV	157	GPP	GPP	GPP	79	GPR	GPR	GPR	1
314	GLP	GLP	GLP	236	GPP	GPP	GPP	158	GPP	GPP	GPP	80	GPP	GPP	GPP	2
315	GPP	GPP	GPP	237	GSN	GSN	GSN	159	GEG	GEG	GEG	81	GPQ	GPQ	GPQ	3

Continued on next page.

Table 1. Continued.

COLLAGEN 1			COLLAGEN 2			COLLAGEN 3			COLLAGEN 4			COLLAGEN 5							
A1	A1	A1	A1	A1	A1	A1	A1	A1	A1	A1	A1	A1	A1	A1					
316	GPS	GPS	GPS	238	GNP	GNP	GNP	160	GKP	GKP	GKP	82	GAT	GAT	GAT	4	GPP	GPP	GPP
317	GDQ	GDQ	GDQ	239	GFP	GFP	GFP	161	GDQ	GDQ	GDQ	83	GPL	GPL	GPL	5	GPA	GPA	GPA
318	GAS	GAS	GAS	240	GPP	GPP	GPP	162	GVP	GVP	GVP	84	GPK	GPK	GPK	6	GAP	GAP	GAP
319	GPA	GPA	GPA	241	GPS	GPS	GPS	163	GEA	GEA	GEA	85	GQT	GQT	GQT	7	GPQ	GPQ	GPQ
320	GPS	GPS	GPS	242	GKD	GKD	GKD	164	GAP	GAP	GAP	86	GEP	GEP	GEP	8	GFQ	GFQ	GFQ
321	GPR	GPR	GPR	243	GPK	GPK	GPK	165	GLV	GLV	GLV	87	GIA	GIA	GIA	9	GNP	GNP	GNP
322	GPP	GPP	GPP	244	GAR	GAR	GAR	166	GPR	GPR	GPR	88	GFK	GFK	GFK	10	GEP	GEP	GEP
323	GPV	GPV	GPV	245	GDS	GDS	GDS	167	GER	GER	GER	89	GEQ	GEQ	GEQ	11	GEP	GEP	GEP
324	GPS	GPS	GPS	246	GPP	GPP	GPP	168	GFP	GFP	GFP	90	GPK	GPK	GPK	12	GVS	GVS	GVS
325	GKD	GKD	GKD	247	GRA	GRA	GRA	169	GER	GER	GER	91	GEP	GEP	GEP	13	GPM	GPM	GPM
326	GAN	GAN	GAN	248	GEP	GEP	GEP	170	GSP	GSP	GSP	92	GPA	GPA	GPA	14	GPR	GPR	GPR
327	GIP	GIP	GIP	249	GLQ	GLQ	GLQ	171	GAQ	GAQ	GAQ	93	GPQ	GPQ	GPQ	15	GPP	GPP	GPP
328	GPI	GPI	GPI	250	GPA	GPA	GPA	172	GLQ	GLQ	GLQ	94	GAP	GAP	GAP	16	GPP	GPP	GPP
329	GPP	GPP	GPP	251	GPP	GPP	GPP	173	GPR	GPR	GPR	95	GPA	GPA	GPA	17	GKP	GKP	GKP
330	GPR	GPR	GPR	252	GEK	GEK	GEK	174	GLP	GLP	GLP	96	GEE	GEE	GEE	18	GDD	GDD	GDD
331	GRS	GRS	GRS	253	GEP	GEP	GEP	175	GTP	GTP	GTP	97	GKR	GKR	GKR	19	GEA	GEA	GEA
332	GET	GET	GET	254	GDD	GDD	GDD	176	GTD	GTD	GTD	98	GAR	GAR	GAR	20	GKP	GKP	GKP
333	GPA	GPA	GPA	255	GPS	GPS	GPS	177	GPK	GPK	GPK	99	GEP	GEP	GEP	21	GKA	GKA	GKA
334	GPP	GPP	GPP	256	GAE	GAE	GAE	178	GAS	GAS	GAS	100	GGV	GGV	GGV	22	GER	GER	GER
335	GNP	GNP	GNP	257	GPP	GPP	GPP	179	GPA	GPA	GPA	101	GPI	GPI	GPI	23	GPP	GPP	GPP
336	GPP	GPP	GPP	258	GPQ	GPQ	GPQ	180	GPP	GPP	GPP	102	GPP	GPP	GPP	24	GPQ	GPQ	GPQ
337	GPP	GPP	GPP	259	GLA	GLA	GLA	181	GAQ	GAQ	GAQ	103	GER	GER	GER	25	GAR	GAR	GAR
338	GPP	GPP	GPP	260	GQR	GQR	GQR	182	GPP	GPP	GPP	104	GAP	GAP	GAP	26	GFP	GFP	GFP

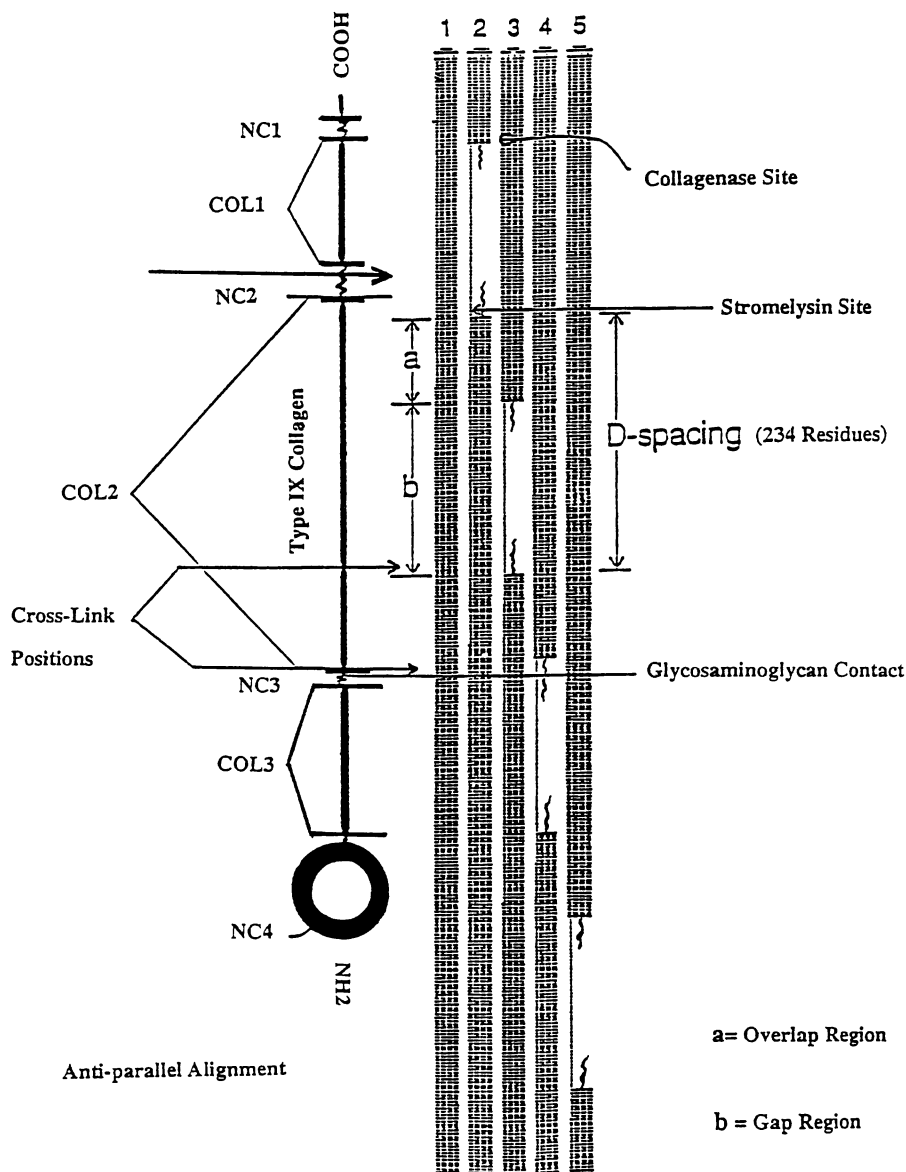


Figure 3. Arrangement of type IX collagen relative to the type II microfibril from Figure 2. Regions in type IX collagen able to form triple-helical structures are labelled "COLn" ($n = 1-3$); non-helical regions are labelled "NCx" ($x = 1-4$). "Cross-Link Positions" designates covalent cross-links between collagen types II and IX; enzyme cleavage sites are labelled as "Collagenase Site" and "Stromelysin Site;" telopeptide segments are labelled as curved lines ("~").

may not correspond to the actual structure that is generated when this single telopeptide chain is complexed with two other identical telopeptide chains, as they exist under native conditions. Analogous to the concept of "Structure-Based Rational Drug Design" where small molecules or peptides are "docked" into the active-site cleft of a specific enzyme in order to find the best fitting molecules, the type II microfibril models provide the pocket-like binding sites for the telopeptide chains. Rather than having to evaluate all the possible conformations of the "free" telopeptide chains, the microfibril models provide information such that the conformational search is restricted to analyzing only the limited set of structures which meet the criteria for: 1) fitting into the binding-site clefts found within the gap spacings; and 2) having the ability to form specific interactions with the collagen sidechain groups found within these gap spacings. Thus, our collagen models are valuable for molecular modeling studies pertaining to the prediction of bioactive telopeptide structures. An example of the N-terminus telopeptides' binding site region is depicted in Color Plate 9. Clearly seen is the Gly1 position of a single collagen triple-helix within the microfibril model. Extending from the Gly1 position of each collagen, in the amino-terminal direction, would be 3 identical telopeptides (telopeptide structures were not incorporated into this figure) whose sequence is given below:

t1 t2 t3 t4 t5 t6 t7 t8 t9 t10 t11 t12 t13 t14 t15 t16 t17 t18 t19
NH₂-Gln-Met-Ala-Gly-Gly-Phe-Asp-Glu-Lys-Ala-Gly-Gly-Ala-Gln-Leu-Gly-Val-Met-Gln-....

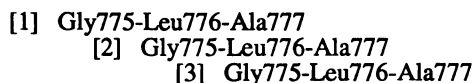
Clearly, the telopeptides are not comprised simply of repeating tripeptide sequences of the form (Gly-X-Y)_n which are so important for allowing the formation of the collagen triple-helical structure. It may be important, though, that under native conditions three telopeptide segments are required for the formation of a non-helical tripeptide complex which then packs within the binding site regions of the gap spacing. Color Plate 9 displays the three-dimensional gap spacing forming the binding site pocket in the fibrillar structure. The approximate dimensions of the binding site pocket are about 8Å x 22Å x 8Å ± 2Å (width x length x depth). Boundaries indicating the predominant types of interactions (i.e., hydrophobic, electrostatic) are color coded adjacent to the fibrillar regions where the telopeptides should pack. Important amino acid sidechains which may contribute to specific telopeptide interactions are also labelled.

3.4.2. Site of Collagenase Interaction. Matrix metalloproteinases (MMPs) are zinc enzymes which have been implicated in the regulation of normal matrix remodeling, cartilage turnover, and in pathologies such as osteoarthritis, tumor metastasis, etc. (51). Some of the most specific of these proteases are collagenases. Since collagenase has been implicated in the degradation of matrix components, therapeutics which target the collagenase enzyme has been a major effort in the pharmaceutical industry. This enzyme cleaves a single peptide bond per α1 chain in the triple helical collagen molecule. Of the 1014 amino acid residues [or 338 (Gly-X-Y) tripeptide repeats] in each chain, the specific sequence in type II collagen which is recognized and cleaved by collagenase is the Gly775-Leu776 peptide bond, producing the classic "one-quarter/three-quarters" cleavage products (52). Although there are many Gly-Leu dipeptide sequences per triple-helical collagen polypeptide chain, only the 775-776 position is known to be cleaved by collagenase. The catalytic site of collagenase contains an essential zinc atom required for bond hydrolysis; the carbonyl oxygen of Gly775 is thought to be chelated to zinc in the substrate-bound state. Although the specific mechanism of hydrolysis is yet to be

NOTE: The color plates can be found in a color section in the center of this volume.

defined, it is believed that a single water molecule acts as a nucleophile to attack the glycine carbonyl center and, hence, triggers the peptide bond hydrolysis.

Extensive work has been published concerning model compounds designed to mimic the Gly-Leu collagenase cleavage site (53). But the three-dimensional structure of ligands as recognized by collagenase has yet to be established. One reason for the difficulty in identifying the bioactive conformation(s) stems from the fact that many ligands which are highly active for collagenase are linear peptidomimetics; their chemical structures only differ from each other to a limited extent (53). The differences in chemical structure give rise to problems in conformational analysis. Perhaps the dilemma is that, under native conditions, collagenase recognizes at least a partial substrate surface which is a complex of several collagen polypeptide chains. Ligands based on a single linear chain are actually attempting to mimic a native tripeptide complex, and, perhaps, additional regions on adjacent collagen molecules. To explain this discrepancy, one must think in terms of the three-dimensional structure of the collagen triple helix. In collagen, it is known that the linear spacing or alignment of each of the three chains with respect to one another is staggered by one amino acid position as shown below. For the sequence cleaved by collagenases, the tripeptide complex of collagen would appear as:



In the above scheme, each polypeptide chain is constrained to a collagen-like conformation which inherently differs from a flexible single polypeptide chain. Furthermore, the tripeptide complex shown above has a well-defined surface contour created from both the folding of three helical polypeptide backbone chains and the nature of the exposed sidechains. As opposed to a single chain model, the peptide system as shown below should be a better and more realistic representation of the (partial) structure which is recognized by collagenase.

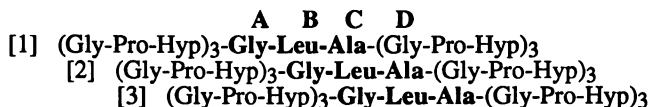


Figure 4 shows the three dimensional collagen structure of the above tripeptide scheme in a "relaxed" stereoviewing mode. Individual polypeptide chains are labelled 1, 2 and 3; each leucine sidechain is shown as a van der Waals dot surface (SYBYL, v.5.5). In the three-dimensional structure of this triple-helical complex in Figure 4, one of the three collagen chains is packed within the microfibril interior and not exposed. It is interesting that the enzyme should recognize a surface consisting of at least two polypeptide chains in this region of a collagen molecule. Furthermore, rather than "seeing" only one nonpolar leucine sidechain as represented by a single linear chain, the enzyme actually senses three regions (labelled B, C and D in above diagram) which contain the Leu and Ala nonpolar sidechains.

Figure 5 depicts a surface in the type II microfibril structure surrounding the Gly775-Leu776 collagenase cleavage site. The fibril surface is composed of the inter-helical packing of several native triple-helical collagens in the arrangement as described above and shown in Figures 2 and 3 (labelled Collagenase Site). A single collagen molecule containing Leu776 is shown as ribbons (wide lines) and the rest of the microfibril (this fibril segment is only 18 residues in length) is shown as lines. The significance of Figure 5 is that it may represent a realistic surface that the

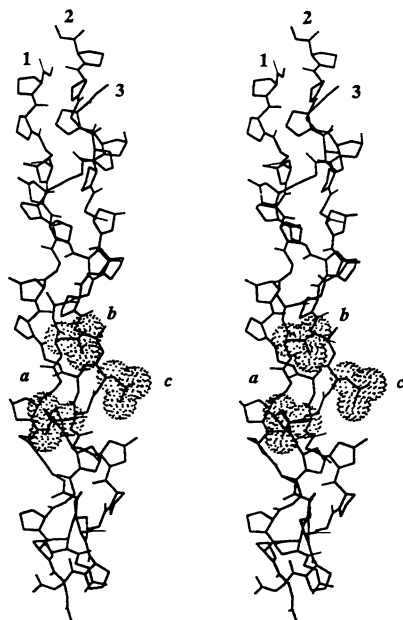


Figure 4. Collagenase cleavage site in type II collagen. Collagen polypeptide chains are labelled 1, 2 and 3. The tripeptide complex is shown in a "relaxed" stereoviewing mode. Each Leucine sidechain is shown as a van der Waals dot surface (SYBYL, v.5.5) and each is labelled a, b and c.

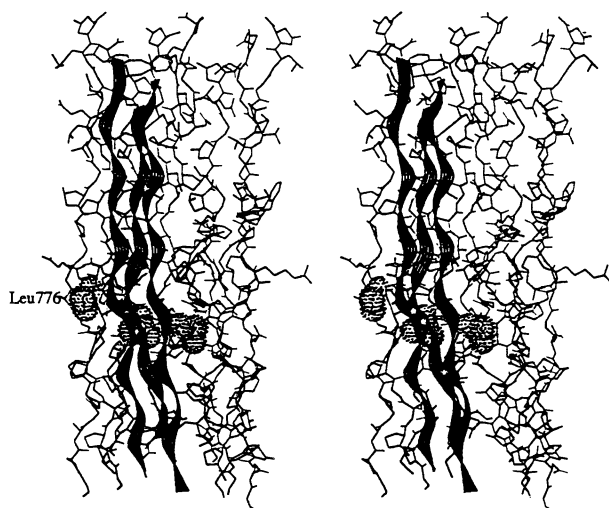


Figure 5. Type II microfibril surface surrounding the Gly775-Leu776 collagenase cleavage site. All three leucines 776 are labelled. A ribbon represents the backbone tracings of a single collagen molecule containing Leu776, with the remainder of the microfibril shown as lines for clarity. The microfibril complex is shown in a "relaxed" stereoviewing mode.

enzyme actually interacts with under native conditions. The implications derived from this type of collagen model can greatly assist both ligand design and in the understanding of collagen-collagenase interactions. We predict that the region in the type II collagen models localized around the Gly775-Leu776 bonds should reflect the contours and amino acid components of the collagenase active-site domain once its x-ray crystallographic structure is solved and published.

3.4.3. Structure-Function Analysis of Proline/Hydroxyproline Distribution in the Microfibril. Structure-function relationships pertaining to specific amino acid residues found in collagen sequences are extremely important. Therefore, since the proline/hydroxyproline content within collagen sequences is approximately 25%, structure-function relationships concerning the properties of the imino acids within fibrous collagens are also very important. Most recent studies pertaining to imino acids concern their structure-function relationships within a single triple-helical collagen molecule. The cyclic ring of prolines and hydroxyprolines, which constrains the ϕ torsional backbone angle to approximately -75 degrees (54), contributes to stabilizing the triple-helical conformation of collagen. Since a single fibrous collagen molecule has dimensions of approximately 13Å in diameter by 3000Å in length, structurally, native collagen is a long thread-like molecule. In order to maintain its three-dimensional characteristics such as its semi-rigidity and rod-like shape, without developing structural "kinking" which would result in the production of sharp bends and knot-like features, the linear sequence of collagen contains a high content of the imino acids. Figure 6 shows a molecular dynamics calculation simulating the possible atomic motions of a triple-helical segment of collagen. Several collagen helices from different time points in the molecular dynamics data set are superimposed upon each other. This triple-helical motif consist of three chains of (Gly-Pro-Hyp)₆ twisted into a right-handed helical complex known for fiber-forming collagens. Although the length of this segment corresponds to only 2% of the actual length of a native collagen molecule, the molecular dynamics simulation depicts interesting features of a triple-helical segment in motion. The total length of the simulation was 500 picoseconds using a distance dependent dielectric constant to implicitly simulate solvent effects. What is evident from this simulation is that the overall cylindrical shape is maintained without any "kinking" effects to produce sharp molecular bends. All internal backbone hydrogen bonds known to stabilize collagen are maintained during the dynamics simulation. Another important stabilizing interaction, not shown in Figure 6, is the "stacking" arrangement of adjacently packed prolines and hydroxyprolines. These van der Waals interactions contribute to stabilizing the triple-helical complex and the constrained ϕ torsional angle (i.e., the ϕ angle is constrained resulting in fewer torsional degrees of freedom for the polypeptide chain) of the imino acids keeps the collagen molecule in a semi-rigid rod-like structure.

The functional role of hydroxyprolines within collagen has been studied further. It has been implicated that the hydroxyl group of hydroxyproline adds additional stability to the collagen molecule through inter- and intra-chain hydrogen bonds; inter-chain bonds are such as those formed between hydroxyproline -OH groups of one chain with the polypeptide backbone carbonyl oxygens of an adjacent chain. Although it has not been shown whether these hydrogen bonds are capable of forming directly, it has been proposed that water molecules may act as "water-bridges" so that these hydroxyproline-based hydrogen bonds can actually form (8). Others have suggested that regions within the collagen sequence which have higher contents of hydroxyprolines, i.e., such as the carboxy-terminal segments, may contribute significantly to nucleating the folding of the collagen triple-helical

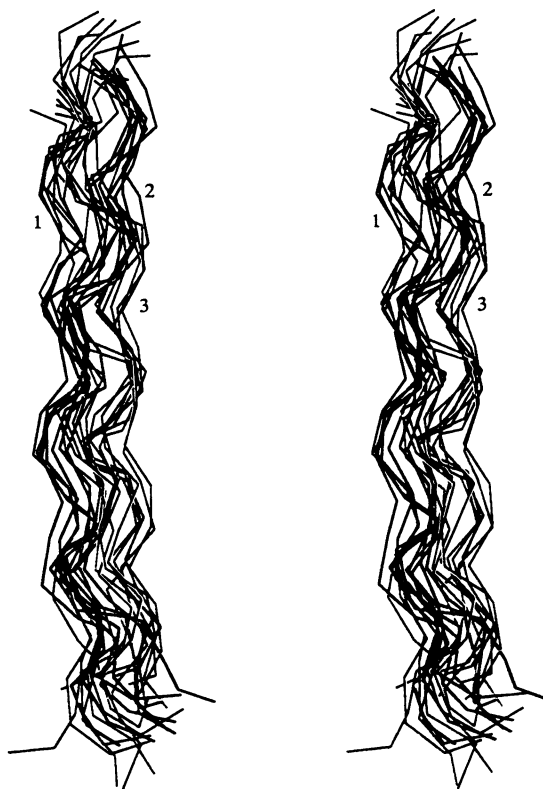


Figure 6. Molecular dynamics simulation of a triple-helical 3(Gly-Pro-Hyp)₆ collagen segment. Each collagen polypeptide chain is shown as a line. Several collagen helices from different time points in the dynamics data set are superimposed in order to define, pictorially, the range of movements of the collagen helical segment. Regions of molecular fluctuation for each chain are labelled 1, 2 and 3.

complex (55, 56). It has been hypothesized that since native collagen folds in the C-terminal to N-terminal direction and the largest hydroxyproline cluster is found in the C-terminal region of fibrous collagen sequences, the hydroxyproline cluster must nucleate the assembly of three collagen chains in a "zipper"-like mechanism. Earlier studies concerning the linear amino acid sequences of fiber-forming collagens found that the positions of hydroxyprolines are not randomly distributed, but are actually found in specific clusters along the collagen primary sequence. These hydroxyproline clusters are thus thought to initiate and maintain the folding of the collagen tripeptide complex throughout the length of the molecule. Since type II collagen molecules exist under native conditions as packed units, the structure-function correlations of specific amino acid sidechains as found and distributed within the three-dimensional microfibril model are both very interesting and important, and may also offer a more realistic understanding of native collagen interactions. Color Plate 10 is a set of space-filling models of the type II microfibril. Color Plate 10 (top) shows the backbone features of the microfibril model indicating the boundaries of two overlap regions and one gap region. Color Plate 10 (middle) shows the distribution of hydroxyprolines and prolines, respectively, within the microfibril unit. Color Plate 10 (upper middle) indicates the importance of the -OH hydroxyl group of hydroxyprolines in providing polar interactions for microfibril stabilizations. It is evident in Color Plate 10 (bottom) that these imino acids are well distributed throughout the microfibril, emphasizing the importance of maintaining overall rigidity throughout the entire length of the microfibril. Imino acids which are only found as small separated clusters in the primary sequence apparently reinforce each other once the individual collagens are packed together within the fibers. Furthermore, also evident in Color Plate 10 (bottom) are specific regions in the microfibril that are depleted of imino acids; these segments in the microfibril model may indicate regions which require some flexibility and/or may indicate sites which are highly susceptible to degradative proteases.

4. Conclusion

In summary, three-dimensional energetic models have been developed for microfibril systems pertaining to fiber-forming collagens (see 23); type II collagen is described herein. These molecular models represent the basic repeating unit in the collagen fiber. Essentially, an entire fiber can be produced simply by stacking these repeating units along the longitudinal axis, for increasing fiber length, and laterally, for increasing fiber width. These three-dimensional energy-minimized "Smith" models contain most of the geometric constraints known for collagen molecules as established from earlier x-ray diffraction studies; they also contain the constraints for lateral packing interactions and possible molecular arrangements as deduced from recent modeling studies and electron microscopy, respectively (37). These microfibril models furthermore are able to reproduce and explain both the negative and positive staining patterns as seen in electron micrographs. Therefore, the three-dimensional microfibril model developed and described here represents a native three-dimensional map of both the surface interactions and interactions between individual collagen molecules.

The collagen microfibril models allow one to study structure-function relationships described previously through experimentation. One example is the identification of naturally occurring cross-links found in collagens which can now be studied three-dimensionally in these modeled fiber systems (23, 24, 37). Formation or modification of specific cross-links due to either aging or certain diseases can be analyzed for their effects on collagen packing. Additionally, cross-linked regions can be studied for their stabilizing effect on the fiber. It is also known that other

collagen types (i.e., type II and type IX in cartilage) interact with fibrous collagens and are covalently attached through naturally occurring cross-links (57). Hence, the fiber models presented here are very useful for gaining three-dimensional insights into specific biological interactions described previously through experimental research.

Another important use for these collagen microfibril models is to trigger new ideas and to better understand native collagen systems. As stated previously, the three-dimensional properties (i.e., stereochemistry) of the microfibril models allow for a more precise study of collagen chemistry based on its folding and sidechain distribution. This model has potential uses in both industrial and academic research (37).

For industrial research pertaining to the development of reagents which modify collagen structures and/or inhibit undesirable enzyme reactions, the analysis of specific collagens is very important. Cross-linkers that are highly specific and potent will have better collagen modifying properties as opposed to reagents which non-specifically interact with different collagen types; in addition, mechanism-based ligands are easier to control than those with unknown mechanism of reaction. The concept of higher specificity and binding affinity is also very important in *target based* drug design. Specificity and potency are essential for decreasing negative side effects such as *in vivo* toxicity of potential drug candidates. Drug design for diseases relevant to collagen environments or systems are very important. One example is the inhibition of extracellular matrix proteases (51) for treating arthritic diseases; by understanding the size, shape and chemistry of the native substrate, one can more effectively identify the important parameters for designing better and more effective ligands.

For basic research interest, the study of native collagen fiber systems should bring about a better understanding of structure-function correlations for collagen(s). Aspects of protein folding, stability, sidechain interactions and relationships can be considered. Structure-function correlations of protein binding sites and proteolytic cleavage sites can be understood in terms of the matrix environment. Prediction of the possible structure of the terminus telopeptides may also be reasonable since their conformation may be dependent on the structural constraints provided by the fibril model. Naturally occurring cross-link sites which have not yet been identified can be proposed. Furthermore, the effects of experimentally manipulated amino acid mutations can be predicted prior to obtaining experimental information. The examples given here represent some of the many possible applications for these native, collagen microfibril models.

Acknowledgments

We would like to thank Drs. Matthew R. Pincus and Eleanor M. Brown for their valuable comments. Appreciation is also given to Dr. Joseph Y. Chang and Mr. Nelson Campbell for helpful suggestions.

5.0. Literature Cited

1. Piez, K. A. In *Extracellular Matrix Biochemistry*; Piez, K. A. and Reddi, A. H., eds.; Elsevier: New York, 1984; pp. 1-39.
2. Martin, G. R., Timpl, R., Muller, P. K. and Kuhn, K. *TIBS* **1985**, *10*, 285-287.
3. Miller, E. J. *Ann. N. Y. Acad. Sci.* **1985**, *460*, 1-13.
4. Gordon, M. K., Gerecke, D. R., Dublet, B., Van Der Rest, M., Sugrue, S. P. and Olsen, B. R. *Ann. N. Y. Acad. Sci.* **1990**, *580*, 8-16.

5. Veis, A., Miller, A., Leibovich, S. J. and Traub, W. *Biochim. Biophys. Acta* **1979**, *576*, 88-98.
6. Na, G. C., Butz, L. J. and Carroll, R. J. *J. Biol. Chem.* **1986**, *261*, 12290-12299.
7. Na, G. C., Butz, L. J., Bailey, D. G. and Carroll, R. J. *Biochemistry* **1986**, *25*, 958-966.
8. Suzuki, E., Fraser, R. D. B. and MacRae, T.P. *Int. J. Biol. Macromol.* **1980**, *2*, 54-56.
9. Chapman, J. A. and Hulmes, D. J. S. In *Ultrastructure of the Connective Tissue Matrix*; Ruggeri, A. and Motta, P. M., eds.; Martinus Nijhoff Publishers: Boston, MA, 1984; pp. 1-33.
10. Brodsky, B. and Eikenberry, E. *Ann. N. Y. Acad. Sci.* **1985**, *460*, 73-85.
11. Eikenberry, E. F. and Brodsky, B. *J. Mol. Biol.* **1980**, *144*, 397-404.
12. Brodsky, B., Eikenberry, E. F., Belbruno, K. and Sterling, K. *Biopolymers* **1982**, *21*, 935-951.
13. Chew, M. W. K. and Squire, J. M. *J. Biol. Macromol.* **1986**, *8*, 27-36.
14. Ripamonti, A., Roveri, N., Braga, D., Hulmes, D. J. S., Miller, A. and Timmins, P. A. *Biopolymers* **1980**, *19*, 965-975.
15. Miller, A. and Wray, J. S. *Nature* **1971**, *230*, 437-439.
16. Miller, A. and Parry, D. A. D. *J. Mol. Biol.* **1973**, *75*, 441-447.
17. Parry, D. A. D. and Craig, A. S. *Nature* **1979**, *282*, 213-215.
18. Squire, J. M. and Freundlich, A. *Nature* **1980**, *288*, 410-413.
19. Smith, J. W. *Nature* **1968**, *219*, 157-158.
20. Veis, A. and Yuan, L. *Biopolymers* **1975**, *14*, 895-900.
21. Woodhead-Galloway, J. In *Connective Tissue Matrix: Topics in Molecular and Structural Biology*; Hukins, D. W. L., ed.; Verlag Chemie: Weinheim, **1984**, Vol. 5; pp. 133-160.
22. Meek, K. M., Chapman, J. A. and Hardcastle, R. A. *J. Biol. Chem.* **1979**, *254*, 10710-10714.
23. Chen, J. M., Fearheller, S. H. and Brown, E. M. *J. Am. Leather Chem. Assoc.* **1991**, *86*, 475-486.
24. Chen, J. M., Fearheller, S. H. and Brown, E. M. *J. Am. Leather Chem. Assoc.* **1991**, *86*, 487-498.
25. Hodge, A. J. and Petruska, A. J. In *Aspects of Protein Structure*; Ramachandran, G. N., ed.; Academic Press: London, 1963; pp. 289-300.
26. Piez, K. A. and Trus, B. L. *J. Mol. Biol.* **1977**, *110*, 701-704.
27. Piez, K. A. and Trus, B. L. *J. Mol. Biol.* **1978**, *122*, 419-432.
28. Traub, W. *FEBS Lett.* **1978**, *92*, 114-120.
29. Giraud-Guille, M.-M. *Mol. Cryst. Liq. Cryst.* **1987**, *153*, 15-30.
30. Torchia, D. A. and Vanderhart, D. L. *J. Mol. Biol.* **1976**, *104*, 315-321.
31. Jelinski, L. W., Sullivan, C. E. and Torchia, D. A. *Nature* **1980**, *284*, 531-534.
32. Torchia, D. A., Hiyama, Y., Sarkar, S. E. and Sullivan, C. E. *Biopolymers* **1985**, *24*, 65-75.
33. Hulmes, D. J. S., Jesior, J.-C., Miller, A., Berthet-Colominas, C. and Wolff, C. *Proc. Natl. Acad. Sci.* **1981**, *78*, 3567-3571.
34. Hulmes, D. J. S. and Miller, A. *Nature* **1981**, *293*, 239-240.
35. Miller, A. *TIBS* **1982**, *7*, 13-18.
36. Piez, K. A. and Trus, B. L. *Biosci. Rep.* **1981**, *1*, 801-810.
37. Chen, J. M., Kung, C. E., Fearheller, S. H. and Brown, E. M. *J. Prot. Chem.* **1991**, *10*, 535-552.
38. Weiner, P. K. and Kollman, P. A. *J. Comp. Chem.* **1981**, *2*, 287-303.
39. Weiner, S. J., Kollman, P. A., Case, D. A., Singh, U. C., Ghio, C., Alagona, G., Profeta, S. and Weiner, P. A. *J. Am. Chem. Soc.* **1984**, *106*, 765-784.

40. Baldwin, C. T., Reginate, A. M., Smith, C., Jimenez, S. A. and Prockop, D. J. *Biochem. J.* **1989**, *262*, 521-528.
41. Lee, D. D. and Glimcher, M. J. *J. Mol. Biol.* **1991**, *217*, 487-501.
42. Eyre, D. R., Paz, M. A. and Gallop, P.M. *Ann. Rev. Biochem.* **1984**, *53*, 717-748.
43. Pope, F. M. and Nicholls, A. C. In *Molecular Medicine*; Malcolm, A. D. B., ed.; IRL Press: Oxford, 1984, Vol. I; pp. 117-175.
44. Jones, E. Y. and Miller, A. *Biopolymers*, **1987**, *26*, 463-480.
45. Chen, J. M., Sheldon, A. and Pincus, M. R. *J. Biomol. Structure and Dynamics* **1993**, *10*, 1067-1089.
46. Wu, J. J., Woods, P. E. and Eyre, D. R. *J. Biol. Chem.* **1992**, *267*, 23007-23014.
47. Otter, A., Scott, P. G. and Kotovych, G. *J. Am. Chem. Soc.* **1987**, *109*, 6995-7001.
48. Otter, A., Scott, P. G. and Kotovych, G. *Biochemistry* **1988**, *27*, 2291-2313.
49. Otter, A., Scott, P. G. and Kotovych, G. *Biochemistry* **1989**, *28*, 8003-8010.
50. Otter, A., Scott, P. G. and Kotovych, G. *Biopolymers* **1993**, *33*, 1443-1459.
51. Woessner, J. F. *FASEB J.* **1991**, *5*, 2145- 2154.
52. Miller, E. J., Harris, E. D., Chung, E., Finch, J. E., McCroskery, P. A. and Butler, W. T. *Biochemistry* **1976**, *15*, 787.
53. Johnson, W. H., Roberts, N. A. and Borkakoti, N. *J. Enz. Inhib.* **1987**, *2*, 1-22.
54. Miller, M. H., Nemethy, G. and Scheraga, H. A. *Macromol.* **1980**, *13*, 470-478.
55. Roth, W. and Heidemann, E. *Biopolymers* **1980**, *19*, 1909-1917.
56. Germann, H. P. and Heidemann, E. *Biopolymers* **1988**, *27*, 157-163.
57. Bailey, A. J., Light, N. D. and Atkins, E. D. T. *Nature* **1980**, *288*, 408-410.

RECEIVED April 14, 1994

Chapter 11

Calculations of Association Free Energies Separation of Electrostatic and Hydrophobic Contributions

Gregory King and Robert A. Barford

Eastern Regional Research Center, Agriculture Research Service,
U.S. Department of Agriculture, 600 East Mermaid Lane,
Philadelphia, PA 19118

Relative association free energies of six sulfonamide/ β -cyclodextrin inclusion complexes are calculated using a thermodynamic cycle that separates these reactions into purely electrostatic and hydrophobic components. Electrostatic free energy differences are calculated using slow-growth thermodynamic integration, and hydrophobic free energy differences are obtained using an empirical relationship based on the difference between the solvent-accessible surface areas of the solute species in their associated and dissociated states. Two sets of calculations are performed: one in which the model system includes solvent, and the other in which the model system does not include solvent. The calculations performed with the solvated model are accurate to roughly ± 1.5 kcal/mol, and correctly select the preferred of the two possible binding conformations in three out of the four cases examined.

The non-covalent association of two or more solute species in solution to form complexes, aggregates, micelles, and other entities is a common phenomenon. Associations involving hydrophobic or amphiphilic solutes in water are especially interesting because entropic (hydrophobic) effects play a large role, and biological molecules such as proteins and lipids are often involved (*1*). Experimental studies can accurately determine the equilibrium constants of these associations, but often do not provide any more than this basic thermodynamic information. It is therefore of interest to develop computational methods for the study of association processes, since such methods offer the possibility of examining these processes in greater detail.

When considering the association of solute molecules X and Y to form the non-covalent complex X:Y in a given solvent, the most important quantity a computational method must be able to produce accurately is the association free energy, ΔG , (or equivalently, the equilibrium constant, K_{eq}) of this process, since knowledge of this quantity allows one to calculate the concentrations of the involved species at equilibrium conditions.

One way to calculate K_{eq} via computer simulation would be to prepare a very large model system containing many molecules (hundreds) of both X and Y immersed in the solvent of interest (we will call this Hypothetical Method 1). The

This chapter not subject to U.S. copyright
Published 1994 American Chemical Society

system would be brought to equilibrium either through Monte Carlo (MC) or molecular dynamics (MD) simulation, using an appropriate potential energy function. At equilibrium the system would contain representative numbers of bound and unbound species, and K_{eq} could be determined simply as $K_{eq} = [X:Y]/([X][Y])$, where square brackets denote molar concentrations (for simplicity, we will assume in this work that activities of species are equal to their molar concentrations).

Another method (Hypothetical Method 2) would be to simulate a very long MC or MD trajectory of a smaller model system. After an initial equilibration period, each system configuration generated in such a simulation may be thought of as a member of a statistical mechanical ensemble. The equilibrium constant could then be calculated based on the number of bound and free species in the ensemble. (Hypothetical Method 1 uses information from an instantaneous "snapshot" of a very large system to calculate K_{eq} . Hypothetical Method 2 uses many "snapshots" of a smaller system.)

The problem with these two methods, however, is that energy barriers may inhibit the interconversion of species from reactant state to product state (or *vice versa*). For instance, if an energy barrier of 3 kcal/mol separates reactant from product states, there is a relative probability of 1 in 148 (based on the Boltzmann distribution of energies at 300° K: $\exp\{-3/0.6\}$) that solute species will possess enough energy to reach the top of the barrier, where they could possibly convert from one state to the other. For a 5 kcal/mol barrier (the energy of a typical hydrogen bond), the relative probability drops to just 1 in 4160. It may therefore be difficult in Hypothetical Method 1 to prepare a system that is large enough, or in Hypothetical Method 2 to conduct a simulation that is long enough to generate truly accurate statistics for a given system, and thus the K_{eq} calculated from the data obtained from such simulations could be in error by orders of magnitude.

It turns out to be more productive (from a computational point of view) to focus on ΔG rather than K_{eq} . The main reason for this is that various schemes – known as thermodynamic cycles – may be devised to take advantage of the fact that free energy is a thermodynamic state function. This means that the free energy change accompanying a reaction is independent of the path the system takes to get from the reactant state to the product state (provided that the transformation is carried out reversibly). Choosing an appropriate thermodynamic cycle often enables one to circumvent difficulties (such as energy barriers) that would be encountered if one attempted to simulate an actual reaction directly. In other words, a thermodynamic cycle allows one to replace the actual reaction coordinate with a fictitious one that is more efficient to simulate (see (2 – 6) for descriptions and applications of thermodynamic cycles in computational studies).

In this work we use a thermodynamic cycle that was developed by Nicholls *et al.* (6) specifically for treating non-covalent association or dissociation reactions. This cycle separates these reactions into purely electrostatic and hydrophobic components. Contributions to association free energies from other sources (e.g. differences in van der Waals interactions) are quite small compared to the electrostatic and hydrophobic terms, and are therefore neglected. This method yields relative free energies rather than standard free energies, and thus calculated free energies differ from the corresponding standard free energies by a constant.

Our study consists of two sets of calculations carried out with different model systems. In the first set of calculations (Method 1), the models include solute and water molecules whereas in Method 2 only solute atoms are included in the models. The main purpose of this work is to test the reliability of Methods 1 and 2 for calculating association free energies. The reliability of the two methods is assessed by comparing the relative association free energies obtained from the calculations to the standard free energies obtained from experiments (allowing for experimental and calculated values to differ by a constant).

We chose a set of six sulfonamide/ β -cyclodextrin inclusion complexes on which to test our methodology, because the possible use of β -cyclodextrin (β CD) or β CD derivatives as extraction agents is being explored by our laboratory's drug residue detection methods unit. These complexes were also chosen because the sulfonamides may conceivably bind to β CD in one of two distinct geometries, and thus we can also test whether our method correctly selects the preferred conformation of each complex.

^1H NMR studies (7) of four of the six sulfonamide/ β CD inclusion complexes were performed in order to determine the actual conformations of the complexes, because previous experimental work (8, 9) did not provide information at this level of detail. (NMR studies of the other two sulfonamide/ β CD complexes were not performed because these two sulfonamides were not available at the time of this study.) The NMR studies also yielded the association constants and stoichiometries of the complexes, establishing a three-way consistency check between the calculations, the current experimental results (7), and the previous experimental results (8, 9).

The remainder of this paper is organized as follows. In the Methods section we describe the methods and models used in our calculations. In the Results and Discussion section we list the results of the calculations and compare them to the corresponding experimental values.

Methods

The sulfonamide and β CD molecules used in this study were constructed using Tripos Associates' SYBYL software package. Residual charges were assigned to the β CD atoms using the partial equalization of orbital electronegativity (PEOE) algorithm of Gasteiger and Marsili (10), and to the sulfonamide atoms using a linear combination of Gasteiger-Marsili charges and charges obtained from MNDO (11) calculations (these two methods of assigning charges are available as part of the SYBYL package). The combination of charges used was

$$q = 0.55q_{\text{GM}} + 0.47q_{\text{MNDO}} \quad (1)$$

where q_{GM} and q_{MNDO} denote the Gasteiger-Marsili and MNDO charge sets, respectively. The relationship expressed in equation 1 was found through a least-squares regression analysis in which the dipole moments obtained with charge set q were fit to the actual dipole moments of a series of 85 small, rigid molecules (King, G., unpublished results).

SYBYL was also used to build the sulfonamide/ β CD complexes. The sulfonamides were docked within the β CD cavity using the program's interactive graphics capabilities, and the complexes' energies were then minimized to nearby potential energy minima. Better structures (in terms of proximity to the complexes' absolute global energy minima) were obtained in an annealing process described later in this section.

We conducted our study using two different model systems. In both models, systems are described at the atomic level (i.e. atoms are the entities that interact with one another through an empirical potential energy function). The potential energy function (or "force field", as it is sometimes called), which includes bond stretching, angle bending, out-of-plane bending, torsional, van der Waals, and electrostatic terms, is as described in (12). Instead of using an explicit hydrogen bonding term, atoms involved in hydrogen bonds are allowed to approach one another more closely by scaling the van der Waals radii of the participating atoms by the factor 0.8. Electrostatic interactions between atoms i and j are calculated as $332q_iq_j/\epsilon_{ij}r_{ij}$, where q_i and q_j are the atomic charges (in units of electrons), r_{ij} is the distance between the

two atoms (in Å), ϵ_{ij} is the dielectric function (unitless), and the factor 332 converts energies from these units to kcal/mol. In the first model (Method 1), water molecules are included, and a constant dielectric function of unity is used in the calculation of electrostatic interactions. In the second model (Method 2), only the solute atoms are explicitly considered, and a distance-dependent dielectric function $\epsilon_{ij} = \epsilon(r_{ij}) = 1 + r_{ij}$ (13) is used in the calculation of electrostatic interactions. This dielectric function is intended to mimic the screening effect the missing water would have on the solute-solute electrostatic interactions. Except for the difference in dielectric functions, the Method 1 and Method 2 force fields are the same. Method 1 simulations are much more time-consuming than the corresponding Method 2 simulations, because Method 1 systems contain many more atoms.

The difference between our method of calculating electrostatic free energies and the method used in (6) is worth noting. In (6) a Finite Difference Poisson-Boltzmann (FDPB) method (14) is used. In this approach the system is represented as a set of points on a regular three-dimensional lattice, each assigned a charge density and dielectric constant. The charge densities of the lattice points are obtained by projecting the charges of the solute atoms onto them. Lattice points within the van der Waals radii of solute atoms are assigned a dielectric constant of 2, and lattice points outside of the solute van der Waals radii (where water would be) are assigned a dielectric constant of 80. The FDPB equation is solved iteratively, yielding the electrostatic potential at each lattice point. The electrostatic potentials may then be used to find the electrostatic free energies. This method is computationally efficient, but oversimplifies the dielectric characteristics of both solute and solvent. Also, the method does not take into account thermal fluctuations, which in some cases may have a significant effect on solute structures.

In this work we use the well-known thermodynamic integration method (15 - 17) to calculate electrostatic free energy differences. The thermodynamic integrations are carried out over the course of MD simulations. An MD simulation may be thought of as a way to sample the configuration space available to the molecules in a given model system. This is in contrast to the calculation performed in (6), in which only one solute configuration is used.

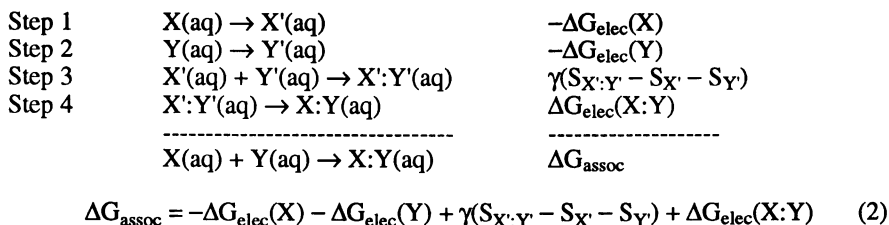
The hydrophobic contributions to association free energies are calculated using the method described in (6). This method is based on the observation that the free energy of transferring a nonpolar solute from a nonpolar solvent to a polar, hydrogen-bonding solvent such as water scales linearly with the accessible surface area presented to the solvent by the solute (6). This predominantly entropic increase in free energy is a result of the reduced number of hydrogen-bonding partners available to solvent molecules located at the surface of a nonpolar solute. (This reduction in hydrogen-bonding partners is the same phenomenon responsible for the interfacial energy (surface tension) that exists between two immiscible solvents.) The solvent-accessible surface area of each solute species is simply calculated and multiplied by a scale factor (we use the recommended value of $0.05 \text{ kcal mol}^{-1} \text{ \AA}^{-2}$), and the difference between the energy values obtained for the associated and dissociated states is taken as the hydrophobic contribution to the association free energy.

The SCAAS model (18) (with several new features) was employed in all of the free energy calculations, using the atomic charge parameters as described above, and the SYBYL force field (12) for bonded and van der Waals interactions. The charges and van der Waals parameters for water were taken from (18). A cutoff of 8 Å was used for all nonbonded interactions. In order to prevent spurious charges from being created by the application of the cutoff, the nonbonded interaction lists were based on electrically-neutral groups of atoms. A time step of 2.0 fsec was used in propagating the MD simulations, with gentle velocity scaling applied to keep the system temperature near the specified value.

In the SCAAS model the solute is immersed within a spherical droplet of solvent molecules (water, in this case). For Method 1 calculations the radii of the solvent spheres were set to the radial distance of the solute atom farthest from the origin plus 6 Å. For Method 2 calculations no solvent was included.

The fact that the model systems used in computer simulations must necessarily be greatly truncated versions of much larger systems means that artificial boundaries are present in these systems. Any abnormal effects these boundaries create in the model systems must be addressed. In the SCAAS model, solvent molecules are subjected to a radial constraint that maintains the proper solvent density and prevents loss of solvent due to evaporation. There is also an important polarization constraint (18) on the distribution of orientations of dipole moment vectors for solvent molecules near the surface of the system. This constraint counteracts the tendency (due to the presence of the solvent-vacuum interface) of solvent molecules near the surface to adopt orientations different from those expected in a sphere excised from an actual system. These constraints obviously apply to Method 1 calculations only, because Method 2 calculations do not include solvent.

The thermodynamic cycle used in the calculations is illustrated in Figure 1. In this cycle, the association reaction $X(\text{aq}) + Y(\text{aq}) \rightarrow X:Y(\text{aq})$ is decomposed into the following four steps:



where X represents a sulfonamide, Y is β CD and $X:Y$ is the non-covalent complex.

In Step 1, the residual atomic charges of a single X molecule (isolated from other solute species) are all gradually reduced to zero to produce X' . In Step 2, a single Y molecule is similarly discharged in the absence of any other solute species to yield Y' .

In Step 3, the chargeless entities X' and Y' are brought together to form the complex $X':Y'$. There are two contributions to the change in free energy connected with this reaction. The first contribution, $\Delta G_{\text{hphobic}}$, is due to the so-called hydrophobic effect. The term "hydrophobic effect" is used to describe various phenomena caused by the disruption of water's hydrogen-bonding network. Such an effect occurs where water comes into contact with nonpolar species. The number of hydrogen bonds effected by the presence of such interfaces is proportional to the water-accessible surface area of the nonpolar species, and thus the free energy penalty for the disruption of the hydrogen-bonding network is proportional to this surface area. $\Delta G_{\text{hphobic}}$ is thus proportional to the change in solvent-accessible surface area that occurs when isolated X' and Y' molecules are brought together to form the complex $X':Y'$. $\Delta G_{\text{hphobic}}$ is calculated empirically as $\Delta G_{\text{hphobic}} = \gamma(S_{X':Y'} - S_{X'} - S_{Y'})$, where a microscopic surface tension $\gamma = 0.05 \text{ kcal mol}^{-1} \text{ \AA}^{-2}$ is used (6) and S_A is the solvent-accessible surface area of solute species A .

The other contribution to the free energy difference in Step 3 is due to changes in van der Waals interactions. Upon complexation, some of the solute-water van der Waals interactions are replaced by solute-solute and water-water van der Waals interactions. The total number of interactions remains roughly the same, however, and since deviations in the magnitudes of van der Waals interaction

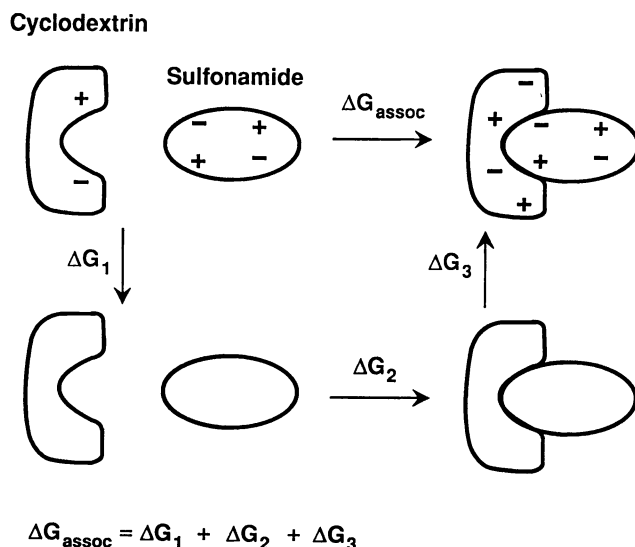


Figure 1: The thermodynamic cycle employed in the calculation of association free energies. In the first stage of the cycle (ΔG_1), the atomic charges of the reactants (which are simulated separately, and are thus much farther apart than is indicated in the figure) are removed. In the second stage (ΔG_2), the discharged (hydrophobic) reactants are brought together to form a complex. In the third stage (ΔG_3), the original atomic charges are restored. The free energy differences for the first and third stages involve only changes in electrostatic variables, and the free energy difference for the second stage involves only the change in the solvent-accessible surface area of the solutes that occurs upon complexation.

energies are very small, the contribution to the free energy change from van der Waals terms is negligible.

In Step 4, the charges of the hydrophobic complex X':Y' are gradually restored to their original values to yield the complex X:Y.

Steps 1 and 2 are shown in Figure 1 as taking place in a single step with a free energy difference $\Delta G_1 = -\Delta G_{\text{elec}}(X) - \Delta G_{\text{elec}}(Y)$. In practice, separate calculations are performed to obtain $-\Delta G_{\text{elec}}(X)$ and $-\Delta G_{\text{elec}}(Y)$. ΔG_2 in Figure 1 corresponds to the hydrophobic free energy change $\Delta G_2 = \Delta G_{\text{hphobic}} = \gamma(S_{X':Y'} - S_X - S_{Y'})$. ΔG_3 in Figure 1 corresponds to $\Delta G_{\text{elec}}(X:Y)$, the step that completes the thermodynamic cycle. The complete equation for the reaction, ΔG_{assoc} , is shown in equation 2.

The charging and uncharging free energies (ΔG_{elec}) are calculated using slow-growth thermodynamic integration. In these slow-growth simulations, the system is gradually transformed from state r (reactants) to state p (products) by incrementally changing a coupling parameter λ from 0 to 1 over the course of an MD simulation, and the free energy difference ΔG_{pr} is obtained by evaluating the integrodifferential equation:

$$\Delta G_{\text{pr}} = G_p - G_r = \int_0^1 d\lambda (\partial G / \partial \lambda) = \int_0^1 d\lambda \langle U_p(\Omega) - U_r(\Omega) \rangle_\lambda \quad (3)$$

where $U_r(\Omega)$ and $U_p(\Omega)$ are the potential energy functions of states r and p, respectively, Ω represents the entire set of coordinates used to describe the system, and $\langle \theta \rangle_\lambda$ denotes an ensemble average of θ on potential surface $U(\Omega, \lambda)$, which is specified by the relation:

$$U(\Omega, \lambda) = U_r(\Omega) + \lambda[U_p(\Omega) - U_r(\Omega)] \quad (4)$$

Steps 1, 2, and 4 in the above reaction scheme each entail the performance of slow-growth calculations in separate systems. Sulfonamide+water systems typically contain about 390 water molecules, while the β CD+water and sulfonamide/ β CD+water systems contain about 460 water molecules.

Before a given slow-growth calculation is initiated, each system is taken through the following four-step annealing process: a 15 psec MD simulation during which the temperature is raised (linearly) from 0 K to 500 K, followed by 5 psec at 500 K, followed by 10 psec during which the temperature is lowered from 500 K to 300 K, followed by a 5 psec equilibration period at 300 K. This annealing procedure allows each system to escape from local potential energy minima and to find more favorable regions of configuration space. After this annealing/equilibration process is completed, slow-growth integration calculations are carried out over 40 psec simulations at 300 K.

The reproducibility of slow-growth free energy calculations is usually checked by performing both "forward" and "backward" calculations, where "forward" means evaluating equation 3 as it is written, and "backward" means exchanging the limits of integration in equation 3. The amount of CPU time required for Method 1 calculations made it unfeasible to perform more than one slow-growth integration for each electrostatic free energy. However, based on forward and backward integrations performed on smaller systems, the error range for electrostatic free energies of uncharged solute species is about 0.5 kcal/mol with this methodology.

The average solvent-accessible surface areas $S_{X':Y'}$, S_X , and $S_{Y'}$ may all be calculated during the electrostatic free energy calculations of the relevant systems, and therefore $\Delta G_{\text{hphobic}}$ may be determined without any additional simulations.

Results and Discussion

The six sulfonamides examined in the computational portion of this study are illustrated in Figure 2. The four sulfonamides examined in our ^1H NMR study (7) are designated in Figure 2 as compounds 1, 2, 3, and 4. The results of the calculations are summarized in Tables I, II and III. In Table I, the relative association free energies of the six sulfonamide/ β CD complexes calculated using the two different model systems (Methods 1 and 2) are reported. Association free energies obtained from solubility studies (8), HPLC studies (9), and our ^1H NMR studies (7) are also given. (In the last three columns, standard association free energies, ΔG° , were obtained from the corresponding association constants, K_{eq} , with the relation $\Delta G^\circ = -RT \ln K_{\text{eq}}$, where R is the ideal gas constant and T is the absolute temperature.) The Method 1 and Method 2 columns each have two rows of data. The first of the two rows corresponds to the complex conformation with the common end (the aniline ring) inserted into the β CD cavity, and the second row corresponds to the conformation with the variable end inserted into the β CD cavity.

All calculated free energies are relative, rather than standard, since the solute species in the model systems are not in their standard states, and correcting to standard state concentrations in our model systems is not straightforward. The differences between free energy values within a given column of the table, however, should be comparable to the differences in adjacent columns.

The first test of the two computational methods is whether they produce the six association free energies in the same order as the corresponding experimental values. As can be seen from Table I, neither of the two methods was able to satisfy this test. It is of interest to note, however, that by shifting each of the Method 1 values up or down by 0 to 1.5 kcal/mol, the proper order can be obtained. This means that the error range associated with Method 1 may be as small as, but is probably not less than ± 1.5 kcal/mol. (Note: this is not a definitive estimate of the error bars associated with these calculations. The error bars of free energy calculations are notoriously difficult to obtain. More well-defined estimates of the errors could be obtained if each calculation were repeated a number of times, as is done for normal laboratory experiments. The amount of time these additional calculations would have required prevented us from attempting this.) Method 1 in its current form is therefore not useful if one needs to discriminate amongst association free energies that differ by less than 1.5 kcal/mol, as is the case in this study. A better test of the computational methods would have been to examine a set of reference complexes having a wider range of association free energies.

Least-squares straight line fits of the calculated free energies to the experimental values yield basic trends in the calculated results. The least-squares fit for Method 1 results is: $\Delta G_{\text{expt}} = 2.56(\Delta G_{\text{calc},1} + 18.6)$, and the fit for Method 2 results is: $\Delta G_{\text{expt}} = -1.59(\Delta G_{\text{calc},2} + 21.1)$. The correlation coefficients are 0.171 and 0.307 for Methods 1 and 2, respectively, neither of which is very good. We expected to obtain fits of the form $\Delta G_{\text{expt}} = a_1(\Delta G_{\text{calc}} + a_0)$ with the coefficient a_1 close to unity. The negative correlation between ΔG_{expt} and $\Delta G_{\text{calc},2}$ values warns against the use of Method 2 in any future work.

The second test of the computational methods is whether they correctly select the preferred conformation of each sulfonamide/ β CD complex. Calculations were performed for both of the two possible conformations. The conformation that yielded the lower free energy was taken as the preferred conformation. The actual preferred conformations of four of the six complexes were determined in our ^1H NMR study (7). A comparison of the predicted and actual preferred conformations is given in Table II. Entries designated as "Ring 1" indicate that the preferred conformation is the one with the common aniline ring inserted within the β CD

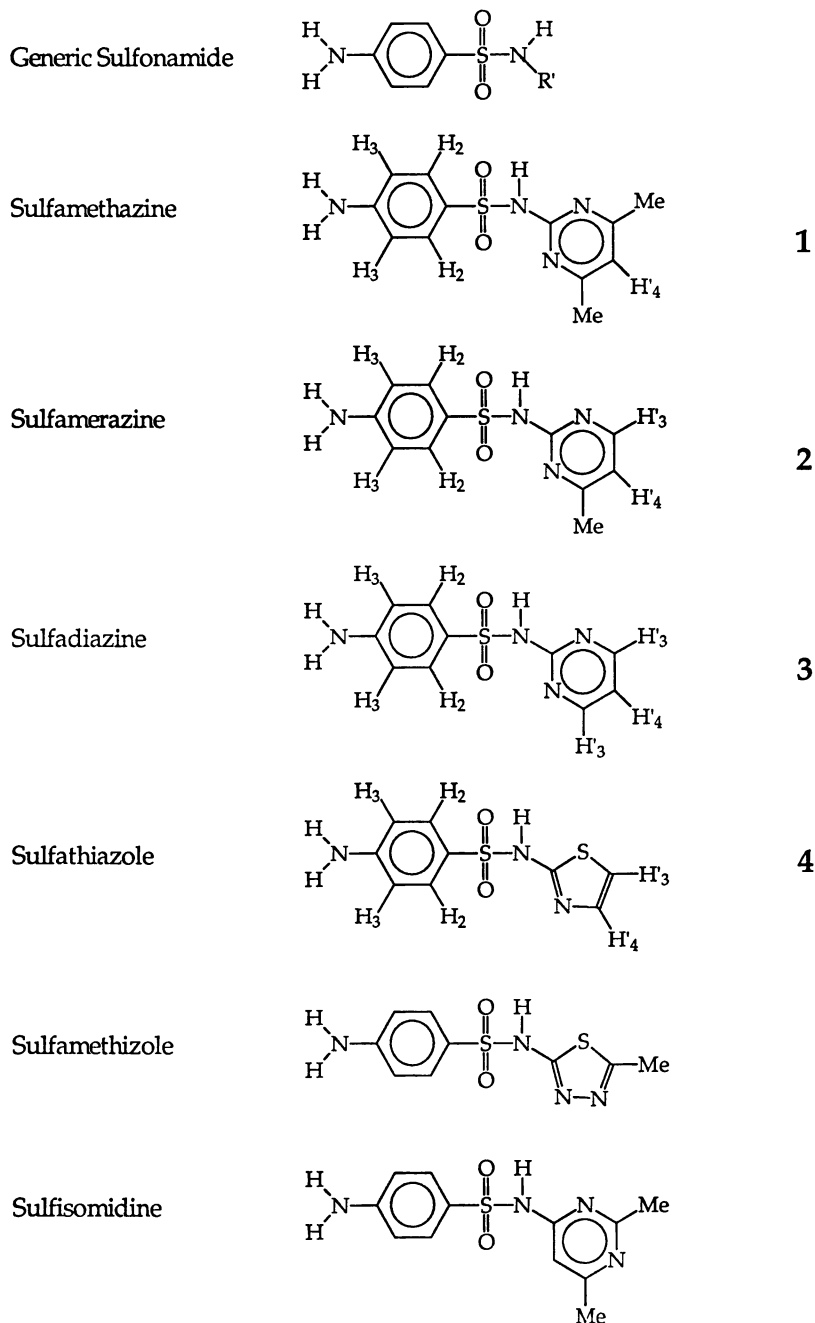


Figure 2: Schematic structures of a generic sulfonamide molecule and the six specific sulfonamides examined in our computational study. The four sulfonamides examined in our ^1H NMR study (7) are labeled 1, 2, 3, and 4.

Table I. Association Free Energies of Sulfonamide/ β CD Inclusion Complexes

Sulfonamide	Method 1 ^a	Method 2 ^b	Solubility ^c	HPLC ^d	¹ H NMR ^e
sulfathiazole	-21.1 -14.9	-18.3 -17.4	-4.44	-4.46	-4.45±0.02
sulfamethizole	-16.8 -18.8	-18.5 -18.3	-4.13	-4.23	—
sulfadiazine	-19.8 -21.8	-17.9 -15.7	-3.45	-3.45	-3.44±0.07
sulfamerazine	-18.1 -16.1	-21.3 -21.3	-2.97	-3.19	-3.07±0.06
sulfamethazine	-14.9 -13.1	-21.1 -14.4	—	—	-2.97±0.14
sulfisomidine	-20.2 -17.9	-18.3 -14.2	-2.88	-2.93	—

^aModel system contains solvent; values calculated in this work. ^bModel system does not contain solvent; values calculated in this work. ^cRef. 8. ^dRef. 9. ^eRef. 7. All free energy values are given in kcal/mol. In the Method 1 and Method 2 columns, the first of the two rows corresponds to the complex conformation with the common (aniline) end inserted into the β CD cavity, and the second row corresponds to the conformation with the variable end inserted. The conformation with the lower free energy (the apparent preferred conformation) is indicated with bold-face type.

Table II. Prediction of Binding Conformations by Methods 1 and 2

Sulfonamide	Method 1	Method 2	NMR
sulfathiazole	Ring 1 ^a	Ring 1	Ring 1
sulfadiazine	Ring 2	Ring 1	Ring 2
sulfamerazine	Ring 1	Tossup	Ring 2
sulfamethazine	Ring 1	Ring 1	Ring 1

^a"Ring 1" indicates that the conformation with the common (aniline) ring inserted into the β CD cavity is the apparent preferred conformation, while "Ring 2" indicates that the conformation with the variable ring inserted is preferred. Method 1 agrees with the NMR results in three out of the four cases, while Method 2 agrees with the NMR results in two or three out of the four cases.

Table III. Breakdown of Electrostatic and Hydrophobic Association Free Energy Differences (Method 1)

Sulfonamide	ΔG_{elec}	$\gamma(S_{X':Y'} - S_{X'} - S_{Y'})$
sulfathiazole	-1.6	-19.5
	-0.8	-14.1
sulfamethizole	+2.1	-18.9
	+3.0	-21.8
sulfadiazine	-2.2	-17.6
	-2.0	-19.8
sulfamerazine	-1.7	-16.4
	-0.3	-15.8
sulfamethazine	+4.6	-19.5
	+5.7	-18.8
sulfisomidine	-1.8	-18.4
	+1.0	-18.9

Energies are expressed in kcal/mol. As in Table I, the first of the two rows corresponds to the complex with the common end (Ring 1) inserted into the β CD cavity, and the second row corresponds to the complex with the variable end (Ring 2) inserted. $\Delta G_{\text{elec}} = \Delta G_{\text{elec}}(X:Y) - \Delta G_{\text{elec}}(X) - \Delta G_{\text{elec}}(Y)$. $\gamma = 0.050 \text{ kcal mol}^{-1} \text{ \AA}^{-2}$ is the microscopic surface tension, and S_A is the average solvent-accessible surface area of solute species A.

cavity, while entries designated as "Ring 2" indicate that the conformation with the variable ring inserted is the preferred one. Method 1 selects the correct conformation in three out of the four cases, and is thus fairly reliable in determining the more stable of two conformations. In the Method 2 calculation for the sulfamerazine/ β CD complex, nearly identical association free energies were obtained for the two conformations, so a preferred conformation based upon a lower free energy could not be predicted. Thus, Method 2 selects the correct conformation in either two or three out of the four cases. Method 2 calculations are quite fast compared to those of Method 1. Each Method 2 value reported in the table was obtained in only a few CPU hours on an Iris 4D/35 workstation. In Method 1, the solvated sulfonamide systems are typically 14 \AA in radius and contain about 390 water molecules. The solvated CD and sulfonamide/CD systems are typically 15.5 \AA in radius and contain about 460 water molecules. Method 1 calculations typically require 250 CPU hours to obtain a single association free energy value.

The experimental association energies of the six reference complexes all fall within 1.6 kcal/mol of one another, and thus Method 1 calculations are not capable of reproducing these values due to the calculations' 1.5 kcal/mol error range. However, the results of the calculations still provide some interesting insights. A decomposition of Method 1 association free energies into electrostatic and hydrophobic contributions is presented in Table III. If sulfonamide/ β CD complex formation were driven solely by the hydrophobic effect (i.e. the tendency of water to "corral" nonpolar solutes), all of the electrostatic free energies — the essentially enthalpic portion of the free energies — would be greater than zero. This, however, is not the case. The fact that electrostatic free energies are negative in four out of the six cases examined indicates that the binding of sulfonamides (and perhaps other types of

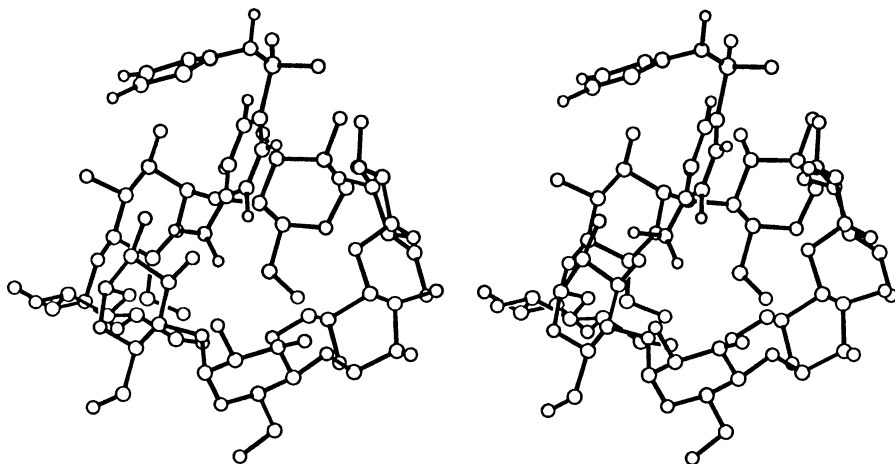


Figure 3: Stereo view of a ball-and-stick representation of the sulfathiazole/ β CD inclusion complex, in its preferred conformation with the aniline ring (Ring 1) inserted within the β CD cavity. Water molecules are not shown, nor are the β CD hydrogen atoms.

guest molecule) to CD should not be thought of as a purely hydrophobic phenomenon. The idea that guest/CD inclusion complexation is not due purely to solute hydrophobicity can also be inferred from a recent study (19) in which both ΔH and ΔS were found to be negative for the formation of the fenoprofen/ β CD complex.

It is interesting to examine the geometries of the sulfonamide/ β CD complexes after they have been allowed to relax via MD simulations of the Method 1 model. Figure 3 displays a ball-and-stick stereo view of the sulfathiazole/ β CD complex in its preferred conformation with the aniline ring (Ring 1) inserted into the β CD cavity. The water present in this system is not shown in the figure, nor are the β CD hydrogen atoms. This sulfathiazole/ β CD structure is typical of the complexes we studied. From the figure we see that "inserted" is a relative term, since the β CD molecule adopts a structure which is more open than structures obtained from gas phase methods such as Method 2. There is, therefore, less of a cavity for the sulfonamide to bind to. The "lock-and-key" effect exhibited by many ligand-receptor complexes is not at work here. Although hydrogen bonds and other attractive electrostatic interactions exist between sulfathiazole and β CD, these interactions are not strong enough to immobilize sulfathiazole within the cavity. The MD simulations show that the sulfonamides are still fairly labile even when bound. This finding is verified by our 1H NMR studies (7), in which it was found that the T_{2s} of the protons of the ring not in the cavity did not change appreciably upon complexation to β CD.

Literature Cited

1. Tanford, C.; *The Hydrophobic Effect: Formation of Micelles and Biological Membranes*; John Wiley & Sons: New York, NY, 1973.
2. Russell, S. T.; Warshel, A. *J. Mol. Biol.* **1985**, *185*, 389.
3. Bash, P. A.; Singh, U. C.; Langridge, R.; Kollman, P. A. *Science* **1987**, *236*, 564.
4. Rao, S. N.; Singh, U. C.; Bash, P. A.; Kollman, P. A. *Nature* **1987**, *328*, 551.

5. Gerber, P. R.; Mark, A. E.; van Gunsteren, W. F. *J. Comp.-Aided Mol. Design* **1993**, *7*, 305.
6. Nicholls, A.; Sharp, K. A.; Honig, B. *Proteins*, **1991**, *11*, 281.
7. King, G.; Pfeffer, P. E.; Irwin, P. L.; Brewster, J. D.; Barford, R. A. *J. Carb. Chem.* (submitted for publication).
8. Cohen, J.; Lach, J. L. *J. Pharm. Sci.* **1963**, *52*, 132.
9. Uekama, K.; Hirayama, F.; Nasu, S.; Matsuo, N.; Irie, T. *Chem. Pharm. Bull.* **1978**, *26*, 3477.
10. Gasteiger, J.; Marsili, M. *Tetrahedron* **1980**, *36*, 3219.
11. Dewar, M. J. S.; Thiel, W. *J. Am. Chem. Soc.* **1977**, *99*, 4899.
12. Clark, M.; Cramer III, R. D.; Van Opdenbosch, N. *J. Comp. Chem.* **1989**, *10*, 982.
13. Gelin, B. R.; Karplus, M. *Biochemistry* **1979**, *18*, 1256.
14. Sharp, K. A.; Honig, B. H. *Ann. Rev. Biophys. Biophys. Chem.* **1990**, *19*, 301.
15. Mezei, M.; Swaminathan, S.; Beveridge, D.L. *J. Am. Chem. Soc.* **1978**, *100*, 3255.
16. Straatsma, T. P.; Berendsen, H. J. C.; Postma, J. P. M. *J. Chem. Phys.* **1986**, *85*, 6720.
17. Singh, U. C.; Brown, F. K.; Bash, P. A.; Kollman, P. A. *J. Am. Chem. Soc.* **1987**, *109*, 1607.
18. King, G.; Warshel, A. *J. Chem. Phys.* **1989**, *91*, 3647.
19. Uccello-Barretta, G.; Chiavacci, C.; Bertucci, C.; Salvadori, P. *Carbohydrate Research* **1993**, *243*, 1.

RECEIVED March 3, 1994

Chapter 12

Structure–Function Analysis of Amino Acid Substitutions in Proteins

Nilofer G. Jiwani¹ and Michael N. Liebman^{1,2}

¹Department of Molecular and Cellular Biochemistry, Loyola University of Chicago, Maywood, IL 60153

²Bioinformatics Program, Amoco Technology Company, Mail Code F-2, 150 West Warrenville Road, Naperville, IL 60563–8460

In evolution, the three-dimensional structure of proteins appears to be preserved longer than the amino acid sequence. Consequently, many evolutionarily related (homologous) proteins exhibit significant variation in amino acid sequence, but still adopt similar structures. A general hypothesis suggests that proteins evolve to maintain physico-chemical properties of functionally and/or structurally important parts of the protein molecule. Thus properties that have been conserved during evolution should be related to structural and/or functional conservation. Our attempts to correlate physical properties in structurally similar regions revealed: (1) Amino acids (especially the charged and/or polar residues) cannot adequately be described by a single value property; and (2) Amino acids in a particular conformational state exhibit a set of properties which define its equivalence with other amino acids. This could explain how different sequences can generate the similar conformations. Our preliminary data thus provide insight into the relationship between conformation and environment in protein.

Understanding the biological function of macromolecules is of fundamental interest to researchers in biochemistry, biophysics, protein engineering and biomedicine. It is an accepted view that the three-dimensional structure of a protein is related to its function, and thus is considered as the focus for analyzing function. Various experimental techniques such as X-ray crystallography (1), NMR (2, 3) have been used to determine protein structures. The three-dimensional structural database for proteins has grown

0097–6156/94/0576–0185\$08.72/0
© 1994 American Chemical Society

substantially since the first determination of the X-ray structure of a protein (4) with more than 1110 proteins maintained in the Protein Data Bank (1). Although this growth in information is substantial, the understanding of the relationship between structure and function, *in vitro* or *in vivo*, remains largely unresolved. Thus the extraction of knowledge about the structure-function relationship from structural information remains the primary objective of present day structural analysis.

One of the specific aims in studying protein structure is to determine and understand how the three-dimensional structure is encoded in the one-dimensional sequence of its amino acids. The accepted hypothesis (5) is that adequate information exists within the amino acid sequence to yield the correct functional three-dimensional structure. However, recent discovery of chaperone proteins which are found to be essential for correct folding of polypeptide chains may challenge the completeness of this assumption (6-9). Based on comparative studies of homologous proteins, tertiary structure appears to be conserved longer than amino acid sequence (i.e. to say sequence is degenerate with respect to structure) in evolution. This may reflect our ability to quantify structural comparison and inability to adequately compare sequence. Consequently, many evolutionarily related (homologous) proteins exhibit significant variation in amino acid sequence, but still adopt similar structures (10). A general hypothesis suggests that proteins evolve to maintain physicochemical properties of functionally or structurally important parts of the protein molecule. Therefore, properties that have been conserved during evolution i.e. common to all members of a family of homologous sequences, should be related to structural or functional similarity.

The availability of protein sequence information has increased dramatically in the past few years. With the large investment of resources in the Human Genome Project, the sequence information is likely to grow even faster. To maximize the knowledge that can be derived from this huge database, it is essential to understand the relationship between sequence and structure. Many studies have examined this relationship by studying various physico-chemical properties (11, 12). Both computational and experimental approaches have been used to analyze the correlation of the amino acid sequence or sequence-based properties with the secondary and tertiary protein structure. Hydrophobicity, an example of a sequence-based property, has been frequently utilized for structure analysis because of its potential role in the interface of protein molecules with solvent (13). It is widely assumed (14-17) that hydrophobicity is a major factor in maintaining the specific conformation of native proteins. Numerous attempts have been made to predict tertiary structures of proteins on the basis of hydrophobicity of the primary

structure (17-19). The hydrophobicity of an amino acid residue is not a property that can be defined easily or measured. Several groups have attempted to derive numerical hydrophobicity scales using a variety of experimental and computational methods. The resulting values typically correspond to the free energy of transfer of the side chain of the amino acid from water to a non-polar environment. Each of the 20 commonly occurring amino acids has thus been assigned a single hydrophobicity value within a scale, although several such scales exist (20). Given the sequence of a protein of unknown structure, the hydrophobicity profile, a graph of hydrophobicity successively averaged over a window of neighboring residues, has been used to predict turns in peptide chain (21), interior/surface regions (22, 23), antigenic sites (24-26), and membrane-spanning segments (23).

Amino acids are categorized based on their individual hydrophobicity values (Table I). The observation of similarity in hydrophobicity is generally assumed to reflect structure similarity, and replacement with amino acid from the same category is considered not to introduce structural perturbation. For example, Leucine and Isoleucine being hydrophobic are considered to be able to be incorporated within similar secondary structure. If this assumption of hydrophobicity equivalence can be used to predict structure equivalence, then we should observe that the reverse should also be true. We have evaluated this hypothesis by comparing the hydrophobic nature of homologous proteins whose three-dimensional structures are known, using serine proteases as a prototype system.

The results of our analyses reported here suggest that the equivalence of hydrophobicity is not generalizable. This is consistent with the evidence of its limited predictive capability (60% - 70%) (27-30). The lack of correspondence raises the question of ability of a single-valued property to describe each amino acid. An amino acid is observed in multiple conformations and each can potentially exhibit a set of properties unique to that conformation. We have examined this hypothesis by analyzing the validity of assignment of single-valued property. In this paper, we present some evidence which implies that (1) An amino acid (especially the charged and/or polar residues) cannot adequately be described by a single value property; and (2) An amino acid in a particular conformational state exhibits a set of properties which impact its interchangeability with other amino acids.

Experimental Procedures

Data. The three-dimensional structures of the proteins used in this study are available as atomic coordinates provided by the Brookhaven Protein Data Bank (PDB) (1). The group of proteins that were selected for the present

Table I : Hydrophobicity values of each of the 20 amino acids (25)

HYDROPHOBIC	HYDROPHILIC	NEUTRAL
THR -0.4	ARG 3.0	GLY 0.0
ALA -0.5	ASP 3.0	PRO 0.0
HIS -0.5	GLU 3.0	
CYS -1.0	LYS 3.0	
MET -1.3	SER 0.3	
VAL -1.5	ASN 0.2	
ILE -1.8	GLN 0.2	
LEU -1.8		
TYR -2.3		
PHE -2.5		
TRP -3.4		

study are listed in Table II. We have renumbered the sequences using sequential numbering beginning with residue 1. This eliminates the common bias imposed by a sequence relationships drawn on a comparison with chymotrypsinogen.

Computational Resources. The algorithms described here are written as FORTRAN programs. All computations have been performed on a Microvax II.

Structure Alignment. This procedure identifies regions of structural similarity between two proteins regardless of the presence of insertions or deletions within their respective amino acid sequences (31, 32). Two approaches were used in this analysis : (a) *Superposition of whole proteins* : This analysis was used for a set of serine proteases (marked with an asterisk in Table II) where one of the serine proteases (eg. bovine trypsin, 1TPO) was used as a reference molecule to which other serine proteases were superimposed to determine structurally similar regions. The method has been described previously (31). Briefly, in this method an algorithm is used which first obtains an initial set of secondary structure equivalences by screening protein 1 against protein 2 using the linear distance plots of both molecules (11). These regions are then compared in tertiary structure by first superimposing the centers-of-gravity of the respective coordinate sets and then applying a minimization procedure to obtain the rotation-translation matrix that achieves the best superposition of the equivalenced atomic coordinates. This is monitored by evaluating the root-mean-square-difference (rms) of equivalenced regions. The rotation-translation matrix that results from this superposition of the initial set of equivalenced amino acid residues is applied to the entire protein structure and residues in tertiary structure alignment are identified. The resultant augmented list of structural equivalences are then used to refine the superposition procedure. This procedure is iteratively applied until convergence is reached to a unique set of topographical equivalences. Convergence is established by a statistical test which compares the distribution of the equivalenced residue coordinates with that of a statistically random three-dimensional sample (57). The local rms values were calculated for the obtained set of topographical equivalences; (b) *Superposition of individual regions* : For this analysis, the region rms was computed by superimposing, individually, each of the structurally equivalenced regions [obtained by approach (a)] of a pair of proteins.

Linear Distance Plot (LDP). The linear distance (LD) representation has been described in detail previously by Liebman (11). Briefly, the linear distance value of each amino acid was computed by summing the series of distances from the N-terminal alpha carbon to each of four successive

Table II : List of proteins used in the study**

PDB	PROTEIN	INHIBITOR	SOURCE	# OF RESIDUES	MOLECULE(S)	RESOLUTION (Å)
1TPO	Beta-Trypsin*		Bovine pancreas	223	1	1.7
1EST	Tosyl-Elastase*		Porcine pancreas	240	1	2.5
3RP2	Rat Mast Cell* Protease		Rat small intestine	224	2	1.9
2KA1	Kallikrein A		Porcine pancreas	232	1	2.5
2ALP	Alpha-lytic protease	Bovine pancreatic trypsin inhibitor	Bovine pancreas	58	1	
5CPA	Carboxypeptidase A		Lysobacter	198	1	1.7
1SBC	Subtilisin Carlsberg		Enzymogenes			
1RHD	Rhodanese		Bovine pancreas	307	1	1.54
5RSA	Ribonuclease A		Bacillus subtilis	275	1	2.5
2LZM	Lysozyme		Bovine liver	293	1	2.5
2PRK	Proteinase K		Bovine pancreas	124	1	2.0
3FAB	Lambda Immuno- globulin Fab (Prime)- Heavy chain		Escherichia coli	164	1	1.7
2SGA	Light chain		Fungus(Tritrachium Album Limber)	279	1	1.5
1ITEC	Proteinase A Thermolysin		Human			2.0
2CAB	Carbonic Anhydrase form B	Eglin-C	Streptomyces griseus	208	1	1.5
3TLN	Thermolysin		Thermoactinomyces vulgarius	220	1	2.2
2CGA	Chymotrypsinogen A*		Human erythrocytes	70	1	2.0
3EST	Native Elastase*		Bacillus thermoprot- colyticus	261	1	1.6
2PKA	Kallikrein A*		Bovine pancreas	316	1	1.8
4ADH	Alcohol dehydrogenase		Porcine pancreas	245	2	1.65
3SGB	Proteinase B		Porcine pancreas	240	1	2.05
3ADK	Adenylate kinase	Ovomucoid inhibitor third domain	Porcine liver	232	2+	2.4
1LZT	Lysozyme, tritclinc		Streptomyces griseus	374	1	1.8
			Turkey	185	1	
			Porcine muscle	56	1	
			Hen egg white	195	1	2.1
				129	1	1.97

** Source : Protein Data Bank (1)

+ only one molecule was used in the study

alpha carbons. Each successive amino acid of the protein sequence serves as an origin for the computation. For a protein N amino acids long, there are (N-4) LD values. The sum of the distances (S) between the alpha carbon origin of this neighborhood and each of its subsequent neighbors yields a characteristic value that reflects the local conformation of the polypeptide chain within that segment. A spectrum-like decomposition of the overall conformation is obtained by plotting S for each neighborhood versus the residue number of the origin atom of the segment. This plot displays a detailed profile of the local folding of a protein molecule which can be used to identify elements of secondary structure, and enables the structural comparison at secondary structure level of two proteins using one-dimensional analysis of the three-dimensional data (Figure 1).

Difference Linear Distance Analysis. This analysis determines similarity at secondary structure level. The analysis involves computation of mean of the difference between LD values of each of the structurally equivalent regions (obtained using Structure Alignment algorithm) of a pair of proteins.

Hydrophobicity Analysis. This analysis was performed on a selected set of serine proteases (marked with an asterisk in Table I) using three different hydrophobicity scales (33). This analysis sequence-based property was carried out using two approaches : (a) *Sequence-based analysis* : For this analysis, the average hydrophobicity was calculated using a scheme described by Kyte and Doolittle (23) for prediction of secondary structures. A hydrophobicity index, h, was attributed to each of the amino acid residues and then the hydrophobicity as a moving average over seven neighboring residues (for a given sequence) was plotted in a manner analogous to the LD; (b) *Structure-based analysis* : For this analysis, the mean hydrophobicity difference was calculated by computing, individually, the mean of the difference between the hydrophobicity values of amino acid residues of each of the structurally aligned regions of a pair of proteins.

Dipole moment Computation. Dipole moment and hydrophobicity, both being a measure of polarity, it seems a valid approach to carry out a comparative analysis between the two. Thus dipole moment computation was undertaken to evaluate the performance of hydrophobicity analysis. The dipole moment computation was performed on a group of proteins of known three-dimensional structures (Table II). The analysis involves the deconvolution of a protein structure into its constituent peptide and side-chain dipoles, using vector addition of the individual bond moments for each of the amino acids assigned from the bond moment values derived from spectroscopic measurements

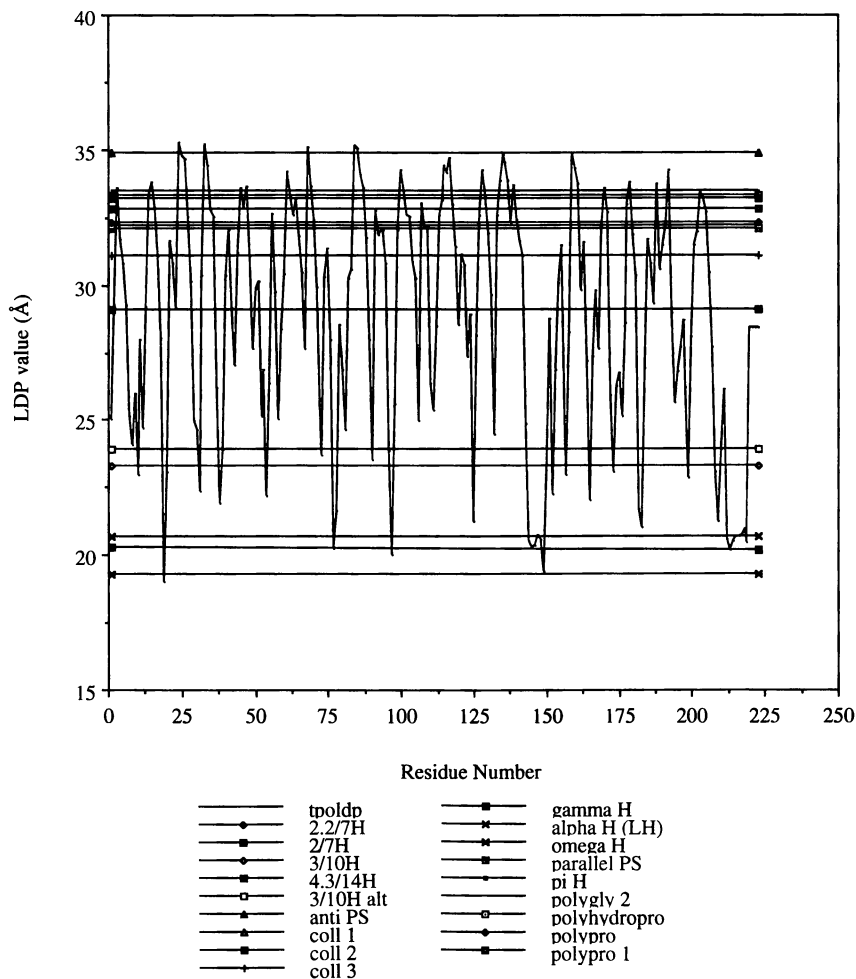


Figure 1. Linear distance plot of trypsin showing standard conformations.

(34). Computing the dipole moment for the trans-peptide bond in this manner yields an average value of 3.81 debyes, while the dipole moment for the individual side-chains varies with the observed conformational flexibility of that particular side-chain, for example, glutamine ranges from 3.03 to 4.11 debyes in native trypsin (1TPO). The conformation produced by the presence of a proline residue causes the peptide bond dipole moment to exhibit a computed dipole moment of 2.71 debyes, while cis-peptides exhibit peptide dipoles between 2.2 - 2.4 debyes. In this manner, it is possible to compute the composite dipole moment for the peptide bonds and separately for the side-chain atoms as well as for the overall protein. We have analyzed these dipole moments computed from the observed structures to introduce the details of the actual conformation into the analysis.

van der Waals volume and Accessible surface area Computations. The steric accessibility of key functional groups in a molecule will determine, in part, the availability of that part of the molecule to attack by other molecules (35, 36). In order to evaluate the total surface area surrounding a particular atom and quantitate its accessibility, a computer program based on a Monte Carlo simulation of space filling within a box of enclosure (37), was used. The box is randomly filled with points with a fixed density of 50 points per cubic angstrom, as previously optimized. The ratio of the number of points within the van der Waals radius of an atom to the total number of points in the box can be used to compute the van der Waals volume (vdw). Accessible surface area is computed in an analogous manner. Random points are generated at uniform density over a sphere around each atom constructed with the van der Waals radius. Points within the van der Waals radius of any other atom are considered buried, otherwise they are exposed. The ratio of the number of exposed points to total points is then computed to approximate the percent of exposed surface. Amino acids with frequently acquired dipole moment values were used for these computations. vdw volume and accessible surface areas were calculated for all 20 naturally occurring amino acids. The list of proteins used in this analysis are shown in Table II.

Results

This section summarizes the results of property analyses of a selected group of proteins with known three-dimensional structures. Each of the property analyses will be presented individually. In the present study, the emphasis is on examination of validity of the use of a single property value to describe an amino acid and subsequently define the structure.

Comparison of Hydrophobicity and Tertiary structure Similarities. Table III displays the results of applying the structure alignment procedure to and computing the mean hydrophobicity difference of a selected set of serine proteases using trypsin and elastase as an examples. The table shows the regions in trypsin structurally equivalent to elastase and their respective mean hydrophobicity difference values. These results indicate the high degree of structural similarity between the regions of trypsin and elastase as revealed by the root-mean-square (rms) values. For example, first seven residues (#1 - 7) in trypsin are structurally equivalent to first seven residues (#1 - 7) in elastase with a local and region rms deviation of 0.63 Å and 0.26 Å respectively. The hydrophobic nature of these structurally equivalent regions was analyzed by computing the mean hydrophobicity difference of the equivalent residues (refer Methods section). It can be observed that the hydrophobic nature of several structurally equivalent regions is markedly different. For example, the mean hydrophobicity difference between the region I (residues 1 - 7) in trypsin and the region I (residues 1 - 7) in elastase is calculated to be 8.71 (as shown below) which suggests that the region of trypsin is more hydrophilic than that of elastase.

1TPO	I	V	G	G	Y	T	C
3EST	V	V	G	G	T	E	A
hydrophobicity difference	3	0	0	0	-19	-34	-5
mean hydrophobicity difference					-8.71		

Comparison of Hydrophobicity and Secondary structure Similarities. Table I lists the common categorization of amino acids based on their individually assigned hydrophobicity values (25). Figure 2 displays the results of structure alignment and LDP analyses of the selected set of serine proteases using trypsin and elastase as an examples. The figure displays the result of the comparison of secondary and tertiary structures of bovine trypsin (1TPO) and porcine elastase (3EST). It was observed that the regions of trypsin and elastase equivalent at tertiary level display a significant equivalence at secondary structure levels as indicated by similarity between their respective LDP profiles. This is also evidenced by mean ldp difference values shown in Table III. Figure 3 shows the hydrophobicity profiles of these same regions of trypsin and elastase equivalenced at tertiary level. Comparison of Figure 2 and Figure 3 revealed that regions with similar secondary and tertiary structures can exhibit significantly different hydrophobicity profiles. For example, the equivalenced regions (residue #s 45 - 76) in trypsin and (residue #s 52 - 83) in elastase are structurally similar. However, the hydrophobicity profiles of these regions differ significantly (compare respective regions in Figures 2 and

Table III : Comparison of rms, ldp difference and hydrophobicity difference values of structurally equivalent regions of bovine trypsin and porcine elastase

Structural equivalences of		Local rms (Angstroms)	Region rms (Angstroms)	Mean Ldp difference(abs) (Angstroms)	Mean Hydrophobicity difference (kcal/s)
Trypsin	Elastase				
1 - 7 (7)	1 - 7 (7)	0.628	0.258	0.134	- 8.710
11 - 20 (10)	11 - 20 (10)	0.769	0.381	0.167	1.500
21 - 42 (22)	26 - 47 (22)	0.567	0.380	0.381	- 1.270
45 - 76 (32)	52 - 83 (32)	0.689	0.569	0.610	3.880
82 - 107 (26)	91 - 116 (26)	0.688	0.636	0.542	4.350
114 - 125 (12)	125 - 136 (12)	0.772	0.479	0.299	2.250
132 - 145 (14)	142 - 155 (14)	0.870	0.643	0.540	7.140
155 - 157 (3)	167 - 169 (3)	1.065	0.098	0.578	- 10.330
159 - 164 (6)	171 - 176 (6)	0.957	0.628	0.023	- 1.830
168 - 183 (16)	179 - 194 (16)	0.634	0.514	0.305	2.380
185 - 195 (11)	200 - 210 (11)	0.805	0.625	0.356	4.910
197 - 223 (27)	214 - 240 (27)	0.687	0.581	0.415	2.150

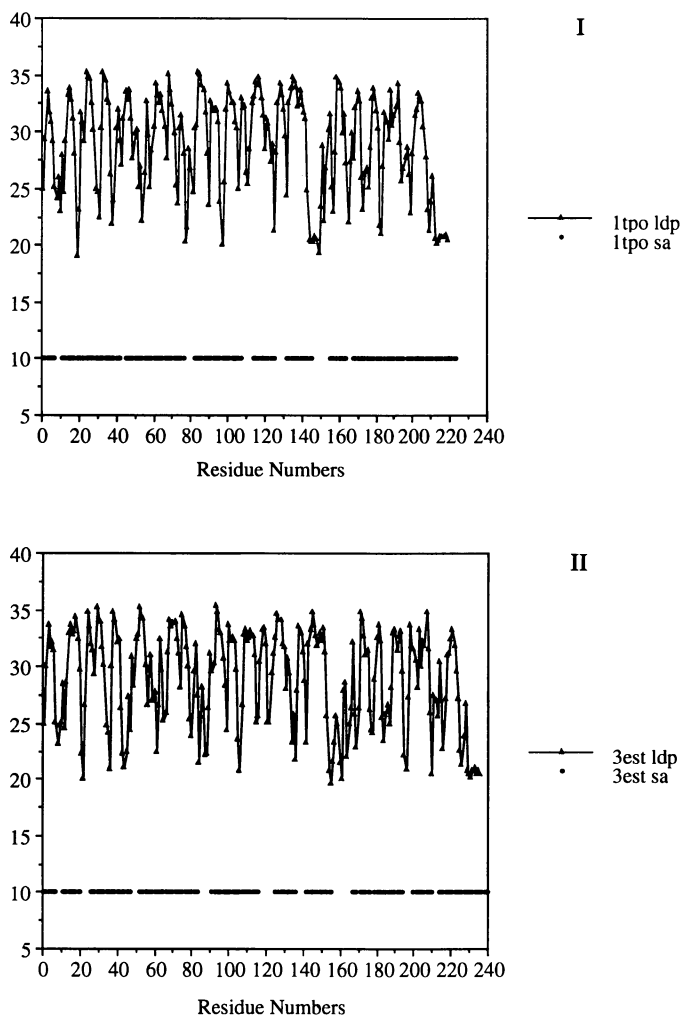


Figure 2. Comparison of linear distance plots and structure alignments of trypsin and elastase: (I) This displays (a) LD plot of trypsin. It is generated using the sum of series of distances from the origin alpha carbon to each of four successive alpha carbons (ref. 11). (b) The horizontal bars represent regions of trypsin which structurally aligned with those of elastase. (II) This displays (a) LD plot of elastase generated as described above. (b) The horizontal bars represent regions of trypsin which structurally aligned with those of trypsin.

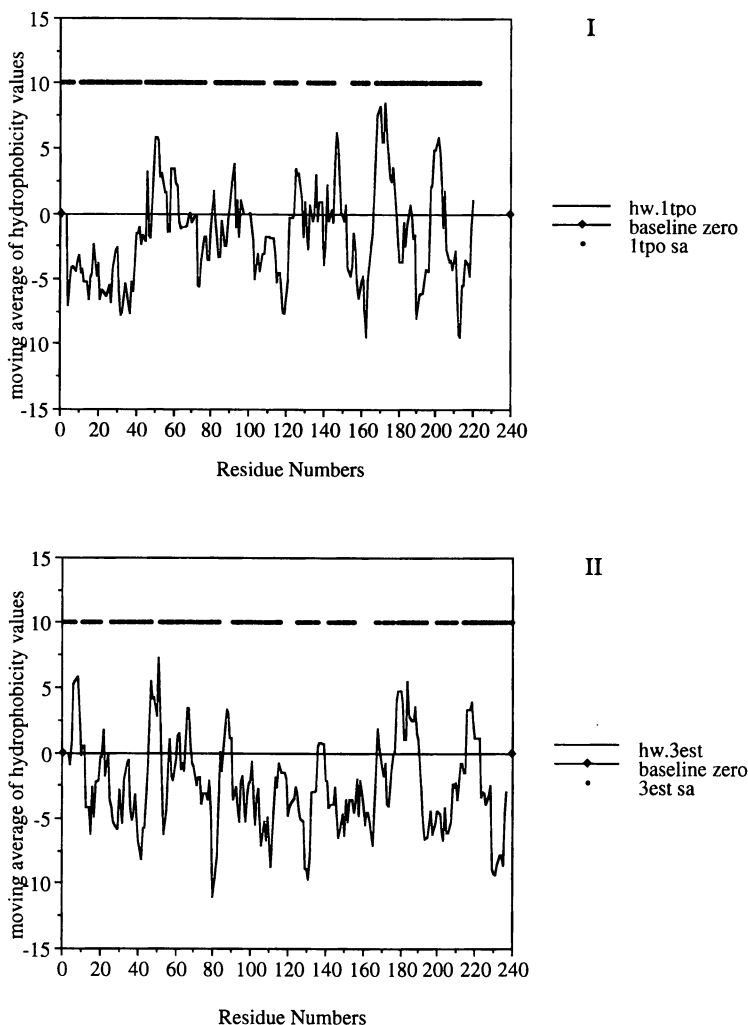


Figure 3. Comparison of hydrophobicity profiles and structure alignments of trypsin and elastase: (I) This displays (a) Hydrophobicity profile of trypsin–hydrophobicity values plotted on y-axis represents an average over seven neighboring residues with values plotted at the center residue of the window. The horizontal line corresponds to the value zero. (b) The horizontal bars represent regions of trypsin which structurally aligned with those of elastase (as shown in Figure 2). (II) This displays (a) Hydrophobicity profile of elastase–hydrophobicity values plotted on y-axis represents an average over seven neighboring residues with values plotted at the center residue of the window. The horizontal line corresponds to zero-value line. (b) The horizontal bars represent regions of trypsin which structurally aligned with those of trypsin (as shown in Figure 2).

3). Similar results were observed using the other two hydrophobicity scales (data not shown).

Structure-derived Property Analyses.

Dipole moment Analysis. Figure 4 shows plot of distribution of computed dipole magnitudes for amino acid, Serine, as observed in 29 proteins. This represents a total of 620 serine residues. The figure displays two prominent peaks (2.4 and 2.8 debyes) with each peak representing the magnitude of dipole moment frequently acquired by serine (which is used as an example) in tertiary conformations of a group of proteins. A frequency distribution of magnitudes of dipole moments for each of the 20 amino acids is shown in Figure 5. It can be seen that each amino acid (except certain hydrophobic and neutral amino acids) displayed a range of dipole moment values with several maxima observed for each. Hydrophobic amino acids such as Isoleucine, Leucine, Valine, Alanine and Phenylalanine were observed to acquire a very similar magnitude (0.6 - 0.7 debye); Glycine, a neutral amino acid, displayed the value of 0.6 debye.

van der Waals volume (vdw) and Accessible surface area Analyses. Figures 6 and 7 display respectively the values of vdw volume and accessible surface area of the amino acid conformations with frequently acquired values of dipole moment magnitude. It was observed that most amino acids also exhibit a range of vdw volume and accessible surface area values. However, compared to dipole moment magnitude, less variability is observed for a given amino acid for these properties. This indicates the limited utility of vdw volume and accessible surface area, alone, to define structure similarity.

Discussion

Reliability of Single value Property. How a specific protein conformation is determined by the amino acid sequence remains a constant source of fascination and speculation. Although the amino acid sequence of a protein is believed to contain the information necessary to determine its three-dimensional structure, the ability to predict the fold of a protein from its sequence alone is still unresolved. Similarities between the sequence and sequence-based properties of a protein of unknown structure and those of a homologous protein of known three-dimensional structure remains most common way to derive structural information. Similarity in hydrophobicity, one of the commonly used sequence-based properties, is generally assumed to indicate structure similarity. It is widely assumed (2-5) that hydrophobicity is the major factor in maintaining the specific structure

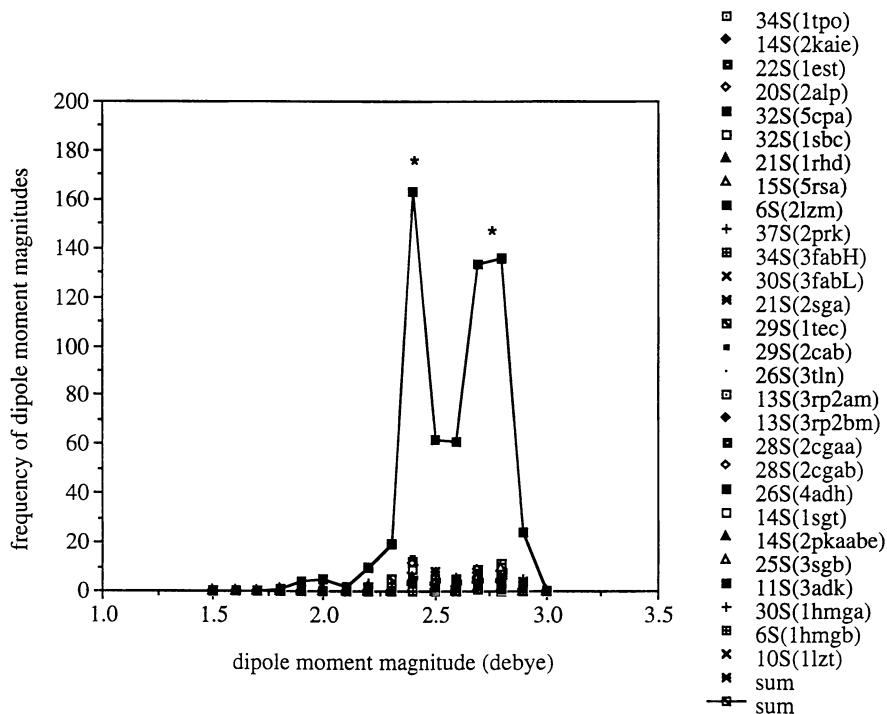


Figure 4. Dipole moment frequency plot of Serine: The plot displays the magnitude (in debye units) and the frequency of magnitudes of dipole moment on x-axis and y-axis respectively. Each peak represents the conformation of frequent occurrence of Serine residue. The dipole moment magnitudes of two prominent peaks are 2.4 and 2.75 debye. The total number of Serine residues observed in each of the proteins used for the computation are shown.

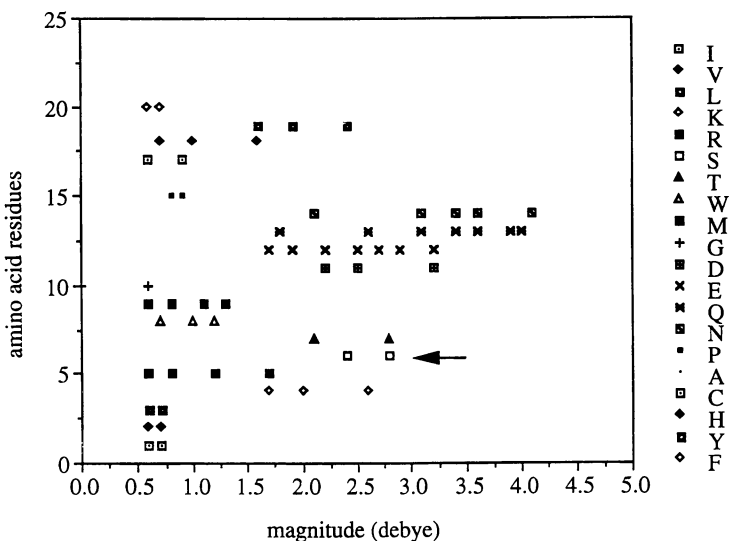


Figure 5. Dipole moment frequencies of 20 amino acid residues: The plot displays the magnitude of dipole moment (in debye units) on x-axis. Each symbol represents an amino acid. The multiple occurrence of each symbol corresponds to the different dipole moment values which each amino acid can acquire frequently. The arrow shows the values of prominent peaks of Serine shown in Figure 4.

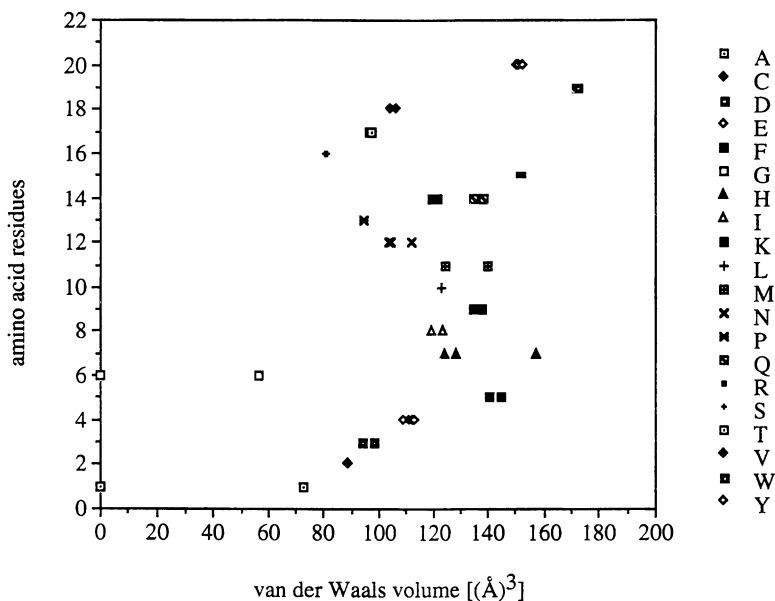


Figure 6. van der Waals volume frequencies of 20 amino acid residues: The plot displays the values of van der Waals volume [in $(\text{\AA})^3$ units] on x-axis. Each symbol represents an amino acid. The multiple occurrence of each symbol corresponds to the volumes of conformations of frequent occurrence of an amino acid shown in Figure 5.

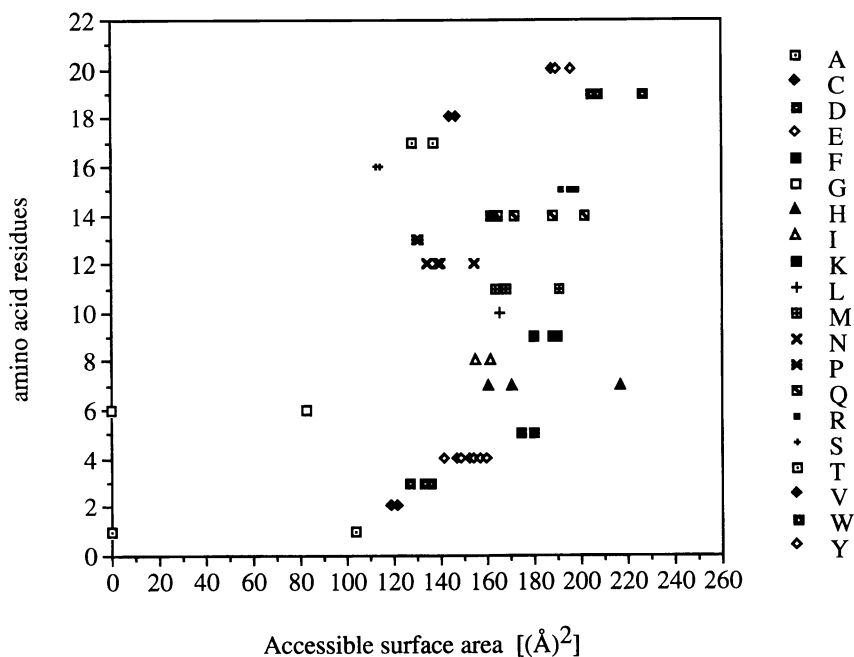


Figure 7. Accessible surface area frequencies of 20 amino acid residues: The plot displays the values of accessible surface area [in (Å)² units] on x-axis. Each symbol represents an amino acid. The multiple occurrence of each symbol corresponds to the accessible surface areas of conformations of frequent occurrence of an amino acid shown in Figure 5.

of native proteins. This has led to numerous attempts to predict tertiary structures of proteins using hydrophobicity as a measure of homology (5-7). Amino acids are categorized based on their individual hydrophobicity values (Table I). The assumption used in structure prediction studies is that the hydrophobicity equivalence reflects structure equivalence as an interchange of amino acids within the same category is considered not to perturb its secondary structure. For example, Serine and Threonine being polar are considered to adopt similar secondary structure. Based on this assumption, all the amino acids within a category are considered structurally equivalent in local environment of proteins. If hydrophobicity equivalence reflects structure equivalence, then our assumption was that the reverse should also be equally true. This was evaluated by analyzing hydrophobic nature of proteins whose three-dimensional structures are known. Examination of the hydrophobic nature of identified structurally similar regions of structurally aligned set of serine proteases displayed that these equivalenced regions do not always exhibit similar hydrophobicity. For example, the region (residues 1 - 7) in trypsin is structurally equivalent to the region (residues 1 - 7) in elastase as evidenced by local and region root-mean-square-difference (rms) deviation of 0.63 Å and 0.26 Å respectively (Table III). However, the mean hydrophobicity difference between these regions is calculated to be -8.71 which suggests that the region of trypsin is more hydrophilic than that of elastase. This implies that hydrophobicity by itself may not be a predictive measure of structure. Also, differences between hydrophobicity profiles obtained for a set of serine proteases and our results of structure alignment study and linear distance plot (LDP) analysis for the same set of proteins was observed (compare Figure 2 with Figure 3). The comparison of Figure 2 with Figure 3 revealed that the regions with similar secondary and tertiary structures exhibited a significantly different hydrophobicity profiles. For example, the secondary and tertiary structures of regions (residues 45 - 76) in trypsin and (residues 52 - 83) in elastase are structurally similar. However, the hydrophobicity profiles of these regions are significantly different. This suggests that structurally equivalent regions do not necessarily have equivalent hydrophobic nature. Thus the assumption that hydrophobicity plots can be utilized for structure prediction may not be generally true. The reason for this is that although hydrophobicity equivalence reflects structure equivalence, hydrophobicity difference does not necessarily mean structural difference.

This lack of correspondence between structure similarity and hydrophobicity similarity suggests the limitation in assigning single property values to an amino acid residue. This is because an amino acid in a given conformation can potentially exhibit properties unique to

the conformation. The conformation of an amino acid within the three-dimensional structure of a protein is a result of interaction within its local environment in the protein. The hydrophobicity of a molecule is a measure of its polarity. This reflects distribution of charge; dipole moment is a measure of charge distribution. It would be expected that some correspondence (not equivalence) should exist between these two parameters. Also, the computed dipole moment will reflect the respective conformation of the amino acid in the local environment of protein. Thus a measure of property of an amino acid in its respective conformation, appears to be a better approach. Therefore we selected dipole moment, a conformation-based, computable property, to examine the validity of assigning of single property value to each amino acid residues in proteins.

The frequency distribution of magnitudes of dipole moments for each of the 20 amino acid showed that each amino acid (except certain hydrophobic and neutral amino acids) exhibits a range of dipole moment values. This indicates that while an amino acid in tertiary structures of proteins occur in multiple but finite number of possible conformations, each of its conformations is associated with a unique set of physico-chemical properties. Some of the hydrophobic amino acids such as Isoleucine, Leucine, Valine, Alanine and Phenylalanine were observed to acquire a very similar magnitudes (0.6 - 0.7 debye); Glycine, a neutral amino acid, displayed the value of 0.6 debye. The observed absence of variability in dipole moment values is expected due to the nature of the side chains of these amino acids. However, it was observed that charged and/or polar residues exhibited a range of dipole moment values with several maxima observed for each. For example, Serine frequently acquired dipole moment values of 2.4 and 2.8 debye; Histidine displayed dipole moment magnitudes of 0.7, 1.0 and 1.6 debye. These observations thus suggest that general trend of describing an amino acid residue by a single property value may not always be appropriate (or applicable).

This has been further supported by computations of van der Waals volume and accessible surface area of these multiple conformations. It was observed that each conformation is associated with variable properties (Figures 5, 6, 7). For example, different conformations of Asparagine have different dipole moment magnitudes (2.1, 3.1, 3.4, 3.6 and 4.1 debyes) as well as different van der Waals volume (111.7, 104.5, 104.1 and 103.7 Å). Also, a significant overlap in a property was observed among amino acid residues (Figure 5). For example, Tyrosine and Glutamic acid exhibits the same dipole moment value (1.9 debyes) (Figure 5). Thus potentially, tyrosine could be replaced by glutamic acid in a sequence if the significant conserved property of that position in the sequence is dipole moment magnitude. This implies that an amino acid in a particular conformational state exhibits a particular

set of properties which may mimic a different amino acid. This could explain how different sequences can generate similar conformations. It was also observed that identical amino acid at identical position in the selected set of proteins, display different dipole moment magnitude. This complies with the evidence that identical residues can adopt different conformations in different proteins (38, 39).

Another interesting observation was that serine residue of catalytic triad essential for enzymatic activity of serine proteases acquires similar magnitude (2.7 - 2.8 debyes). The conservation of this property implies the potential role of dipole moment magnitude in biochemical function of these enzymes.

Thus the property analyses of amino acid residues in protein implies that single property value may be inadequate to describe an amino acid. The generalization of structural similarity based on a single property value (which is the basis of hydrophobicity-based structure prediction studies) is inadequate. This provides an explanation for non-correspondence between structural and hydrophobic equivalences as all conformations of an amino acid are represented by a single hydrophobicity value (Figure 8). Also, it should be noted that the hydrophobicity value represents the partitioning tendency of an amino acid between polar and nonpolar phases. Thus it is a single value of a free amino acid which does not correspond to its value in local environment of protein. On the other hand, dipole moment of each amino acid residue in proteins is computed from its structural coordinates in tertiary conformation. Thus it represents a property acquired by an amino acid in local environment of protein. And thus unlike hydrophobicity value which represents property due to sequence, it is property both due to structure and sequence. Therefore the analysis of dipole moment instead of hydrophobicity will provide information of a molecule which reflects its local environment.

The dipole moment analysis presented here evidences three novel observations : (i) the range of dipole moment values indicating the potentiality of an amino acid to occur in multiple conformations; (ii) describing an amino acid residue in terms of single property value may not always be appropriate (or applicable); and (iii) significant overlap of dipole moment magnitude values among amino acid residues suggesting its ability to predict interchange of amino acids which are "silent" or "non-equivalent". This observation will be of utmost importance and can potentially aid to understand the observed dysfunctional natural mutants. The evaluation of this hypothesis using natural mutation database of Factor IX as protocol system will be discussed elsewhere (Jiwani, N.G. et al; manuscript under preparation).

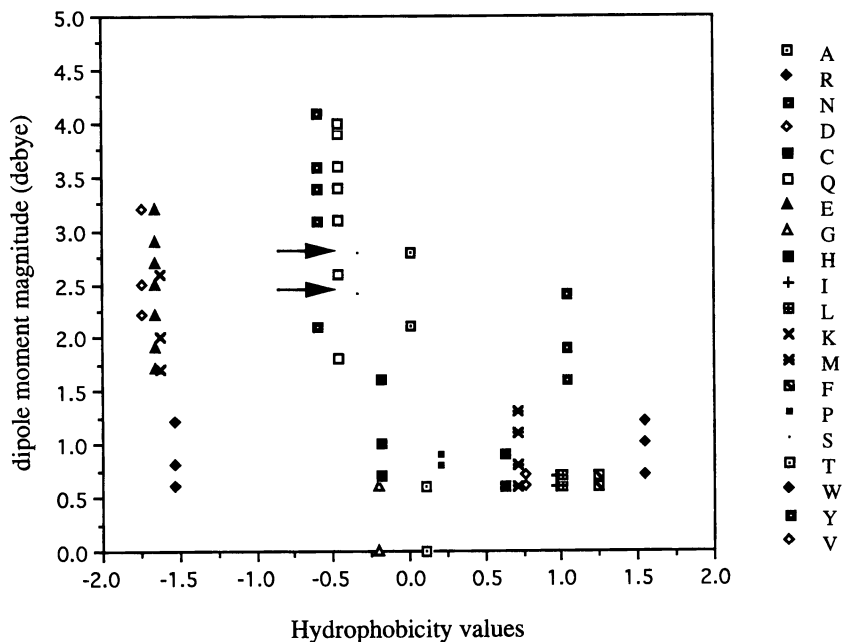


Figure 8. Hydrophobicity versus dipole moment: The plot displays the hydrophobicity values and the magnitude of dipole moment (in debye units) of each of the 20 amino acids on x-axis and y-axis respectively. The arrows display the values of prominent peaks of Serine shown in Figure 4.

Conclusion

In conclusion, (a) an examination of hydrophobic nature of structurally aligned proteins reveal that the regions of structural similarity does not always exhibit hydrophobicity equivalences, Thus hydrophobicity value cannot reliably be used for predicting structure similarity; (b) frequency distribution of magnitudes of dipole moments for each of the 20 amino acids in a group of proteins indicates that : (i) while an amino acid can occur in more than one conformation, each of its conformations is associated with a set of unique physico-chemical properties including van der Waals volume and accessible surface area. Thus an amino acid in a particular conformational state exhibits a particular set of properties which may mimic a different amino acid. This could explain how different sequences can generate the similar conformations, although, further analysis is needed and is underway; (ii) single property value may not always provide adequate information to describe an amino acid residue; (iii) significant overlap of dipole moment magnitude values among amino acid residues suggest its ability to predict interchange of amino acids which are "silent" or "non-equivalent". This observation has a potentiality to interpret the observed dysfunctional natural mutants. Our preliminary data thus provide insight into conformation and environmental effects in protein. They also indicate that the complexity of the phenomenon of molecular organization cannot be explained by conventional use of single numerical value in analysis of structure-function relationship due to its inability to provide sufficient information necessary to obtain complete understanding of this fundamental and the most intriguing question which is as yet to be solved.

It should be noted that although, these observations are based on analyses of serine proteases, the information thus obtained can be readily generalizable to other proteins.

Literature Cited

1. Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Mayer, E.F., Jr., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T. and Tasumi, M. *J. Mol. Biol.* **1977**, 112, 535-542.
2. Markley, J.L. and Ulrich, E.L. *Ann. Rev. Biophys. Bioeng.* **1984**, 13, 493.
3. Jarletzky, O. and Roberts, G.C.K. *NMR in Molecular Biology*, New York, Academic Press, **1981**.
4. Kendrew, J.C., Bodo, G., Dintzis, H.M., Parrish, R.G. and Wyckoff, H. *Nature* **1958**, 181, 662-666.
5. Anfinsen, C.B. *Science* **1973**, 181, 223-230.
6. Ellis, R.J. *Sem. Cell Biol.* **1990**, 1, 1-9.
7. Georgopoulos, C. *Trends Biochem. Sci.* **1992**, 17, 295-309.

8. Collier, D.N. *Adv. Prot. Chem.* **1993**, 44, 151-193.
9. Hendrick, J.P. and Hartl, F.U. *Ann. Rev. Biochem.* **1993**, 62, 349-384.
10. Orcutt, B.C. and Dayhoff, M.O. *Protein Sequence Database*, National Biomedical Research Foundation, Washington, D.C. **1982**.
11. Liebman, M.N., Venanzi, C.A. and Weinstein, H. *Biopolymers* **1985**, 24, 1721-1758.
12. Kidera, A., Konishi, Y., Ooi, T. and Scheraga, H.A. *J. Protein Chemistry* **1985**, 4, 265-297.
13. Tanford, C. *Science* **1978**, 200, 1012-1018.
14. Kauzmann, W. *Nature* **1987**, 325, 763-764.
15. Baldwin, R.L. *Proc. Natl. Acad. Sci., USA* **1986**, 83, 8069-8072.
16. Lipman, D.J., Pastor, R.W. and Lee, B. *Biopolymers* **1987**, 26, 17-26.
17. Eisenberg, D. and McLachlan, A.D. *Nature* **1986**, 319, 199-203.
18. Cid, H., Bunster, M., Arriagada, E. and Campos, M. *FEBS Lett.* **1982**, 150, 247-254.
19. Miller, S., Lesk, A.M., Janin, J. and Chothia, C. *Nature* **1987**, 328, 834-836.
20. Cornette, J.L., Cease, K.B., Margalit, H., Spouge, J.L., Berzofsky, J.A. and DeLisi, C. *J. Mol. Biol.* **1987**, 195, 659-685.
21. Rose, G.D. *Nature* **1978**, 272, 586-590.
22. Rose, G.D. and Roy, S. *Proc. Natl. Acad. Sci., USA* **1980**, 77, 4643-4647.
23. Kyte, J. and Doolittle, R.F. *J. Mol. Biol.* **1982**, 157, 105 - 132.
24. Both, G.W. and Sleigh, M.J. *Nucleic Acids Res.* **1980**, 8, 2561-2575.
25. Hopp, T.P. and Woods, K.R. *Proc. Natl. Acad. Sci., USA* **1981**, 78, 3824-3828.
26. Novotny, J. and Auffray, C. *Nucleic Acids Res.* **1984**, 12, 243-255.
27. Qian, N. and Sejnowski, J. *J. Mol. Biol.* **1988**, 202, 865-884.
28. Bohr, H., Bohr, J., Brunak, S., Cotterill, R., Lautrup, B., Nørskov, L., Olsen, O. and Petersen, S. *FEBS Lett.* **1988**, 241, 223-228.
29. Holley, H. and Karplus, M. *Proc. Natl. Acad. Sci., USA* **1989**, 86, 152-156.
30. Taylor, W. and Thornton, J. *J. Mol. Biol.* **1984**, 173, 487-514.
31. Liebman, M.N. *Enzyme* **1986**, 36, 115-140.
32. Rao, S.T. and Rossmann, M.G. *J. Mol. Biol.* **1973**, 76, 241-256.
33. Kidera, A., Konishi, Y., Oka, M., Ooi, T. and Scheraga, H.A. *J. Protein Chemistry* **1985**, 4, 23-55.
34. Pethig, R. *Dielectric and Electronic Properties of Biological Materials*, John Wiley & Sons, New York **1979**.
35. Weinstein, H., Namboodiri, K., Osman, R., Liebman, M.N.

- and Rabinowitz, J. In *QSAR in Toxicology and Xenobiochemistry*, Tichy, M., Ed.; Elsevier : Amsterdam, **1985**; pp 451-463.
36. Pearlman, R.S. In *Physical Chemical Properties of Drugs*; Yalkowsky, S.H., Sinkula, A.A., Valvani, S.C., Eds.; Marcel Dekker : New York, **1980**; pp 321-347.
37. Liebman, M.N. *Biophys. J.* **1980**, 32, 213-215.
38. Kabsch, W. and Sander, C. *Proc. Natl. Acad. Sci., USA* **1984**, 81, 1075-1078.
39. Argos, P. *J. Mol. Biol.* **1987**, 197, 331-348.

RECEIVED July 11, 1994

Chapter 13

Comparison of Spiral Structures in Wheat High Molecular Weight Glutenin Subunits and Elastin by Molecular Modeling

Donald D. Kasarda¹, Gregory King², and Thomas F. Kumosinski²

¹Western Regional Research Center, Agriculture Research Service, U.S. Department of Agriculture, 800 Buchanan Street, Albany, CA 94710

²Eastern Regional Research Center, Agriculture Research Service, U.S. Department of Agriculture, 600 East Mermaid Lane, Philadelphia, PA 19118

The high-molecular-weight glutenin subunits, which contribute importantly to the elasticity of wheat flour doughs, have a large central domain composed of repeating amino acid sequences rich in glutamine, proline, and glycine. Although a β -spiral conformation, similar to that proposed for the polypeptide of elastin, has been suggested for these repeats, results of molecular modeling by secondary structure prediction, energy minimization, and chemical dynamics calculations indicate that assignment of several inverse γ turns to the consensus repeats produces a highly stable spiral structure. This spiral is stabilized by extensive interturn hydrogen bonding involving the glutamine side chains.

Doughs made from bread wheat flour are cohesive and elastic, yet have considerable extensibility. These viscoelastic properties are mainly contributed by the gluten proteins—storage proteins found in the starchy endosperm of the wheat kernel, which is the source of white flour obtained in the milling process. Gluten proteins have traditionally been divided by solubility into two roughly equal main fractions, gliadins and glutenins. Today, the more soluble gliadins are more likely to be defined as monomeric proteins, whereas the less soluble glutenins are known to consist of polymeric structures formed through intermolecular disulfide crosslinking of protein subunits. The elasticity of doughs is contributed by these glutenin polymers (1).

Both the gliadin and glutenin fractions are made up of many similar, but distinguishable, proteins having large percentages of glutamine (30-55 per cent on a molar basis) and proline (11-30 per cent on a molar basis) in their amino acid compositions. The amide side chain of glutamine can serve as both hydrogen bond donor and acceptor and the high percentage of glutamine in gluten probably accounts for much of its cohesive nature. Hydrogen bonding is apparently the primary secondary bonding force responsible for determining the structure and interactions of gluten proteins (1).

Glutenin subunits are usually divided into high-molecular-weight glutenin subunits (HMW-GS) and low-molecular-weight glutenin subunits (LMW-GS) on the basis of subunit molecular weights (2). The gluten proteins are coded by genes that have considerable allelic variation and consequently wheat varieties usually have complements of gluten proteins that differ among varieties, sometimes widely. Certain HMW-GS, recognized by their mobilities in SDS-polyacrylamide gel electrophoresis,

This chapter not subject to U.S. copyright
Published 1994 American Chemical Society

were the first gluten proteins shown to correlate with bread making quality characteristics (3), and although other types of subunits have since been demonstrated to correlate with quality, the importance of HMW-GS to dough elasticity and bread-making quality is generally accepted. Flours derived from wheats with no HMW-GS form water-flour doughs that have almost no elasticity and relatively poor cohesiveness (4,5).

How do these contributions to dough properties by HMW-GS arise? Speculations have centered on two characteristics, 1) the number and arrangement of cysteine residues in their primary structures, which might determine the type and extent of their polymerization through intermolecular disulfide bond formation, and 2) their unusual shape. Here, our focus is on the latter possibility.

Complete amino acid sequences based on DNA sequences have demonstrated that HMW-GS are made up of three domains (6). The N-terminal domain (consisting of about 100 amino acid residues) and the C-terminal domain (consisting of about 50 residues) contained most of the cysteine residues, whereas the large central domain (ranging from about 500 to 700 amino acids in length for the different HMW-GS) consisted of repeating sequences made up largely of glutamine, glycine, and proline. Although the repeats are imperfect, they are generally similar to 6-, 9-, and 15-residue consensus sequences, interspersed in no obvious order throughout the repeating sequence domain. This can be recognized by analysis of the overall domain—although the repeating sequences may also be thought of as 3-, 6-, and 9-residue repeats (6). In Figure 1, we show three sequences in HMW-GS 1DX5 (7) that happen to correspond exactly to the 6-, 9-, and 15-residue consensus sequences.

Tatham et al (8), noting that predictions of secondary structure by the Chou-Fasman method (9) indicated frequent β -turns in the repeating sequence domain of the HMW-GS, suggested that these repeats might form a β -spiral structure based on β turns that was similar to the structure proposed by Urry and coworkers (10) for certain repeating sequences found in elastin. In this chapter, we shall refer to any shallow-pitched, helical arrangement of a polypeptide chain, similar to that formed by the elastin polypentapeptide (10), as a spiral. Field et al. (11) analyzed the shape of a HMW-GS by viscometric methods and concluded that it had a rod-like shape with dimensions of about 50 nm x 1.75 nm (500 Å x 17.5 Å) in acetic acid, which supported the likelihood of a spiral structure for the repeating sequences. Miles et al. (12) obtained scanning tunneling micrographs of a HMW-GS that supported a rod-like structure and concluded that the rodlike subunits had an apparent diameter of 19 Å (although the spacing of the array on the substrate actually corresponded to about 30 Å and the 19 Å spacing might correspond to a minimum diameter), and a repeating character (pitch) spaced at 15 Å along the rods that might correspond to the turns of a spiral. A highly simplified idea of the structure of HMW-GS arising from these studies is illustrated in the drawing of Figure 2, which also illustrates the differences in cysteine residues between the two main types (designated x and y).

In the work we report here, we have attempted to define, through the use of computerized approaches to molecular modeling, molecular structures that are in accord with the limited available physical data. In addition, we make some comparisons of our results to the model of Urry (10) for elastin, which was the basis for the suggestion by Tatham et al. (8) of a β -spiral-like conformational structure for the repeat domain (repeat region) of HMW-GS. Some earlier molecular modeling studies focused on possible β turns of a single 6-residue consensus repeat (13) and on the structures surrounding the cysteine residues of HMW-GS (14), which are located mainly in the N- and C-terminal domains.

Methods

Hardware, Software, and Conformational Prediction. Modeling was carried out mainly with an Evans and Sutherland PS390 interactive computer graphics display

6-residue repeat:	449	454	-Gln-Pro-Gly-Gln-Gly-Gln-
9-residue repeat:	455	463	-Gln-Pro-Gly-Gln-Gly-Gln-Gln-Gly-Gln-
15-residue repeat:	118	132	-Gln-Pro-Gly-Gln-Gly-Gln-Gln-Gly-Tyr-Tyr-Pro-Thr-Ser-Pro-Gln-

Figure 1. Illustration of the 6-, 9-, and 15-residue consensus repeating amino acid sequences as represented by actual sequences found in HMW-GS 1DX5.

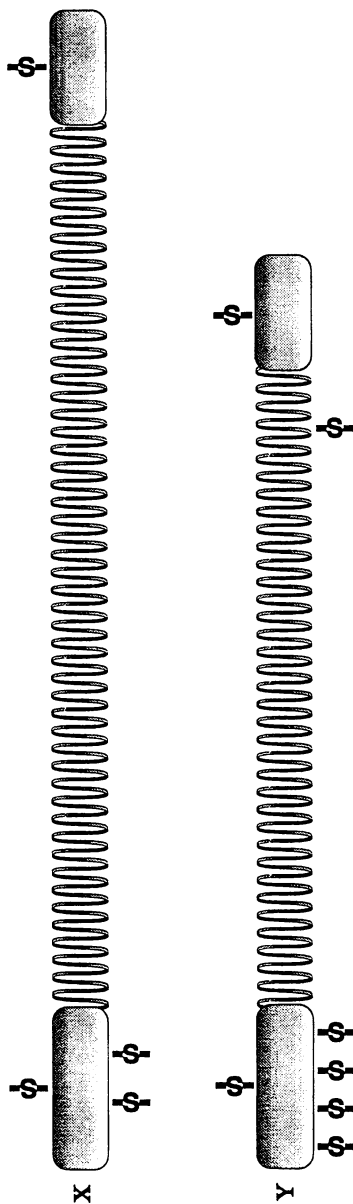


Figure 2. Schematic diagram of x-type and y-type high-molecular-weight glutenin subunits.

driven by the Sybyl software package from Tripos, Inc., but also with a Silicon Graphics, Inc., Personal Iris workstation running software from Molecular Simulations, Inc. (MSI), that included Quanta and CHARMM (15).

Sequence-based prediction of conformation for the 6-residue consensus sequence -QPGQGQ- was carried out by the method of Garnier et al. (16), but the conformations applied were based primarily on past experience (17) and various possibilities were tested by a trial and error approach in which success was measured by achievement of an acceptably stable spiral structure for a polypeptide chain corresponding to 60, 120, or 240 amino acids. The 9-residue and 15-residue consensus sequences were not modeled independently, but were fit to the template resulting from modeling of the 6-residue sequence.

The model of the polypeptide chain corresponding to the -QPGQGQ- repeats was constructed using the software packages, which include a dictionary of geometric parameters, such as bond lengths, bond angles, and Van der Waals radii, all of which were compatible with parameters determined by X-ray crystallographic analysis. The polypeptide chain was constructed amino acid residue-by-residue and then each residue assigned ϕ and ψ angles characteristic of the conformational structures being tested. All ω angles were assigned to the conventional *trans* configuration.

Energy Minimization. When the constructed polypeptide chain formed a helical (spiral) structure, the energy of the structure was analyzed and it was subjected to energy minimization to relieve strain and Van der Waals overlaps and to seek important local energy minima, again using the software included in either of the two packages. In this study, empirical energy functions were used to model the structures. The potential energy model is described as a collection of overlapping balls (with radii based on Van der Waals radii) for the atoms, which are connected to one another by springs that mimic the vibrational character of the bonds. In the calculation of energy, the atoms are assigned Van der Waals attractive and repulsive forces, along with electrostatic forces, for bonded and nonbonded interactions. Modeling by means of the Sybyl software was based on the force field of Kollman (18, 19), otherwise the force field of CHARMM (15) was used. Solvent was not usually included in the models except for one modeling experiment that included water molecules. Minimization calculations were carried out for structures *in vacuo* and involved either a conjugate gradient algorithm (Sybyl) or the method of steepest descents (CHARMM) (15).

Dynamics. In order to extend rigid geometry minimization and to take into account the dynamic state of the molecules at room temperature, heat (kinetic energy) was applied to the system in small steps (randomly assigning velocities to each atom) to bring the temperature up to 300 °K. Integration of the Newtonian equations of motion allowed the trajectories of the accelerated atoms to be determined. In some cases (Sybyl), bond distances for the polypeptide backbone were constrained. Molecular dynamics calculations were carried on for a number of steps to equilibrate the system until statistical properties became stable with time (15). The resulting structures were then subjected once again to energy minimization.

Results

Conformational Assignments. Sequence-based prediction methods for turns were examined for the 6-residue consensus sequence, but were of minimal value. The method of Garnier et al. (16) gave indication that only turns would be expected in a polypeptide chain based on repeats of the QPGQGQ sequence. The type of turn was not predicted by the method of Garnier et al. (16). Accordingly, we tested several different types of turns.

β turns. Because Tatham et al. (8) had based their model of HMW-GS repeats on the β spiral, which incorporates β turns (10), we tried this type of turn in various combinations—for example, assigning $N + 1$ and $N + 2$ of a type II β turn to the Pro-Gly combination. Although a spiral structure was obtained, it was highly distorted, having a flattened, ribbon shape as a consequence of the severe change in backbone chain direction mandated by the β turns. In addition the ribbon had a severe twist. Upon minimization, the energy remained positive. Various other combinations of β turns were tried, but with similarly poor results. The energies of the resulting structures were not particularly good after energy minimization.

γ turns. We found, however, that a particular combination of inverse γ turns was highly effective in producing stable spiral structures. The γ turn is a 3-residue turn and here we assigned ϕ, ψ values of -75° and 59° to the central residue of the inverse γ turn (17). This type of turn is similar to the conformation known as a C_7 structure when a hydrogen bond connects residues N and $N + 3$ (20). The best results were obtained for assignment of inverse γ turns to Pro 2, Gln 4, and Gly 5 of the 6-residue repeating sequence. The Gln 4, Gly 5 assignments resulted in a short stretch of 2-7 ribbon structure. The combined effect of the assignments to a 6-residue sequence was an S-shaped curve in which part of the S was truncated as illustrated in Figure 3, which shows backbone structure only. When a second and then a third consensus sequence were added to the first, the 18-residue structure formed was three-lobed in shape with a strong tendency to form the first turn of a spiral (Figure 3). When many repeats were added, the three-lobed structure was maintained, but with a slight rotational displacement of each successive lobe (Figure 4). Changes in this structure upon energy minimization were relatively small; the initially positive energy became negative rapidly upon energy minimization. The initial structure is illustrated in side (Color Plate No. 11) and top (Figure 5) views of a 120-residue spiral. In this study, we placed our emphasis on an inverse turn with ϕ, ψ values of -75° and 59° because these were deemed common for γ turns (particularly those involving a proline residue) in proteins for which structural information was available. Preliminary studies indicated, however, that normal γ turns (signs reversed) also produced a spiral, as did small variations in the ϕ, ψ angles of both types of turn, when applied to the same residues in the repeats.

Dynamics. Application of dynamics to the system for 30 picoseconds (ps), by which time, energy and the various other trajectories became stabilized, followed by a re-minimization of the energy, resulted in considerable displacement of atoms from their positions in the energy minimized structure and an enhancement of hydrogen bonding by Gln side chains—to the backbone and to other Gln side chains. This displacement is illustrated in Color Plate No. 12, which is a side view of the spiral (compare with equivalent view of the initial structure in Color Plate No. 11 and in Figure 6, which is a view down the major axis of the spiral after dynamics. The trajectory for the energy of the system became stable rapidly during the dynamics simulation indicating the inherent stability of the structure. Despite the distortions of the structure introduced by dynamics simulations, the overall spiral shape was only slightly changed in appearance and in dimensions. The diameter of the spiral was approximately 24 Å and each turn was displaced from the previous turn by about 10 Å. A Ramachandran plot (Figure 7) after dynamics showed only one clear clustering of ϕ, ψ torsional angles ($N-C_\alpha$ and $C_\alpha-C_{\text{carbonyl}}$, respectively)—in the region corresponding to inverse γ turns ($-75^\circ, 59^\circ$), although this was less evident in some other modeling experiments, where the torsional angles were generally more characteristic of non-ordered structure. The γ turns evidently survived the moderately extensive structural displacements resulting from the dynamic simulations to some extent in this particular calculation, but a regular repetition of γ turns does not seem characteristic of the resulting spiral. Inclusion of solvent water in one model of the 6-residue repeats resulted in some hydrogen bonding of water to glutamine side chains, but did not disrupt the spiral.

NOTE: The color plates can be found in a color section in the center of this volume.

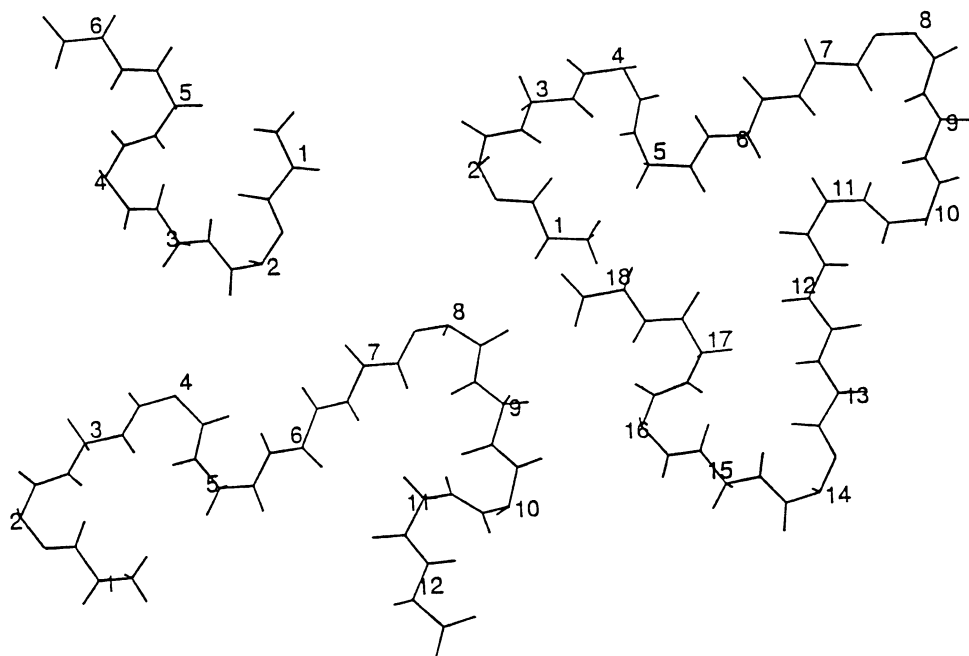


Figure 3. The structure of 6, 12, and 18 residues corresponding to the 6-residue consensus sequence after application of γ turns. The peptide chains are oriented as they would be in a view looking down the spiral axis and backbone structure only is shown in stick form.

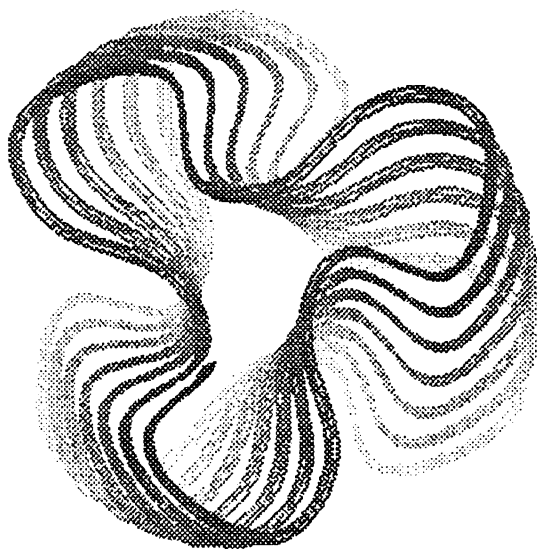


Figure 4. Model of the peptide chain (backbone only; represented by a ribbon) for 120 residues (6+ turns) of the 6-residue repeating sequence. View looking down the spiral axis.

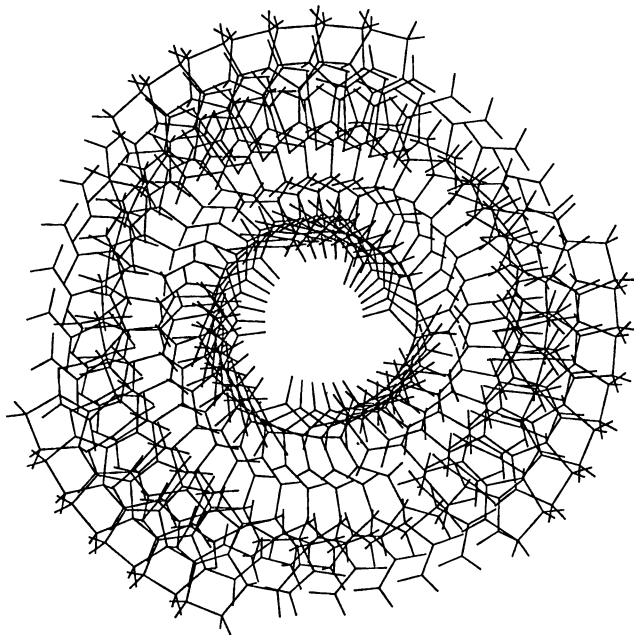


Figure 5. Stick model of structure from Color Plate No. 11. View looking down spiral axis. All bonds shown.

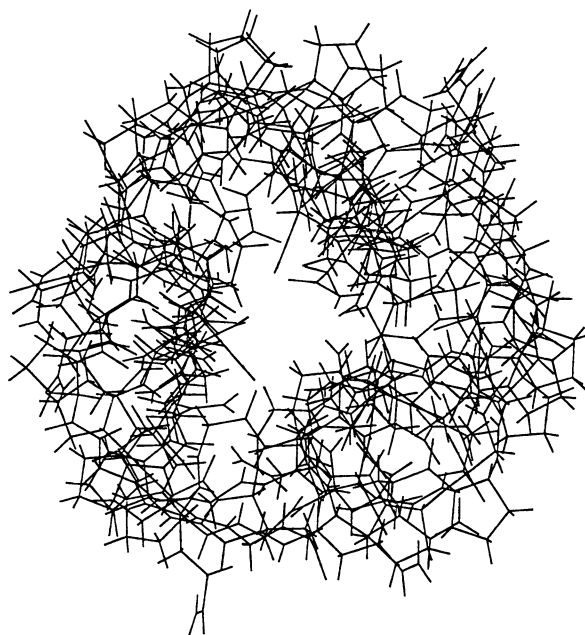


Figure 6. Structure as in Figure 5, but after 30 ps of dynamics calculations.

Modeling of peptides corresponding to the 9-residue and 15-residue consensus sequences. The 9- and 15-residue consensus sequences were fitted as to a template to the spiral that resulted from modeling of the 6-residue consensus sequence. The appropriate amino acids were substituted into the spiral—proceeding from the first 6-residue sequence (common to all 3 sequences). The result showed no serious overlaps as indicated by reasonable energy values after minimization. Dynamics resulted in distortion of the highly regular structure resulting from minimization, but, again, energy values were reasonably negative, although clustering of ϕ, ψ angles in the region of the Ramachandran plot corresponding to inverse γ turns was less for the 9-residue sequence and considerably less for the 15-residue sequence as compared with the 6-residue sequence.

This template approach resulted in an alternating arrangement of the tyrosine-tyrosine residues that correspond to positions 9 and 10 of the 15-residue repeat. The first two projected from the surface of the spiral (sometimes hydrogen bonding to one another after dynamical simulation) whereas the next two had one tyrosine projecting out from the surface and the other projecting in toward the center of the spiral, hydrogen bonding to the backbone chain in the energy minimized structure. This arrangement was continued throughout the spiral. The decreased number of glutamines in the 15-residue repeat resulted in less hydrogen bonding between turns of the spiral as compared with the 6- and 9-residue repeats.

Elastin polypentapeptide β -spiral. Modeling of the polypentapeptide structure of elastin, which consists of repeats of the sequence -VPGVG-, by similar approaches to those we used for the HMW-GS repeats, did not result in the structure proposed by Urry and coworkers (10). When we applied a type II β turn to the Pro-Gly residues in the repeats, with other residues in the extended form, a spiral structure resulted in which valine residues projected in towards the center of the spiral and proline rings projected out from the surface of the spiral. In contrast, the structure proposed by Urry and coworkers (10) has valine residues projecting out from the surface of the spiral and the rings of proline residues lying largely in the surface of a cylinder defining the surface of the spiral. Application of the ϕ, ψ angles provided by C.-H. Luan and D. W. Urry (personal communication), based on physical studies, yielded their proposed β spiral structure in which valine residues project out from the spiral surface. Both models of the polypentapeptide had reasonable stabilities according to molecular modeling, but the Urry structure was not stable when we subjected it to molecular dynamics simulation at 300 °K. Wasserman and Salemme (21) carried out dynamics calculations on the Urry structure that included solvent, but they constrained end atoms in their study. We were unable to reproduce the Urry structure using sequence-based algorithms for secondary structure prediction, followed by energy minimization, and molecular dynamics simulations without application of constraints. Possible reasons for the failure of the modeling software to predict the structure for the repeating polypentapeptide as proposed by Urry (10) will be discussed in the following section.

Discussion

Our modeling studies indicate that a spiral structure for the repeating sequences of HMW-GS is highly compatible with energy and dynamics calculations when particular turn conformations are applied to the appropriate consensus sequences. Surprisingly, we have found the lowest energy after dynamics and minimization for a spiral based on initial definition of inverse γ turns rather than β turns as originally suggested by Tatham et al. (8). Additionally, the dimensional fit to existing physical data was moderately good, better than for any other structures we examined. The diameter of the spiral

predicted by our best model is about 24 Å, which may be compared with predictions of 18 Å from viscosity studies (11) and 20 Å from scanning tunnelling micrographs (12). The scanning tunnelling micrographs also show a spiral repeat of 15 Å. The turns in our model repeat at about 9 Å.

Analysis of ϕ , ψ angles of our model after energy minimization and dynamics calculations had been applied indicated no significant number of β -turns of any type, but a residual cluster of γ turns remained (Fig. 7), although this was not so evident in some other analyses. Our model might be called a γ spiral, but even γ turns did not make up a regular repeating motif after dynamics simulations when replicate analyses were considered.

The principal stabilization of the spiral we modeled appears to result from hydrogen bonding of glutamine side chain amide groups to the backbone amide groups and to other glutamine side chain amide groups. Glutamine makes up about 40% of the amino acids on a molar basis in HMW-GS repeats, 50% in our 6-residue repeat model, 56% in the 9-residue repeat model, and 33% in the 15-residue repeat model. As a consequence of the predominance of glutamine side chains at the surface, the γ spiral appears to be highly hydrophilic. This is in accord with the very early elution of HMW-GS from C₁₈ columns in reverse phase HPLC (22). Although nonpolar interactions and ionic bonding may contribute to the stabilization of conformational structure in the repeating sequence domains of HMW-GS, the contribution is likely to be minor in comparison with hydrogen bonding. Even the presence of the two adjacent tyrosine residues that occur in the 15-residue repeats interspersed occasionally throughout the repeating region is unlikely to modify strongly the general hydrophilicity contributed by the glutamine side chains. The decreased number of glutamine residues in the 15-residue repeats will likely decrease the amount of interturn hydrogen bonding for these repeats and, as a consequence, may diminish spiral stability where these repeats occur in the repeating sequence domains of HMW-GS. If a spring-like expansion of the spiral domain were to occur under mixing stress in doughs, it might occur selectively at the 15-residue repeats.

It has been indicated that γ turns are probably of low stability in aqueous solution because the hydrogen bond that stabilizes the turn can readily be broken by interaction with water (23). Because our modeling did not include interactions with solvent, it might be considered that interactions with water would break up the hydrogen bonding that we propose stabilizes the spiral in the absence of water and call into question the importance of γ turns to the spiral structure of the repeating sequence domain of the HMW-GS. Exchange of side chain hydrogen bonds with water is likely to some extent. Increased exchange of water into the protein rich phase with increase of temperature has been reported in NMR studies of the HMW-GS. Nevertheless, the extensive network of hydrogen bonding that can be seen in our models is likely to be stabilized through cooperative interactions to a considerable degree. We point out that gluten proteins are cohesive and insoluble in neutral aqueous solutions. This is likely to result from extensive inter- and intramolecular hydrogen bonding. The exchange with water, although it results in considerable swelling of gluten, does not result in solubility. In an actual dough, some of the hydrogen bonds we show in our model of the HMW-GS repeats would be broken by interaction with water and others would interchange with glutamine (and some other) side chains of neighboring protein molecules. We do not consider that this necessarily poses any serious problems in relation to our contention that hydrogen bonding, combined with some natural tendency for the proline-containing repeats to form a spiral structure, is primarily responsible for stabilization of the spiral characteristic of HMW-GS.

The failure of the same approach we took for modeling of HMW-GS repeats to correctly predict the Urry structure of the elastin polypentapeptide brings into question the validity of our model in the absence of any detailed supporting physical evidence from NMR or X-ray crystallography. We put forward the following arguments to

suggest that the discrepancy is not as troubling as might be perceived initially. Our failure to produce the Urry structure for the elastin polypentapeptide might result from a failure to adequately sample conformational space. We point out, however, that elastin is a system in which hydrophobic bonding plays a key role—hydrophobic bonding of the type described by Kauzmann (24) in which entropic effects resulting from the tendency of water molecules to be more highly ordered in the vicinity of hydrophobic groups favor the interaction or clustering of such residues in such a way as to minimize the exposure of these groups to water molecules. The two systems, elastin and HMW-GS, are strongly contrasted by their NMR properties as discussed in the work of Belton et al. (25). The modeling programs we used are reasonably competent in dealing with electrostatic interactions, including hydrogen bonds, but are unable to deal with hydrophobic or other types of entropy-determined interactions, especially considering that water was not included in our elastin polypentapeptide model.

Consequently, we think that there is a reasonably good possibility that our modeling results for the sequence repeats approach reality—despite the failure of the approach to deal adequately with the elastin polypentapeptide system. It might also be noted that our 6-residue consensus sequence is homologous in the positions of Pro2, Gly3, and Gly5 to the polyhexapeptide with repeating sequence -VPGVGV- studied by Urry (10), who suggested that this particular elastin repeat might form a relatively rigid structure in contrast to the elastomeric polypentapeptide. Thus, the polyhexapeptide repeat of Urry might be a better analog to the 6-residue repeat of the HMW-GS than the polypentapeptide. Obviously, further physical studies, which we are pursuing, are needed to settle eventually the question of the structure of the repeating sequence domain of the HMW-GS and its relationship to those formed by the various repeats of elastin.

Literature Cited

1. Kasarda, D. D. *In* Wheat is Unique, Pomeranz, Y., Ed., American Association of Cereal Chemists, St. Paul, MN, **1989**, pp 277-302.
2. Bietz, J. A.; Shepherd, K. W.; Wall, J. S. *Cereal Chem.* **1975**, *52*, 513-532.
3. Payne, P. I.; Corfield, K. G.; Holt, L. M.; Blackman, J. *Sci. Food Agric.*, **1981**, *32*, 51-60.
4. Lawrence, G. J.; MacRitchie, C. W.; Wrigley, C. W., *J. Cereal Sci.*, **1988**, *7*, 109-112.
5. Gao, L; Bushuk, W., *Cereal Chem.*, **1993**, *70*, 475-480.
6. Shewry, P. R.; Halford, N. G.; Tatham, A. S., *Oxford Surveys of Plant Molecular & Cell Biology*, **1989**, *6*, 163-219.
7. Anderson, O. D.; Greene, F. C.; Yip, R. E.; Halford, N.G.; Shewry, P. R.; Malpica-Romero, J.-M., *Nucleic Acids Res.*, **1989**, *17*, 461-462.
8. Tatham, A. S.; Shewry, P. R.; Milfin, B. J., *FEBS Lett.*, **1984**, *177*, 205-208.
9. Chou, P. Y.; Fasman, G. D., *Ann. Rev. Biochem.*, **1978**, *47*, 251-276.
10. Urry, D. W., *Methods in Enzymology*, **1982**, *82*, 673-716.
11. Field, J. M.; Tatham, A. S.; Shewry, P. R., *Biochem. J.*, **1987**, *247*, 215-221.
12. Miles, M. J.; Carr, H. J.; McMaster, T. C.; I'Anson, K. J.; Belton, P. S.; Morris, V. J.; Field, J. M.; Shewry, P. R.; Tatham, A. S., *Proc. Natl. Acad. Sci. USA*, **1991**, *88*, 68-71.
13. Brock, C. J. *In* Gluten Proteins 1990, Bushuk, W., and Tkachuk, R., Eds., American Association of Cereal Chemists, St. Paul, MN, **1990**, pp. 441-446.
14. Greene, F. C., and Anderson, O. D. *In* Gluten Proteins 1990, Bushuk, W., and Takchuk, R., Eds., American Association of Cereal Chemists, St. Paul, MN, **1990**, pp. 362-375.
15. Brooks, B. R.; Bruccoleri, R. E.; Olafson, B.D.; States, D. J.; Swaminathan, S.; Karplus, M., *J. Comput. Chem.*, **1983**, *4*, 187-217.

16. Garnier, J.; Osguthorpe, D. J.; Robson, B., *J. Mol. Biol.*, **1987**, *120* , 97-120.
17. Kumosinski, T. F.; Brown, E. M.; Farrell, H. M., Jr., *J. Dairy Sci.*, *in press*.
18. Kollman, P., *Ann. Rev. Phys. Chem.* **1987**, *38* , 303-316.
19. Weiner, P. K.; Kollman, P., *J. Comput. Chem.*, **1981**, *2* , 287-303.
20. Richardson, J. S., *Adv. Protein Chem.*, **1981**, *34* , 167-339.
21. Wasserman, Z. R.; Salemme, F. R., *Biopolymers*, **1990**, *29*, 1613-1631.
22. Burnouf, T.; Bietz, J. A., *J. Chromatog.*, **1984**, *299* , 185-199.
23. Rose, G. D.; Gierasch, L. M.; Smith, J. A., *Adv. Protein Chem.*, **1985**, *37* , 1-109.
24. Kauzmann, W., *Adv. Protein Chem.*, **1959**, *14* , 1-64.
25. Belton, P. S.; Colquhoun, I. J., Field, J. M.; Grant, A.; Shewry, P. R.; Tatham, A. S., *J. Cereal Sci.*, **1994**, *19*, 115-121.

RECEIVED April 14, 1994

Chapter 14

Modeling Biological Pathways

A Discrete-Event Systems Approach

Venkatramana N. Reddy^{1,4}, Michael L. Mavrovouniotis^{1,2},
and Michael N. Liebman³

¹Institute for Systems Research, University of Maryland,
College Park, MD 20742

²Chemical Engineering Department, Northwestern University,
Evanston, IL 60208–3120

³Bioinformatics Program, Amoco Technology Company, Mail Code F-2,
150 West Warrenville Road, Naperville, IL 60563–8460

A discrete-event systems approach is proposed for the modeling of biochemical reaction systems. The approach is based on Petri nets, which are particularly suited to modeling stoichiometric transformations, i.e., the interconversion of metabolites in fixed proportions. Properties of Petri nets and methods for their analysis are presented, along with their interpretation for biological systems.

The modeling, simulation, and analysis of biological pathways require the integration of a large volume of diverse data from the biological, chemical and physical sciences, if one attempts a complete and quantitative analysis. Usually, the problem of modeling a complex biochemical system involves data that are incomplete and unreliable.

With a traditional quantitative model, such as a system of differential equations of the form $dx/dt=f(x)$, which describes the change of a vector (x) of metabolite concentrations based on enzyme kinetics, one cannot tolerate even a single unknown parameter in the function $f(x)$: The model cannot be simulated to obtain any results, unless all the parameters appearing in the kinetic expressions (in the right hand differential equations, $f(x)$) are known.

In other cases, many parameters are not constant; thus, in order to draw a general conclusion about the behavior of a pathway, one would have to simulate many combinations of values for these parameters. Suppose, for example, that the concentrations of the enzymes participating in a pathway are known to vary within certain ranges, and we are interested in the general characteristics of the pathway within these ranges. In simulating the quantitative model, a large number of combinations of values for the enzyme concentrations might be tested to establish the general pattern.

⁴Current address: Department of Chemical Engineering, Northwestern University, 2145 Sheridan Road, Evanston, IL 60201–3120

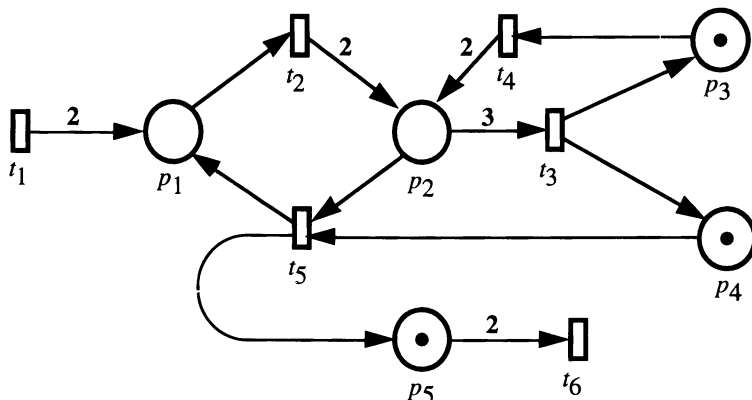


Figure 1. A Petri net graph with *places*, *transitions* and *arcs*. (Reproduced with permission from ref. 10. Copyright 1993 AAAI.)

While various techniques (such as sensitivity analysis) exist to facilitate work with quantitative models in the face of uncertain parameters, one way to address these problems is to begin with a practical qualitative model of the system and avoid the quantitative analysis - in fact, avoid the construction of a complete quantitative model altogether.

A qualitative analysis allows us to draw preliminary conclusions about the biological pathway such as the influence of particular reactions, metabolites or pathway segments on the overall system behavior. In drawing any *a priori* conclusions about the behavior of a pathway, a qualitative method that excludes conducting detailed simulations and analysis of the system is essentially independent of the detailed parametric information (i.e., rate constants, cooperativity indices, etc.) of the pathway.

Our approach to the qualitative modeling of a pathway incorporates the use of a discrete event methodology for the representation and simulation of bioreaction networks. The reactions and other biological processes are modeled as discrete events and analyzed by applying *Petri net* modeling and analysis techniques to this representation of the pathway. Introductory presentations of the theory of Petri nets, the properties of these nets, and methods for their analysis are given below.

Petri Nets

Petri nets are a mathematical and computational tool for the modeling and analysis of discrete event systems. Petri nets offer a formal way to represent the structure of a discrete-event system, simulate its behavior, and draw certain types of general conclusions on the properties of the system. The methodology has wide applications in a number of varied fields such as Computer Science (13), Control Engineering (4, 12), Manufacturing Systems (7) and Information Science (1, 2).

The essential concepts in Petri net theory are outlined in this section. Further details on the theory and applications of Petri nets are available elsewhere in literature (8, 9, 11).

Definitions. A Petri net is a graph (Figure 1) formed by two kinds of nodes, called *places* and *transitions*. Directed edges, called *arcs*, connect places to transitions, and transitions to places. For the sake of convenience, the presence of multiple arcs between a single place and a single transition is represented by a single arc with an arc-weight: We associate a non-zero integer equal to the number of implied connecting arcs with this one weighted arc.

A positive integer number of *tokens* may be assigned to each place; these numbers of tokens form the state of the Petri net (which will be defined as a *marking*, below). The state of the net changes to another (after each event) as the discrete event system operates.

Pictorially, places are represented by circles, transitions by boxes, arcs by lines ending in an arrow, and tokens as black dots placed in the circles. Generally if there is no arc-weight explicitly specified on the graph it is assumed to be equal to one.

Execution. Each transition is associated with a finite number of input places and output places. From the perspective of modeling a discrete event system, it is necessary to satisfy a set of pre-conditions (defined by the input places) before an event (transition) may occur; the event results in a set of post conditions (output places).

In a Petri net a transition is *enabled* when the number of tokens in its input places is greater than or equal to the weights on the arcs connecting the places to the transition. A transition with no input places, called a *source transition*, is always enabled. In Figure 1, the transitions t_1 and t_4 are enabled, while the rest are not.

An enabled transition can *fire*, consuming tokens from its input places and depositing tokens in its output places; the numbers of tokens consumed and produced are determined by the arc-weights. The firing of transitions can be understood as the movement of tokens, from one place to another, through the transitions. A transition with no output places, called a *sink transition*, can fire when enabled consuming the tokens from its input places.

Figure 2 shows the same Petri net graph from Figure 1 after the firing of several transitions. The firing of one enabled transition may deposit tokens in the input places of another transition - thus enabling that transition to fire in turn. In Figure 1 t_1 and t_4 are enabled, hence one possible firing sequence could begin with transition t_1 firing and depositing two tokens in place p_1 ; then t_4 firing, consuming one token from p_3 and depositing two tokens in p_2 (Figure 2(a)). Similarly, the firing sequence t_2, t_3, t_5 and t_6 , starting from a marking of Figure 2(a) will result in the marking of Figure 2(b).

Marking. The state of a Petri net is determined by the number of tokens present in each place of the net. The marking \mathbf{M} of a Petri net is a vector whose elements correspond to the number of tokens present at each place of the Petri net. Thus, the size of the vector \mathbf{M} is equal to the number of places in that net. The execution of the Petri net changes the marking by decreasing tokens in certain places and increasing them in other places. The initial state of a Petri net before execution is called the *initial marking* \mathbf{M}_0 .

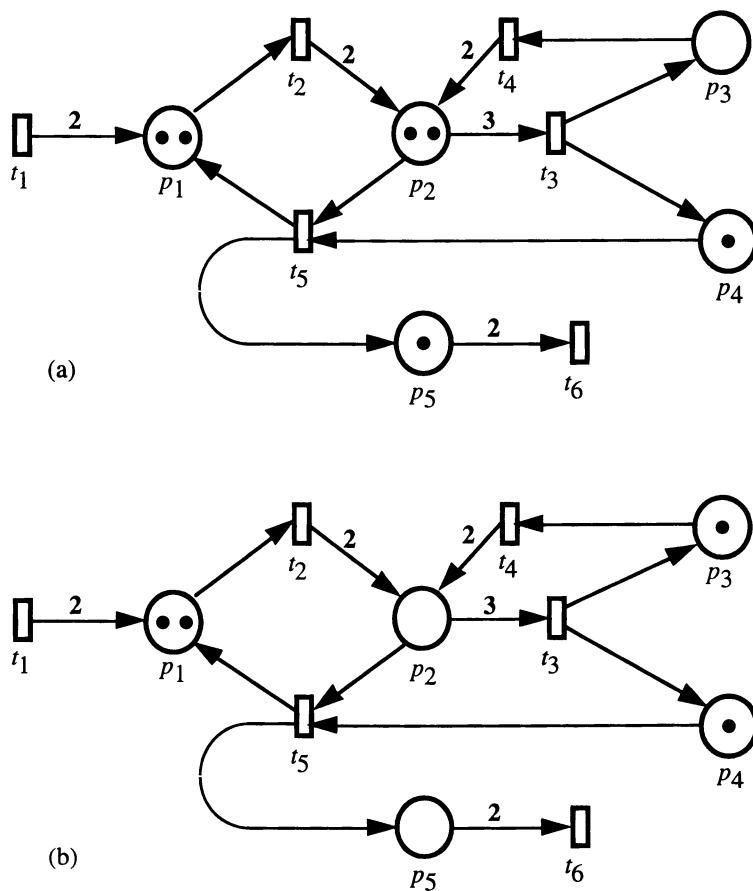


Figure 2. (a) Marking after firing enabled transitions t_1 and t_4 from Figure 1. (b) Marking after firing in sequence $t_2, t_3, t_5,$ and t_6 from Figure 2(a). (Reproduced with permission from ref. 10. Copyright 1993 AAI.)

Mathematical Definition

A Petri net is represented as $PN = (\mathbf{P}, \mathbf{T}, \mathbf{E}, \mathbf{W}, \mathbf{M}_0)$, where:

$\mathbf{P} = \{p_1, p_2, p_3, \dots, p_m\}$ is a finite set of places

$\mathbf{T} = \{t_1, t_2, t_3, \dots, t_n\}$ is a finite set of transitions

$\mathbf{E} \subseteq (\mathbf{P} \times \mathbf{T}) \cup (\mathbf{T} \times \mathbf{P})$ is a set of arcs

$\mathbf{W} : \mathbf{E} \rightarrow \{1, 2, 3, \dots\}$ is a weight function

$\mathbf{M}_0 : \mathbf{P} \rightarrow \{0, 1, 2, \dots\}$ is the initial marking

\mathbf{P} and \mathbf{T} being disjoint sets

Since the Petri net is presented as a model for a discrete event system, it is helpful to have a system of equations that can be used to specify and manipulate the state of the system.

For a Petri net with m places and n transitions, we can formulate (10) a state equation of the type

$$\mathbf{M}_k = \mathbf{M}_{k-1} + \mathbf{A}^T \mathbf{u}_k, \quad k = 1, 2, 3, \dots \quad (1)$$

The index k represents a point in a firing sequence. For each k , \mathbf{M}_k represents an $m \times 1$ vector, the marking after the k th firing; \mathbf{u}_k an $n \times 1$ vector, the *control vector* indicating the transition fired at the k th firing; and \mathbf{A} an $n \times m$ matrix, the *incidence matrix* whose elements a_{ij} denote the change in the number of tokens in place j due to the firing of transition i . The control vector \mathbf{u}_k is simply the unit vector, containing the entry of 1 in the position corresponding to the transition that fired, and 0 everywhere else. The matrix \mathbf{A} describes the weights on the arcs, with an entry of $a_{ij} = 0$ describing the absence of an arc altogether between transition i and place j .

If a particular marking \mathbf{M}_n is reached from the initial marking \mathbf{M}_0 , through a firing sequence $\sigma = \{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3, \dots, \mathbf{u}_n\}$, and the state-equations are summed for all the firings in this σ , we obtain

$$\mathbf{M}_n = \mathbf{M}_0 + \mathbf{A}^T \sum_{k=1}^n \mathbf{u}_k \quad (2)$$

We define $\mathbf{x} = \sum_{k=1}^n \mathbf{u}_k$, as an $n \times 1$ vector, called the *firing count vector*. The element i in \mathbf{x} indicates the number of times transition i must fire to transform \mathbf{M}_0 to \mathbf{M}_n . Substituting the definition of \mathbf{x} in Equation 2, we obtain

$$\mathbf{M}_n - \mathbf{M}_0 = \mathbf{A}^T \mathbf{x} \quad (3)$$

or

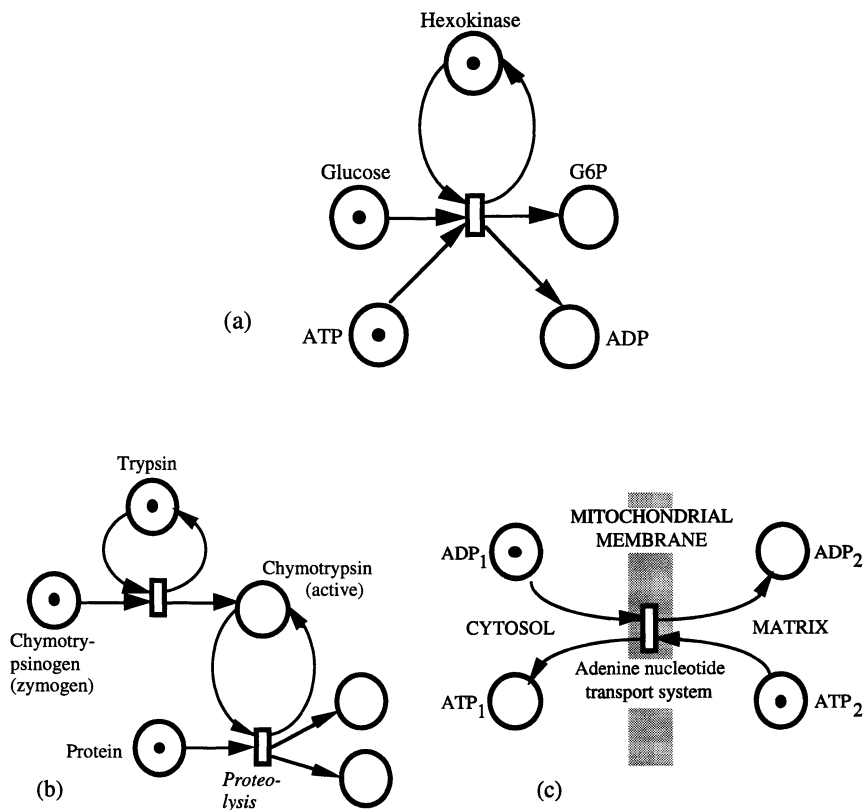


Figure 3. (a) Each place represents a biological compound. (b) & (c) Two different places represent the same compound distinguished by their activities and/or location in the organelle. (Adapted from ref. 10.)

$$\mathbf{A}^T \mathbf{x} = \Delta \mathbf{M} \quad (4)$$

For example, from the Petri net graph of Figure 1 we obtain the incidence matrix \mathbf{A} and the initial marking \mathbf{M}_0 as follows

$$\mathbf{A} = \begin{bmatrix} 2 & 0 & 0 & 0 & 0 \\ -1 & 2 & 0 & 0 & 0 \\ 0 & -3 & 1 & 1 & 0 \\ 0 & 2 & -1 & 0 & 0 \\ 1 & -1 & 0 & -1 & 1 \\ 0 & 0 & 0 & 0 & 2 \end{bmatrix} \quad \mathbf{M}_0 = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$

At $k = 1$, t_1 fires ...

$$\therefore \mathbf{u}_1 = [1 \ 0 \ 0 \ 0 \ 0 \ 0]^T \text{ and } \mathbf{M}_1 = [2 \ 0 \ 1 \ 1 \ 1]^T$$

At $k = 2$, t_4 fires ...

$$\therefore \mathbf{u}_2 = [0 \ 0 \ 0 \ 0 \ 1 \ 0]^T \text{ and } \mathbf{M}_2 = [2 \ 2 \ 0 \ 1 \ 1]^T$$

and so on.

Representation of Biological Pathways

The representation of the essential components in a biological pathway, using Petri net terminology, is the first step in modeling the metabolic network as a discrete event system (10).

For biological pathways, places would represent compounds (such as metabolites, enzymes, cofactors, etc.) participating in the biochemical system. Tokens indicate the presence of a compound.

Instead of having just one place represent each compound, it may be necessary to prescribe two or more places to represent the compound, when there are alternative physical attributes, changes in the activity of the compound, or distinct biological functions.

For example, in Figure 3(a), each place represents one biological compound, whereas in Figure 3(b), two places represent the same compound because a difference arising in activities: One place represents the inactive zymogen and the other the active enzyme Chymotrypsin. As another example, if we would like to distinguish between compounds based on their location in the cell, we could have different places represent the same compound. For instance, in Figure 3(c), the ATP pools inside and outside the mitochondrion in a cell are different and their relative concentrations are determined through a selective transport process. Hence, we could have two places, one representing the compound inside and the other representing the compound outside the mitochondrion.

As would be natural, we assign transitions to represent individual reactions. A series (chain) of forward reactions could be represented as a single transition, if desired, provided that the intermediary compounds are not of primary interest. Arc-weights represent the stoichiometry of reactions, and the direction of an arc is based on the thermodynamic feasibility of the reaction.

The properties if Petri nets are useful in drawing qualitative conclusions about the behavior and structure of biological pathways. The properties are divided in two categories which are defined, along with some of the most relevant properties, in the next two sections.

Behavioral Properties of Petri nets

Behavioral properties of Petri nets are those properties which depend on the initial marking \mathbf{M}_0 of the Petri net. Below, we follow the definition of each behavioral property with a brief comment on the interpretation or application of that property in the context of biochemical pathways.

Liveness. A Petri net is said to be live, if from any marking reachable from M_0 it is possible to fire any transition in the net through some further firing sequence.

In a biological pathway this is a condition that, in any state of the transformation, all individual reactions are potentially active.

Reachability. A marking M_n is reachable from the initial marking M_0 if there exists a firing sequence that can accomplish the change. A necessary (although not sufficient) condition for a marking M_n to be reachable from the initial marking M_0 is the existence of a non-negative solution to Equation 4.

The reachability of a marking from some other marking, in the Petri net representation of a biological pathway, determines the possibility of formation of a specified set of reactant metabolites, by some sequence of reactions that is dictated by the Petri net's firing sequence(s).

Boundedness. A Petri net is bounded if the number of tokens in each place is finite for any marking reachable from the initial marking.

Boundedness of a Petri net model can be used to test for the possibility of excessive accumulation of compounds in the reaction network.

Reversibility. A Petri net is said to be reversible if the initial marking M_0 is reachable from all other possible markings in the set of markings reachable for M_0 .

This property corresponds to reversibility of biotransformations. Most biological pathways are thermodynamically irreversible. This means that they cannot be literally undone one by one. However, an alternate set of reactions may produce the precursor metabolites that were consumed, and thus undo the consequences of the forward reactions.

Fairness. If a firing sequence σ is finite or if every transition in σ occurs infinitely often, then σ is globally fair. A Petri net is globally fair if every σ in all possible markings from M_0 is globally fair.

A pathway that has the property of global fairness suggests the existence of a state of continuous operation, starting from an initial state, without outside intervention. This could result in the formation of certain compounds in unrestricted amounts.

Structural Properties

The properties that do not depend on the marking M_0 of the net, but only on the structure or connectivity of the net come under this category. Again, we follow the definition of each property with some comments that assess the property in the context of biochemical pathways.

S-invariants. These are defined by the solutions to the equation

$$Ay = 0, \quad y \geq 0 \quad (5)$$

Here, \mathbf{y} is an $m \times 1$ vector. The non-zero entries in \mathbf{y} , constitute the *support* of an S-invariant. The support is the set of places whose token count does not change with any firing sequence from \mathbf{M}_0 .

In the Petri net model of a pathway the support of an S-invariant determines the set of compounds whose total net concentrations remain unchanged in the course of a biotransformation. This may happen with compounds that act in a catalytic capacity. For example, the unbound form on enzymes, along with all its bound (or inactivated) forms, may collectively represent an S-invariant. This will occur if there is no production of new enzyme, but merely association/dissociation and activation/inactivation events. S-invariants may also occur for currency metabolites (such as ATP, ADP, and AMP) if, in the system being modeled, consumption of one member of the family is always accompanied by production of another (in an equal number of moles).

T-invariants. These invariants are the solutions to the equation

$$\mathbf{A}^T \mathbf{x} = 0, \quad \mathbf{x} \geq 0 \quad (6)$$

The support of a T-invariant, defined by \mathbf{x} , is the set of transitions that have to fire, from some \mathbf{M}_0 , to return the Petri net to the same \mathbf{M}_0 .

T-invariants give us an insight into the sequences of pathway steps that are necessary to form a cyclic transformation.

Features of Petri net models

We discuss below a few additional features of Petri net models that are relevant in their analysis or in their application to bioreaction systems.

Extendibility. If a high level of abstraction is initially adopted, to construct a simpler Petri net, the net can be subsequently extended upon the initial structure with the modification of relevant sections of the net. This modification need not involve changing the structure of the complete net. For instance, a transition can be visualized as the representation of a Petri sub-net (Figure 4) and any modification to this sub-net is reflected in the behavior of the original transition. This feature is particularly useful in cases where the present knowledge is incomplete, and we would like a representation that can be extended upon the present state of knowledge without significant deviation from the existing structure.

Abstraction. Petri nets allow a certain level of abstraction in the representation of biological reaction systems. This corresponds, in many cases, to a process which is the reverse from that mentioned in the previous paragraph. For example, if a part of a pathway is not of primary interest to us or is not of direct consequence to our analysis, it is possible to collapse this information to a smaller representation without the loss of behavior from the original net. This feature can be useful in the comparison of different pathways for structural links that can result in similar behavior in the network.

Classification. Ordinary Petri nets can be classified according to their structural properties to facilitate the analysis of the model. Structural classification includes the subclasses *State Machines* (SM), *Marked Graphs* (MG), *Free-Choice* (FC) nets, *Extended Free-Choice* (EFC) nets, and *Asymmetric-Choice* (AC) nets (8).

In each of the restricted forms of Petri nets, a few particular properties can be established conclusively.

Methods of Analysis

Reachability Tree. The Reachability tree enumerates the possible states of a Petri net starting from a specified initial marking M_0 . Each node of the tree denotes the reachable marking M starting from M_0 . The branches from a node denote the possible transition firings from that node. The branches in the tree are terminated if a marking results in a deadlock (i.e., no enabled transitions exist) or if a marking is repeated.

The Reachability tree for the Petri net graph in Figure 1 is shown in Figure 5. The root of the tree is the node (0 0 1 1 1) representing the initial marking, t_1 and t_4 are the possible firings leading to the respective markings (2 0 1 1 1) and (0 2 0 1 1), and so forth. The node (2 2 0 1 1)* is a repeated marking and branching from that node is terminated.

We can also choose to terminate branching when a marking supersedes one of its ancestors, i.e., it has elements greater than or equal to its ancestor marking (on an element by element comparison). This would be indicated through special notation that flags some of the entries in the marking as unbounded.

It is possible to answer many questions regarding the properties of Petri nets, such as reachability, liveness, boundedness, etc., by an extended search of the reachability tree.

Often, complex systems result in a Reachability tree that is very large requiring exponential time and memory to solve the problem. However, the analysis of smaller nets is simple and convenient with this method.

Structural Reduction. Large Petri nets can be reduced to smaller nets by the substitution of certain combinations of places and transitions (Figure 6) without sacrificing the original properties of the net (8). This can be viewed as a method for model reduction that lowers the complexity of the original network. Reducing the size of the net is of importance not only for reducing the complexity of the system, but also in achieving a reduction in the memory usage on a computer.

A different approach to the structural reduction of a Petri net is by partitioning the net into smaller nets. The sub-nets preserve the characteristics of the original net (3), hence the analysis of the sub-nets would lead to similar results but with lesser effort.

Matrix Methods. The matrix representation of a Petri net model can be used in the algebraic analysis of properties such as reachability. A necessary condition for the reachability of a marking M from the initial marking M_0 is the existence of a solution to Equation 4, the elements of x being non-negative integer numbers. The sufficiency condition is a more difficult task and the reachability for a general Petri net can conclusively be deduced from the Reachability Tree.

The liveness of a Petri net is a difficult problem (9). The non-existence of deadlock markings -- a marking where there are no more enabled transitions -- is pertinent to the analysis of biological pathways since it addresses the issue of the non-existence of active reactions. An algebraic method for determining the non-existence of deadlock markings has been illustrated in (5). These methods of solution are easier to implement on a computer than the Reachability Tree method and therefore are more appealing.

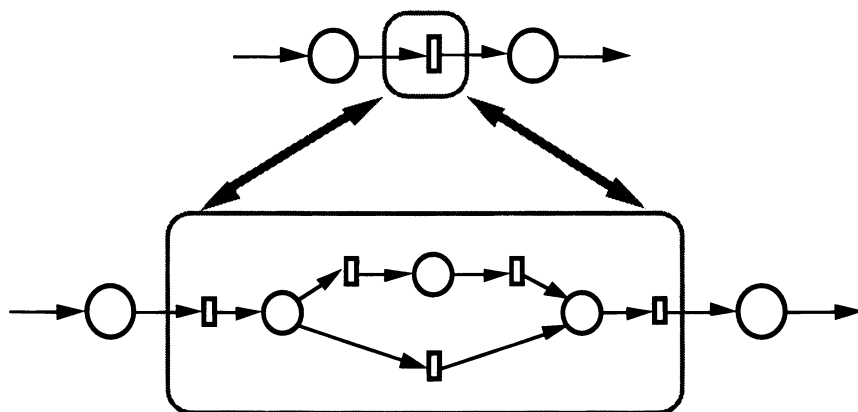


Figure 4. A transition as a representation of a sub-net.

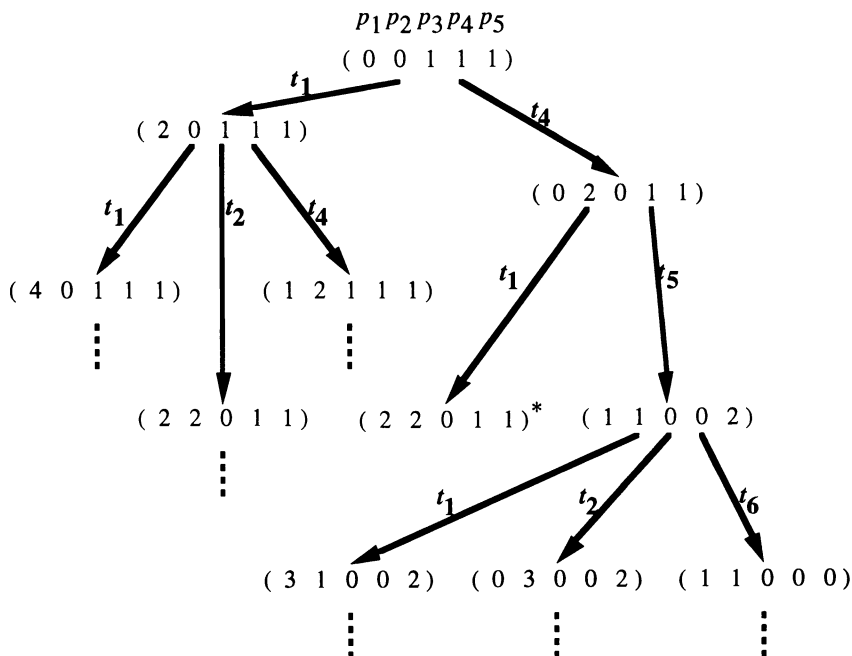


Figure 5. Reachability Tree for the Petri net from Figure 1.

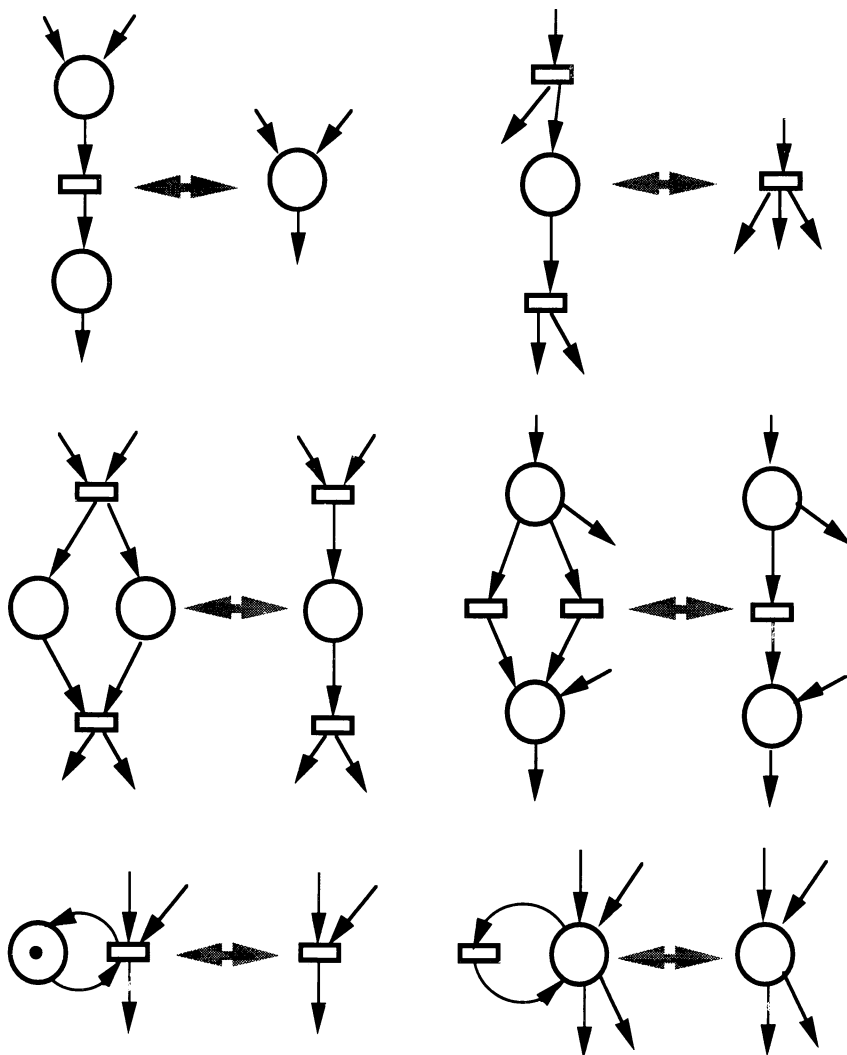


Figure 6. Some examples of structural reductions that are possible in Petri net graphs. (Reproduced with permission from ref. 10. Copyright 1993 AAI.)

Structural Mapping. Structural mapping is the systematic comparison of structures of the Petri net models. It is possible to identify structural similarities of biological pathways with this method. Intuitively, we expect Petri nets with the same structure to exhibit similar dynamic behavior, hence common structural links between Petri net models of pathways would predict similar behavior among the pathways. Pathway models can be compared to elucidate the behavior of one path-

way from the known behavior in the other pathway, if they have common structural links.

Concluding Remarks

The methodology of Petri nets demonstrates the use of qualitative methods as useful preliminary analysis tools for biological pathways. The method is easy to implement and visually comprehensible. Although most properties can be determined for certain classes of Petri nets, the solution to the liveness, reachability, and boundedness problems are still very complex for the general Petri net structure (9). The issue of complexity and decidability of problems in the general class of Petri nets is a matter of concern; for practical purposes, one may strive to model the system (or parts of the system) through one of the restricted categories of Petri nets, reducing the complexity issues.

Petri nets can also be extended by suitable modifications to the definition to enhance the modeling power of the net. *Timed Petri nets* (8) and *Colored nets* (6) are results of such extensions. In fact, the extension of Petri nets by the inclusion of *inhibitor arcs* (i.e., a transition is only enabled when the input place connected by this arc does not contain any tokens) can increase the modeling power of Petri nets to that of a Turing machine (9). Naturally, the extension of the modeling power exacerbates undecidability and complexity obstacles. Thus, an important task is the identification of appropriate narrow classes of nets that combine reasonable expressive power for the domain of biochemical pathways without undue complexity burdens.

Acknowledgements

This work was supported in part by the National Library of Medicine (grant LM 05278) and by Amoco.

Literature Cited

1. Berthelot, G.; Terrat, R. *IEEE Trans. Commun.* 1982, vol. COM-30(12), pp. 2497--2505.
2. Diaz, M. *Comput. Networks* 1982, vol. 6, pp. 419--441.
3. Fahmy, H. M. A. *Theo. Comp. Sci.* 1990, vol. 77, pp. 321--330.
4. Ichikawa, A.; Hiraishi, K. *Discrete Event Systems: Models and Application; Lecture Notes in Control and Information Sciences 103*; Springer-Verlag: New York, NY, 1987.
5. Ivanov, N. N. *Automation and Remote Control.* 1991, vol. 52, pp. 986--989.
6. Jensen, K. In *Colored Petri Nets*; Brauer, W.; Rozenberg, G.; Salomaa, A. Eds.; Monographs on Theoretical Computer Science; Springer-Verlag: New York, NY, 1992; Vol. 1.

7. Martinez, J.; Alla, H.; Silva, M. *Modeling and Design of Flexible Manufacturing Systems*; Elsevier Science Publ.: New York, NY, 1986.
8. Murata, T. *Proc. of the IEEE* 1989, vol. 77, pp. 541--580.
9. Peterson, J. L. *Petri Net Theory and the Modeling of Systems*; Prentice-Hall, Inc.: New Jersey, NJ, 1981.
10. Reddy, V. N.; Mavrovouniotis, M. L.; Liebman, M. N. In *Proc. of the First Intl. Conf. on Intell. Sys. for Mol. Biol.*; Hunter, L.; Searls, D.; Shavlik, J., Eds.; AAAI/MIT Press: Menlo Park, CA, 1993.
11. Reisig, W. *Petri Nets: An Introduction*; Springer-Verlag: New York, NY, 1985.
12. Yamalidou, E. C.; Kantor, J. C. *Computers Chem. Eng.* 1991, vol. 15, pp. 503--519.
13. Yau, S. S.; Caglayan, M. U. *IEEE Trans. Soft. Eng.* 1983, vol. SE-9, pp. 733--745.

RECEIVED June 6, 1994

Chapter 15

Inhibitor-Induced Structural Changes in Serine Proteases Monitored by Fourier Transform Infrared Spectroscopy

Rina K. Dukor and Michael N. Liebman

Bioinformatics Program, Amoco Technology Company, Mail Code F-2,
150 West Warrenville Road, Naperville, IL 60563-8460

Serine proteases are an important family of enzymes whose members participate in a variety of biological activities such as digestion and coagulation. The three-dimensional structure of some of the digestive serine proteases, its inhibitors and their complexes have been determined to a very high atomic resolution. In this paper, conformational perturbations in enzymes (trypsin subfamily) caused by binding of synthetic and natural inhibitors and solvent changes are examined. The changes are monitored by comparison of second-derivative FTIR spectra of inhibited and uninhibited enzymes in solution (H_2O). The results are compared to those obtained from X-ray crystallography studies.

Infrared absorption spectroscopy was first introduced as a tool for the study of peptide and protein conformations more than 40 years ago (1). In the IR absorption spectrum, amide groups are strong chromophores that give rise to nine strong characteristic bands, named amide A, B and I-VII. Among these bands, amide I band (which is due mostly to the C=O stretching vibrations of the peptide backbone) has been primarily used for protein secondary structure determination studies due to its high sensitivity to small changes in molecular geometry and hydrogen bonding of the peptide group. It was recognized early on that for proteins in solution, amide I band is broad and featureless and for proteins that contain segments with different conformations, the results are not always easy to interpret

0097-6156/94/0576-0235\$08.00/0
© 1994 American Chemical Society

(2). Therefore, the early experiments were used to estimate the predominant conformation and to qualitatively follow conformational changes due to changes in solution environment (3-5). Some attempts have been made to derive quantitative information based on curve-fitting the original spectrum (6).

The field of application of IR spectroscopy to the protein structural studies has erupted in the early 80's with commercial availability of Fourier transform infrared spectrometers. Due to inherent advantages of Fourier transform infrared spectroscopy (FTIR) such as high S/N ratios and high wavenumber precision it now became possible to obtain spectra that can be easily manipulated by computer. This led to access to such mathematical treatments as second derivative analysis (7), Fourier self-deconvolution (FSD) (8) and interactive spectral subtraction (9).

Interactive spectral subtraction is especially useful for correcting absorption spectra of proteins in water. Water has a strong absorption band (HOH bending mode) at ~ 1650 cm^{-1} , exactly at the frequency of amide I band. This problem is overcome experimentally by using D_2O as a solvent, which is non-absorbing in this region. But this in turn creates a problem because of incomplete H-D exchange that can lead to band displacements and small structural changes (10). Therefore, all studies in this laboratory are performed for proteins in H_2O -based solvent systems and data analysis has emphasized the careful correction for water bands in the observed spectra.

In the recent years, FTIR studies of proteins have concentrated on extracting quantitative secondary structure information from the spectra. The methods currently used are either based on band narrowing and curve fitting of the amide I band ('frequency-based' approach) or on the principle of 'pattern recognition' (such as factor analysis and partial least squares method) (there is a large variety of good work on these topics; for reviews see ref. 11-14).

Ability to obtain quantitative secondary structure information from FTIR spectra is of particular value when no other technique is available for similar analysis (e.g. membrane-bound proteins) or when a fast approximate result is of interest. We believe, however, that the real power of applying FTIR spectroscopy to proteins lies not in analysis of protein conformation but to follow discrete conformational changes induced by a variety of perturbations (e.g. solvent, inhibitor, site-directed mutation, etc.).

The goal of the work that is partially described here is to find a consistent, reliable method for correlating spectral information to structural knowledge. To date, the highest resolution structural information on proteins comes from x-ray diffraction and therefore we decided to first examine spectroscopically those proteins for which high

resolution x-ray data exists. To optimize the system under analysis to evaluate the limit of resolution of structural detail which the method can detect we chose to examine a series of highly homologous proteins, serine proteases. Two subsets are chosen: one consisting of trypsin and a wide range of its forms and the other containing other trypsin-like serine proteases. In the first subset trypsin and its zymogen, trypsinogen, are subjected to different environmental perturbations such as change of solvent, pH and inhibition by natural and synthetic inhibitors. To demonstrate the methods developed, we discuss the results of some of this subset. The results for the rest of the proteins will be published separately.

Materials And Methods

Sample Preparation. Trypsinogen (TG) and trypsin inhibitor (PTI), both from bovine pancreas, were purchased from Worthington Biochemical Corp. Some trypsinogen (T-1143, bovine pancreas) was purchased from Sigma Chemical Co. for comparison. All samples were used without further purification. Other chemicals used were of reagent grade.

Proteins were prepared at approximate concentrations of 20-60 mg/ml in H₂O based solutions under the same environmental conditions as used in the crystal structure determinations (i.e. pH, solvent, inhibitors). Table I lists crystallization conditions for trypsinogen subset along with Brookhaven Protein Data Bank (pdb) (15) name and atomic resolution of x-ray data.

Spectroscopy Measurements and Data Analysis. Infrared absorption spectra were collected with a Bomem MB-100 FTIR spectrometer equipped with SiC source and DTGS detector. All spectra were recorded at 2 cm⁻¹ resolution by co-adding 1000 scans. Solutions were placed in a cell with CaF₂ windows and 6 μ mylar spacer (Graseby Specac). For each protein measurement, a single beam spectrum of empty cell and of buffer was collected. Buffer spectrum was subtracted from that of protein following procedure of Pérolet et al (16). The instrument was continuously purged with dry air (FTIR Purge Gas Generator from Balston, Inc.) to reduce atmospheric absorption, but nonetheless, before any mathematical treatments, a water vapor absorbance spectrum, collected under the same conditions as the sample, was subtracted. This was done in such a way as to eliminate sharp residual rotational-vibrational peaks of vapor phase water in the 1800-1730 cm⁻¹ region, i.e. region where water absorbs but sample does not.

For determination of protein concentrations UV-VIS spectra were collected on Unicam UV-2 spectrophotometer on the same protein samples as used in FTIR study. Samples were placed in a quartz demountable rectangular cell of 10 μ path length (Starna Cells, Inc.). Absorption was

Table I. Subset of Crystallization Conditions for Trypsinogen and Complexes

<i>pdb name</i> (<i>res Å</i>)	<i>inhibitor</i> (<i>molar ratio</i> ^a)	<i>solvent</i>	<i>pH</i>	<i>complex with</i>
1. 2tga (1.8)	BPTI (1/4)	MgSO ₄	6.9	-
2. 2tgp (1.9)	-	MgSO ₄	6.9	BPTI
3. 3tpi (1.9)	-	MgSO ₄	6.9	BPTI + ile-val
4. 1tgc (1.8)	BPTI (1/4)	50%CH ₃ OH	7.0	-
5. 1tgn (1.65)	benzamidine	30%C ₂ H ₅ OH	7.5	-
6. 1tgb (1.8)	BPTI (1/4)	30% PEG	7.6	-

^aMolar ratio of inhibitor to trypsinogen
(moles inhibitor / moles trypsinogen)

measured at 280 nm and concentration calculated using the literature epsilon values at 280 nm (17).

Protein FTIR absorption data is converted into the units of molar extinction coefficients so that different samples could be compared. The exact pathlength of IR cell is determined by an interference fringe method of empty cell (18). Other methods of normalization of amide I band were attempted such as normalizing to area or to tyrosine band but were found to be inadequate. Second derivative spectra are calculated using Maximum Likelihood derivative algorithm (Spectrum Square Associates, Inc.) with a 7 cm⁻¹ half-width for all proteins.

Computational Methods. All computational methods used for analysis of crystallographic structure data are based on algorithms of Liebman and co-workers (19-22). In short, three-dimensional protein structures are represented by two geometric parameters: α -carbon distances (linear distance values - LD) and α -carbon torsion angles (backbone dihedral angles - BDA). The LD value of an amino acid in a protein sequence, i , is computed by summing up the distances from its α -carbon to each of the four successive residue α -carbons. Therefore, each LD value describes the conformation of five residues, and for a protein of N residues there are $(N-4)$ LD values. A plot of LD values versus sequence position gives a detailed profile of local folding. To compare two forms of the same protein, their respective LD's are subtracted pairwise and a plot will highlight regions that undergo change (20). The BDA is a dihedral angle between two planes, one containing i , $i+1$, $i+2$ and the second containing $i+1$, $i+2$, $i+3$. One BDA value describes the orientation of four residues. This parameter gives representation of handedness of polypeptide chain and is used to distinguish between ambiguity in distance-based analysis alone.

To determine if fragments of polypeptide backbone in different protein structures are equivalent, an algorithm, based on similarity in the LD and BDA values, has been developed (21) and was used here with some modifications. The basis for the algorithm is described in detail by Prestrelski et al. (21) and additional modifications (Klein, G. C., and Liebman, M. N., unpublished) are summarized here. First, we define a substructure as any octapeptide conformation that occurs in more than one protein in the set of proteins studied. Substructures are established by screening the LD and BDA values of all potential substructures in the proteins in the set using cost function. To accomplish this, the proteins are considered sequentially and fragments of a defined length ($N=8$) are compared with all other fragments of the same length. The 'screened' file contains all the substructures and their matches. One resultant table contains substructure number, frequency of its occurrences in the set and corresponding LD and BDA values. A second lists

the order of substructures assigned to each residue as one progresses along the amino acid sequence. A hierarchical ordering of substructures is then generated, containing 5 LD values, to find equivalent regions in the proteins in the set.

All calculations described here were performed on a Sun computer.

Results and Discussion

The trypsin-like subfamily of serine proteases is ideal for the study of conformational changes that take place upon environmental perturbations due to a large number of high resolution crystallographic data available for under a variety of conditions. Figure 1 shows a flow-chart (similar to a minimum spanning tree) of the available x-ray data (pdb name) and the corresponding perturbations. As a demonstration of the analysis method, FTIR data is presented for trypsinogen (2tga), trypsinogen-BPTI complex (2tgp) and trypsinogen in different solvents (1tgb, 1tgc). The substructure library calculation was carried out using all structures shown in Figure 1.

Figure 2 shows difference linear distance plots for (a) 1tgb, (b) 1tgc and (c) 2tgp as each is compared to 2tga. The crystal structure 1tgb is of trypsinogen in 30% polyethylene glycol (24). Polyethylene glycol is quite different from MgSO₄ (the solvent used for 2tga (25) structure) in salt concentration and dielectric constant. Based on the difference LD plot, the main difference between the two structures is in the region between residues 168-174 (positions are referred to based on the ordered residue list in β -trypsin, i.e. 1-223 for 223 residues in β -trypsin (20). The seventh residue in trypsinogen, isoleucine, assumes position number 1.). This region is part of the "activation domain" (26), i.e. regions which are conformationally different in trypsinogen as compared to trypsin. In trypsinogen, the activation domain is disordered (26) but it is disordered in both environments (24). Based on the magnitude of the change in linear distance values (2.2 Å) and comparison with changes upon binding of pancreatic trypsin inhibitor (Figure 2c, see below) it is reasonable to conclude that the changes seen upon a solvent change are conformational in nature and most likely are due to rigidification or ordering of a 'disordered' chain.

The crystal structure trypsinogen 1tgc (27) is in 50% methanol-water mixture. Based on the difference LD plot, the 'largest' difference observed is in the same region as that described above, residues 168-174, but the change is about 4 times smaller. Therefore, it is concluded that there is a very small effect (if any at all) of alcohol on the conformation of trypsinogen.

Bovine pancreatic trypsin inhibitor (PTI) is a small protein of 58 amino acids that binds trypsinogen with the

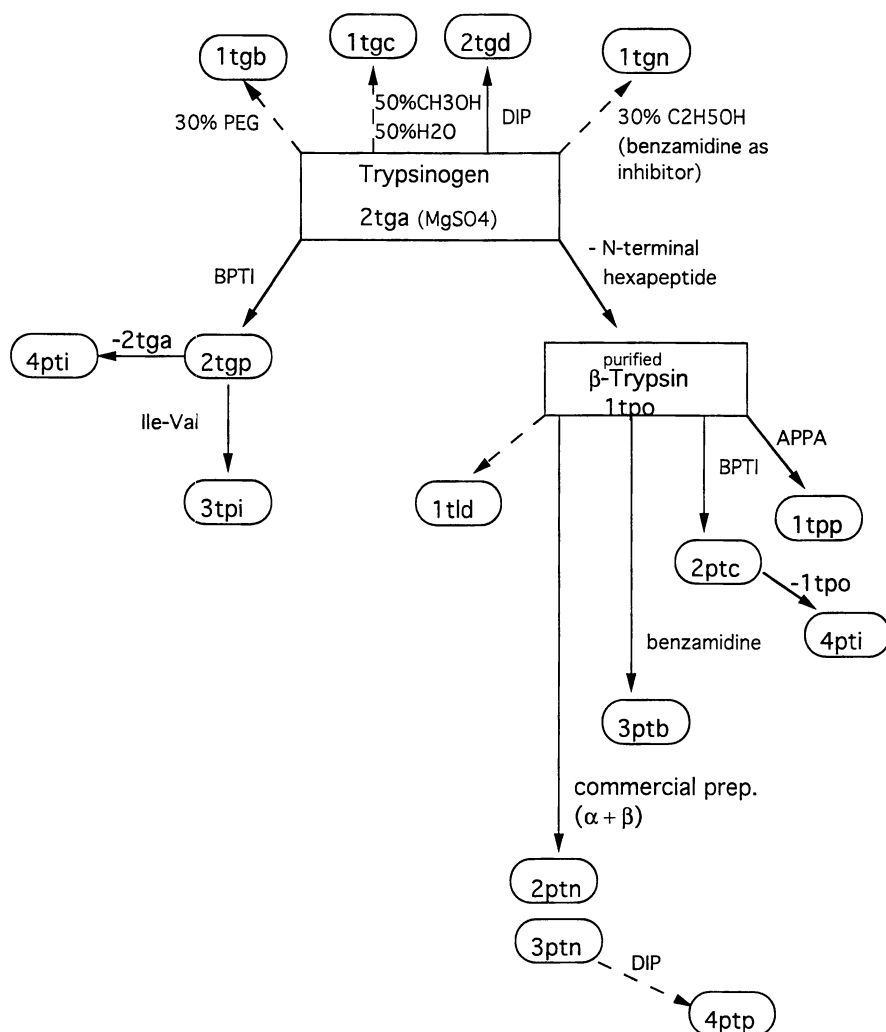


Figure 1. Flow-chart of crystallographic data for trypsin and its zymogen and the corresponding perturbations. Each structure is identified by its Protein Data Bank (pdb) name.

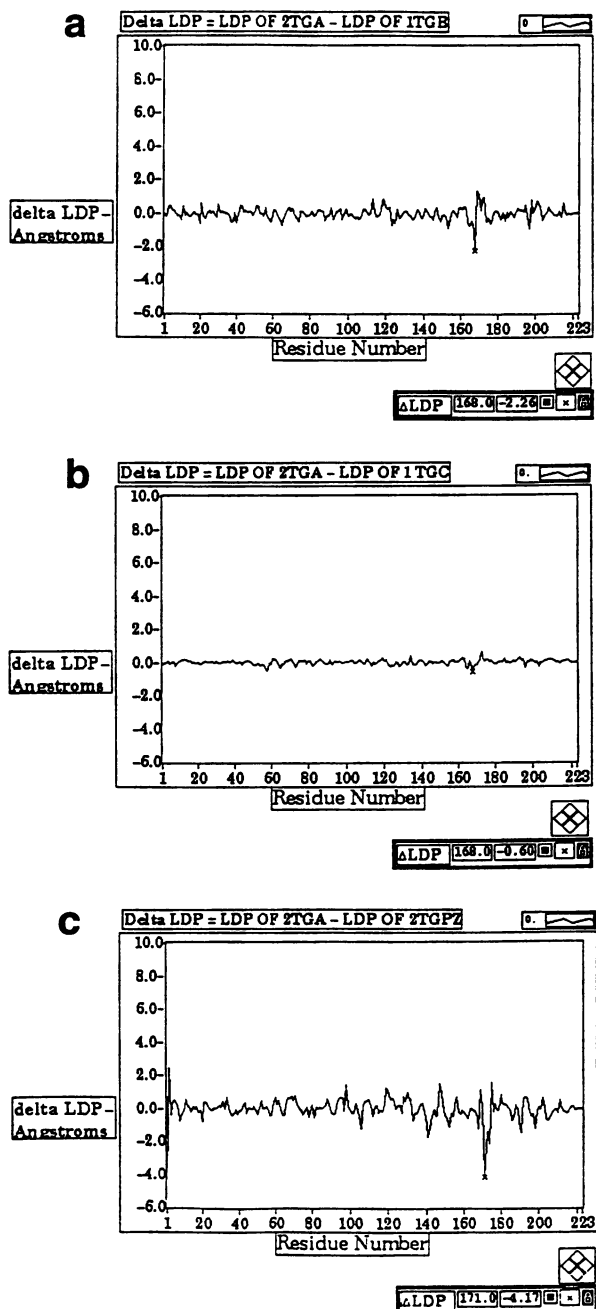


Figure 2. Difference linear distance plots for (a) 1tgb (trypsinogen in 30% polyethylene glycol), (b) 1tgc (trypsinogen in 50% methanol-water mixture) and (c) 2tgp (trypsinogen-bovine pancreatic inhibitor complex) as compared to 2tga (trypsinogen in $MgSO_4$).

binding constant of 10^6 M^{-1} (28). Upon binding of PTI, the activation domain becomes structured (26) and overall the trypsinogen component adopts a more trypsin-like conformation (28). The difference LD plot (Figure 2c) shows the difference between LD values for free trypsinogen and a trypsinogen-PTI complex. Again the change is observed in the same activation domain region, residues 168 to 175, but the sign of the difference plot for residues 170 to 175 is opposite to that seen in 2tga-1tgb plot (Figure 2a). This difference in sign for the two plots probably implies that if there is ordering of the activation region upon a complex formation or solvent change, then these conformational changes that result from the two perturbations are different.

The difference FTIR derivative spectra for trypsinogen in 12 -13% polyethylene glycol (PEG), 20% CH_3OH and a complex with PTI as compared to free trypsinogen are shown in Figures 3, 4 and 5 respectively. All three spectra show a difference in the amide I region ($1600 - 1690 \text{ cm}^{-1}$) and no differences in amide II ($1480 - 1575 \text{ cm}^{-1}$) region are observed. The amide II region results primarily from NH bending and CN stretching. It is not as sensitive to secondary structural changes as amide I and when the changes are small, it is not expected to reveal differences. The largest differences seen in amide I for all three spectra are in the same 'region', i. e. $1633 - 1643 \text{ cm}^{-1}$ (others assign this region as β -sheet (11-14, 29, 30), possibly indicating that the change is within the same substructure and potentially even in the same position of protein. The LD results, as described above, suggest that it is safe to assume that the conformational changes that take place in solution are in fact in the same part of the polypeptide chain in all three cases. Other small difference peaks observed at higher frequencies ($1670 - 1690 \text{ cm}^{-1}$) are due to small changes in loops and turns. It is important to point out that the results presented here are preliminary and several sample preparation techniques require further refinement before further analysis can be made. Also, the spectrum of the complex of trypsinogen with PTI contains contributions from amides of both polypeptides and since it is not determined exactly how much PTI is bound to trypsinogen it is not clear how much of its spectrum to subtract from that of the complex.

The protein substructure library used in this analysis was created for the protein set shown in Figure 1. Analysis of the library indicates that for trypsinogen crystal structures 2tga, 1tgc and 2tgp starting with residue number 166 the octapeptide substructure is of the same type through residue number 173. The library average LD values for this substructure are: 27.4, 29.7, 27.8, and 31.8 \AA which are characteristic of LD values for more 'extended' structures, similar to 2.2₇ helix (LD=29.08) (21). For 1tgb, the octapeptide substructure

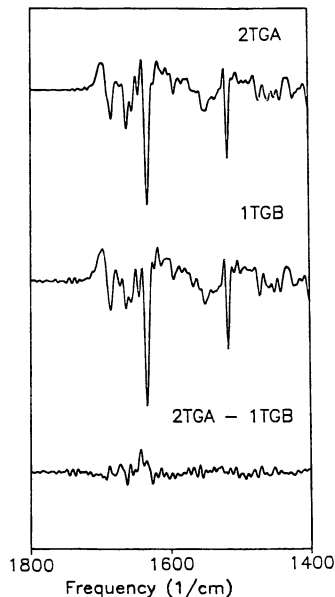


Figure 3. Difference FTIR derivative spectra for trypsinogen in 12-13% polyethylene glycol as compared to trypsinogen in MgSO_4 .

starts at residue 165 and has an average LD values of 22.2, 27.4, 29.8, and 27.9 Å which indicate that the first 5 residues are involved in a turn (probably deformed) and then folds into an 'extended' structure (*I*). At residue 173, the folding pattern becomes the same for 2tga, 1tgb and 1tgc but not for 2tgp. At this point the conformation changes for 2tgp and then changes again at residues 174 and 175. At residue 176 all four structures adopt the same conformation. These results are consistent with the difference LD plots but give more detail and insight on the nature of the exact changes. The hierarchical ordering of substructures indicates that for all other crystal structures in Figure 1 except 1tgn, 2tgd and 4pti the region starting with residue 166 can be defined by the same substructure (as described above this region is part of the 'activation domain').

Conclusion

In this paper, we describe a method for analyzing structural changes in proteins using a combination of experimental, i.e. FTIR spectroscopy and computational techniques. The method involves the comparison of proteins in different states using second derivative analysis of

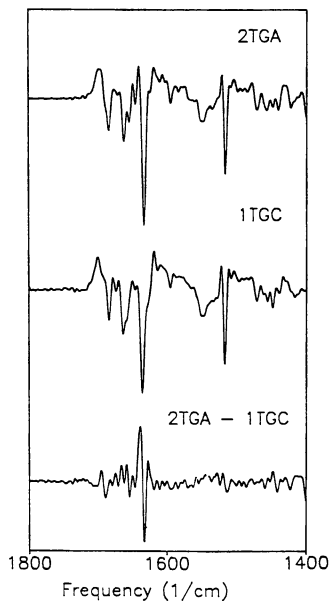


Figure 4. Difference FTIR derivative spectra for trypsinogen in 20% methanol as compared to trypsinogen in MgSO_4 .

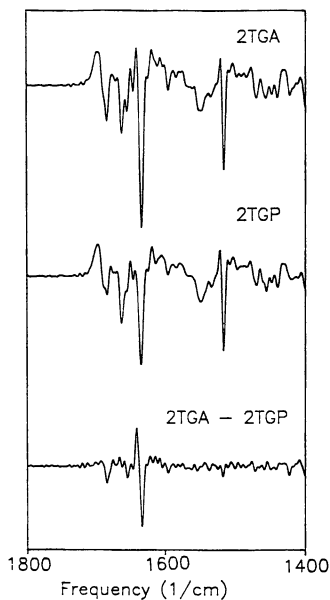


Figure 5. Difference FTIR derivative spectra for trypsinogen-bovine pancreatic trypsin inhibitor complex as compared to trypsinogen in MgSO_4 .

FTIR spectra and difference LD values computed from structures observed by x-ray crystallography. Collection and analysis of data for a family of proteins, e.g. serine proteases, enables the assignment of observed spectral bands to protein substructure-based features and further enables analysis of spectra where no x-ray structure is available. It is important to note that although we report the initial results of this study it is apparent that the method we describe does not rely on bandfitting the spectra or on developing frequency-based band assignments. However, as the methods for data analysis from FTIR spectra continue to evolve (see other chapters in this volume) FTIR spectroscopy promises to be a powerful method for evaluating protein structure and monitoring protein conformational response in solution.

Acknowledgments

We wish to thank Mr. Gary Klein for his help in developing programs to create protein substructure library.

Literature Cited

1. Elliott, A., and Ambrose, E. J. *Nature*, **1950**, *165*, 921
2. Susi, H. *Methods Enzymol.* **1973**, *26*, 455
3. Timasheff, S. N., Susi, H., and Stevens, L. *J. Biol. Chem.* **1967**, *242*, 5467
4. Susi, H., Timasheff, S.N., and Stevens, L. *J. Biol. Chem.* **1967**, *242*, 5460
5. Gratzer, W.B., Bailey, E., and Beaven G. H. *Biochem. Biophys. Res. Commun.* **1967**, *28*, 914
6. Rüegg, M., Metzger, V., and Susi, H. *Biopolymers* **1975**, *14*, 1465
7. Kauppinen, J. K., Moffatt, D. J., Mantsch, H. H., and Cameron, D. G. *Anal. Chem.* **1981**, *53*, 1454
8. Kauppinen, J. K., Moffatt, D. J., Mantsch, H. H., and Cameron, D. G. *Appl. Spec.* **1981**, *35*, 271
9. Therrein, M., Lafleur, M., and Pézolet, M. *Proc. SPIE Int. Soc. Opt. Eng.* **1985**, *553*, 173
10. Englander, S. W., and Kallenbach, N. R. *Q. Rev. Biophys.* **1984**, *16*, 521
11. Susi, H., and Byler, D. M. *Methods Enzymol.* **1986**, *130*, 290
12. Surewicz, W. K., and Mantsch, H. H. *Biochim. et Biophys. Acta* **1988**, *952*, 115
13. Surewicz, W. K., Mantsch, H. H., and Chapman, D. *Biochemistry* **1993**, *32*, 389
14. Arrondo, J. L. R., Muga, A., Castresana, J., and Goñi, F. M. In *Prog. Biophys. Molec. Biol.* **1993**, *59*, 23
15. Bernstein, F. C., Koetzle, T. F., Williams, G.J.B., Meyer, Jr., E. F., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T., and Tasumi, M. *J. Mol. Biol.* **1977**, *112*, 535

16. Dousseau, F., Therrien, M., and Pézolet, M. *Appl. Spec.* **1989**, *43*, 538
17. *Practical Handbook of Biochemistry and Molecular Biology*; Fasman, G. D., Ed.; CRC Press, Inc.: Boca Raton, FL., 1989
18. Ingle, Jr., J. D., and Crouch, S. R. *Spectrochemical Analysis*; Prentice Hall, Inc.: Englewood Cliffs, NJ, 1988; p.418
19. Liebman, M. N., Venanzi, C. A., and Weinstein, H. *Biopolymers* **1985**, *24*, 1724
20. Liebman, M. N. *Enzyme* **1986**, *36*, 115
21. Prestrelski, S. J., Williams, Jr., A. L., and Liebman, M. N. *Proteins: Structure, Function, and Genetics.* **1992**, *430*
22. Prestrelski, S. J., Byler, D. M., and Liebman, M. N. *Proteins: Structure, Function, and Genetics.* **1992**, *440*
23. Prestrelski, S. J. *Ph.D. Thesis* **1990**, The City University of New York
24. Bode, W., and Huber, R. *FEBS Letters* **1978**, *90*, 265
25. Bode, W., Fehlhammer, H., and Huber, R. *J. Mol. Biol.* **1976**, *106*, 325
26. Huber, R., and Bode, W. *Accts. Chem. Res.* **1978**, *11*, 114
27. Singh, T. P., Bode W., and Huber, R. *Acta Cryst.* **1980**, *B36*, 621
28. Bode, W., Schwager, P., and Huber, R. *J. Mol. Biol.* **1978**, *118*, 99
29. Kumosinski, T. F., and Unruh, J. J. *present publication*
30. Dong, A., Huang, P., and Caughey, W. S. *Biochemistry* **1990**, *29*, 3303

RECEIVED June 9, 1994

Chapter 16

Spectroscopy and Molecular Modeling of Electrochemically Active Films of Myoglobin and Didodecyldimethylammonium Bromide

James F. Rusling¹, Alaa-Eldin F. Nassar¹, and Thomas F. Kumosinski²

¹Department of Chemistry, P.O. Box U-60, University of Connecticut,
Storrs, CT 06269-3060

²Eastern Regional Research Center, Agricultural Research Service,
U.S. Department of Agriculture, 600 East Mermaid Lane,
Philadelphia, PA 19118

Water-insoluble coatings of didodecyldimethylammonium bromide (DDAB) on solid supports incorporate the protein myoglobin from solutions at pH 5.5-7.5 to form stable Mb-DDAB films. We previously found that electron transfer involving the heme Fe(III)/Fe(II) redox couple in 20 μm thick Mb-DDAB films on electrodes was 1000-fold faster than for Mb in aqueous solutions. The present work examines the supramolecular structure of Mb-DDAB films by reflectance-absorbance infrared, visible linear dichroism, and electron spin resonance spectroscopies. Molecular dynamics of Mb-DDAB models provided information on hydrophobic and coulombic interactions between Mb and DDAB. When combined with earlier thermal and electron transfer studies, results suggest that Mb-DDAB films feature lamellar liquid crystal DDAB arranged in bilayers with tilted hydrocarbon tails as in biological membranes. Mb in DDAB films has a secondary structure close to its native state, attains a preferred orientation in the films, and has Fe(III)heme in the high spin state. Mb electron transfer may be enhanced by adsorbed surfactant on the electrode which inhibits macromolecular impurities from adsorbing and blocking interfacial charge transfer.

Biomembranes in living organisms are typically about half protein and half lipid. Many proteins and enzymes fulfill their biological functions as integral parts of membranes. The lipids are arranged in bilayers. Proteins can reside on the surface of, imbedded partly within, or extending across

these bilayers (1). In this paper we discuss films of the protein myoglobin (Mb) and a surfactant which forms lamellar bilayer structures, didodecyl-dimethylammonium bromide (DDAB) (2). Myoglobin is a small oxygen carrying muscle protein with MW ca. 17,000. It contains a ferriheme prosthetic group that can be reduced electrochemically (3).

Mb inserts spontaneously and rapidly from solution into cast DDAB films to form films stable for a month or more in buffers containing 50 mM NaBr (2). These films are in a lamellar liquid crystal phase at room temperature. When prepared on electrodes, the Fe(III) heme in Mb-DDAB films in the liquid crystal phase can be converted to Fe(II) heme at rates 1000-fold larger than for Mb in solution (2). Such films feature stacks of surfactant bilayers similar to lipid bilayers, and might provide useful experimental models for membrane-bound proteins.

Our specific interest in Mb-DDAB derives from the desire to develop catalytic films to reductively dehalogenate organohalide pollutants at electrodes. In previous work, catalytic films were prepared from water insoluble cationic surfactants and negatively charged metal macrocyclic complexes such as phthalocyanine tetrasulfonates or corrin hexacarboxylates (4). DDAB is particularly suited for such applications since its films are in the liquid crystal phase at room temperature, facilitating efficient mass and charge transport necessary for catalytic applications.

Mb reduces organohalides in its Fe(II) form in solution (5) and in DDAB films (2). Thus, Mb-DDAB films might serve as the basis for practical systems to dehalogenate or detect organohalide pollutants. They also might provide a model for reductive dehalogenation of pollutants in mammalian livers (6a) and anaerobic bacteria (6b). Kunitake et al. incorporated Mb into multibilayer surfactant films (7) prior to our work. The protein was reported to achieve a specific orientation in cast films of a double chain surfactant with an anionic phosphate head group (7).

Possible explanations for the remarkable increase of electron transfer rate for Mb in DDAB films include (2) (i) the influence of strongly adsorbed surfactant in protecting the electrode from adsorption of macromolecules which can block electron transfer, and (ii) preferential orientation of Mb in a way favorable for electron transfer. Such orientation effects have been claimed for electrode surfaces chemically tailored to promote fast electron transfer to redox proteins (8). However, we are aware of very little direct structural evidence to support such views.

In this paper, we describe structural studies on Mb-DDAB films combining ESR, UV-VIS linear dichroism, and infrared reflectance-absorbance spectroscopy with molecular modeling and dynamics. The specific aim of this work is to characterize the supramolecular structure of the films. Such studies also provide insight into the influence of film structure on electron transfer kinetics of Mb in the film.

Experimental Section

Materials. Lyophilized myoglobin (Mb) from horse skeletal muscle was from Sigma. Buffered myoglobin solutions were filtered through a YM30 filter (Amicon, 30,000 cutoff) to remove high molecular weight impurities (2). *Tris*-hydroxymethylaminomethane•HCl was used for pH 7.5 and acetate for pH 5.45 buffers, respectively. Buffers were 0.01 M in the conjugate base and contained 50 mM NaBr.

Didodecyldimethylammonium bromide (DDAB) (>99%) was from Eastman Kodak. Water was purified with a Barnstead Nanopure system to a specific resistance >15 Megohm-cm. All other chemicals were ACS reagent grade.

Apparatus and procedures. Cast DDAB films were prepared on solid substrates appropriate for each type of experiment. Briefly, a volume of 0.1-0.01 M DDAB in chloroform measured to give film thicknesses of 2-40 μm , as necessary for a given experiment, was placed on the solid substrate with a microsyringe and spread evenly. Chloroform was evaporated gradually overnight (2). The cast films were then equilibrated in buffer solutions of 0.1-0.5 mM Mb for several hours. Uptake of Mb into the films monitored by cyclic voltammetry (CV) showed that steady state concentrations of Mb are achieved in <30 min. Steady state concentrations of Mb in the films were estimated at 0.35-0.45 mM by integrating under slow scan voltammetric curves. Mb-DDAB films prepared on carbon electrodes and transferred to buffers containing 50 mM NaBr but no Mb gave stable CV signals for MbFe(III) reduction for a month or more (2).

Electronic absorption spectra were obtained by using a Perkin-Elmer $\lambda 6$ UV-VIS spectrophotometer. Films were prepared on quartz slides and soaked in 0.2 mM Mb in buffer solutions containing 50 mM NaBr, then dried in air before obtaining spectra. Matched dichroic sheet polarizers (Melles Griot) were used for UV-VIS linear dichroism. Spectra were obtained with the incident light beam normal to the film plane.

ESR experiments were done with a Varian E-g spectrometer using field modulation at 100 kHz, with modulation amplifier set at 20 G. Temperature was controlled at 100 K. Films for ESR were prepared on a 5 cm. long flattened face of a 2 mm radius quartz rod which could be oriented with respect to the magnetic field. Some films were also prepared by mixing 10 mM DDAB in water with 0.1 mM Mb and depositing on the quartz face. Spectra were similar for films prepared by both methods.

Infrared spectra were obtained by using a Mattson Galaxy 6020 FT-IR spectrometer with a liquid nitrogen cooled MCT detector at 4 cm^{-1} resolution. Reflectance Absorbance Infrared Spectroscopy (RAIR) was done by using a SPECAC variable incident angle accessory with a wire grid polarizer of KRS-5 to obtain p-polarized light, as described previously (9). Films were cast onto glass slides coated with aluminum by vapor deposition. Spectra of these aluminum films were subtracted as background.

Computations. Molecular modeling and dynamics were done on an Evans and Sutherland (St. Louis, Mo.) PS-390 interactive computer driven by Sibyl Molecular Modeling software on a Silicon Graphics (Mt. View, CA) W-4D35 Processor. All computations were done by using a Tripos (10a) force field at 300 K. Only essential Mb hydrogens were included, and Mb atomic charges were those reported by Coleman et al (10b). Charges on DDAB were twice those obtained from a Gasteiger-Huckel calculation. Non-bonded interactions were cut off at 5 Å. Initial testing showed that a van der Waals attraction 73 times the usual value for DDAB hydrocarbon tails successfully emulates the hydrophobic effect (10c) and holds together DDAB bilayers. This large van der Waals attraction was used for the DDAB hydrocarbon tails, only, for all computations.

Results

Reflectance-Absorbance FT-IR Spectroscopy (RAIR). This technique provided information about orientation of DDAB (9), and about the secondary structure of Mb in the films. Previous analysis of CH₂ stretching bands showed that the surfactant hydrocarbon tails in Mb-DDAB films tilt at average angles of 30° to the normal to the film plane (2).

The shapes of the amide I and amide II infrared absorbance bands of proteins are sensitive to the secondary structure of the polypeptide chain (11,12). These bands are seen in the 1500-1700 cm⁻¹ region in the RAIR spectrum of a film prepared from Mb only, with no surfactant (Figure 1a). The spectrum is similar, although the bands are broader, to spectra of Mb in water. The peaks shown under the main bands are the results of a deconvolution/regression analysis, which can provide detailed information on secondary protein structure. Briefly, a Fourier deconvolution is done on the raw data and both the original and deconvoluted spectra are fit by nonlinear regression analysis to a common model of *n*-peaks representing the sum of all α-helix, β-sheet, turn and other structural components of the polypeptide backbone that overlap to produce the observed amide I and II bands. Regression parameters are the heights, widths and positions of each component peak. Initial choice of *n* and the parameter values is guided by the second derivative of the spectrum. The final *n* is obtained as the value that gives the best fit among a series of alternatives. Convergence of this analysis is reached when regression results on the original and deconvoluted spectra agree within a preconceived tolerance (12).

Films of Mb-DDAB had RAIR spectra in the amide region (Figure 1b) similar to those of films of Mb (Figure 1a). Focussing on the amide I region, the first 10 corresponding resolved peaks have frequencies within ±2 cm⁻¹ and similar areas for the two films (Table I). An additional peak was found for Mb-DDAB. General similarities in Figures 1a and 1b suggest that Mb retains the essential features of its native secondary structure in DDAB films. Similar results were obtained at both pH values. Denatured

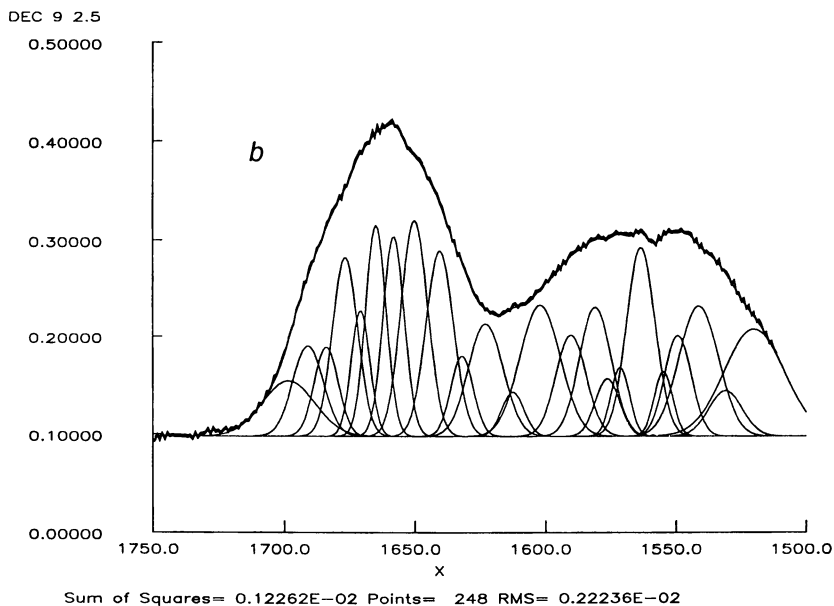
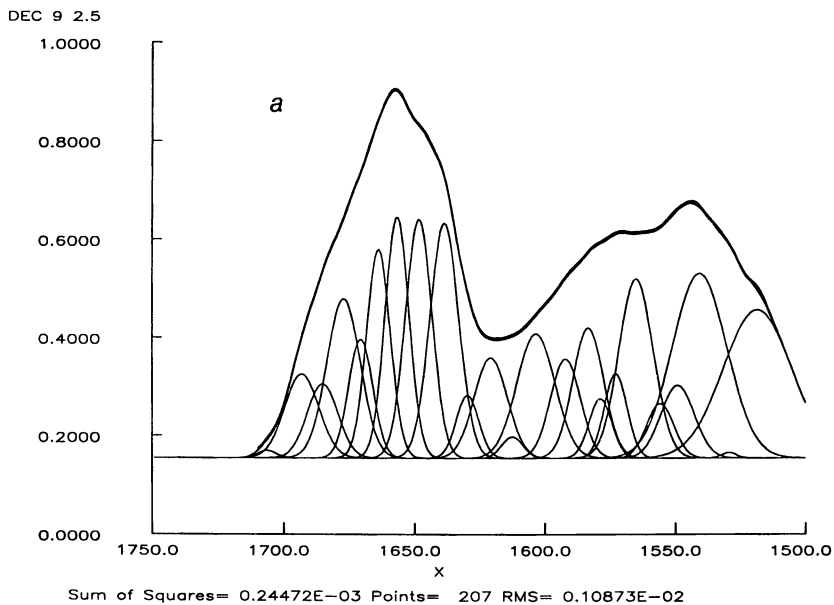


Figure 1. Reflectance-absorbance infrared (RAIR) spectra at 60° incidence of Mb films prepared by using pH 5.5 acetate buffer containing 50 mM NaBr: (a) Mb only; (b) Mb-DDAB. Outer envelopes show agreement of experimental and nonlinear regression results. Underlying peaks are those found by best fit deconvolution/regression (see text).

Table I. Analysis of amide I bands in Mb films^a by deconvolution/regression

Mb film			Mb-DDAB film	
peak no.	freq., cm ⁻¹	rel. area, %	freq., cm ⁻¹	rel. area, %
2	1693	7.4	1691	7.4
3	1685	5.9	1684	5.7
4	1677	13.7	1677	12.1
5	1671	7.5	1671	6.3
6	1664	12.8	1665	11.8
7	1657	14.8	1658	11.6
8	1648	16.5	1650	15.6
9	1639	17.8	1640	14.0
10	1630	3.7	1631	4.9
11			1623	10.6

^aFilms prepared using pH 5.5 acetate buffer containing 50 mM NaBr.

Mb showed very different spectral shapes and severely decreased absorbances in amide I and II regions.

Second derivative RAIR spectra of Mb and Mb-DDAB films were also similar (Figure 2), but not identical. Spectra of Mb films prepared with and without 50 mM NaBr in solution and a Mb-DDAB film prepared in pH 5.5 acetate buffer with 50 mM NaBr show nearly all second derivative peaks at the same frequencies. However, several of these peaks have different shapes and/or relative intensities in the various samples. Differences in the spectra between Mb and Mb-NaBr suggest that salt may have some effect on structure or orientation of Mb in films.

RAIR using a polarized source is subject to surface selection rules that depend on the orientation of transition dipole moments in the sample (9). Thus, small differences in the RAIR spectra of Mb and Mb-DDAB films can be caused by differences in the secondary structure of Mb or by differences in orientation of Mb in these two films. These spectral nuances are being explored further by angle-resolved polarized RAIR and regression/deconvolution analysis.

Electron Spin Resonance. The Fe(III) in Mb is bound within the four-coordinate planar heme ring, and has two axial ligand sites. One of these axial sites is occupied by a histidine contributed from the protein, and the other by a labile ligand that controls the spin state of the complex. The Fe(III)heme in Mb has an unpaired electron, and the spin state can be determined experimentally from electron spin resonance spectra (ESR) (13). Also, by changing the angle of the film plane to the magnetic field, the orientation of Mb in the films can be investigated (7).

ESR spectra of Mb-DDAB show major peaks close to 1000 G (Figure 3) and smaller peaks at 3200 G at pH 5.5 and 7.5. The peak at 1000 G is the $g_{\perp} = 6$ peak and the 3200 G peak corresponds to the $g_{\parallel} = 2$ peak. These peaks

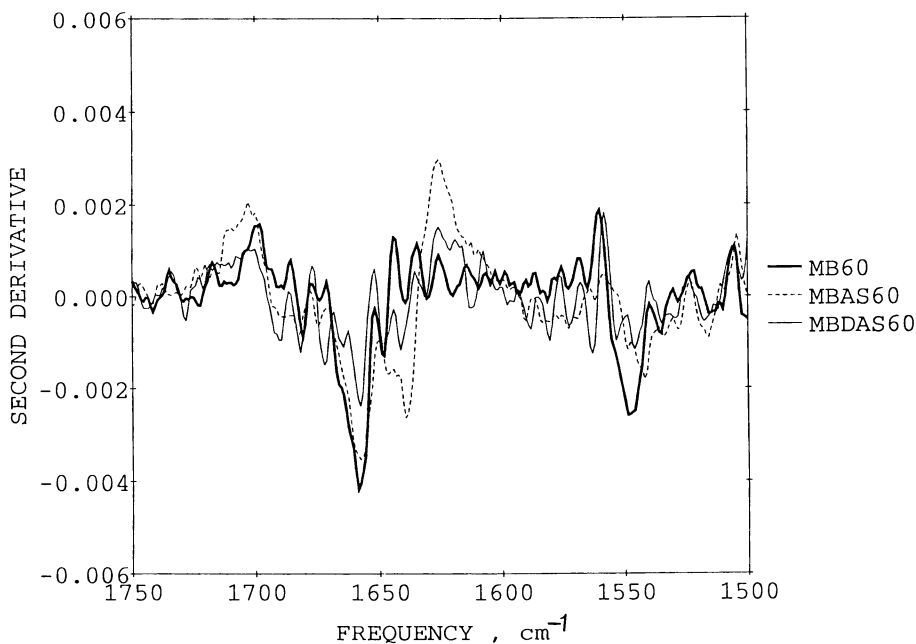


Figure 2. Second derivative RAIR spectra at 60° incidence of Mb films prepared by using pH 5.5 acetate buffer: (MB60) Mb cast from the buffer; (MBAS60) Mb cast from the buffer containing 50 mM NaBr; (MBDAS60) Mb-DDAB prepared with buffer containing 50 mM NaBr.

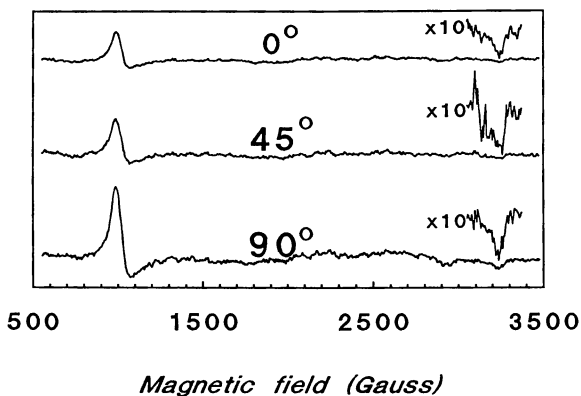


Figure 3. ESR spectra of Mb-DDAB film prepared at pH 5.5 showing changes in spectra at different angles of the film plane with respect to the magnetic field.

are characteristic of the high spin state of Fe(III)heme in myoglobin (13). This spin state features a weak axial ligand on the Fe(III).

A dependence of relative peak size on the angle between the magnetic field and the normal to the film plane is observed (Figure 3). Mb orientation in anionic surfactant films has been demonstrated previously by measuring a changing ratio at 4 K between the small peak at 3200 G and the main 1000 G peak (7). Our spectra show similar trends, with ratios g_{\perp}/g_{\parallel} in the order $90^{\circ} > 0^{\circ} > 45^{\circ}$. However, in the Mb-DDAB spectra, the signal to noise ratio at 100 K is not sufficient for quantitative measurement of the smaller peak. Although the data suggest orientation of Mb in the DDAB films, reliable estimates of orientation angles cannot be made.

Electronic Spectra and Linear Dichroism. Absorbance of visible light by the ferriheme in Mb is responsible for a strong band around 410 nm, called the Soret band. A smaller band from the aromatic peptide side chains is found at 280 nm (Figure 4). The complexing strength of the labile axial ligand governs the position of the Soret band (14).

Spectra of Mb in pH 5.5 and 7.5 buffers had the Soret band at 409 nm. Mb denatured with urea gave a band at 400 nm in buffer solutions, and at 405 nm in a DDAB film. Films cast from Mb alone gave Soret bands at 410 nm at pH 5.5 and 414 nm at pH 7.5, respectively. Values of λ_{\max} in the Mb-DDAB film were shifted to 413 and 415 nm at pH 5.5 and 7.5, respectively (2). These values are closer to the 411 nm for crystalline high spin aquoMb than to those of 426 nm for CNMb crystals and 422 nm for azideMb crystals, both of which have strong axial ligands and are low spin (14). Thus, in agreement with IR and ESR data, electronic absorption spectra of Mb-DDAB films suggests that Mb is in its native, high spin form.

Spectra had different intensities when obtained with parallel or perpendicularly polarized light (Figure 4). This observed linear dichroism was used to find an average order parameter $S = (1 - 3 \cos^2 \phi) / 2$ by using a simplified expression (15a):

$$\Delta A / A = 3 S \quad (1)$$

where $\Delta A = A_{\parallel} - A_{\perp}$ is the difference between the peak absorbances with parallel and perpendicularly polarized light, $A = (A_{\parallel} + 2A_{\perp}) / 3$ is the absorbance of a randomly organized sample, and ϕ is the angle between the transition moment for the absorption and the normal to the film plane. Thus, the order parameter is found from:

$$S = (A_{\parallel} - A_{\perp}) / (A_{\parallel} + 2A_{\perp}) \quad (2)$$

Dichroic ratios $A_{\parallel} / A_{\perp}$ for the Soret band as well as order parameters were significantly larger for Mb-DDAB films than for Mb films (Table II). These results suggest preferred average orientation for Mb in the DDAB films in agreement with conclusions from ESR. Since the transition

Table II. Linear Dichroic Ratios and Order Parameters from Soret Bands

pH ^a	Mb films			Mb-DDAB films	
	sample no.	$A_{ }/A_{\perp}$	S^b	$A_{ }/A_{\perp}$	S^b
7.5	1	1.32	0.097	1.51	0.145
	2	1.16	0.051	1.59	0.165
	3	1.16	0.050	1.53	0.151
	4	1.27	0.085	1.75	0.200
	5			1.80	0.211
7.5	(avg± s.d.)	1.23±0.08	0.071±0.024	1.64±0.13	0.17±0.03 ^c
5.5	1	1.11	0.036	1.37	0.110
	2	1.14	0.045	1.34	0.103
	3	1.13	0.043	1.39	0.116
	4	1.13	0.040	1.40	0.118
	5	1.19	0.060	1.59	0.166
	6			1.60	0.168
5.5	(avg± s.d.)	1.14±0.03	0.045±0.009	1.45±0.12	0.13±0.03 ^c

^aBuffers contained 50 mM NaBr. ^bOrder parameter from eq 2. ^cAverage values of ϕ for Mb-DDAB were $62\pm 2^\circ$ at pH 7.5 and $61\pm 2^\circ$ at pH 5.5, respectively. These values are not corrected for optical errors which could lead to bias on the order of 10% (15b).

moment of heme is in the plane of the molecule (14), ϕ represents the angle between the heme plane and the normal to the film plane. Values of ϕ in Table II (footnote c) are only approximate. More sophisticated experiments which correct for certain optical errors (15b) are underway.

Molecular Modeling and Dynamics. The aim of this part of the work is to assess the theoretical viability of supramolecular models for Mb-DDAB films. Molecular models for bilayer films of DDAB were stable in the classic bilayer form with hydrocarbon tails together and charged head groups facing outward. The major fluctuations during dynamics were formation and disappearance of kinks and bends in the hydrocarbon tails. This is reminiscent of the mechanism for fluidity of the liquid crystal phase of bilayer membranes (1).

Models of Mb alone featured many charge-paired partners of amino acid side chains on the globular protein surface. These charge pairs are reported (16) to impart stability to the secondary structure of Mb in water between pH 5 and 9. The globular structure of Mb also contains about 75% helix (17).

A model of Mb-DDAB films was constructed from a bilayer of 48 DDABs surrounding a molecule of Mb. The model of Mb with unprotonated histidine residues (Color Plate 13) is roughly equivalent to its solution structure at pH 7.5. Side views of this model show that Mb extends about 10 Å beyond the head group planes on both sides of the DDAB bilayer. When this model was allowed to undergo dynamics for 40

NOTE: The color plates can be found in a color section in the center of this volume.

ps, its essential stability was evident. Mb stayed within the cavity in the DDAB bilayer and only a sort of breathing motion was observed around an average Mb-DDAB structure. Equilibrium was achieved within 40 ps, as shown by a time invariant radius of gyration.

The charge paired amino acid partners of Mb can also be seen as pairs of red (negative) and purple (positive) amino acid residues close together on the Mb surface (Color Plate 13). Green hydrophobic residues remain within the interior, as with native Mb in water. This indicates that no large differences in secondary structure exist between Mb in water and Mb in the DDAB bilayer model.

Similar results were found when the Mb histidine residues were protonated (Color Plate 14), representing the structure at pH 5.5. Again, many charge pairs were found on the Mb surface. This Mb-DDAB model was also stable during 40 ps dynamics runs.

Since it is difficult to assess interactions of individual DDAB head groups with surface amino acid residues on Mb from the above models, a simpler model was constructed for this purpose. Two bilayer structures of four DDAB molecules each were docked with one set of head groups close to carboxylate side chains on the Mb surface (Figure 5a). During 40 ps dynamics computations, the surfactant head groups closest to Mb moved slightly away from the docking position, and the hydrocarbon chains of the DDABs waved about somewhat (Figure 5b). Furthermore, one of the bromide counterions moved well away from the protein surface. This suggests a weak electrostatic interaction between the head groups and the protein surface.

A second set of similar models were constructed, but with the two DDAB mini-bilayers arranged with both sets of head groups and hydrocarbon tails close to the Mb surface (Figure 6a). This structure was also unstable. One set of hydrocarbon tails and one of the bromide ions moved away from the protein surface in 40 ps of dynamics (Figure 6b), although the DDAB clusters remained together.

Discussion

Film Structure and Stability. We first review previous findings on Mb-DDAB films (2), then integrate these with the results in this paper. Gel-to-liquid crystal phase transition temperatures (T_C) of Mb-DDAB films were 12 °C at pH 5.45 and 15 °C at pH 7.5. T_C was 15 °C for bilayer vesicles of DDAB, and 11 °C for pure lamellar DDAB films. These T_C values similar to those of known DDAB bilayer systems suggest that DDAB in the films has a lamellar bilayer structure when Mb is present, and confirm that the films are liquid crystalline at ambient temperature. Our film thicknesses in the μm range would represent thousands of bilayers in an ideal stacked lamellar film (4a).

Scanning electron microscopy of freeze fractured cross sections of DDAB films suggest that these layered structures are somewhat wavy (4c), and contain significant defects. A wavy multilamellar structure has also

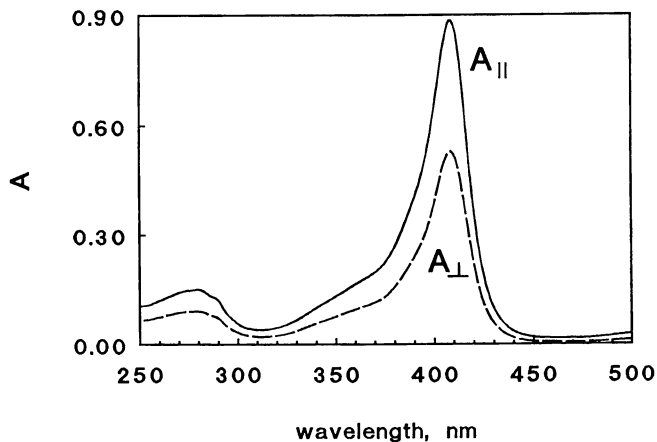


Figure 4. UV-VIS spectra of Mb-DDAB film prepared by using pH 5.5 buffers containing 50 mM NaBr showing the influence of \parallel and \perp plane polarized light.

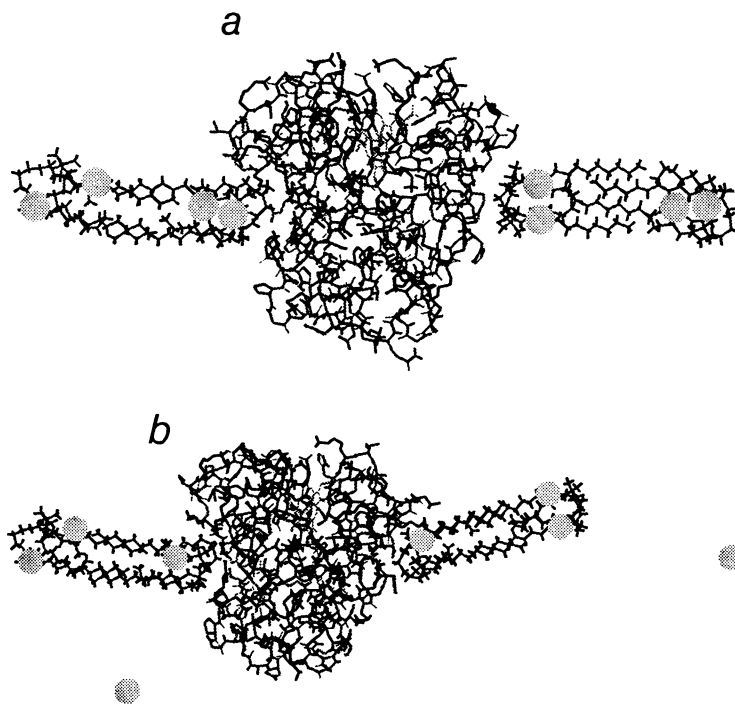


Figure 5. Model with two aggregates of 4 DDABs each docked with one set of head groups at the Mb surface: (a) initial energy minimized structure; (b) structure after 40 ps dynamics. Spheres are bromide ions.

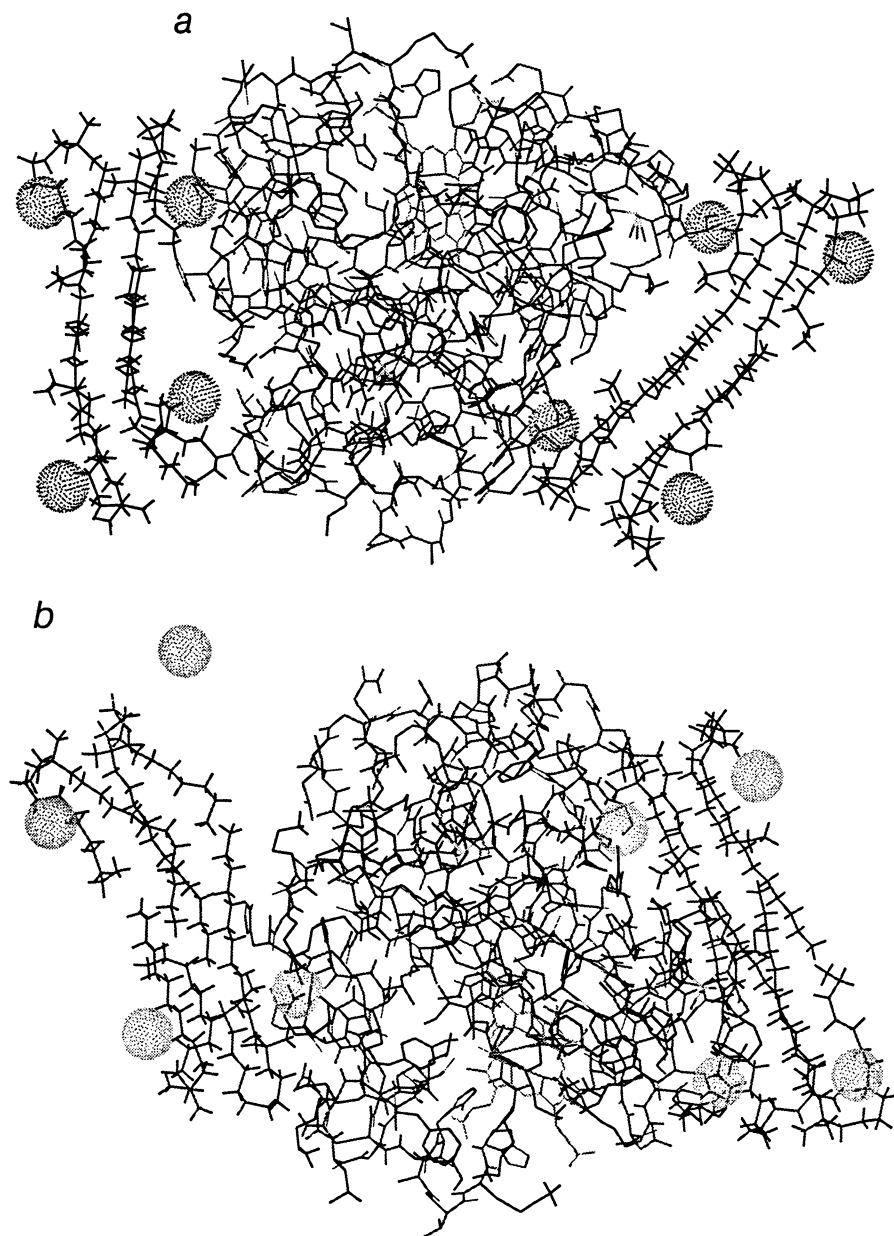


Figure 6. Model with two aggregates of 4 DDABs each docked with both sets of head groups at the Mb surface: (a) initial energy minimized structure; (b) structure after 40 ps dynamics. Spheres are bromide ions.

been proposed for films of a double chain phosphate surfactant containing Mb (7).

Mb-DDAB films are considerably more stable and retentive of the introduced moiety than DDAB films containing ferrocyanide ions (4a) and a cobalt(III) corrinhexacarboxylate (4e). Up to 70% of these electroactive anions are leached from DDAB films in aqueous 0.1 M KBr in several hours, as monitored by cyclic voltammetry (CV). About 10-20% of the CV signal is lost over several weeks when the same experiment is done for metal phthalocyaninetetrasulfonate-DDAB films (4a,b). Mb-DDAB films are the most stable, with a loss of only 10-20% of the CV signal over one month (2).

The metal complex multi-anions discussed above can clearly enter the DDAB films by an ion exchange mechanism. The complex multi-anions replace Br⁻ in the films. The driving force for partition of Mb into DDAB films is less apparent. Mb has a +6 surface charge at pH 5.5 and a +1 charge at pH 7.5 (16), which would not seem conducive to coulombic interactions with the cationic DDAB head groups. We showed previously that buffer solutions alone do not replace large amounts of Br⁻ in the films in the absence of protein. On the other hand, DDAB films treated with Mb in the same buffers lost 65-85% of their original Br⁻ ions (2). This cannot be explained by interactions of anionic amino acid residues of Mb with DDAB head groups. Based on amounts of DDAB and Mb in the films, less than 5% of the head group charge can be compensated by Mb carboxyl groups. Furthermore, such interactions would require disruption of stabilizing charge pair interactions of amino acid residues on the Mb surface.

The present spectroscopic results extend understanding of the Mb-DDAB film structure. Spectra were similar at pH 5.5 and pH 7.5. RAIR (2) showed that the hydrocarbon tails in Mb-DDAB films are tilted to the normal to the film plane, consistent with residing in a bilayer (1). IR, ESR, and VIS spectra confirmed that Mb was incorporated into the films without denaturation.

At present, small differences in RAIR spectra for Mb and Mb-DDAB films (Figures 1 and 2) are interpreted with difficulty. They may reflect either different orientations (9) or slightly different secondary structures (12) of Mb in the different films. Tools to distinguish between these two possibilities are not yet fully developed. However, linear dichroism showed that Mb is more ordered in Mb-DDAB films than in films of Mb alone. Thus, it is possible that IR differences between Mb and Mb-DDAB may be caused only by differences in Mb orientation.

ESR and Soret absorption bands clearly show that the Fe(III)heme in myoglobin is in the high spin state. This was also found for Mb in films of double chain phosphate surfactants (7). Both linear dichroism (Table II) and ESR peak ratio differences at different film-to-magnetic field angles (Figure 3) suggest that Mb is preferentially oriented in the DDAB films.

A more sophisticated theory and experimental protocol for linear dichroism is being explored to accurately estimate Mb orientation (Cf. Table II, footnote c). This theory predicts that no dichroism should be

obtained for a uniaxial sample when the incidence angle of source light is 90° to the film plane. A uniaxial sample is one that is uniform in the x - y plane, with the only structural differences in the z or normal direction. Observance of linear dichroism with normal source incidence suggests that the Mb-DDAB films are not fully uniaxial. This is in accord with the proposal of a wavy lamellar structure and the presence of significant defects in the films.

Molecular dynamics of Mb-DDAB models revealed the possibility of weak coulombic interactions of the Mb surface with DDAB head groups. Horse Mb has 20 carboxylate residues (18,19) which could possibly interact with cationic head groups in the film. Interaction of some of these residues with DDAB head groups might replace the charge pairs with positive amino acid residues mentioned previously. However, this was not indicated as a strong possibility in the dynamics computations (Figures 5 and 6). Since Mb is known to bind ions and neutral buffer species on its surface (18), it is possible that Mb may bring bound buffer components from the solution along with it into the film. When ionized, these buffer anions may be shared with or donated to the head groups, explaining the observed loss of Br^- from the films. Because of the large computation times involved, we have not yet considered buffer components bound to Mb in molecular dynamics. However, recent dynamics studies by Klein and coworkers suggest that counterions can be shared between head groups in monolayer surfactant structures (20).

The possibility that Mb can exist within hydrophobic regions of DDAB bilayers was suggested by molecular dynamics (Figures 5 and 6). Little interaction of DDAB head groups with the Mb surface was apparent. It is of interest that α -helix predominates in the regions of transmembrane proteins that cross lipid membranes (1). Mb is about 75% α -helix, which might facilitate interactions with hydrophobic regions of DDAB bilayers.

Film structure and electrochemical kinetics. Electrochemical studies (2) of Mb-DDAB films cast onto pyrolytic graphite (PG) electrodes provided rate constants for electron transfer involving the heme Fe(III)/Fe(II) redox couple, as well as standard potentials (E°), and diffusion coefficients for charge transport (D_{ct}) through the films. Electrochemical studies also showed that Mb diffuses through the DDAB films.

The bottom three rows in Table I give electrochemical parameters for Mb in aqueous buffer solutions. Electron transfer was not detected for Mb in solution at PG electrodes, but at InSnO_2 a very small apparent standard heterogeneous rate constant (k^0) of $7 \times 10^{-6} \text{ cm s}^{-1}$ was found. Low spin CN-Mb has a rate constant about 100-fold larger.

The most striking observation in Table III is that rate constants for electron transfer of high spin Mb in DDAB films are 1000-fold larger than those for high spin aquo-Mb in water. In the realm of heterogeneous electron transfer at electrodes, rates in the DDAB films are moderately fast, but not exceedingly so (e.g. $>1 \text{ cm s}^{-1}$). However, they are quite respectable

Table III. Summary of Electrochemical Parameters for Myoglobin at 25 °C.

pH	sample - electrode	$10^6 D_{ct}$ cm ² /s	E° , V/NHE	$10^3 k^{\circ}$, cm/s	ref.
5.5 ^a	Mb-DDAB-PG	0.54	0.093	7±1	2
7.5 ^a	Mb-DDAB-PG	0.41	0.055	8±1	2
5.5 ^a	Mb-CTAB-PG ^b	0.55	0.025	3±2	2
5.5 ^a	Mb-SDS-PG ^b	0.30	-0.030	2±1	2
5.5-7.5 ^a	H ₂ O, Mb/bare PG	—	—	ND ^d	2
7.0	H ₂ O, Mb/bare InSnO ₂	0.5 ^c	0.05	0.007	3
7.0	aq. CN-Mb/InSnO ₂	0.5 ^c	-0.385	0.7	3

^aSolutions contained 50 mM NaBr. Results are averages from ref. 2 of values found by cyclic and normal pulse voltammetry. ^bPG electrodes had adsorbed films of Mb and soluble surfactants sodium dodecylsulfate (SDS) and cetyltrimethylammonium bromide (CTAB). ^cDiffusion coefficient of Mb in solution. ^dElectron transfer not detected.

for redox proteins, which often have difficulties exchanging electrons directly with electrodes (8). PG electrodes with coatings of adsorbed water soluble surfactants sodium dodecylsulfate (SDS) and cetyltrimethylammonium bromide (CTAB) also provide relatively fast electron transfer (Table III).

Comparison of standard potentials in the pH 7.5 film and in water at pH 7 (Table III) suggests that the labile axial ligand on the Mb Fe(III)heme may be water. Charge transfer diffusion coefficients (D_{ct}) are also listed in Table III. These are measures of the rate of transport of charge through the film when electrons are injected into it from the electrode. This is usually treated as a diffusion process (21). Values of D_{ct} for Mb-DDAB are just slightly smaller than those of DDAB films loaded with ferrocyanide, cobalt(III)corrin hexacarboxylate, or metal phthalocyanine tetrasulfonates (4) which gave $0.6-1 \times 10^{-6}$ cm² s⁻¹ for liquid crystal films. D_{ct} for Mb-DDAB in the gel state is 10-20 fold smaller (2).

Films of ionic polymers loaded with electroactive counter ions are conceptually similar to DDAB films. D_{ct} for Mb-DDAB films are significantly larger than values of 10^{-8} to 10^{-10} cm² s⁻¹ for typical ionic polymers (21). Thus, even with the relatively large Mb molecule present, charge transport through liquid crystal DDAB films is faster than through most ionic polymer films.

A common question for electroactive films is whether D_{ct} corresponds to physical diffusion of the electroactive species or to electron-exchange between active sites, so called "charge hopping" (21). Measured times for breakthrough of Mb across DDAB films were consistent with those predicted for a molecule with $D_{ct} = 4 \times 10^{-7}$ cm² s⁻¹ (cf. Table III) to achieve a root mean square displacement equivalent to film thickness. This suggests that Mb diffuses physically within Mb-DDAB films.

Likely reasons for enhanced electron transfer rates in Mb-DDAB include the following: (i) strongly adsorbed surfactant on the electrode may inhibit competitive adsorption of macromolecular impurities in solution (3,8) which might otherwise block electron transfer; and (ii) Mb may be oriented in the film in a way favorable for electron transfer, for example with the heme group close to the electrode. Similar rate enhancements found with films of different types of surfactants (Table III) suggest that surfactant adsorption on the electrode must play a role. However, this does not prohibit an influence for Mb orientation. ESR and linear dichroism suggest orientation of Mb in the DDAB films. These data do not as yet provide detailed structural information on the orientation of the key Mb molecules closest to the electrode surface where the electron transfer events occur.

If molecular orientation plays a critical role in electron transfer kinetics, Mb molecules would have to achieve preferred orientations dynamically at the electrode-film interface during the voltammetric measurement. The electrode is most likely covered by a strongly adsorbed surfactant bilayer (22), before as well as after Mb is introduced into the film. If Mb resides on the outer surface of a bilayer of DDAB at the PG-film interface, it would be at least the width of a DDAB bilayer (30 Å) away from the electrode. Transfer of an electron across such a distance is slow (23). These considerations suggest that Mb might be able to pass through the fluid DDAB bilayers during the dynamic electrochemical experiments.

Studies of diffusion through biomembranes suggest that passage of proteins through intact bilayers may be slow (1). However, biomembranes of many organisms exist close to their phase transition temperatures, and may even contain small solid-like regions of gel-state lipids (24). On the other hand, the Mb-DDAB transition (T_c) to the fluid liquid crystal phase occurs more than 10 °C below room temperature. Thus, Mb-DDAB films may be in a more highly fluid state than typical biomembranes. High fluidity may permit a small protein like Mb to pass through more easily than if the film were close to its T_c . Consequently, the solid-like gel state retards charge and mass transport significantly, while the liquid crystal state facilitates these processes (2).

The dynamic nature of surfactant bilayers also needs to be considered. Out of plane bending leading to transient pores or defects in biomembranes is now rather well accepted (1). These defects become larger in electric fields similar to those generated in voltammetry and can lead to bilayer rupture. Whether or not such processes can assist charge transport in DDAB films remains speculative at this stage of our work.

Summary and Conclusions

Spectroscopic and molecular modeling results presented herein along with earlier thermal and electron transfer studies can be combined to give a structural picture of Mb-DDAB films. The following features are established: (i) lamellar liquid crystal DDAB arranged in bilayers with tilted

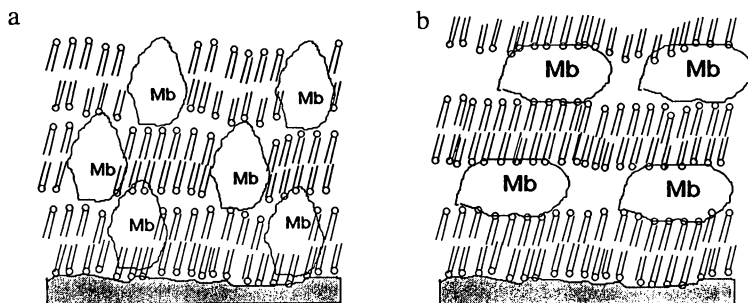


Figure 7. Conceptual models of several bilayers of static Mb-DDAB films: (a) Mb resides partly within bilayers; (b) Mb resides between bilayers.

hydrocarbon tails as in biological membranes; (ii) Mb in a high spin state with preferred orientation in the films; (iii) electron transfer enhanced by adsorbed surfactant on the electrode; (iv) charge transport facilitated by the fluid liquid crystal phase; and (v) physical diffusion of Mb within the films.

In studies of multibilayer films of double chain phosphate surfactants containing Mb, Kunitake et al. proposed that Mb is intercalated between surfactant bilayers (7). At present we can imagine two limiting static models for the Mb-DDAB films, one with Mb within bilayers, but interacting in some way with adjacent bilayer head groups, and one with Mb between bilayers (Figure 7). An intermediate supramolecular structure between these two extremes is also possible. Both limiting models might increase the stability of the films by electrostatic interactions of the DDAB head groups and the Mb surface amino acid residues or buffer components bound to Mb. However, because Mb has a positive charge at the pHs employed, direct surface interactions would have to be stronger than existing charge pair interactions of amino acid residues on the Mb surface.

There is as yet no definitive experimental evidence in favor of either limiting model concerning the site of Mb residence in DDAB films. However, it is difficult to explain the electron transfer and charge transport results by the between-bilayer model (Figure 7b), unless Mb diffusion through defects in the films plays a dominant role in these processes. Such defects would be expected to be retained in both gel and liquid crystal states of a given film. Thus, the large difference in charge transport rate (2) above and below T_c might not occur if Mb were transported only through defects.

Models featuring the possibility of Mb entry into fluid bilayers at the electrode-film interface go further at present to explain electrochemical results. Nevertheless, further studies are needed to achieve a more complete picture of the supramolecular structure of Mb-DDAB films.

Acknowledgments. This work was supported by U.S. PHS grant No. ES03154 from NIH awarded by the National Institute of Environmental Health Sciences. The authors are grateful to Dr. Veeradej Chynwat for assistance with ESR and linear dichroism.

References and Notes

- (1) Kotyk, A.; Janacek, K.; Koryta, J. *Biophysical Chemistry of Membrane Structure*, Wiley: Chichester, U. K., 1988, pp. 54-73.
- (2) Rusling, J. F.; Nassar, A. F. *J. Am. Chem. Soc.*, **1993**, *115*, 11891-11847.
- (3) King, B. C.; Hawkridge, F. M.; Hoffman, B. M. *J. Am. Chem. Soc.* **1992**, *114*, 10603-10608.
- (4) (a) Rusling, J. F.; Zhang, H. *Langmuir* **1991**, *7*, 1791-1796. (b) Rusling, J. F.; Hu, N.; Zhang, H.; Howe, D.; Miaw, C.-L.; Couture, E. in *Electrochemistry in Colloids and Dispersions*, Mackay, R. A. and Texter, J. (Eds.), VCH Publishers; N, Y, **1992**, pp. 303-318. (c) Hu, N.; Howe, D. J.; Ahmadi, M. F.; Rusling, J. F. *Anal. Chem.* **1992**, *64*, 3180-3186. (d) Zhang, H.; Rusling, J. F. *Talanta*, **1993**, *40*, 741-747. (e) Miaw, C.-L.; Hu, N.; Bobbitt, J. M.; Ma, Z.; Ahmadi, M. F.; Rusling, J. F. *Langmuir* **1993**, *9*, 315-322.
- (5) (a) Wade, R. S.; Castro, C. E. *J. Am. Chem. Soc.* **1973**, *95*, 231-234. (b) Bartnicki, E. W.; Belser, N. O.; Castro, C. E. *Biochemistry*, **1978**, *17*, 5582-5586.
- (6) (a) Pryor, W. A. in Pryor, W. A. (Ed.) "Free Radicals in Biology" Academic: New York, 1976, pp. 1-50. (b) Brown, J. F.; Bedard, D. L.; Brennan, M. J.; Carnahan, J. C.; Feng, H.; Wagner, R. F. *Science*, **1987**, *236*, 709.
- (7) (a) Hamachi, I.; Noda, S.; Kunitake, T. *J. Am. Chem. Soc.* **1990**, *112*, 6744-6745. (b) Hamachi, I.; Honda, T.; Noda, S.; Kunitake, T. *Chem. Lett.* **1991**, 1121-1124. (c) Hamachi, I.; Noda, S.; Kunitake, T. *J. Am. Chem. Soc.* **1991**, *113*, 9625-9630.
- (8) (a) Armstrong, F. A.; Hill, H. A. O.; Walton, N. J. *Accts. Chem. Res.* **1988**, *21*, 407-413. (b) Armstrong, F. A. in *Bioinorganic Chemistry, Structure and Bonding* 72, Springer-Verlag, Berlin, 1990, pp. 137-221.
- (9) Suga, K.; Rusling, J. F. *Langmuir* **1993**, *9*, 3649-3655.
- (10) (a) Press, W. H.; Flannery, B. P.; Teukolsky, S. A.; Vetterling, W. T. *Numerical Recipes* Cambridge Univ. Press: Cambridge, MA: 1988, pp. 301-327. (b) Weiner, S. J.; Coleman, P. A.; Nguyen, D. T.; Case, D. A. *J. Comput. Chem.* **1986**, *7*, 230. (c) Hildebrand, J. H. *Proc. Nat. Acad. Sci.* **1979**, *76*, 194.
- (11) Kauppinen, J. K.; Moffatt, D. J.; Mantsch, H. H.; Cameron, D. G. *Appl. Spec.* **1981**, *35*, 271-276.
- (12) Rusling, J. F.; Kumosinski, T. F. *Intell. Instr. & Computers*, **1992** (July/Aug.) 139-145.
- (13) Konetani, T.; Schleyer, H. *J. Biol. Chem.* **1967**, *242*, 3926.
- (14) Eaton, W. A.; Hochstrasser, R. M. *J. Chem. Phys.* **1968**, *49*, 985-995.
- (15) (a) Breton, J.; Michel-Villaz, M.; Pailotin, G. *Biochim. Biophys. Acta* **1973**, *314*, 42-56. (b) Norden, B.; Lindblom, G.; Jonas, I. *J. Phys. Chem.* **1977**, *81*, 2086-2093.
- (16) (a) Friend, S. H.; Gurd, F. R. N. *Biochemistry* **1979**, *18*, 4620. (b) Shire, S. J.; Hania, G. I. H.; Gurd, F. R. N. *Biochemistry* **1974**, *13*, 2967-2980. (c) Friend, S. H.; Gurd, F. R. N. *Biochemistry* **1979**, *18*, 4612-4619.
- (17) Kumosinski, T. F.; Farrell, H., in Kumosinski, T. F.; Liebman, M. N. (Eds.); *Molecular Modeling*, this volume.

- (18) Antonioni, E.; Rossi-Bernardi, L.; Chinacone, E. (Eds), *Methods in Enzymology*, Vol. 76, Academic: N. Y., 1981, pp. 552-559.
- (19) Antonioni, E.; Brunori, M. *Hemoglobin and Myoglobin in their Reactions with Ligands*, North Holland: Amsterdam, 1971.
- (20) Klein, M. L., Plenary Lecture, 67th Colloid and Surface Science Symposium, Toronto, Canada, June, 1993.
- (21) Charge transport diffusion through electroactive films is discussed by Murray, R. W. in Bard, A. J. (Ed.) *Electroanalytical Chemistry*, Vol. 13, Marcel Dekker, N. Y. 1984, pp. 191-368.
- (22) (a) Head down bilayers of DDAB and CTAB have been inferred^{22b} from surface enhanced Raman spectroscopy in cast films on silver. There is evidence that head down orientation of these surfactants and SDS on electrodes occurs over a wide potential range when adsorption occurs from a relatively concentrated solution.^{22c} (b) Suga, K.; Bradley, M.; Rusling, J. F. *Langmuir*, 1993, 9, 3063-3066. (c) Rusling, J. F. in Bard, A. J., Ed, *Electroanalytical Chemistry*, Vol. 19, 1994, Marcel Dekker: New York, pp. 1-88.
- (23) (a) Closs, G.L.; Miller, J.R. *Science*, 1988, 240, 440.; (b) Mayo, S. L.; Ellis, W. R.; Crutchley, R. J.; Gray, H. B. *Science*, 1986, 233, 948.
- (24) (a) Lee, A. G. *Biochim. Biophys. Acta* 1977, 472, 237-281; 285-344. (b) Nagel; J. F. *Ann. Rev. Phys. Chem.* 1980, 31, 157-159.

RECEIVED December 20, 1993

Chapter 17

Molecular Dynamics and NMR Studies of Concentrated Electrolytes and Dipoles in Water

Ion C. Baianu¹, E. M. Ozu¹, T. C. Wei¹, and Thomas F. Kumosinski²

¹Department of Food Science, Agricultural and Food Chemistry–Nuclear Magnetic Resonance Facility, University of Illinois at Urbana, 580 Bevier Hall, 905 South Goodwin Avenue, Urbana, IL 61801

²Eastern Regional Research Center, Agricultural Research Service, U.S. Department of Agriculture, 600 East Mermaid Lane, Philadelphia, PA 19118

The molecular dynamics of water and selected ions was studied in concentrated electrolyte solutions with, or without, dipolar ions added. Our experimental results by multinuclear spin relaxation techniques were then compared with molecular dynamics computations for water and ions in concentrated electrolyte solutions ($\text{LiCl} \cdot R(\text{H}_2\text{O})/R(\text{D}_2\text{O})$ and $\text{NaCl} \cdot R(\text{H}_2\text{O})/R(\text{D}_2\text{O})$, with $4 < R < 12$ for Li^+ and $6 < R < 16$ for NaCl). Multinuclear spin relaxation data were analyzed with a thermodynamic linkage model of hydrated ion clusters of various sizes and composition. Our results indicate that tetramer clusters of hydrated Li^+ and Cl^- are the preferred structures formed in such concentrated electrolyte solutions as a consequence of dimer-tetramer equilibria that occur at 293 K. Within such clusters water molecules undergo hindered reorientation motions in the hydration shell of the cation. The corresponding correlation time of water (D_2O), determined by ^{17}O NMR, is less than 30 ps for $4 < R < 12$ in all solutions studied at 293 K.

The study of the dynamics of molecules by computer modeling is a relatively recent and rapidly expanding field of research; this field is now known as “**Molecular Dynamics**”. A standard approach to molecular dynamics computations on a fast computer or a supercomputer begins with an arbitrary, or pseudo-random, configuration of a group of molecules or ‘particles’, and computes their subsequent positions and velocities as a function of time with Newton’s equations of motions for these particles. The forces acting on molecules are derived from potential functions for a large number of particles in a “box” which is also set up with a grid for convenience in following the particle configurations and their statistics. Because of the complexity of the potential functions (interaction potentials), the reliability of the method is greatest for the simplest systems, such as “inert” gases and molten salts (1,2), (Figure 1a). Extensive molecular dynamics computations were also carried out for **liquid water** and aqueous solutions of electrolytes (2). In the latter case, the interaction potentials are not really known but they were assumed to involve only slight perturbations of the potentials for liquid water, which in itself is an approximation. For aqueous solutions of electrolytes with ionic radii that are larger than about 0.9 Å, and single charge, this approach might produce results that resemble

0097-6156/94/0576-0269\$14.48/0

© 1994 American Chemical Society

the local structure of electrolyte solutions determined experimentally. For electrolytes with high ionic field strengths, such as Li^+ , Ca^{2+} , Zn^{2+} , Be^{2+} , La^{3+} , F^- , OH^- , etc, the results of this molecular dynamics approach disagree with local structures (Figure 1b) that are emerging from experimental studies (3-5).

Compare, for example, the local structure of molten KCl in Figure 1a (which was derived from molecular dynamics computations) with that in Color Plate 15 for concentrated LiCl solutions in water (which was derived from Nuclear Magnetic Resonance (NMR) studies). The local structures of KCl, NaCl and CsCl in aqueous solutions derived from molecular dynamics computations were all similar to that shown in Figure 1a for molten KCl. Clearly, the local structures (4) of $\text{LiCl} \cdot n\text{H}_2\text{O}$ ($2.2 \leq n \leq 12$) solutions shown in Color Plate 15 and Figures 1b and 1c are quite different from the proposed molecular dynamics structures for the same system. Certain experimental results that are available for $\text{LiCl} \cdot n\text{H}_2\text{O}$ (3-5) indicate that the high-ionic field strength of Li^+ causes **water-bridging** between Li^+ and Cl^- , as well as the formation of **hydrated ion-pair clusters** (Figure 1b and Color Plate 16), whose presence and structure could not be detected in the molecular dynamics results (2).

Monte-Carlo Simulations and Spectroscopy

A more recent approach to the molecular dynamics of aqueous solutions of electrolytes (6) involved the use of **statistical mechanics** combined with **Monte Carlo simulations**. The interparticle potential and the field gradient function employed to calculate the **fluctuations** in solutions were based upon quantum mechanics/quantum chemistry calculations (6). This approach allows one to estimate the **correlation times** related to the computed fluctuations in solution. The Monte Carlo simulation involves the generation of a large number of molecule/particle configurations, starting from an **equilibrium** ensemble; the total potential energy is then assumed to be reasonably approximated by a **sum of pair-interaction energies**. One proceeds then to compute any physical property that can be expressed as a function of the interparticle distances and orientations. Such configurations generated in the Monte Carlo simulation on a mainframe computer (or, preferably, on a **supercomputer**) can be also employed to calculate **hydration numbers**, theoretical **radial distribution functions**, or other interesting properties of electrolyte solutions. The main advantage of this approach is that one can compare directly the predictions of molecular dynamics computations with the analysis of experimental results. For example, one can compare molecular dynamics results with spectroscopic (NMR, IR, Raman, etc) and X-ray/neutron/ electron scattering data for electrolyte solutions. Among the spectroscopic techniques widely used in studies of electrolyte solutions most prominent are laser-Raman scattering and Nuclear Magnetic Resonance (NMR). Nuclear Magnetic Resonance has the advantages of being able to monitor **all** the nuclei present in aqueous solutions of electrolytes, provide directly dynamic information through **relaxation** studies and allow one to derive **hydration numbers** from spectral (high-resolution NMR) studies/measurements of chemical shifts. Since nuclear spin relaxation studies are often able to estimate **correlation times** for solutions by making only a minimum number of assumptions, such results are especially relevant to the evaluation of molecular dynamics computations.

Nuclear Magnetic Resonance Spectroscopy and Relaxation

Nuclear Magnetic Resonance (NMR) is a branch of radio frequency (r.f.) absorption spectroscopy which is specifically concerned with the **resonant** absorption of radiowaves by the nuclei of a sample placed in an intense magnetic field; the

NOTE: The color plates can be found in a color section in the center of this volume.

radiofrequency absorption occurs as a result of transitions between the **nuclear spin** energy levels, in the presence of a very homogeneous, strong and static magnetic field. The stronger the magnetic field, \mathbf{H}_0 , the higher is the frequency required for resonance, and the more intense is the absorption, as indicated in Figure 2.

The transition induced between the nuclear spin energy levels (Figure 2) by the resonant r.f. wave or pulse is recorded as a sharp peak for a liquid; for a pair of nuclear spin energy levels shown in Figure 2, a single peak is recorded whose lineshape is **Lorentzian** in a simple liquid or solution. The linewidth of the NMR absorption peak at half-height is determined by the **lifetime** of the excited nuclear spin state in the presence of only negligible magnetic field inhomogeneities. More precisely expressed, after the occurrence of the NMR absorption the system of nuclear spins relaxes through interactions between the nuclear spins ('spin-spin', or T_2 -relaxation process). The NMR signal is recorded with a coil whose axis, x , is perpendicular to the direction z of the static magnetic field, \mathbf{H}_0 , in Figure 2; therefore, the loss of phase coherence of the nuclear spins in the xy -plane, which is normal to \mathbf{H}_0 , occurs as a result of **transverse (T_2) nuclear spin relaxation** processes (Figure 3a) that involve interactions between nuclear spins. The corresponding **Free Induction Decay (FID)** signal is shown in Figure 3b. On the other hand, the nuclear spin magnetization has a component M_z along the magnetic field direction (z), which is not directly observed by the detector coil whose axis is in the xy -plane, along the x -direction. Immediately after the nuclear spin excitation, the M_z component points against the static magnetic field \mathbf{H}_0 and later relaxes (or comes back) with a characteristic time constant, T_1 , towards the magnetic field direction. This relaxation process which occurs along the z -axis is called the longitudinal, or 'spin-lattice' (T_1) relaxation. The latter name is often used because the T_1 -relaxation process involves interactions of the nuclear spins with the surrounding electrons ("the lattice"); in a crystalline solid, the nuclear spins interact with the electrons in the surrounding **crystal lattice** causing the relaxation of the M_z magnetization component towards the magnetic field direction. The use of the term 'spin-lattice' relaxation is, however, not restricted to T_1 -relaxation in crystalline solids but is employed, in general, for any system, either solid, liquid or gas.

Mechanisms of Relaxation in a Liquid

A. **For a spin $I=1/2$** , (such as ^1H) the major contributions to relaxation are made by:

- spin-spin coupling
- magnetic dipolar interactions
- chemical exchange (of protons).

B. **For quadrupolar nuclei — (with spin $I > 1/2$** , such as ^{17}O), the relaxation is caused by the interactions of the nuclear quadrupole moment with the surrounding, fluctuating **electrical** field. In the case of deuterium (^2H) NMR, however, chemical exchange also contributes to the nuclear spin relaxation.

Nuclear spin relaxation is, therefore, dependent upon the **molecular dynamics** of the liquid, which is characterized by a **correlation time**, or distribution of correlation times: the faster the motions are, the shorter is the correlation time and the longer is the relaxation time. A motion of a molecule, such as a **rotation** around a specific

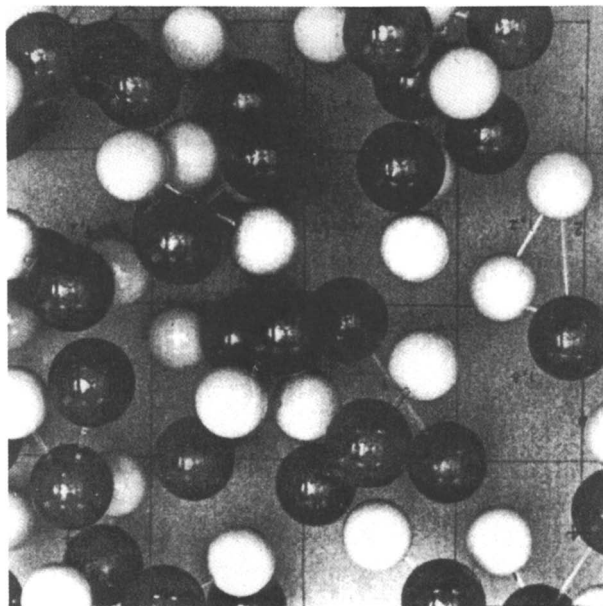


Figure 1a. Computer generated configuration of ions in molten potassium chloride, based on a molecular dynamics program written in FORTRAN-77. (Modified from ref. 1).

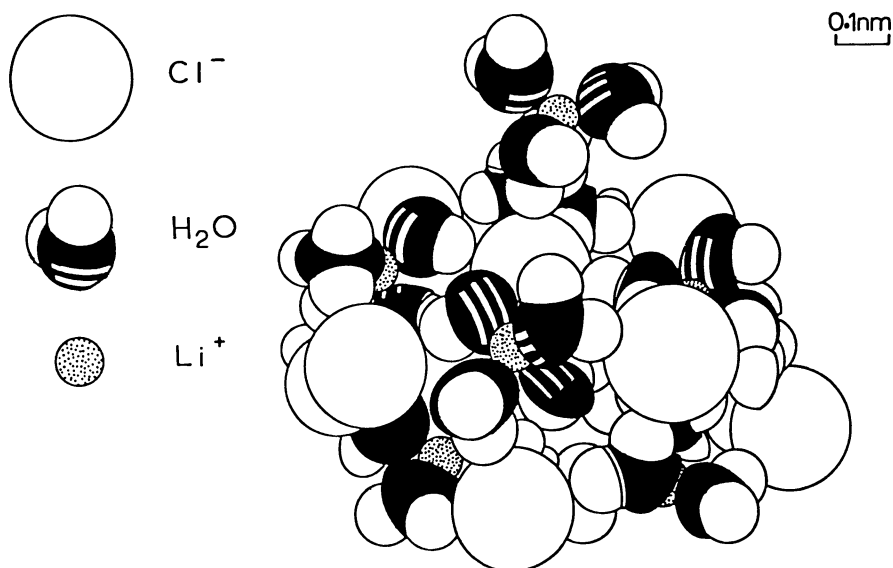


Figure 1b. Local structure of water bridged $\text{Li}^+(4\text{H}_2\text{O})\text{Cl}^-$ clusters in glasses at 100 K, derived from pulsed ^1H NMR data. (Modified from ref. 5).

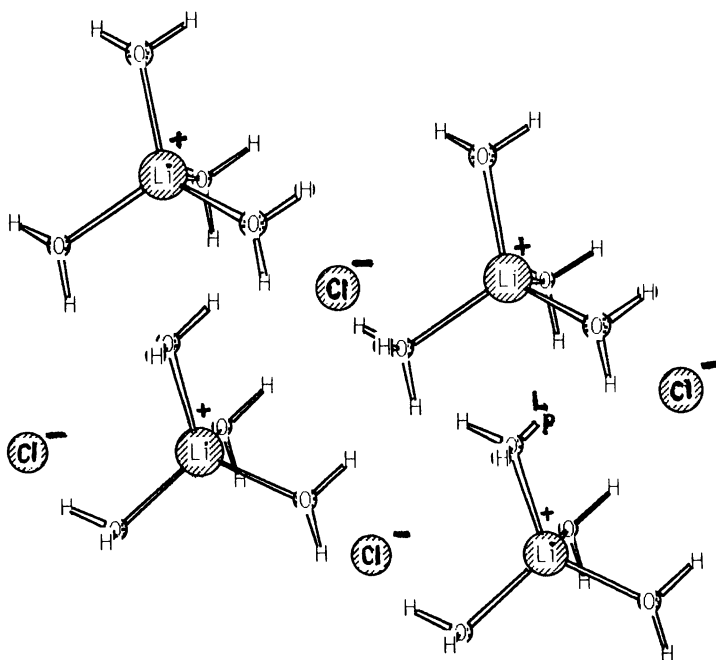


Figure 1c. Schematic model of the local structure of aqueous solutions of LiCl containing hydrated ion-pairs in $\text{Li}^+(\text{nH}_2\text{O})\text{Cl}^-$, water-bridged clusters.

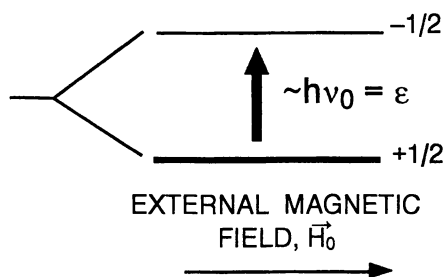


Figure 2. Nuclear spin energy levels for a spin $-1/2$ nucleus (such as ^1H) in a strong static magnetic field, \vec{H}_0 . A resonant radiowave is absorbed by the nuclear spins undergoing transitions from the lower to the upper energy levels and this energy absorption is observed as an NMR signal.

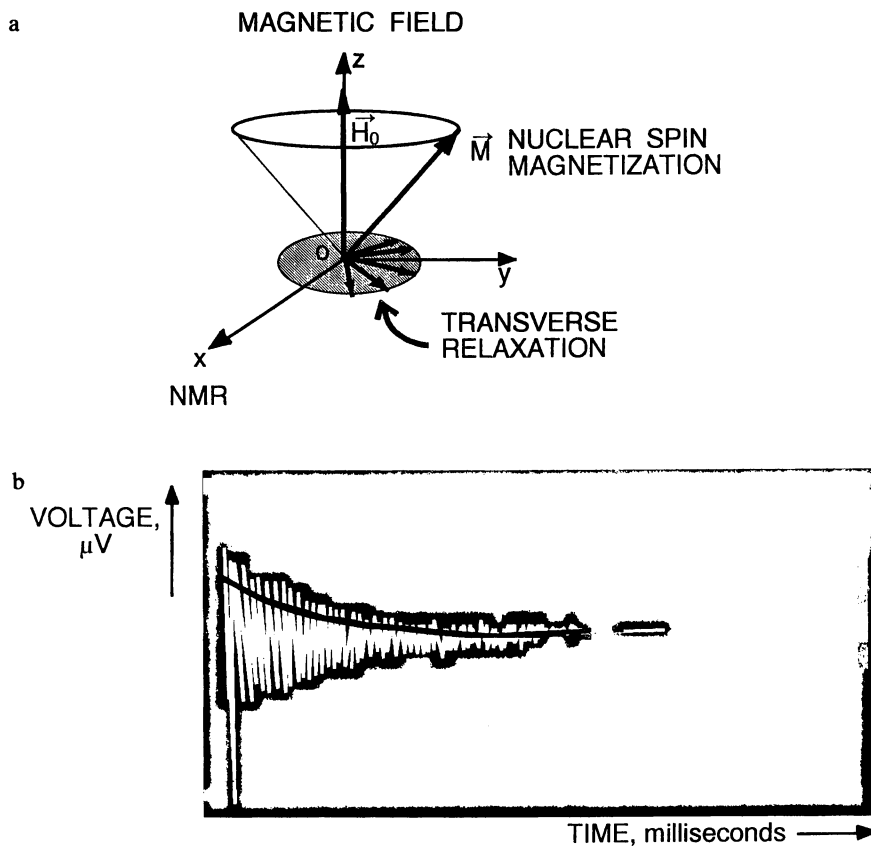


Figure 3. (a) Precession of the nuclear spin magnetization in a strong, external magnetic field. (b) The free induction decay signal observed in single-pulse NMR experiments, in the transverse plane depicted in Fig. 3a.

axis, or a translation in a specific direction, occurs in a determined interval of time which is the correlation time of that particular motion. The random, **isotropic reorientation** of a molecule is, however, the type of motion usually considered in the nuclear spin relaxation studies of liquids. In this case, a single **average** correlation time for the isotropic reorientation, τ_c , provides the simplest estimate of the **molecular mobility** in the liquid.

Basic Deuterium NMR Theory. The discussion of the results in the following sections requires a brief presentation of the basic deuterium NMR theory, with emphasis on the ^2H NMR spectra of solids.

An isolated deuteron with a nuclear spin $I = 1$ in a single crystal rigid lattice would have three equally spaced, Zeeman energy levels if and only if the electrical quadrupole interaction was neglected in the calculation (Figure 4). The presence of an appreciable quadrupolar interaction (which is the real case for **any** deuterium containing system, with only one exception) requires, however, the calculation of [**Zeeman + Quadrupole**] deuteron spin energy levels (according to Abragam (10)) with a quadrupolar Hamiltonian spin operator:

$$\mathbf{H}_Q = [I(2I-1)/4] \cdot (e^2qQ/\hbar) \cdot [(3\cos^2\theta - 1)/2] \cdot [3m^2 - I(I + 1)] \quad (1)$$

where $I = 1$, (e^2qQ/\hbar) is the quadrupole coupling constant, eq is the electric field gradient at the deuteron, eQ is the deuteron quadrupole moment, $m = 1, 0$ or -1 is the magnetic quantum number for the deuterium, θ is the angle which defines the orientation of the single crystal with respect to the static magnetic field and \hbar is Planck's constant. The single deuteron spin energy levels obtained with the \mathbf{H}_Q operator given by Equation 1 are presented in Figure 4b; notably, the presence of the electrical quadrupole interaction causes a rise in the $m = +1$ and $m = -1$ spin energy levels, the separation of the spin energy levels is, thus, **unequal** and it is **orientation dependent** (Figures 4b and 4c, respectively). There are, therefore, only two possible single-quantum transitions ($\Delta m = \pm 1$) for the **isolated** deuteron (that is, a deuteron which is **not** interacting with any deuterons or other nuclei), as shown in the NMR spectrum of the single deuteron in Figure 4b at right. The angular, or orientation dependent, $[(3\cos^2\theta-1)/2]$ term in Equation (1) will cause the doublet of the deuteron to change with the orientation θ of the $X \rightarrow D$ bond vector in the single crystal with respect to the external magnetic field vector \mathbf{H}_0 , in the manner shown in Figure 4c; note that at the "magic-angle" ($\theta = 54.74^\circ$) the angular term $[(3\cos^2\theta-1)/2]$ vanishes, and there is only a single peak located in the center, instead of a doublet. In the case of deuterium NMR of crystalline powders the spectrum is obtained by averaging the angular term over all possible orientations; in the resulting powder pattern, the 90° orientation dominates, whereas the θ peaks appear as edges, or singularities, symmetrically placed with respect to the center of the powder pattern, and with twice the spacing of the 90° splitting (as shown in Figure 4d). As an example, the two deuterons of a water molecule in a D_2O (hexagonal structure, I_h) ice powder give a spectrum such as the one shown in Figure 4d, with a splitting $\Delta\nu_Q$ between the two central peaks of ~ 160 kHz in the rigid lattice (at 100 K). At the other extreme, in liquid D_2O at room temperature (293 K), the quadrupolar splitting disappears completely because of the averaging of the quadrupolar interaction by the extremely fast, isotropic (random) motions of D_2O in the liquid; only a single Lorentzian ^2H

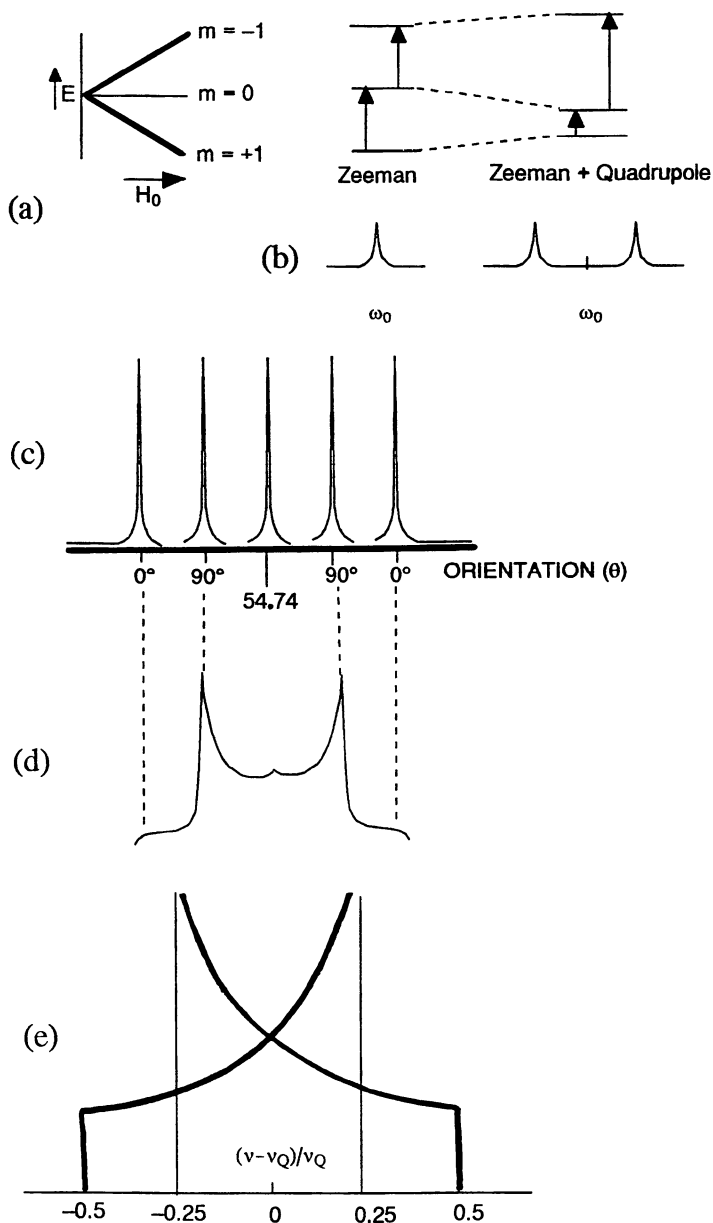


Figure 4. Deuteron spin energy levels in a static magnetic field, in the absence (a) and presence (b) of quadrupolar interactions; (c) orientation dependence of ^2H NMR spectra of a deuteron in a single crystal (d) ^2H NMR spectrum of a deuteron in a polycrystalline powder. The theoretical spectrum (e) is calculated in the absence of dipolar broadening. (Experimental and theoretical ^2H NMR powder spectra were taken with permission from ref. 8 (Copyright 1982, Academic Press); spectrum (d) shows the calculated deuteron NMR peaks for the two transitions in Fig. 4b for a polycrystalline powder).

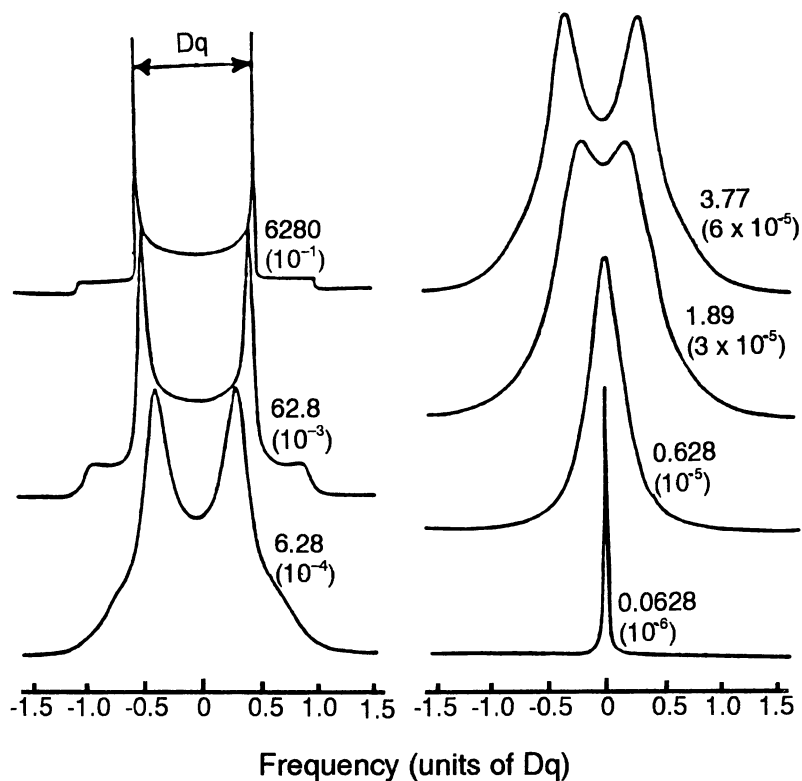


Figure 5. The influence of rotational motions of various rates on the ^2H NMR powder spectrum of a partially ordered solid (residual quadrupole splitting of 10 kHz). The correlation times for overall motion (in seconds per radian) are shown in parentheses, whereas the product $2\pi Dq\tau$, (where Dq is the residual quadrupole splitting and τ is the correlation time), are shown without parentheses. (Reproduced with permission from ref. 8. Copyright 1982, Academic Press).

NMR peak of ~ 0.5 Hz half-height linewidth is observed for both deuterons in the D_2O molecule. Thus, the deuterium NMR relaxation in the liquid is essentially **quadrupolar**. The correlation time of liquid D_2O calculated from this single lorentzian 2H NMR peak is about 4.7 ps at 293 K (20°C). Between these two extremes of a rigid solid and a liquid there are many intermediate cases possible Figure 5, two of which are especially significant to the analysis of the 2H NMR data presented in the next section. The fast rotation of the D_2O molecule about its symmetry axis would substantially reduce the quadrupole splitting from the value measured in the ice rigid lattice; this would also cause changes in the lineshape. Static disorder, caused by variations of the O-D bond orientations throughout the lattice of a solid, would also cause a reduction in the deuteron quadrupole splitting for the D_2O molecule; such variations in molecular orientations of the X→D bond vector can be characterized by one to three order parameters, depending on the types of partial disorder encountered. For D_2O molecules in a lattice, the variation in the orientation of the D_2O molecule symmetry axis throughout the lattice can be represented by a single order parameter, S; the reduction in the quadrupolar splitting caused by such static disorder can be calculated from a powder spectrum with the following equation:

$$\Delta\nu_q = (3/4) \cdot (e^2qQ/\hbar)S \cdot \tau_c \quad (2)$$

where $\Delta\nu_q$ is the residual quadrupolar splitting and τ_c is the average correlation time of the D_2O molecules in the partially disordered lattice. For a completely random, “amorphous” solid, $S = 0$, whereas for a “perfect” crystal lattice, with **no** variation in the orientation of D_2O molecules, $S = 1.0$. It is interesting that even for relatively small values of S, the deuterium NMR powder patterns retain sharp features in the absence of fast motions (8), if the dipolar interactions involving the deuterons are not strong.

Deuterium NMR lineshapes are also markedly affected by exchange processes, such as the rotational jumps or “flips” of deuterons amongst alternate sites. An example is shown in Figure 6 where the 2H NMR lineshapes are presented as a function of the exchange rate (or “flip rate”) of deuterons in an aromatic ring system (10) ($e^2qQ/\hbar = 180$ kHz and $\eta = 0.06$). Large-angle flips yield unique “double-horn” features (Figure 6B) when these occur about an axis which is itself moving. Such a dynamic model is intuitively appealing for aromatic rings trapped within **channel clathrates** (or embedded in highly ordered smectic mesophases) and for **phenyl** group substituents in **proteins**. This type of model can be generalized to include more orientations around the flip axis to simulate the motions of **adsorbates trapped near surfaces** or on catalytically active sites.

Local Structure and Molecular Dynamics in Aqueous Solutions of Electrolytes

X-ray and Neutron Scattering Studies of Local Structure in Aqueous Solutions of Electrolytes. Understanding the local, or **short-range** ($\tau_c < 5 \text{ \AA}$), structure of aqueous solutions of electrolytes is a long-standing problem of electrochemistry. This problem is more difficult than the case of liquid metals or molten salts where the number of distinct atom or ionic species is $n_s < 3$. For a molten salt such as KCl, with

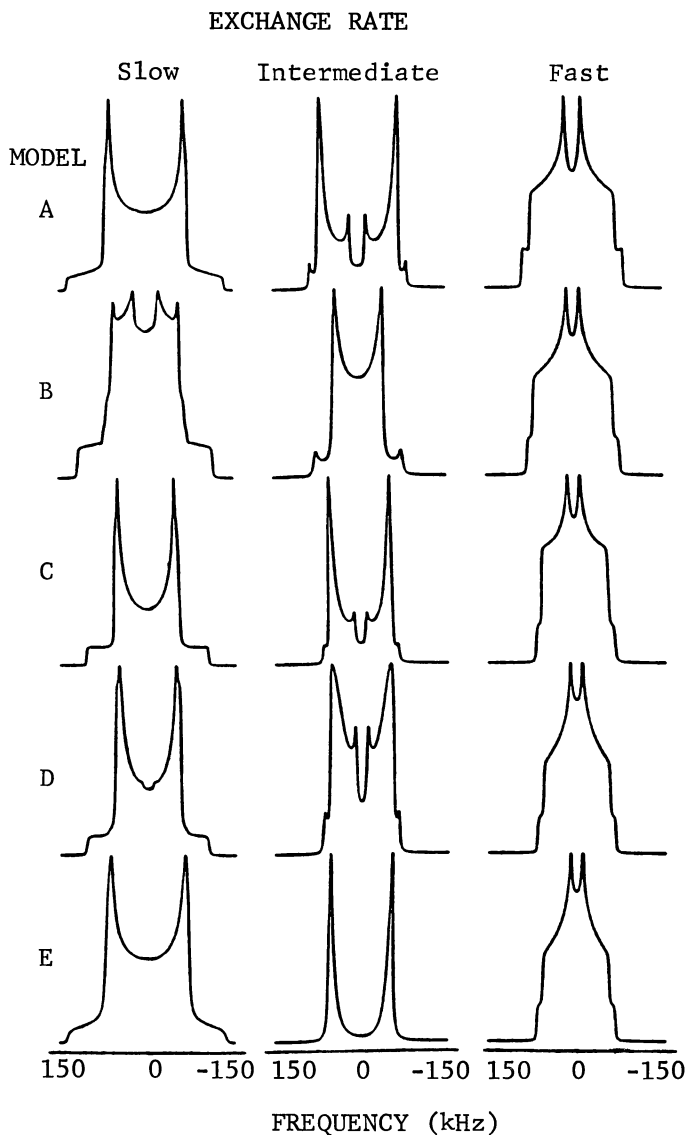


Figure 6. Simulated deuterium quadrupole-echo lineshapes under various exchange conditions. In all cases the spacing between $2 \mu\text{s}$ 90° pulses is $50 \mu\text{s}$, and all spectra have been normalized to unit intensity. The static quadrupole-coupling parameters are $e^2qQ/h = 180 \text{ kHz}$ and $\eta = 0.06$, as expected for ortho- or meta- deuterons on a benzene ring. The motion in model A is simply 180° flips about the C_1 - C_4 axis of the ring, while models B through E all include additional motion as described in the text. "Slow", "intermediate", and "fast" refer to flip rate, which was stepped from 103 to 105 to 107 s^{-1} for each model. A set of experimental lineshapes which exhibited features of models B,C,D, or E could not be adequately reproduced by the two-site model A. (Reproduced with permission from ref. 10. Copyright 1987, Academic Press).

$n_s = 2$, the number of pair correlation functions that characterize completely the local structure is 3, whereas for molten LiOH, $n_s = 3$ and the number of atom-pair correlations required is 6 (Figure 7a); four additional correlation functions and 2D-correlation functions were needed to define the coordination of Li^+ and OH^- in molten LiOH, and to represent the complex medium range structure ($r_c \sim 10 \text{ \AA}$) caused by the anisotropic (directional) nature of the OH^- ion. For an aqueous solution of electrolytes ten atom-pair correlations are needed. Since X-rays are scattered only very weakly by ^1H and ^2H atoms, the atom-pair correlation functions involving these atoms are unobtainable from X-ray diffraction experiments. However, the atom-pair correlation function, $g_{\text{OO}}(r)$ for the oxygen atoms can be reliably obtained from X-ray diffraction measurements for concentrated solutions. To determine the remaining pair correlation functions that involve ^2H (or ^1H) one needs to carry out neutron scattering measurements also. Few high quality/high Q, neutron diffractometers were built because of the very high cost. Both neutron scattering measurements and the analysis/corrections of neutron scattering data are relatively complex, time-consuming and require careful scrutiny (11). As a result, progress has been relatively slow in solving this problem, and a larger number of X-ray scattering than neutron scattering studies were carried out.

An important development in the neutron scattering study of electrolyte solutions was the selective use of **isotopic** substitution to obtain the partial, pair-atom correlation functions (11). Among the most studied systems are the aqueous solutions of LiCl and those of NiCl_2 . Figure 8a shows the X-ray scattering intensity as a function of the modulus of the scattering vector, $s = 4\pi \sin\theta/\lambda \text{ (\AA}^{-1}\text{)}$, for two aqueous solutions of concentrated NiCl_2 . Note the limitation to s -values of less than 10 \AA^{-1} in this data set and the limited resolution in the corresponding total correlation function, $G(r)$, (Figure 8b). The X-ray scattering curve for a metallic glass (12) (**FeNiPB**, Metglass 2826) with the same number of atom species ($n_s = 4$) is shown in Figure 8c for comparison with Figure 8a. Note the apparently simpler X-ray scattering function for the **FeNiPB** metallic glass and the small peak at low angles, near 1 \AA^{-1} in Figure 8c. Annealing of this glass causes slow **structural relaxation** (broadly similar to the case of molten LiOH discussed above) which diminishes the intensity of the 1 \AA^{-1} peak and sharpens up the rest of the pattern. The limit of s in such measurements was extended to 14 \AA^{-1} (moderate resolution); an X-ray scattering curve to a higher s -value of 17.4 \AA^{-1} (high resolution) is shown in Figure 8d for a metallic glass of simpler composition, Co_9P , (13) ($n_s = 2$). Both sets of data in Figures 8c and 8d were obtained by an energy-dispersive technique so that the r.d.f.'s can be directly determined without uncertain corrections of the scattered X-ray intensity for Compton scattering (14).

Neutron scattering curves are shown in Figure 9a for a series of concentrated electrolyte solutions in water. The NiCl_2 scattering curve has two lower peaks near 3 \AA^{-1} , following the main peak (instead of one peak as in the case of the other solutions). From such data, but at higher resolution ($Q \approx 16 \text{ \AA}^{-1}$), the neutron partial structure factor S_{NiNi} was derived for aqueous solutions of NiCl at various concentrations (Figure 9b).

The effect of isotopic substitutions on the neutron scattering intensity at the main scattering peak near 2 \AA^{-1} is shown in Figure 9c, both for the anion (Cl^-) and the

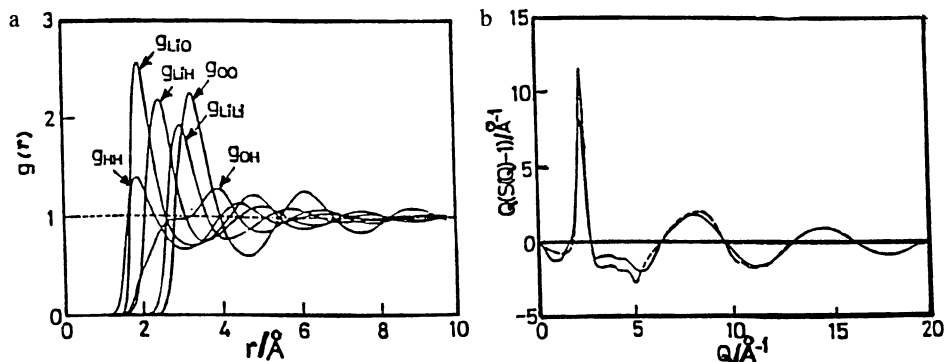


Figure 7. (a) Atom-atom pair correlation functions for molten LiOH at 767 K. (b) Calculated (solid line) and experimental (broken line) neutron weighted structure functions $Q[S(Q)-1]$ for molten LiOD. (Reproduced with permission from ref. 30. Copyright 1990, the American Institute of Physics).

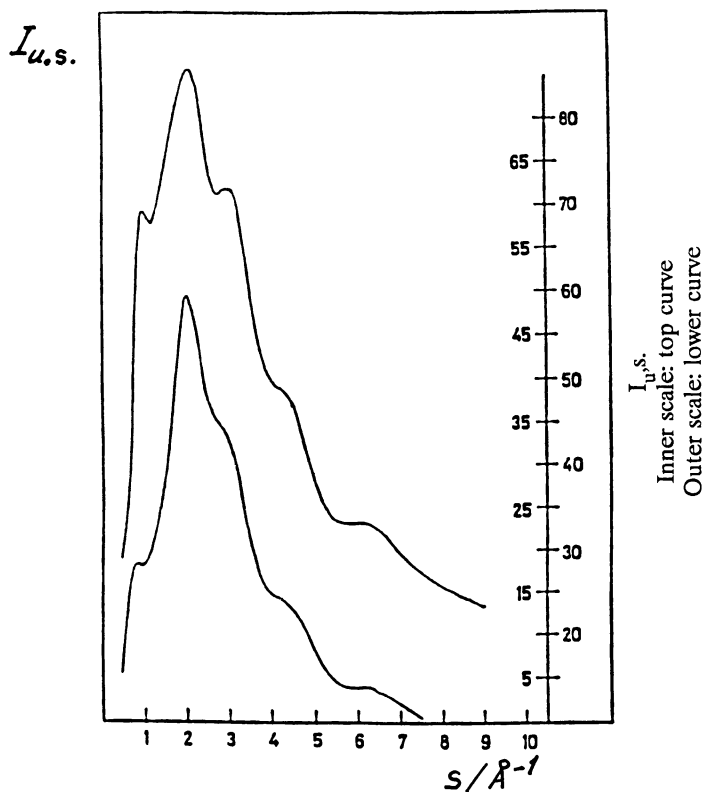


Figure 8a. Intensity curves (e.u.) for two NiCl_2 solutions in water. Top curve is for 4 mol dm^{-3} , and bottom curve is for 2 mol dm^{-3} . (Reproduced with permission from ref. 6. Copyright 1977, the Royal Society of Chemistry: Cambridge, UK).

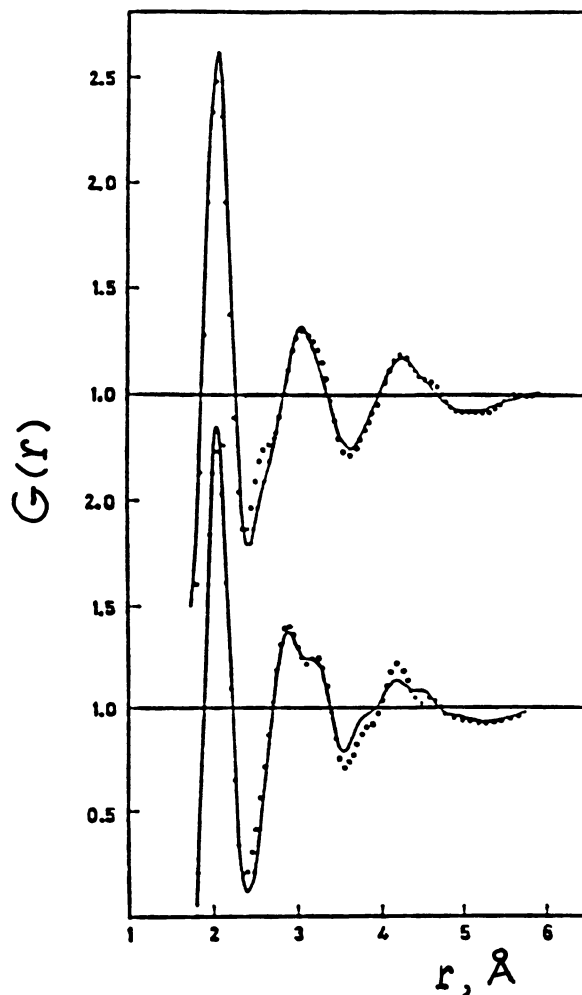


Figure 8b. Radial distribution function derived from the data in Figure 8a for 2 aqueous solutions of NiCl_2 . (Reproduced with permission from ref. 6. Copyright 1977, the Royal Society of Chemistry: Cambridge, UK).

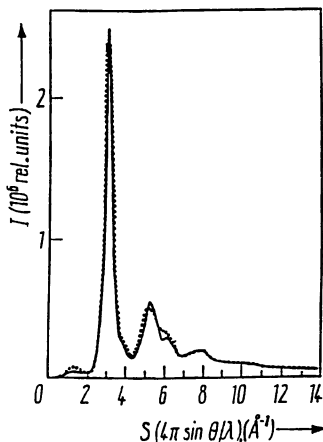


Figure 8c. X-ray intensity functions of Metglas 2826, before and after annealing at 628 K for 30 min (intensity corrections as described in the text); the intensity oscillations between 13 and 17 \AA^{-1} were within the noise level; intensity measurements between 17 and 22 \AA^{-1} were made by a resonance method - to be published ($\text{Ag}_{k\alpha_1}$ rotating anode / Ru fluorescence filter) - showed again a smooth decrease, without marked oscillations as those seen in intensity functions derived from EDXS measurements); --- as received, — annealed.

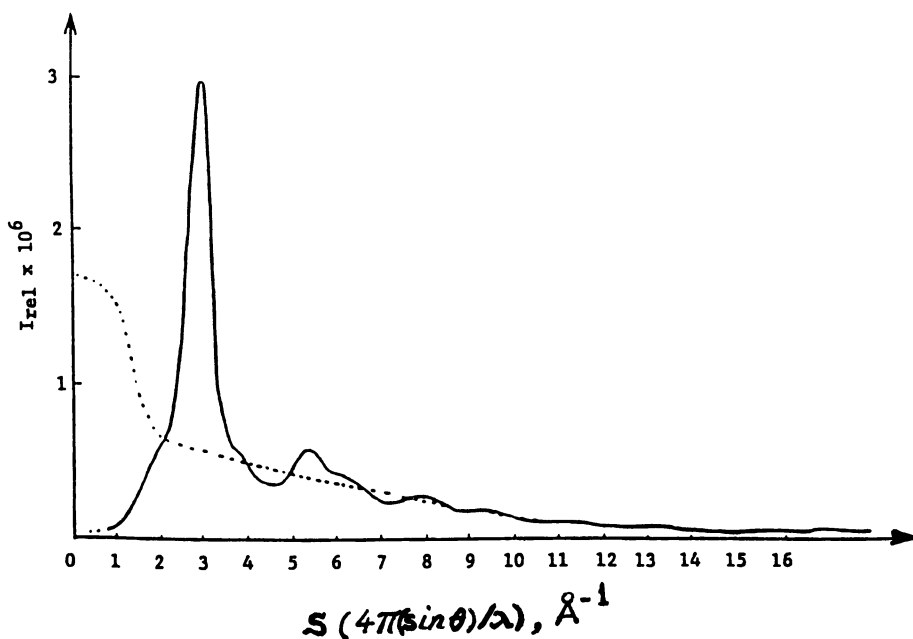


Figure 8d. The X-ray interference function of noncrystalline $\text{Co}_{0.9}\text{P}_{0.1}$ at 293 K, obtained by energy-dispersive techniques.

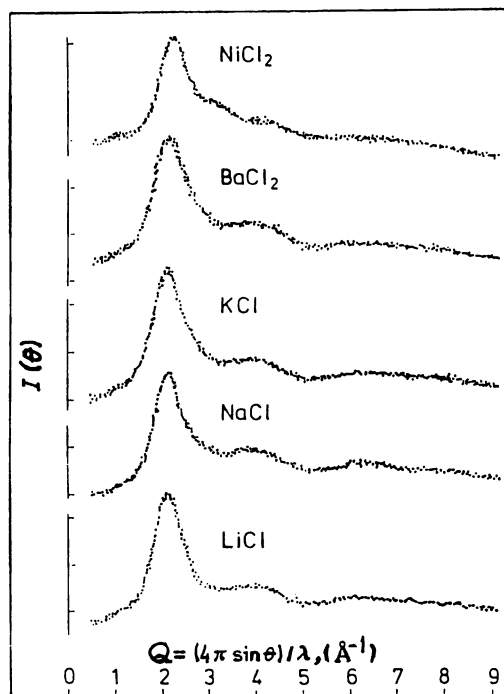


Figure 9a. Intensity (in arbitrary units) versus scattering angle of thermal neutrons for a variety of solutes in heavy water. All the solutions are close to saturation. The patterns are similar because the bulk of the scattering arises from water. (Reproduced with permission from ref. 11. Copyright 1978, IOT Publishing: Bristol, UK).

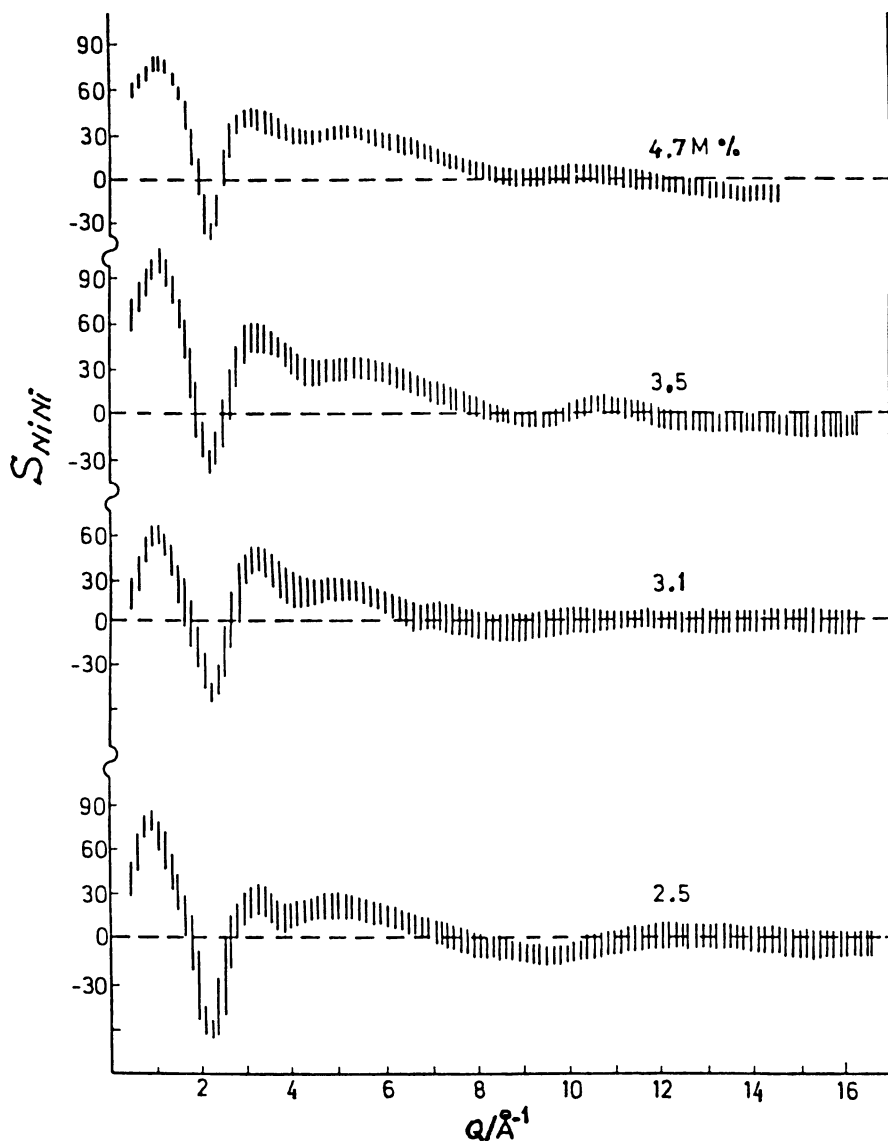


Figure 9b. The variation with Q of the partial structure factor S_{NiNi} for aqueous solutions of $NiCl_2$ of various percentage compositions. (Reproduced with permission from ref. 48. Copyright 1973, Elsevier Science Publishers B.V.: Amsterdam, The Netherlands).

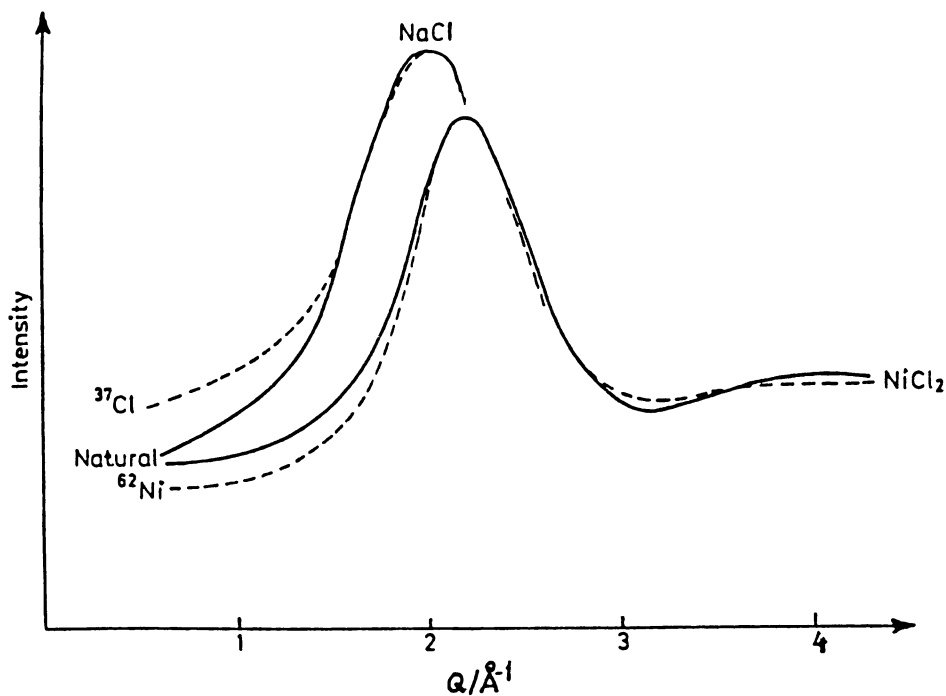


Figure 9c. The isotope substitution effect on neutron scattering intensities for NaCl and NiCl₂. (Dotted lines are for the substituted isotopes ^{37}Cl and ^{62}Ni ; from ref. 20).

cation (Ni^{2+}). This difference in the neutron scattering intensity for different isotopes gives the basis for the 'difference method' of local structure determination of electrolyte solutions by neutron scattering combined with isotopic enrichment. For a salt MX_n dissolved in D_2O , with M^+ being the cation and X^- being the anion, one can calculate the effect of isotopic substitution of the cation on the neutron scattering intensity in absolute units. The difference in neutron scattering cross-sections between two identical solutions but with different isotopes of the cation, M^+ , can be written as (15):

$$\Delta_{\text{M}}(\text{Q}) = \Delta_{\text{M}}^{\text{I}}(\text{Q}) + \text{CORRECTION TERMS} \quad (3)$$

where the correction terms are inelastic and incoherent scattering terms, and $\Delta_{\text{M}}^{\text{I}}(\text{Q})$ is the difference between the **coherent** neutron scattering intensity components of the two D_2O solutions made with different isotopes of M . The function $\Delta_{\text{M}}^{\text{I}}(\text{Q})$ is called the **first-order difference** and can be calculated in terms of the atomic fraction c of M , appropriate (coherent) scattering factors, b_{M} , b_{M}^1 , b_{O} , b_{D} , and partial structure factors $S_{\alpha\beta}(\text{Q})$:

$$\Delta_{\text{M}}^{\text{I}}(\text{Q}) = A_1[S_{\text{MO}}(\text{Q})-1] + B_1[S_{\text{MD}}(\text{Q})-1] + C_1[S_{\text{XM}}(\text{Q})-1] + D_1[S_{\text{MM}}(\text{Q})-1] \quad (4)$$

where A_1 , B_1 , C_1 , and D_1 are expressed in terms of c , n and the appropriate products of **coherent scattering amplitudes**, b (in 10^{-12} cm units). Thus,

$$A_1 = (2/3)c(1-c-n \cdot c) \cdot b_{\text{O}}(b_{\text{M}} - b_{\text{M}}^1), \text{ for the } S_{\text{MO}} \text{ term}, \quad (5)$$

$$B_1 = (4/3) \cdot c(1-c-n \cdot c)b_{\text{D}} \cdot (b_{\text{M}} - b_{\text{M}}^1), \text{ for the } S_{\text{MD}} \text{ term}, \quad (6)$$

$$C_1 = 2nc^2 \cdot b_{\text{X}}(b_{\text{M}} - b_{\text{M}}^1), \text{ for the } S_{\text{XM}} \text{ term}, \quad (7)$$

and

$$D_1 = c^2 \cdot [(b_{\text{M}}^2 - (b_{\text{M}}^1)^2)], \text{ for the } S_{\text{MM}} \text{ term}, \quad (8)$$

For the isotopic substitution of the anion X^- , a **first-order difference**, $\Delta_{\text{X}}(\text{Q})$, can be written in the same form as the above equations, with the appropriate ($b_{\text{X}} - b_{\text{X}}^1$) difference appearing in the products of the coherent scattering amplitudes, instead of ($b_{\text{M}} - b_{\text{M}}^1$). As an example, the isotopic substitution of ^{37}Cl with ^{35}Cl would give a difference ($b_{\text{X}} - b_{\text{X}}^1$) of $(1.18 - 0.26) = 0.92 \times 10^{-12}$ cm, whereas the substitution of ^{62}Ni with natural Ni (mixture) would provide a difference ($b_{\text{M}} - b_{\text{M}}^1$) of $[1.03 - (-0.87)] = 1.90 \times 10^{-12}$ cm. Note that the method is twice as sensitive to the nickel cation substitution than to the Cl anion substitution. The use of the natural Cl isotope mixture and ^{37}Cl yields a difference ($b_{\text{X}} - b_{\text{X}}^1$) of only 0.70×10^{-12} cm.

In order to derive the $S_{XX}(Q)$ and $S_{MM}(Q)$ to test for ordering of the ions in solution one would need to employ **three** different isotopes for either S_{XX} or S_{MM} determination. The cross-term $S_{MX}(Q)$ requires **four** different isotopes. Furthermore, a second-order difference method is required to determine S_{XX} , S_{MM} or S_{MX} .

The ionic hydration is best understood in real, rather than scattering space; therefore, the Fourier transforms (FT) of $\Delta_M^i(Q)$ and $\Delta_X(Q)$ are required to obtain the corresponding distribution function, G . For the cation, M^+ :

$$G_M(r) = 1/2\pi^2 g(r) \cdot \int \Delta_M^i(Q) \cdot Q \sin(Q \cdot r) \cdot dr \quad (9)$$

or

$$G_M(r) = A_1(g_{MD}(r) - 1) + D_1(g_{MD}(r) - 1) \quad (10)$$

Because A_1 and B_1 are much greater than C_1 and D_1 one has that

$$G_M(r) = A_1(g_{MO}(r) - 1) + B_1(g_{MD}(r) - 1) \quad (11)$$

which would allow one to locate the water nuclei (O and D) around the cation M^+ . Examples for Ni^{2+} and Li^+ are shown, respectively, in Figures 10a, 10b and 10c for concentrated $NiCl_2$ and $LiCl$ solutions in D_2O (16,17). A similar determination was carried out for the Cl^- anion in $NaCl$ and $CaCl_2$ solutions in D_2O (17); the weighted distribution $G_{Cl}(r) = A_2(g_{ClO}(r) - 1) + B_2(g_{ClD}(r) - 1) + C_2(g_{ClCl}(r) - 1) + D_2(g_{ClCl}(r) - 1)$ is shown in Figure 11b for a 4.5 mol solution of $CaCl_2$ in D_2O . Related X-ray scattering data and calculations are shown in Figures 11a. The first-order difference, $\Delta_{Ni}(Q)$, from which the weighted distribution $G_{Ni}(r)$ was obtained by FT (Figure 10a), is shown in Figure 12; this difference is well-behaved at high Q (18).

The main features of the weighted distribution function $G_{Ni}(r)$ (Figure 10a) are the Ni-O peak near 2.1 Å, the Ni-D peak near 2.7 Å and the second nearest-neighbor peaks near 4.5 Å. The water coordination in the first hydration shell of Ni^{2+} is shown above the $G_{Ni}(r)$ curve and involves six water molecules that have a tilted DOD bisector axis by an angle $\phi = 34 \pm 8^\circ$ to the Ni-O axis. In solutions of concentrations lower than 1 molal the Ni-D distance increased suggesting a distortion of the hydration shell of Ni^{2+} as the packing fraction of hydrated ions is increased (16). This observation may help with the selection of more realistic choices of the interionic potential than those currently employed in MD or MC simulations. The coordination of water around the chloride anion (Figure 11b) also involves a tilting of the water molecule OD axis from the Cl-O axis by an angle ψ of 5° or less (15, 16). The Cl-O nearest-neighbor peak in Figure 11b had a maximum at 3.20 ± 0.04 Å in agreement with the X-ray scattering data for $CaCl_2$ solutions in H_2O (18). For $NiCl_2$ solutions the Cl^- was apparently absent from the first coordination shell of Ni^{2+} . A second hydration shell around Ni^{2+} was also suggested by both neutron diffraction and quasi-elastic neutron scattering studies (15, 20, 21).

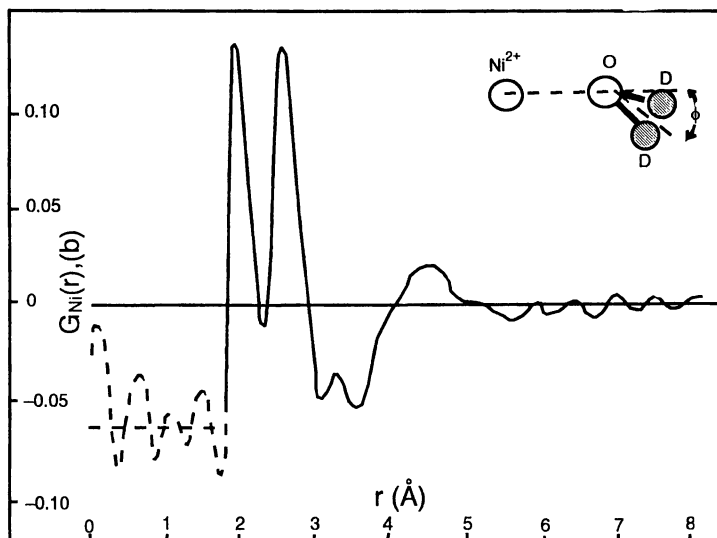


Figure 10a. The weighted distribution function $G_{Ni}(r) = A1(g_{NiO} - 1) + B1(g_{NiO} - 1) + C1(g_{NiCl} - 1) + D1(g_{NiNi} - 1)$ for a 4.41 molal solution of $NiCl_2$ obtained by taking the Fourier transform of the data in figure 9a. At this concentration the coefficients $A1 \dots D1$ are respectively 17.4, 40.0, 5.05 and 0.32 mol. The conformation of $Ni-D_2O$ consistent with this $G_{Ni}(r)$ is shown at the top right. (Reproduced with permission from ref. 11. Copyright 1978, IOT Publishing: Bristol, UK).

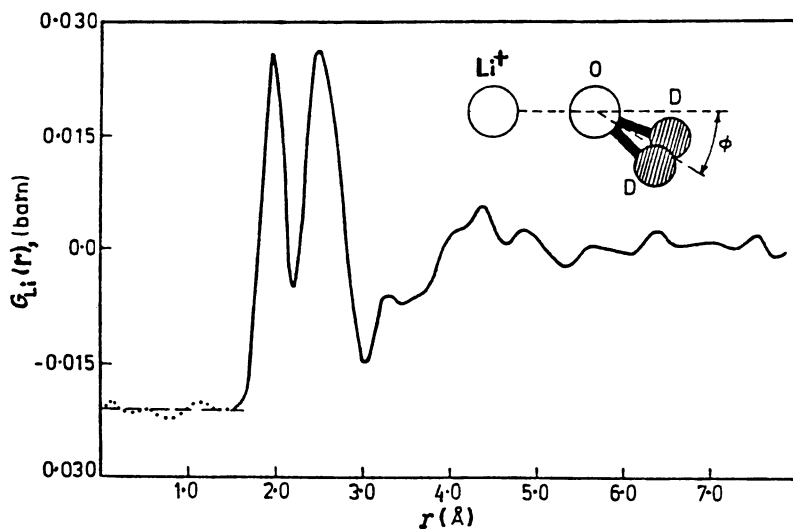


Figure 10b. $G_{Li}(r)$ for a 3.57 molal solution of $LiCl$ in D_2O . Top right: the $Li-D_2O$ conformation consistent with $G_{Li}(r)$ shown in figures 1 and 2 was that with the ϕ values given in table 2 of ref. 15. (Reproduced with permission from ref. 15. Copyright 1980, the Institute of Physics: London, UK).

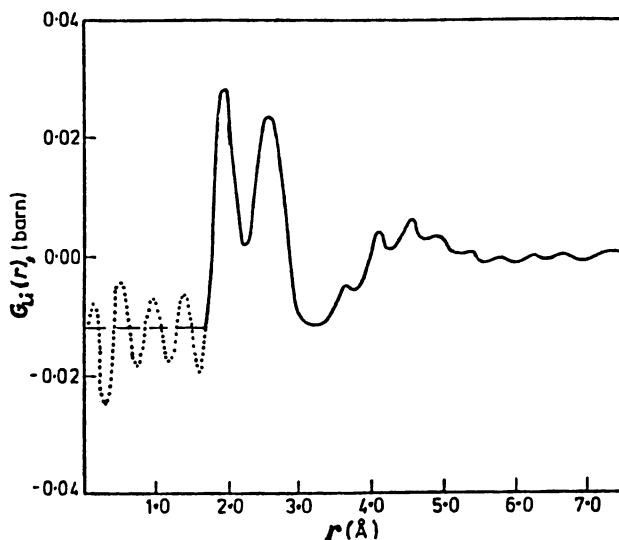


Figure 10c. Hydration structure of Li^+ and Cl^- ions in concentrated solutions in D_2O derived from neutron scattering isotope difference measurements. (Modified from ref. 11).

The analysis of the first-order difference results (11) for concentrated LiCl solutions in D_2O showed that the orientation of D_2O around Cl^- and Li^+ (Figure 10b) is similar to that of D_2O in NiCl_2 solutions (Figure 10a). The hydration number for Li^+ was strongly concentration dependent and increased to 5.5 below 3.6 mol (Table I); these neutron diffraction results did not seem to support tetrahedral coordination of D_2O around Li^+ . The possibility still exists that a hydration of '4+2' is involved, with a first hydration shell which is tetrahedral and a second hydration "shell" of two D_2O molecules differently oriented in comparison with either the water molecules in the bulk or in the first hydration shell of Li^+ . It was also found that **water-bridging** between Li^+ and Cl^- ions occurs at the higher concentrations of LiCl , in agreement with the results of previous work by pulsed ^1H NMR on $\text{LiCl} \cdot n\text{H}_2\text{O}$ glasses at 100 K (3). The first-difference neutron diffraction results for aqueous solutions of concentrated electrolytes disagree with the results of MD simulations of electrolyte solutions in water which employed the ST2 water potential (11); the latter favor a ϕ -tilt angle of water around cations of 55° . The results from other MD calculations were also compared with the local structure of LiCl solutions in D_2O derived from the neutron diffraction data with the first-order difference; partial agreement with some of the MD simulations was found (30), (Table II). Previous X-ray and neutron scattering results for aqueous solutions of LiCl (18) arrived at somewhat different conclusions

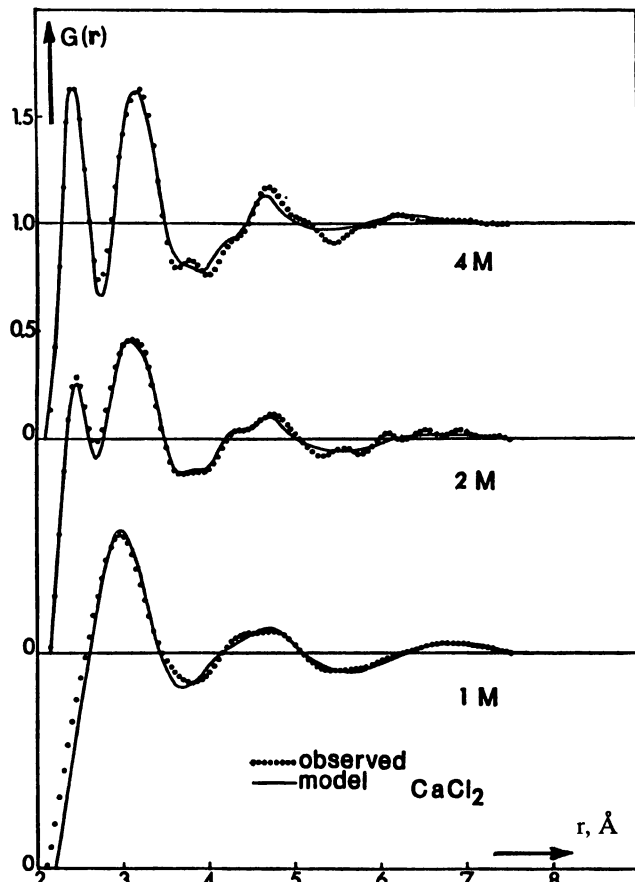


Figure 11a. Radial distribution functions for CaCl_2 , derived from X-ray scattering studies of concentrated (1M, 2M and 4M) CaCl_2 studies in water. (Reproduced with permission from ref. 18. Copyright 1976, the American Institute of Physics).

based on the total distribution functions, $g(r)$, (Figures 13a and 13b, respectively). At high LiCl concentrations ($n \leq 16$) it was shown (43) that the nearest-neighbor tetrahedral coordination of hydrogen-bonded water molecules was absent because there was **no peak** maximum near 2.8 Å in the $g(r)$ derived from the X-ray scattering measurements. The second-neighbor peak for tetrahedrally H-bonded water near 4.5 Å was also **absent** for $n \leq 16$ in $\text{LiCl} \cdot n\text{H}_2\text{O}$ solutions. H-bonded, tetrahedral water species were present, however, in the more dilute $\text{LiCl} \cdot n\text{H}_2\text{O}$ solutions, with $n \geq 33$, suggesting the presence of **two different environments** for water at these LiCl concentrations; the first type of water coordination was that characteristic of the Li^+ and Cl^- (separate) hydration shells, whereas the second type was characteristic of bulk, liquid water.

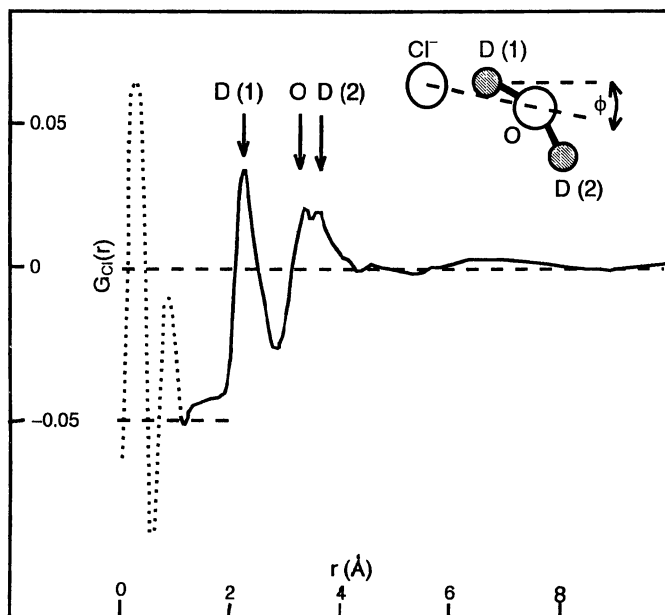


Figure 11b. The weighted distribution $G_{Cl}(r) = A2(g_{ClO} - 1) + B2(g_{ClD} - 1) + C2(g_{ClCl} - 1) + D2(g_{ClCl} - 1)$ for a 4.49 mol solution of $CaCl_2$. At this concentration the coefficients $A2 \dots D2$ are respectively 16.0, 36.8, 1.16 and 3.77 mb. The conformation of $Cl-D_2O$ consistent with this $G_{Cl}(r)$ is shown at the top right. (Reproduced with permission from ref. 11. Copyright 1978, IOT Publishing: Bristol, UK).

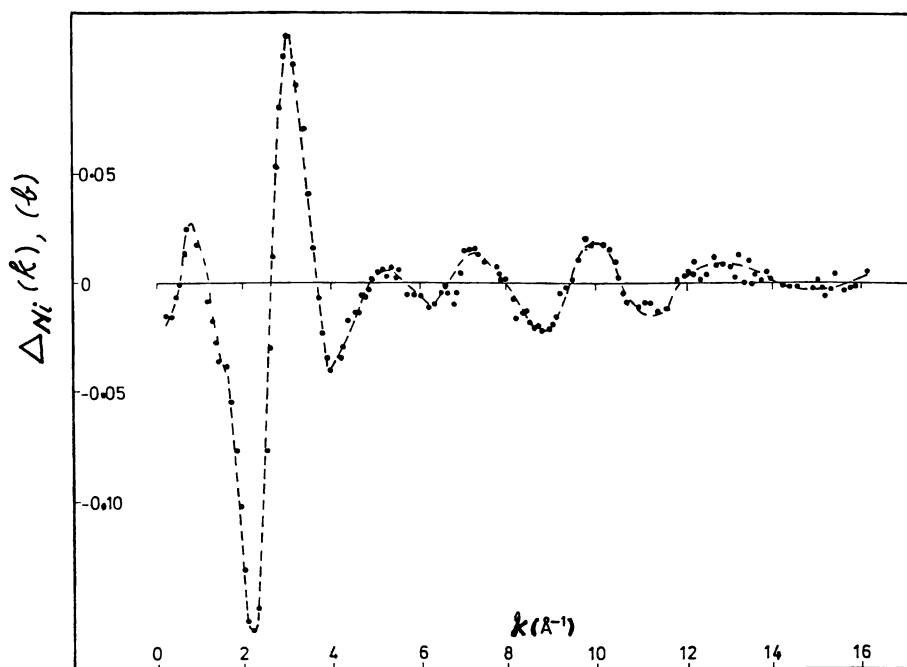


Figure 12. The first order difference for two 4.41 molal NiCl_2 solutions made from ^{62}Ni and $^{\text{nat}}\text{Ni}$ (native mixture). It should be noted that this difference is well behaved at high value of k . (Reproduced with permission from ref. 11. Copyright 1980, IOT Publishing: Bristol, UK).

Table I. Hydration of Li^+

Mol	Lithium-Oxygen Distance	Lithium-Deuterium Distance	ϕ^a	Hydration Number
9.95	1.95 ± 0.02 ; 2.1 ± 0.05	2.50 ± 0.02 ; 2.6 ± 0.05	$52^\circ \pm 5^\circ$; $35^\circ \pm 5^\circ$	3.3 ± 0.5 ; 4 ± 0.5
3.57	1.95 ± 0.02	2.55 ± 0.02	$40^\circ \pm 5^\circ$	5.5 ± 0.3

^a ϕ is the angle between the plane of the water molecule and the Li-O axis. It was assumed that r_{OD} is 1 Å and DOD was 105.5° in ref. 3, whereas DOD was assumed 109.5° in ref. 3 and 5, and r_{OD} was determined to be 0.95 Å.

SOURCE: Reprinted with permission from ref. 15. Copyright 1984.

An **independent-hydration** model was proposed for the hydration of ions and for the analysis of the diffraction data for $\text{LiCl} \cdot n\text{H}_2\text{O}$. The geometry of water coordination in this model is as specified in Figure 14. Additional results obtained with this model from X-ray scattering data for LiCl , LiBr , CaCl_2 and CaBr_2 solutions (19) are summarized in Table III.

The neutron diffraction data for $^7\text{LiCl} \cdot n\text{D}_2\text{O}$ solutions (21) were analyzed with the same model (19) and the corresponding $g(r)$'s are shown in Figure 13b. The first OD peak has a maximum near $r = 1 \text{ \AA}$ in Figure 13b; the comparison of Figures 13a and 13b provides a rough estimate of the Cl-O and Cl-Cl pair-correlation contributions to $g(r)$ of $^7\text{LiCl} \cdot n\text{D}_2\text{O}$ for $n \leq 16$. The peak near $r = 3.2 \text{ \AA}$ in the $g(r)$ of molten $^7\text{LiOD}$ is broader and of lower amplitude than the corresponding peak for $^7\text{LiCl} \cdot 3.0\text{D}_2\text{O}$ or $^7\text{LiCl} \cdot 4.1\text{D}_2\text{O}$ in Figure 13b. The **tetrahedrally** coordinated oxygen atoms around Li^+ are, therefore, only partially responsible for this peak in the $g(r)$ of $^7\text{LiCl} \cdot n\text{D}_2\text{O}$. The Cl-O pair correlation is expected to contribute also to the 3.2 \AA , asymmetric peak in the $g(r)$ of $^7\text{LiCl}$ solutions in D_2O ; note also that the 3.2 \AA peak in the $g(r)$ of molten $^7\text{LiOD}$ is, on the other hand, **symmetric**. The 3.2 \AA distance for the Cl-O pair correlation nearest neighbor peak would be consistent with the randomly distorted octahedral coordination of six water molecules packed around the Cl^- , as suggested schematically in Figure 14c. The "independent-hydration" model, however, neglected the **water-bridging between Li^+ and Cl^-** which becomes important in concentrated ($n \leq 16$) $\text{LiCl} \cdot n\text{H}_2\text{O}$ (or D_2O) solutions; the correlation radius, derived from Figures 13a and 13b extends to about 7 \AA in the concentrated solutions and is about the same as in liquid water (or D_2O) at 293 K.

Molecular Dynamics of Aqueous Solutions of Electrolytes. Monte Carlo Simulation of Ion-Water Clusters. The isolated ion-water clusters are relatively simple models for aqueous solutions of electrolytes. Their study is likely to provide potential functions and parameters for the interactions of ions such as Li^+ , Na^+ , K^+ , F^- , Cl^- , Br^- , I^- and Ca^{2+} with water, which could be of use also in MD or MC simulations of aqueous solutions of such electrolytes. The isolated ion-water clusters for alkali and halide ions were intensively studied experimentally by mass spectrometry (22-23) which provided estimates of the Gibbs free-energies, $\Delta G^\circ(N-1, N)$ for the formation of ion-water complexes, as well as enthalpies of reaction for the ionic hydration. There have been several attempts to calculate these quantities by employing empirical potential function techniques and quantum mechanical methods, most of which were based on two-body potential functions. The neglect of the many-body effects causes the calculated values of the interaction energy to be **too negative** in comparison with the experimental values, and this effect increases with cluster size or ionic field strength.

Recently, non-additive potentials were employed together with methods of statistical mechanics that involve MC computations of the free energy perturbation in terms of a series of moments of the potential energy function (24). The total

Table II. Lithium-Water Configurations in LiCl Solutions

Mol or Ion-Water Ratio	r_{LiO} (Å)	r_{LiH} (Å)	Hydration Number	Method
2.2 mol	2.10	2.60	7	Molecular dynamics (ST2)
2:50	2.20	3.30	5	Cluster calculation
2:200	2.00	2.60	5	Cluster calculation
Single Li ⁺ ion	1.81-1.89	—	—	Ab initio
1:6	—	—	4.0	Electrostatic Model
1:4	—	—	4.0	Experimental, gas phase
1:4	2.1 ± 0.05	2.6 ± 0.05	4 ± 0.5	X-ray/Neutron scattering
1:2 to 1:12	2.1 ± 0.05	2.6 ± 0.05	4 ± 0.5	Pulsed ¹ H NMR

Table III. Mean Distance, r_{CS} , and Mean Square Deviations, σ_{CS} , for Cation-Solvent Interactions in the Groups Cation (H₂O)_n, and Mean Distances, r_{AS} , and Mean Square Deviations, σ_{AS} , for Anion-Solvent Interactions in the Groups Anion (H₂O)_m^a

Solutions	n	m	r_{CS}	σ_{CS}	r_{AS}	σ_{AS}
LiCl • 13.90 H ₂ O	4	6	1.99	0.28	3.04	0.17
LiCl • 8.15 H ₂ O	4	6	1.95	0.25	3.10	0.20
LiCl • 6.44 H ₂ O	4	6	2.04	0.04	3.09	0.22
LiCl • 4.01 H ₂ O	4	6	2.22	0.31	3.18	0.19
LiCl • 24.97 H ₂ O	4	6	2.25	0.25	3.29	0.23
LiBr • 24.97 H ₂ O	4	6	2.14	0.25	3.29	0.25
LiBr • 10.83 H ₂ O	4	6	2.16	0.25	3.29	0.26
LiBr • 8.41 H ₂ O	4	6	2.16	0.25	3.29	0.26
CaBr ₂ • 44.05 H ₂ O	6	6	2.40	0.12	3.32	0.24
CaBr ₂ • 25.95 H ₂ O	6	6	2.44	0.15	3.34	0.26
CaCl ₂ • 55.82 H ₂ O	6	6	2.42	0.14	3.14	0.22
CaCl ₂ • 26.62 H ₂ O	6	6	2.41	0.15	3.14	0.23
CaCl ₂ • 12.28 H ₂ O	6	6	2.42	0.13	3.15	0.21

^aObtained from least-square refinement.

SOURCE: Reprinted with permission from ref. 18.

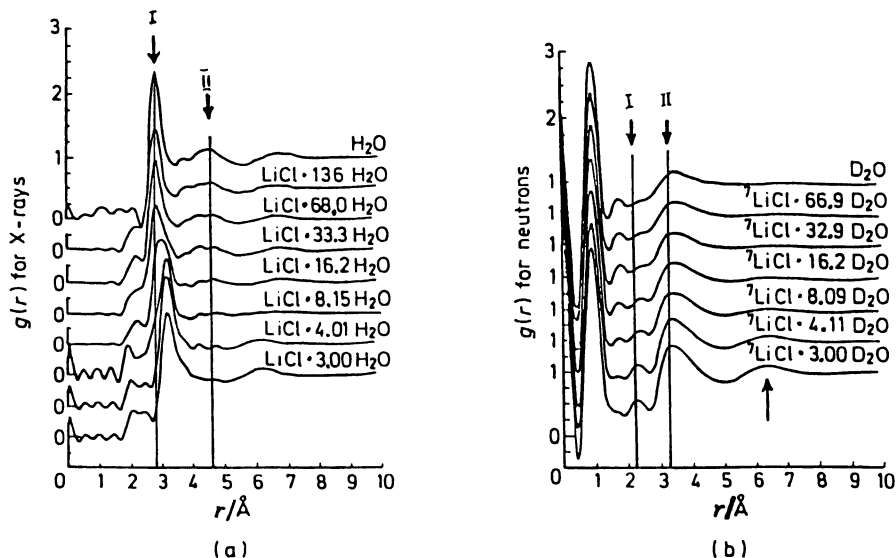


Figure 13. The radial distribution functions $g(r)$ measured with: (a) X-rays and (b) neutrons for aqueous solutions of lithium chloride of various concentrations. The individual curves are displaced vertically to avoid overlap. (Reproduced with permission from ref. 21. Copyright 1973, the American Institute of Physics).

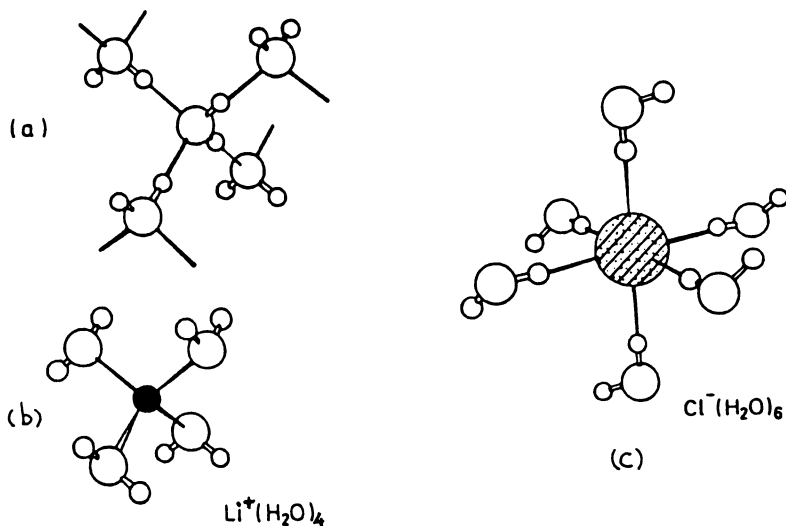


Figure 14. Diagram (a) shows the environment of water molecules in pure water. Diagrams (b) and (c) show, respectively, the four-fold and six-fold coordination by water molecules of the Li^+ and Cl^- ions in concentrated solutions (c). (Reproduced with permission from ref. 21. Copyright 1973, the American Institute of Physics).

interaction energy for the ion-water cluster was expressed (24) as:

$$E_{\text{tot}} = E_{\text{water-water pair}} + E_{\text{water-ion pair}} + E_{\text{ex-ion-water-water}} + E_{\text{pol}} \quad (12)$$

where $E_{\text{water-water pair}}$ is a RWK2 pairwise additive potential, modified to include nonadditive terms. Such a model yields reasonable values for the second virial coefficient and water-dimer energy in the gas phase, as well as lattice densities and energies for ice I_h and ice VII. The polarization term was (44):

$$E_{\text{pol}} = (1/2) \sum_j a_j [E_j \cdot E_j] \quad (13)$$

where a_j are the **polarizabilities** of the polarizable centers within the clusters, and E_j is the electric field created at a center j by the surrounding molecules. MC simulations were carried out at 298 K for Na^+ , K^+ , Mg^{2+} , F^- , and Cl^- ions hydrated with a variable number of water molecules between 1 and 6. The best agreement with experimental values of the enthalpy was obtained for the larger ion, K^+ , the least satisfactory agreement was for Li^+ , although the results were improved over those obtained previously with two-body potentials. The structural features of the ion-water clusters obtained with the nonadditive potentials are summarized in Table II. The first hydration shell of Li^+ was tetrahedral ($n_H = 4.0$) and there were 2 additional water molecules outside the first hydration shell of Li^+ . The Na^+ had a well defined '5+1' structure, whereas the K^+ hexahydrate complex was between the '4+2' (like $\text{Li}^+(\text{H}_2\text{O})_6$) and the '5+1' structure (like $\text{Na}^+(\text{H}_2\text{O})_6$). Fluoride hydration involved 4 water molecules in a single hydration shell, whereas Cl^- had a well defined structure of the type '3+1'. The water coordination around F^- or Cl^- was not, however, tetrahedral but there were instead 'favorable' water-water interactions present only around the anions. It was suggested that in electrolyte solutions the coordination number would be higher than for the isolated ion-water clusters, because in solution the addition of water molecules to the first coordination sphere of ions would be energetically compensated more effectively by polarizing water molecules that are located in the second hydration shell (24). The computation method employed for isolated ion-water clusters could also be applied to study ionic solutions, and it will be interesting to see such results, especially since at the moment the only extensive MD reported for ionic solutions are with the MCY or earlier potentials (25, 26).

Orientational Correlation Functions for Ionic Solutions in Water (42). The second-rank orientational correlation function $C_2^Z(t)$, ($\ell = 2$ in reference (1)) were also calculated in MD simulations with the MCY potential (2) for hydrated ions and showed similar time dependences (Figures 7 and 8) in reference (2)) with that found in liquid water (Figure 1); the exponential decay for $t > 0.1$ ps was, however, slower for all the ions investigated ($[\text{Li}^+]_{\text{aq}}$, $[\text{F}^-]_{\text{aq}}$ and $[\text{Cl}^-]_{\text{aq}}$) in comparison with liquid H_2O , and the τ_r values increased with decreasing ionic radius (increasing ionic field strength), as expected. Calculated reorientation times τ_r for the various hydrated ions (26) are summarized in Table IV corresponding to the $\alpha_2(t)$, ($\alpha = x, y$ or z) orientational correlation functions for ionic solutions in water around 0.8 M. Note

that the value of τ_r^y (or " τ_2^y ") for liquid H₂O (2.01 ps) at 286 K is smaller by about a factor of 2 than the experimental value of τ_c of water determined by NMR relaxation measurements at 286 K. The dipolar correlation functions were also calculated for ionic solutions in water and the values of the dipolar correlation time ("relaxation time"), τ_1^z , are summarized in Table V. Note that $\tau_2^z > \tau_2^y$ for liquid H₂O at 286 K; however, the calculated $\tau_1^z / \tau_2^y = 1.9$, which is less than the expected value of 3 expected for this ratio from the experimental $\tau_{\text{dielectric}} / \tau_c^{\text{NMR}}$ value, as well as from dielectric relaxation theories (27, 28). Both values of calculated τ_1^z and τ_2^y are too short in comparison with $\tau_{\text{dielectric}}$ and τ_c^{NMR} . This may be caused by the fact that **collective** effects are ignored in the MCY potential, which are however, important in the OCF's involved in dipole reorientations and dielectric relaxation in water.

Table IV. Calculated ℓ -2 Reorientation Times^a

Ion	Temperature (K)		Time (ps)		
			X	Y ^b	Z
Li ⁺	297	A ₂ T ₂	3.7	5.8	8.0
Na ⁺	298	A ₂ T ₂	1.4	2.3	2.1
K ⁺	274	A ₂ T ₂	1.1	2.1	1.4
F ⁻	278	A ₂ T ₂	3.9	4.8	5.2
Cl ⁻	287	A ₂ T ₂	1.1	2.1	1.4
H ₂ O	286	A ₂ T ₂	1.03	2.01	1.07

^aDefined as the integral of the normalized correlation function.

^bThe set labeled Y is related to the NMR intramolecular dipole-dipole relaxation time. SOURCE: Reprinted with permission from ref. 26. Copyright 1982.

Table V. Reorientational Correlation Times (in ps)

Run	τ_1^x	τ_1^y	τ_1^z	τ_2^x	τ_2^y	τ_2^z	τ_{NMR}
1	9.0	16.3	10.3	4.3	7.0 (5.7)	4.4	
2	3.2	6.0	3.5	1.8	2.7 (2.1)	1.7	3.6 ^a
3	3.6	6.3	3.6	2.0	2.8 (2.2)	1.8	4.8 ^b
4	3.1	5.7	3.8	1.7	2.7(2.0)	1.8	
5	1.0	1.9	1.2	0.6	0.9 (0.6)	0.6	0.9 ^c

Bracketed numbers are results for A₂^zτ₂^y.

^aH₂O at 283 K; ^bD₂O at 283 K; ^cH₂O at 363 K.

SOURCE: Reprinted with permission from ref. 26. Copyright 1982.

Residence Times of Water in the Hydration Shell of Ions. The residence time, τ_R , is defined as the **mean time** for which a water molecule remains in the first hydration (or coordination) shell of an ion before exchanging with the bulk water population. In the absence of a bulk water population, as in the case of very concentrated solutions, τ_R^{CN} would be determined by the mean time in which a specified water molecule diffuses outside the nearest-neighbors (coordination) peak in the ion-water oxygen pair correlation function, $g_{ion-oxygen}$, discussed in the previous section. For example, a residence time of about 30 ps was estimated for water in the hydration shell of Li^+ based on NMR relaxation measurements (28). The MD simulations can also estimate the residence time either from the van't Hove (distinct) space-time correlation, $G_{ion-oxygen}^d(t, t)$, or by defining (23) a function $Q_j(t, t_n; t^*)$ for a water molecule j ; the latter function takes the value 1.00 if the molecule j is within the first peak of $g_{ion-oxygen}(r)$, (the first hydration shell of the ion) at both timesteps t_n and $(t+t_n)$, and it is zero otherwise. If the water molecule resides outside the hydration shell for only very short times $t^* \ll t_n$, then $Q_j(t, t_n; t^*) = 1.00$. A suitable value for t^* would be around 2 ps, the time interval in which water molecules in the bulk exchange neighbors, (t^* is a sort of " τ_R " in bulk water). The exchange of the water molecules with the bulk water is then characterized by:

$$n_{ion}(t) = 1/N_t \sum_{n=1}^{N_t} \sum_j Q_j(t_n, t; t^*) \quad (14)$$

where there are n time steps in the MD simulation (25). At $t = 0$, $n_{ion}(0)$ is the average coordination (or hydration) number of the ion; $n_{ion}(t')$ would give the mean number of water molecules that are left within the hydration shell at the time t' . For both alkali and halide ions, the function $n_{ion}(t)$ decays exponentially at short times:

$$n_{ion}(t) = n_{ion}(0) \cdot \exp(-t/\tau_R) \quad (15)$$

where $n_{ion}(0) = n_H$ is the average hydration number of the ion. Calculated values of the residence times for alkali halide solutions in water are summarized in Table VI, and were found to be in quite good agreement with values estimated from NMR relaxation measurements.

Velocity Correlation Functions of Ions in Aqueous Solutions. The VCF's of water in various solutions of alkali halides were calculated for water in the first hydration shell of the ions (26). Their time dependence is not much different from that of liquid water. On the other hand, the VCF's of the ions in aqueous solutions (Figure 15) are quite different from the VCF of water (Figure 9a and 9b in reference 1). Upon FT of the VCF's of $[Li^+]_{aq}$ and $[Na^+]_{aq}$ shown in Figure 15 new frequencies appear near 400 cm^{-1} for $[Li^+]_{aq}$, near 260 cm^{-1} for $[Na^+]_{aq}$ (and $[F^-]_{aq}$). The peak frequency seems to be correlated with the reciprocal of the ionic radius. Raman studies by Nash et al. (29) for solutions containing $[Li^+]_{aq}$ reported a band

near 380 cm^{-1} which was assigned to the F_2 mode of the Li^+ ion "in its tetrahedral cage" (of water molecules), and a band near 440 cm^{-1} assigned to the "A1 mode of the coordination complex", (the so called 'breathing' vibration of coordinated water). Somewhat similar results were obtained in very concentrated $\text{LiCl}\cdot n\text{H}_2\text{O}$ solutions ($2 < n < 12$) at 298 K. If these assignments (29) were correct, than the "power" spectra calculated by FT of VCF's for the hydrated ions would be in reasonable agreement, in terms of the peak position for $[\text{Li}^+]_{\text{aq}}$. Such results are discussed next in more detail.

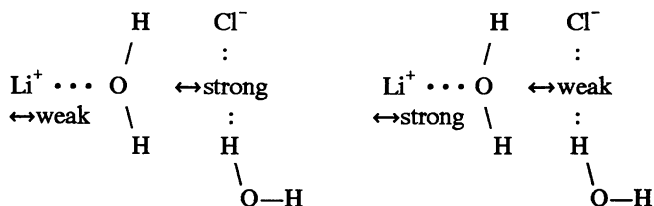
Table VI. Residence Times Calculated for Monovalent Ions and Water.

Ion	T (K)	τ_R (ps)
Li^+	278	33
Li^+	368	6.0
Na^+	282	10
K^+	274	4.8
F^-	278	20
Cl^-	287	4.5
H_2O	286	4.5

Raman Spectra of $\text{LiCl}\cdot n\text{H}_2\text{O}$ Solutions and Glasses. Since the MD simulations with the MCY potential for ionic solutions in water (31) did not consider the intramolecular vibrational modes of water molecules (water molecules were 'rigid'), the important OH-stretching region could not be compared with the corresponding bands in the Raman spectra of ionic solutions in water. Such a comparison between the theory and experiment was, however, made for molten alkali halides, and could also be carried out for the **vibrational correlation function** (26) $C_v(t)$, as well as for the **rotational autocorrelation function** (34), $C_r(t)$. In this context, it will suffice to consider the main characteristics of the Raman spectra of $\text{LiCl}\cdot n\text{H}_2\text{O}$ solutions and glasses; a quantitative analysis of such spectra will be reported elsewhere.

The OH-stretching region of the polarized (VV) Raman spectrum of a $\text{LiCl}\cdot 4.1\text{H}_2\text{O}$ solution at 293 K is presented in Figure 16a; the corresponding region of the VH Raman spectrum is shown in Figure 16b. The OH-stretching (VV) bands of liquid water at 293 K are shown for comparison in Figure 16c. The liquid water has a distinct band near 3200 cm^{-1} , whereas this band is nearly absent in the $\text{LiCl}\cdot 4.1\text{H}_2\text{O}$ solution Raman spectrum. Since the OH-stretching band near 3200 cm^{-1} was assigned to the tetrahedrally H-bonded water (HB) species, this implies that the tetrahedrally H-bonded water structure is 'broken', or absent, in the $\text{LiCl}\cdot 4.1\text{H}_2\text{O}$ solution, in full agreement with the r.d.f.'s derived from X-ray and neutron scattering studies of concentrated LiCl solutions in water as (D_2O), (Figure 13a and 13b). It is

rather interesting that, upon cooling the concentrated $\text{LiCl} \cdot n\text{H}_2\text{O}$ ($2.2 \leq n \leq 11$) solutions to 100 K, these form a glass, without phase separation/ice formation (5). The OH-stretching region of the Raman spectrum of a $\text{LiCl} \cdot 4\text{H}_2\text{O}$ solution at 293 K (Figure 16a) is compared with that of the corresponding glass at 119 K (Figure 16d), close to the 'rigid lattice' condition (in which any large amplitude motions are frozen out). Clearly, several OH-stretching bands increase in intensity at 119 K relative to the non-tetrahedral ("NHB") water bands near 3450 cm^{-1} and 3400 cm^{-1} in Figures 16d and 16e. (Their precise positions will require a Gaussian deconvolution of the Raman spectrum after appropriate intensity corrections.) Unlike the Raman spectrum of glassy water at 90 K (reference 1), the VV spectrum of glassy $\text{LiCl} \cdot 4\text{H}_2\text{O}$ at 119 K (Figure 16d) lacks the 3080 cm^{-1} band assigned to the ν_1^{sb} symmetric OH-stretch of tetrahedrally H-bonded water. This confirms the absence of ice I_b , cubic ice (II/III) and 'glassy' water in the $\text{LiCl} \cdot 4\text{H}_2\text{O}$ glass at 119 K. The intense band near 3200 cm^{-1} which was present in supercooled, liquid water at -33°C (240 K) is also absent in the Raman spectrum of $\text{LiCl} \cdot 4\text{H}_2\text{O}$ glass at 119 K, or it has lower amplitude than the other bands. This suggests very strongly that the local structure of the $\text{LiCl} \cdot n\text{H}_2\text{O}$ glasses is quite different from that of various ices, 'glassy' water, supercooled liquid water, or liquid water. The absence of tetrahedrally, H-bonded water in the $\text{LiCl} \cdot n\text{H}_2\text{O}$ glasses is, therefore, strongly suggested by these Raman scattering observations. Furthermore, significant differences are observed between the $\text{LiCl} \cdot n\text{H}_2\text{O}$ solutions and glasses in the OH-stretching region from 3250 cm^{-1} to 3450 cm^{-1} in the polarized Raman spectra. Such differences seem to suggest that the 'bent-bonds', responsible for the ν_1^{bb} , OH symmetric stretch (3370 cm^{-1} band), occur more frequently in the $\text{LiCl} \cdot 4\text{H}_2\text{O}$ glass at 119 K than in the corresponding solution. Both the $\text{LiCl} \cdot n\text{H}_2\text{O}$ solutions and glasses have in common with liquid water, supercooled water, 'glassy' water and the various ices, the OH-stretch band near $3,450 \text{ cm}^{-1}$ assigned to the ν_1^{wb} asymmetric stretch mode, with one strong and one weak bond. However, the $3,450 \text{ cm}^{-1}$ band had much lower height in 'glassy' water at 90 K than the tetrahedrally, H-bonded water band near 3080 cm^{-1} , whereas in $\text{LiCl} \cdot n\text{H}_2\text{O}$ solutions and glasses the $3,450 \text{ cm}^{-1}$ band is the tallest one and seems to dominate the spectrum. This suggests that the dominating water species in the $\text{LiCl} \cdot n\text{H}_2\text{O}$ solutions or glasses have the following bond scheme:



responsible for the ν_1^{wb} intense band, in agreement with the model of **water-binding**

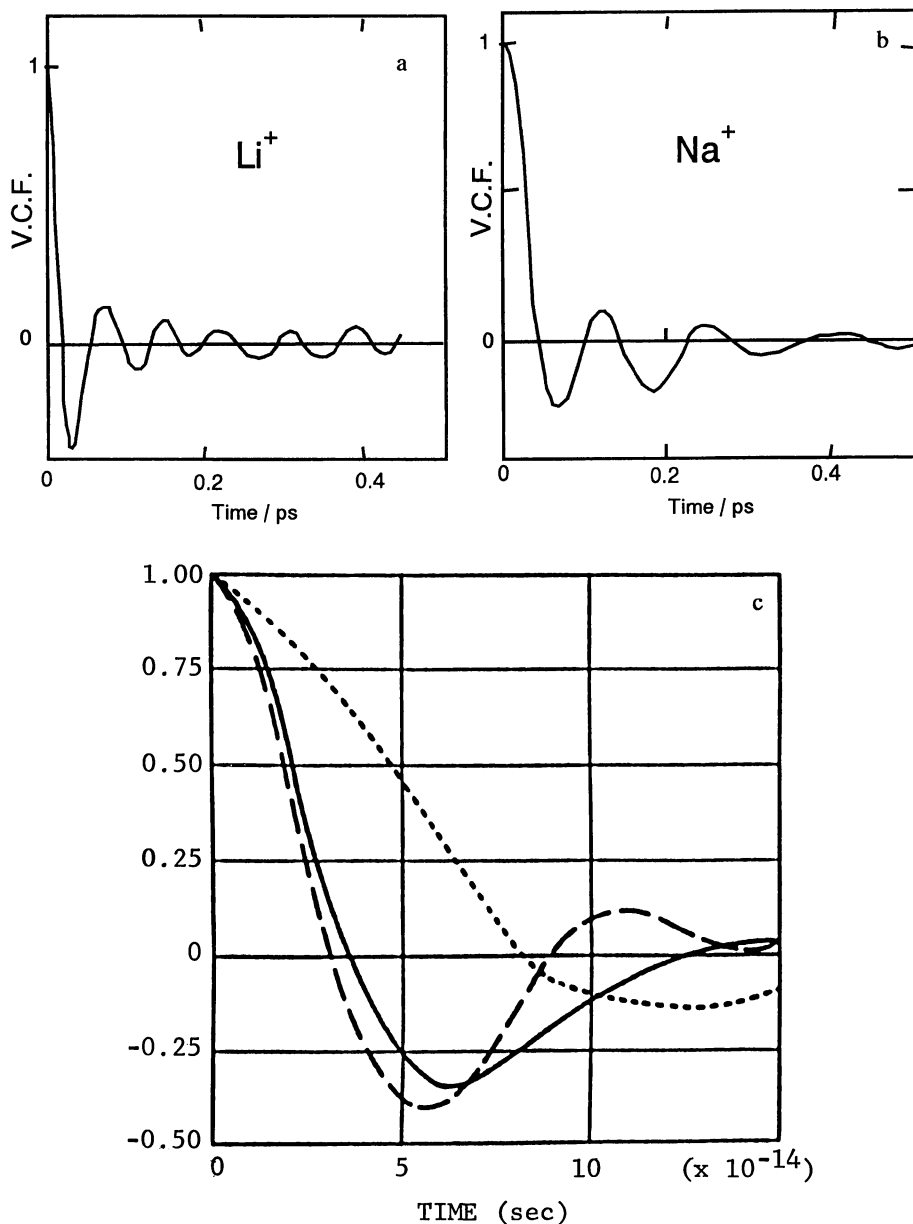


Figure 15. The VCFs of Li^+ , Na^+ , K^+ , I^- and Cl^- ions in aqueous solutions and molten salts. (a) VCF of Li^+ ; (b) VCF of Na^+ (modified from ref. 47); (c) VCF's for molten KCl; (d) comparison of Li^+ and I^- VCF's in molten LiI. (Reproduced with permission from ref. 43. Copyright 1975, Plenum Press).

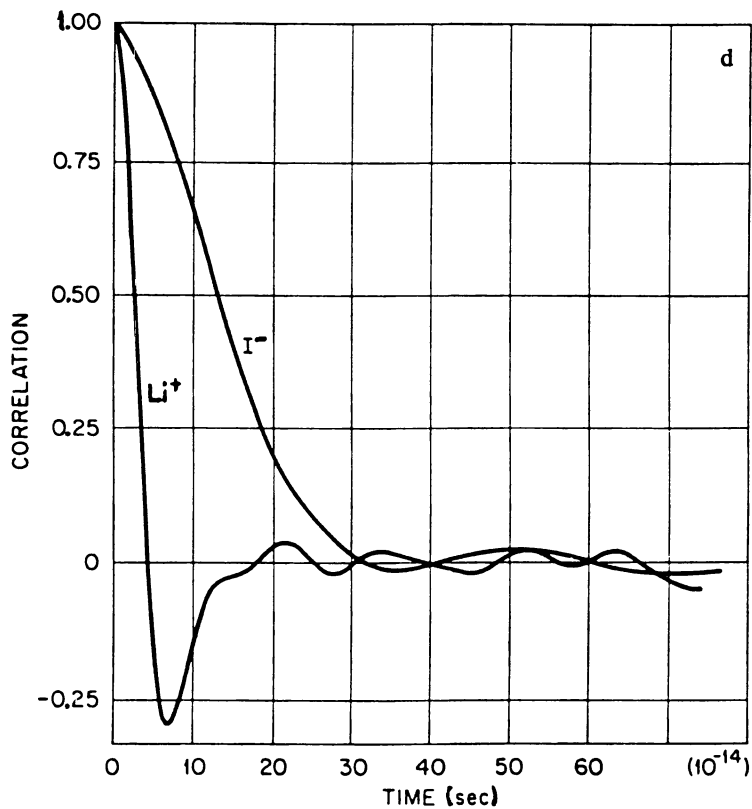


Figure 15. Continued.

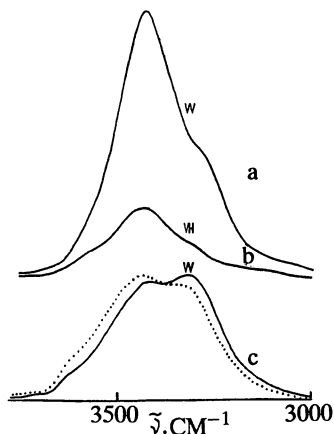


Figure 16a,b&c. Comparison of polarized Raman spectra of $\text{LiCl} \cdot 4\text{H}_2\text{O}$ solutions with corresponding spectra of liquid H_2O at 293 K. (a) VV Raman spectrum of $\text{LiCl} \cdot 4\text{H}_2\text{O}$ at 293 K. (b) VH Raman spectrum of $\text{LiCl} \cdot 4\text{H}_2\text{O}$ at 293 K, (c) VV Raman spectrum of H_2O at 293 K (Baianu, I. C. unpublished results).

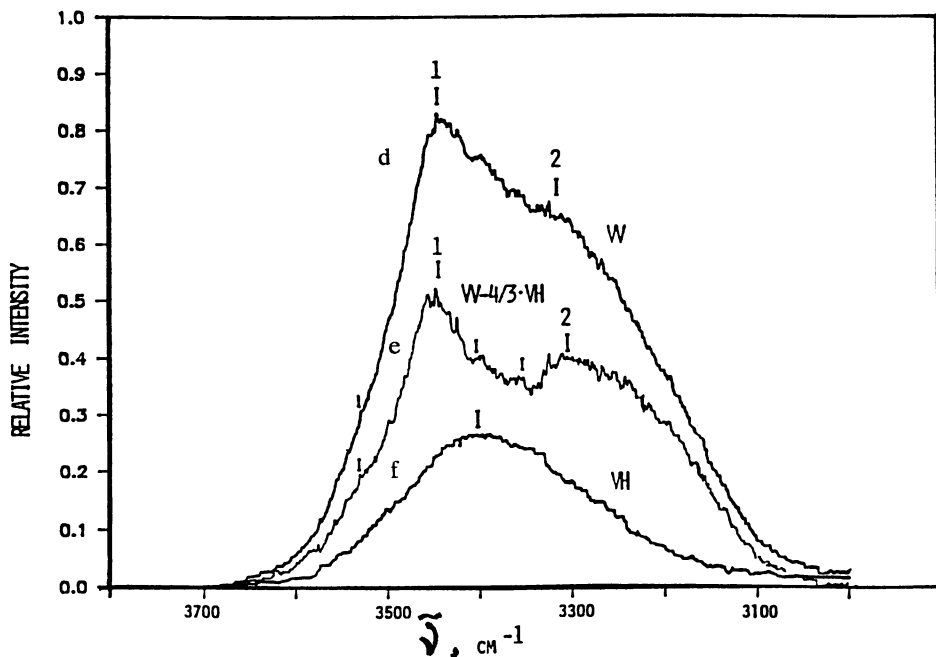


Figure 16d,e&f. (d) VV Raman spectrum of $\text{LiCl} \cdot 4\text{H}_2\text{O}$ at 119 K; (e) difference Raman spectrum, $(\text{VV} - 4/3 \cdot \text{VH})$ at 119 K for $\text{LiCl} \cdot 4\text{H}_2\text{O}$; (f) VH Raman spectrum of $\text{LiCl} \cdot 4\text{H}_2\text{O}$ at 119 K (Baianu, I. C., unpublished results).

between the cation and the anion (3) shown in Figure 1c. The structure of such **hydrated ion clusters** (Figure 1c) would involve **more 'bent'-bonds** in the $\text{LiCl}\cdot n\text{H}_2\text{O}$ glasses than in the corresponding solutions at 293 K (Figure 1b). Such 'bent'-bonds would result in **more compact**, and perhaps more regular, **water-bridged ion-pair clusters** in the $\text{LiCl}\cdot n\text{H}_2\text{O}$ glasses at low temperatures than in solutions at 293 K. An additional, OH-stretching band may also be present near 3610 cm^{-1} , similar to that found in molten LiOH (30). The polarized Raman spectra of $\text{LiCl}\cdot 4.1\text{H}_2\text{O}$ solutions included also a relatively strong band with a maximum at 1650 cm^{-1} (Figure 17) which was previously assigned to the ν_2 HOH bend-mode, in the same position as in liquid water. There was also present a weaker, broad and asymmetric band with a maximum at 3600 cm^{-1} in the VV Raman spectrum of $\text{LiCl}\cdot 4.1\text{H}_2\text{O}$ at 293 K. This 3600 cm^{-1} centered band is assigned to translational modes of water, and may also include the Li^+ libration inside its hydration shell ("F₂ mode" and the "A₁-mode" (29)). The 3600 cm^{-1} band could be related to the VCF of $[\text{Li}^+]_{\text{aq}}$ discussed in the previous section (Figure 15). The conclusion of this section is that water in concentrated ($2.2 \leq n \leq 12$) $\text{LiCl}\cdot n\text{H}_2\text{O}$ solutions and glasses has **unique** local structures, reflected in their polarized Raman spectra that have distinctive OH stretching band structures; such band structures are very different from those of liquid water, supercooled water, 'glassy' water and various ices (including ice I_h), contrary to several theoretical predictions, including MD calculations for the more dilute solutions (26).

Multinuclear Spin Relaxation Measurements in Relation to the Local Structure and Dynamics in Aqueous Electrolyte Solutions and their Glasses at low temperatures. Aqueous solutions of alkali halides were intensely studied by nuclear magnetic relaxation techniques; amongst the nuclei studied were: ^1H , ^2H , ^{17}O , ^6Li , ^7Li , ^{19}F , ^{23}Na , ^{87}Rb , ^{127}I and ^{133}Cs . In general, the relaxation rates for various nuclei increase with the value of the Sternheimer anti-shielding factor ($1 + \gamma_\infty$).

Proton spin-lattice (T_1) measurements at 60 MHz were reported for both ^7Li and ^6Li in LiI solutions in D_2O as a function of temperature (31). The variation of the longitudinal relaxation rate, $(1/T_1)$, with temperature exhibited a maximum at low temperature ($\sim 200\text{ K}$) for these concentrated solutions (Figure 18). Estimates were obtained for the relaxation contributions, such as the magnetic dipole-dipole interaction between ^7Li and ^1H , the H-D interactions, or the $^{127}\text{I}-^1\text{H}$ interaction. The latter estimates were employed in conjunction with a simple pair-distribution function for iodide and the water protons around I^- in 6M LiI solutions at 198 K in order to investigate the hydration of the I-anion. These results suggested an asymmetric hydration water configuration around I^- , with $a = 2.6 \pm 0.2\text{ \AA}$.

The model pair distribution function considered initially for $\text{I}^- \cdots \text{H}_2\text{O}$ was a delta-function; this was then replaced by a Gaussian distribution that provided a "better description of the real situation". The radius b of a second hydration shell around the iodide was estimated to have a value of 4 \AA and a rotational correlation time of water molecules in the first hydration shell of iodide was calculated as

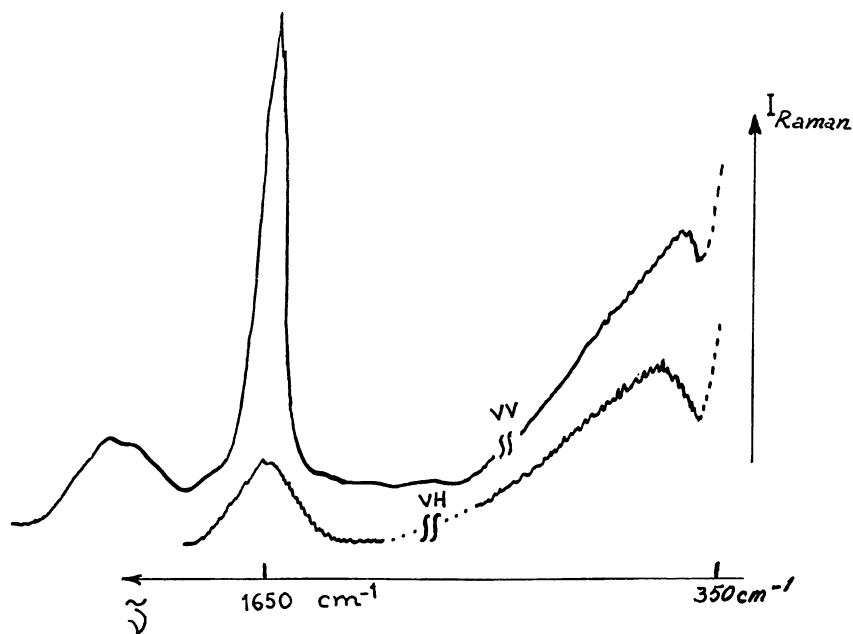


Figure 17. The 1650 cm^{-1} centered band in polarized Raman spectra of $\text{LiCl}\cdot 4.1\text{H}_2\text{O}$ glasses at 119 K.

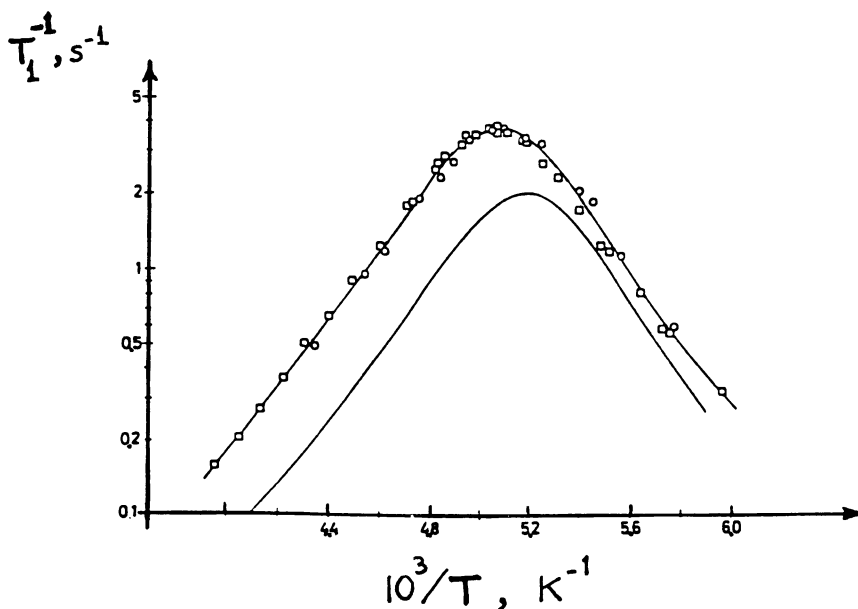


Figure 18. 60 MHz ^1H NMR spin-lattice (T_1) measurements for LiI solutions in D_2O , showing the presence of a T_1 -minimum near 200 K. (Reproduced with permission from ref. 31. Copyright 1976, Plenum Press).

$\tau_c \approx a^2/3D$, where D is the self-diffusion coefficient of D_2O ; for a 6M LiI solution at 197.5 K the value of τ_c was 1.7 ns. The $^{127}I \cdots \cdots ^1H$ contribution to relaxation was found, however, to correspond to much shorter correlation times of the ^{127}I nucleus and, therefore, the $^{127}I \cdots \cdots ^1H$ interaction is not the main source of relaxation. On the other hand, the correlation time of the water protons in the first hydration shell of Li^+ was estimated to be $\tau_c = 2.2$ ns at 197 K for the 6M 7LiI solution in D_2O . The $^7Li^+ \cdots \cdots ^1H$ contribution to T_1 relaxation yielded a value of a_{Li-H} of 2.56 ± 0.14 Å, corresponding to a $Li^+ \cdots \cdots O$ distance of 2.08 Å, in reasonable agreement with the peak values in the $G(r)$ of a 7M LiCl solution in H_2O , previously obtained from X-ray scattering studies (21).

Aqueous Solutions of LiCl and other Alkali Halides. 2H NMR spin-lattice and spin-spin (T_2) relaxation times were reported (31) at 10 MHz for aqueous (D_2O) solutions of lithium chloride as a function of temperature (in the range from 173 to 320 K) and concentration ($R = 3.5$ to 6.3 mol D_2O per mol LiCl). As in the case of the 1H NMR data for LiI solutions discussed above (30), the spin-lattice relaxation rates exhibited a maximum for $LiCl \cdot RD_2O$ solutions at ~ 197 K (corresponding to the T_1 minimum in Figure 19), whereas $1/T_2$ decreased continuously with increasing temperature (32) over the entire range (Figure 19). Such data were analyzed with the model shown in Figure 1b and Color Plate 16 for the $LiCl \cdot 4H_2O$ glass at 100 K. It is possible to estimate with this model and the data in Table VII the intramolecular D-D distance for the interstitial D_2O molecules at 100 K from the e^2qQ/h value of 287 ± 5 kHz; this value is significantly higher than those reported for liquid D_2O at 293 K (222 kHz (32) and 230 ± 10 kHz (32, 34)), and the value of 214 kHz reported for $LiCl \cdot 4D_2O$ at 100 K (Table VII). Assuming a DOD bond angle of 105.5° for D_2O at 100 K and 200 K, one has that $r_{DD} = 1.50$ Å for interstitial water from the data in Table VII, whereas the $LiCl \cdot 4H_2O$ glass model yielded $r_{HH} = 1.55$ Å at 100 K. The corresponding values for the OD and OH bond of D_2O /water are, respectively, $r_{OD} = 0.968$ Å and $r_{OH} = 0.974$ Å. These bond length values are smaller than those obtained for ice I_h (D_2O) from neutron diffraction data (1.02 Å (35) to 1.008 Å (36)), and are significantly larger than the $r_{OD} = 0.957$ Å reported for water vapor. We are tentatively assigning the 3450 cm^{-1} OH stretching band in Figure 16 to this interstitial water population with an r_{OH} value of $\sim 0.97 \pm 0.01$ Å. The hydration water within the $LiCl \cdot 4D_2O$ (H_2O) clusters at 197 K has a value of e^2qQ/h of 214 ± 10 kHz, corresponding to $r_{DD} = 1.59$ Å, and close to the value for liquid D_2O ; the corresponding value of r_{OD} is 1.000 Å for a DOD angle of 105.5° .

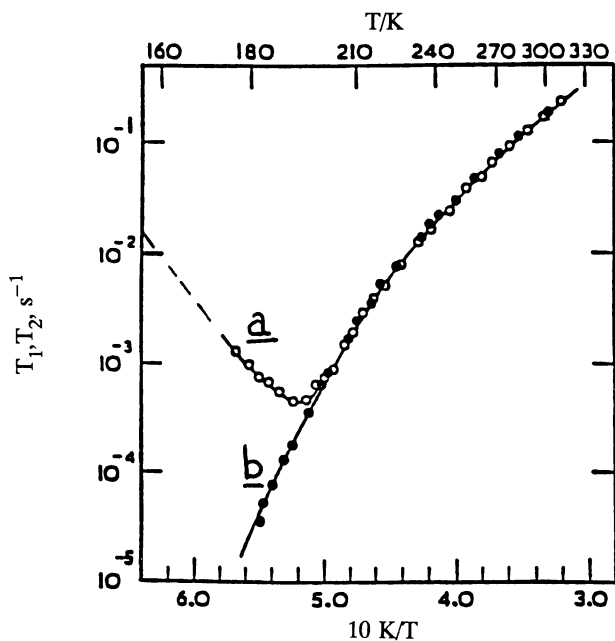


Figure 19. (a) ^2H NMR spin-lattice (T_1) measurements at 10 MHz for $\text{LiCl} \cdot n\text{D}_2\text{O}$ solutions as function of temperature (from 173 to 320 K). (b) Temperature dependence of transverse ^2H NMR relaxation rates for $\text{LiCl} \cdot n\text{D}_2\text{O}$ solutions. (Reproduced with permission from ref. 32. Copyright 1978, the Royal Society of Chemistry: Cambridge, UK).

Table VII. Deuterium Quadrupole Coupling Constants of D₂O in Several Solids Containing ions and for Free HDO Molecules.

System	D Quadrupole Coupling Constant, e^2Qq/h
Free HDO	290 kHz
Interstitial D ₂ O (in LiCl · 8D ₂ O at 100 K)	287 ± 5 kHz
Li ₂ SO ₄ · D ₂ O Crystal Hydrate	260 kHz
Liquid D ₂ O at 298 K	222 kHz; 230 ± 10 kHz
(COOD) ₂ · 2D ₂ O	250 kHz
Ice I _h at 253 K	254 kHz
LiCl · 4D ₂ O at 100 K	214 kHz

The pulsed ¹H NMR data at 100 K for LiCl · R H₂O glasses yielded $\langle r_{\text{HH}} \rangle = 1.59 \pm 0.01 \text{ \AA}$ and $r_{\text{OH}} = 1.000 \pm 0.01 \text{ \AA}$, in excellent agreement with these calculations based on the ²H NMR data at ~200 K. The effect of vibrational/librational motions on the $\langle r_{\text{HH}} \rangle$ values was here neglected; somewhat surprisingly, in spite of this approximation, the results obtained are consistent with those derived from neutron scattering data for liquid D₂O. It is interesting that the interstitial (“NHB”) water population has a molecular geometry intermediate between liquid water and vapor, and a bond length of $1.54 \pm 0.01 \text{ \AA}$, slightly closer to the water vapor value than to the liquid water or ice I_h. Although the water within Li⁺(H₂O)₄Cl⁻ clusters at 100 K has r_{HH} (r_{DD}) and r_{OH} (r_{OD}) values close to the liquid water values, it definitely has a different dynamics and local structure from those of liquid water at 293 K. The absence of a relatively narrow 3100 to 3200 cm⁻¹ OH-stretching band in the spectra of both LiCl aqueous solutions and glasses (Figures 16a and 16d, respectively) is also consistent with the X-ray/neutron scattering results (20) for LiCl · R H₂O (D₂O) solutions. On the other hand, liquid water exhibits the THB band at ~3200 cm⁻¹; both ice I_h and glassy water at 90 K exhibit also a sharp and tall band near 3100 cm⁻¹ that is completely absent in the concentrated LiCl solutions or glasses (Figures 16a,b,c,d). A new band is present in the latter systems near 3300 cm⁻¹ and this is a very broad band, unlike the 3100 cm⁻¹ band of ice I_h. At 119 K, in the glass, such a very broad band may correspond to a wide distribution of disordered water molecule configurations bridging the Li⁺ and Cl⁻ ions in the glass structure. The water distribution is, however, not completely random, as the c.w. ¹H NMR spectra of LiCl · R H₂O glasses at 10 K indicate (Figure 20c).

The anion effect on the local structure of lithium halide · (R H₂O) glasses at 100 K is reflected in the strength of the intermolecular proton-proton dipolar interactions; the larger iodide ion increases the intermolecular $\langle r_{\text{HH}} \rangle$ values significantly, as expected

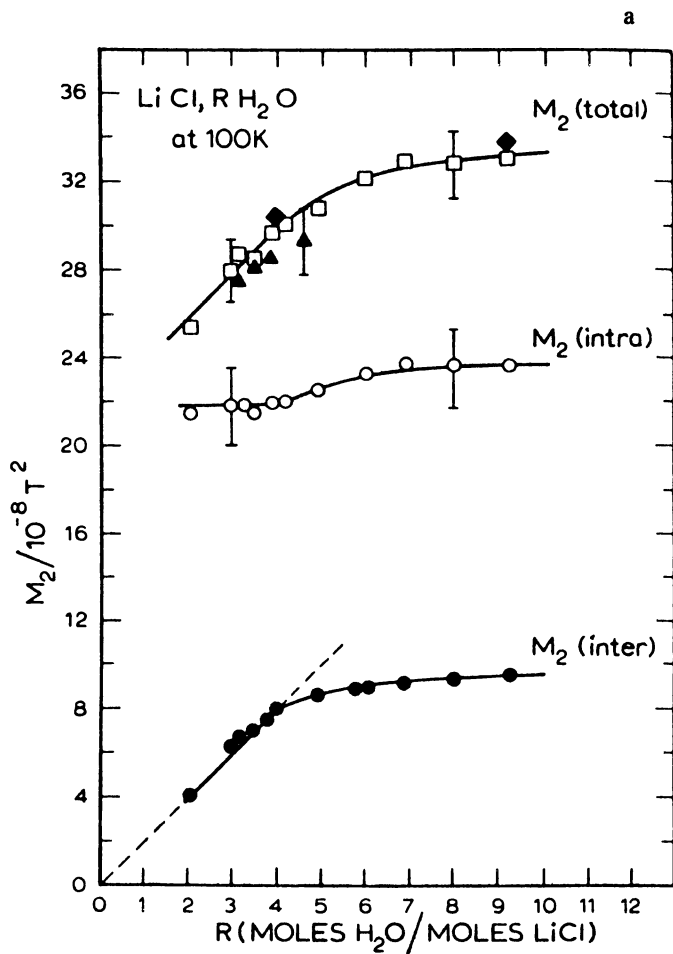


Figure 20a. Concentration dependence of ^1H NMR second moments in the 'rigid' lattice of $\text{LiCl} \cdot n\text{H}_2\text{O}$ glasses at 100 K (from ref. 5).

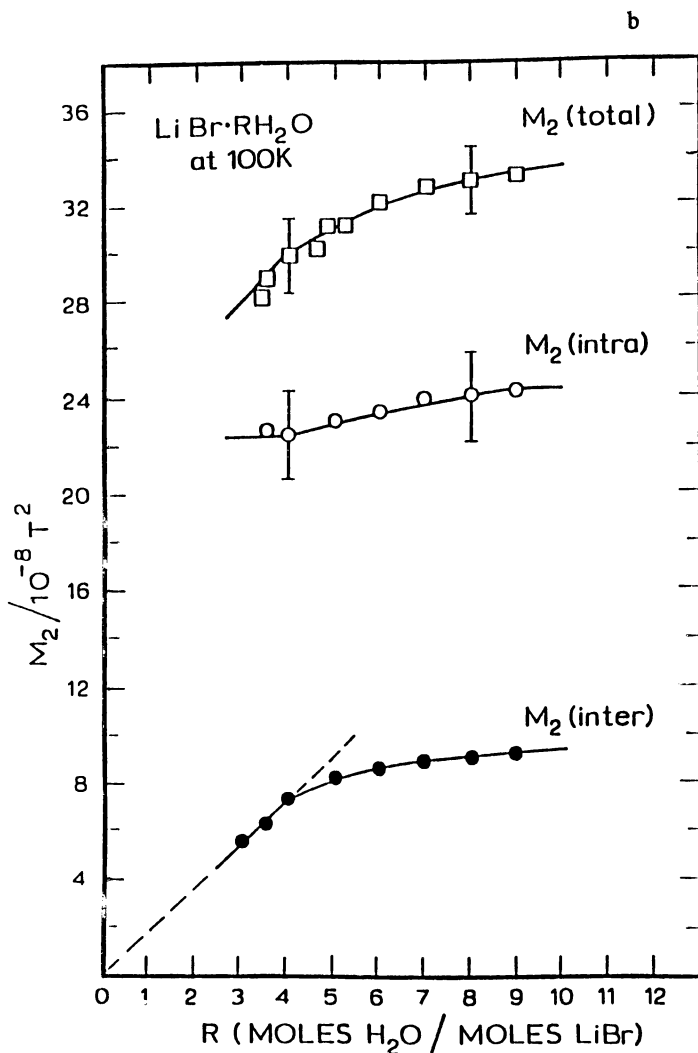


Figure 20b. Concentration dependence of ^1H NMR second moments in the rigid lattice of $\text{LiBr} \cdot n\text{H}_2\text{O}$ glasses at 100 K.

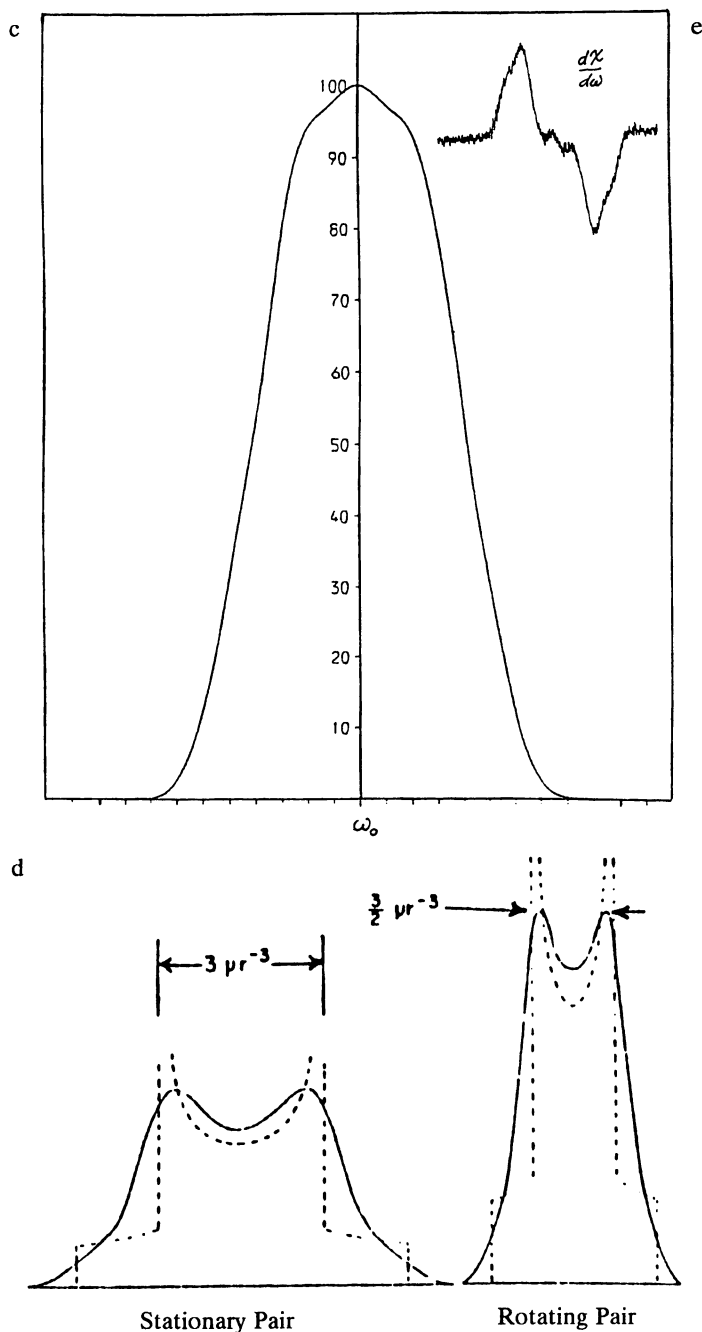


Figure 20c,d&e. (c) Continuous wave (c.w.) ^1H NMR spectra of a $\text{LiCl} \cdot 4\text{H}_2\text{O}$ glass at 10 K (top), compared with (d) the spectra of salt crystal hydrates (bottom); the insert (e) shows the second derivative of the ^1H NMR spectrum of the $\text{LiCl} \cdot 4\text{H}_2\text{O}$ glass at 10 K (Baianu, I. C., unpublished results).

from a random placement model of the anions in the glass structure. Otherwise, the behaviors of the LiCl, LiBr and LiI·RH₂O glasses are quite similar (Figures 20a and 20b).

The onset of structural relaxation processes in the LiCl (Br,I)·RH₂O glasses is observed near 120 K and 140 K, as a two-step decrease of the intermolecular dipolar (H-H) interactions strength; such temperatures are well below the glass transition temperature, T_g = 170 K, of these glasses. The results suggest that the interstitial water has a larger contribution to the structural relaxation processes, presumably through hindered reorientations of the interstitial water molecules. Similar behavior was observed for a number of other glasses, such as: Zn(NO₃)₂·RH₂O, Cd(NO₃)₂·RH₂O, Ca(NO₃)₂·RH₂O, La(NO₃)₃·RH₂O (Lightowers, D.; Boden, N. Unpublished results, Leeds University.) and CsF·RH₂O glasses (Baijanu, I. C. Unpublished results) at low temperatures. Since structural relaxation processes result in rearrangements of water molecules in the glass structure (and probably ions also), it seems reasonable to assume that the glass structure at low temperatures is somewhat different from the glass structures above ~120 K, and, most likely, different also from the local structure in the liquid solutions.

A direct answer to this question could be obtained by comparing the local structure of LiCl·RH₂O solutions in the form of pair-correlation functions with the corresponding g(r)'s for the LiCl·RH₂O glasses at 100 K. Unfortunately, the X-ray and neutron scattering data are yet unavailable for these glasses. We note, however, that the Li⁺···O, Cl⁻···O, <r_{OH}> and <r_{HH}> distances for the glass model in Figure 1e are consistent with the peak maxima in the correlation functions for the LiCl·RH₂O solutions at 293 K. However, a more detailed comparison of the local structures of the glass and the corresponding aqueous solution has not yet been made since it requires additional data. Therefore, a multinuclear spin relaxation approach was adopted for investigating the local structure of LiCl·RH₂O solutions. Relaxation times were measured for the following nuclei: ¹H, ²H, ⁷Li and ¹⁷O. ³⁵Cl and ³⁷Cl NMR relaxation measurements are also planned for these solutions and glasses. A few representative results are presented in Figures 21a-d. Similar experiments were carried out by ¹⁷O and ²³Na NMR for NaCl·RH₂O/D₂O solutions that do not form homogeneous glasses; notably, the concentration dependences of the transverse relaxation rates (R₂ = 1/T₂) were quite different between the LiCl and NaCl solutions in D₂O. The analysis of the ¹⁷O NMR relaxation for LiCl·RD₂O solutions at 293 K (Figure 21a) with a modified Debye (ion activity) model yielded a value of closest approach for Li⁺ and Cl⁻, a_{Li-Cl} = 2.88 ± 0.4 Å, which compares favorably with the expected value of 2.5 Å.

The mean square charge fluctuation term, <Z²>^{1/2} for this model had a value of 1.96 ± 0.4, obtained from a nonlinear regression fit of the data, which is close to the expected value of 2.0. Nevertheless, this simple Debye-modified model neglects the possibility of hydrated clusters that are bigger than the Li⁺···Cl⁻ direct contact ion-pair. The latter possibility was investigated with a thermodynamic linkage model (37, 38) for Li⁺, Cl⁻ and water by considering several possible equilibria for hydrated cluster formation of the following type:

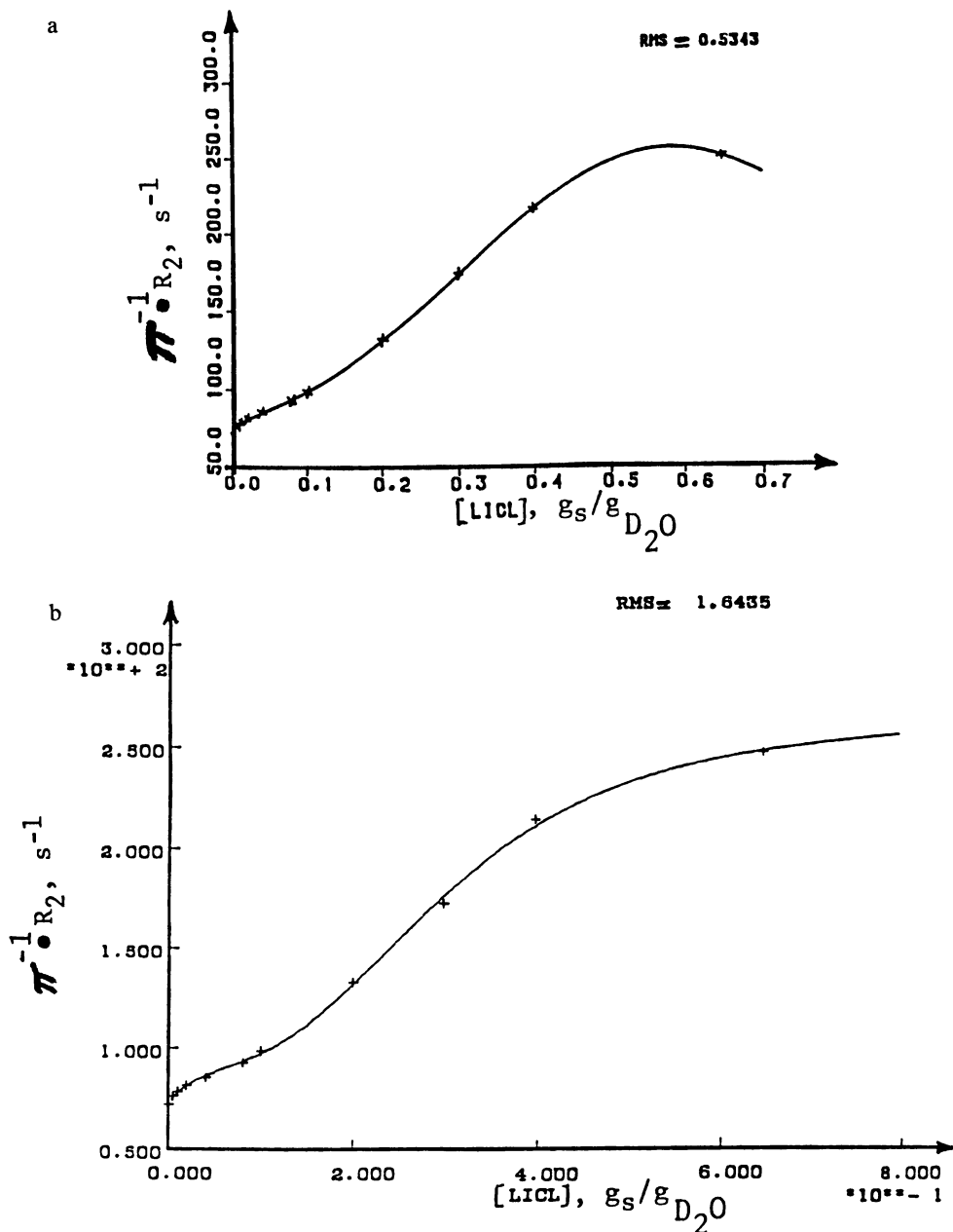


Figure 21a&b. (a) The concentration dependence of ^{17}O NMR transverse relaxation in aqueous solutions of $\text{LiCl} \cdot n\text{D}_2\text{O}$ at 293 K fitted curve is with the Debye-modified, ion-activity, model. (b) Fitted curve of ^{17}O NMR transverse relaxation of water in $\text{LiCl} \cdot n\text{D}_2\text{O}$ solutions at 293 K with the thermodynamic linked-functions (ion cluster) model.

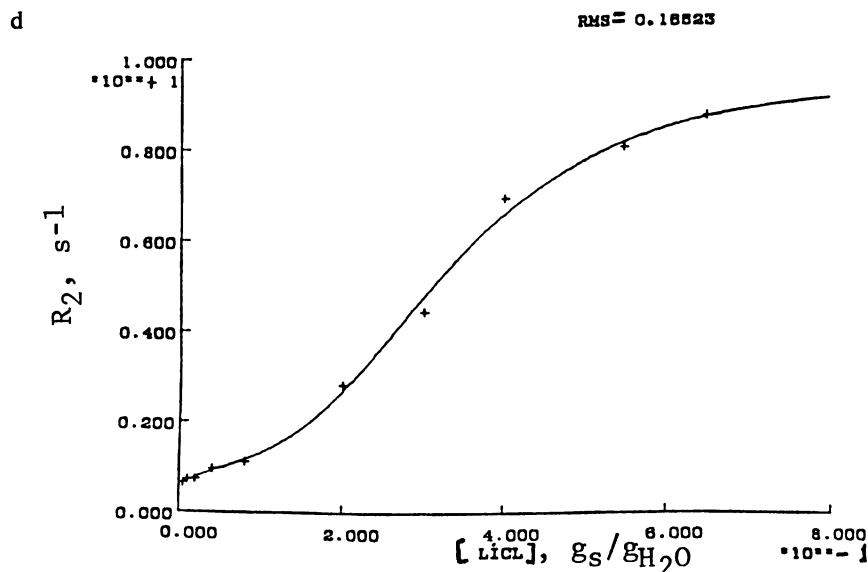
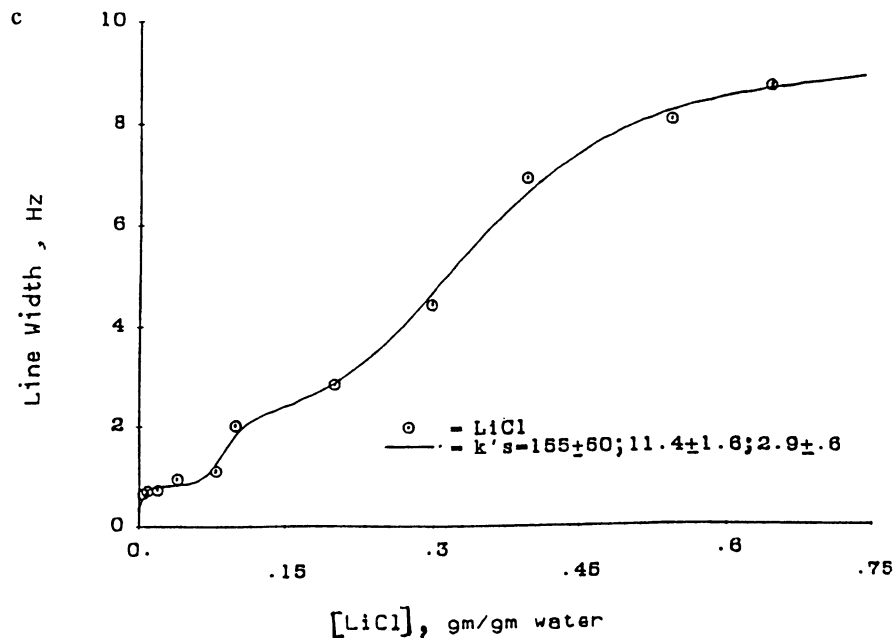
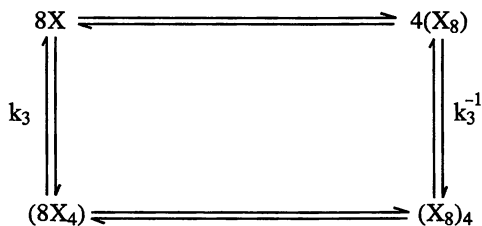


Figure 21c&d. (c) ^7Li NMR transverse relaxation measurements in aqueous solutions of $\text{LiCl} \cdot n\text{D}_2\text{O}$ at 293 K; shown is also the fitted curve with the ion-pairing cluster model. (d) Nonlinear regression analysis of the data in Figure 21c with a Debye-modified, ion-activity model of $\text{LiCl} \cdot n\text{D}_2\text{O}$ solutions at 293 K.



where $\text{Li}^+ + \text{Cl}^- \rightleftharpoons X$ (dimer) and $nX \rightleftharpoons (X_n)$, (n -mer) equilibria are assumed. The nonlinear regression analysis of both ^{17}O and ^7Li NMR relaxation data for LiCl solutions in D_2O yielded a best fit for tetramers ($q = 4$, $m = 8$, $n = 1$) of hydrated Li^+ and Cl^- ions. An idealized CPK model of such a $\text{LiCl} \cdot R\text{D}_2\text{O}$ ($R = 4.00$) cluster is shown in Color Plate 15. The ion-pairing and sequential self-association (clustering) process for these solutions have the apparent association rates, K_i , which can be obtained by a thermodynamic linkage analysis of the NMR data. Since this analysis provides the best fit amongst several simple models we conclude that dimer-tetramer equilibria are likely to occur in LiCl solutions. A transition from dimers towards preferred tetramers appears to occur at $R < 20$. The results may also suggest that the $\text{LiCl} \cdot R\text{D}_2\text{O}$ glasses consist mostly of tetramers (or higher n -mer) clusters that have a minimum correlation radius of about 12 Å. Interstitial water molecules would be close-packed between the tetramer (or higher n -mer) clusters and form relatively few hydrogen bonds amongst themselves, or with water molecules within the clusters. The analysis of both ^{17}O and ^7Li NMR relaxation data yields the result that tetramers are the preferred cluster structure for the very concentrated solutions. A force-field energy minimization calculation for $\text{LiCl} \cdot 8\text{H}_2\text{O}$ at 300 K (Figures 22a and 22c) shows the presence of predominantly dimer H-bonded water molecules (Figures 22a), as well as the typical charge alternation seen in MD simulations of molten salts. Similar results were obtained for $\text{NaCl} \cdot 8\text{H}_2\text{O}$ (Figure 22b). It is also interesting that the dimer-tetramer cluster model fits exceedingly well the deuterium NMR longitudinal relaxation data (Figure 23a) at 293 K for the $\text{LiCl} \cdot R \text{D}_2\text{O}$ solutions (32), as well as the concentration dependence of the glass transition temperature (Figure 23b) for these compositions (39). We have to conclude that a dimer-tetramer model should be adopted for the analysis of the deuterium NMR data for both LiCl solutions and glasses. For $2 < R < 4$ the concentration dependence of the deuterium NMR relaxation rate is, however, different from that observed for ^{17}O or ^7Li . This suggests a jump diffusion mechanism for deuterium exchange which occurs only slowly for $R < 4$ because of the presence of large activation barriers set up by the alternating charge distribution of the ionic clusters in space; on the other hand, in the presence of interstitial water molecules ($R > 4$), fast deuterium exchange can occur via interstitial water molecules that are H-bonded to water inside the ionic clusters.

For values of R close to 3, the ^7Li spin-lattice relaxation at 200 K exhibits two components in a LiCl aqueous (D_2O) solution (40). The two decay constants have a minimum near 180 K, suggesting the presence of anisotropic reorientation motions of the hydrated complex $\text{Li}^+(\text{D}_2\text{O})_n$, ($n < 4$); the latter solutions are in the nonextreme narrowing regime (40). It was concluded that the water molecules surrounding the

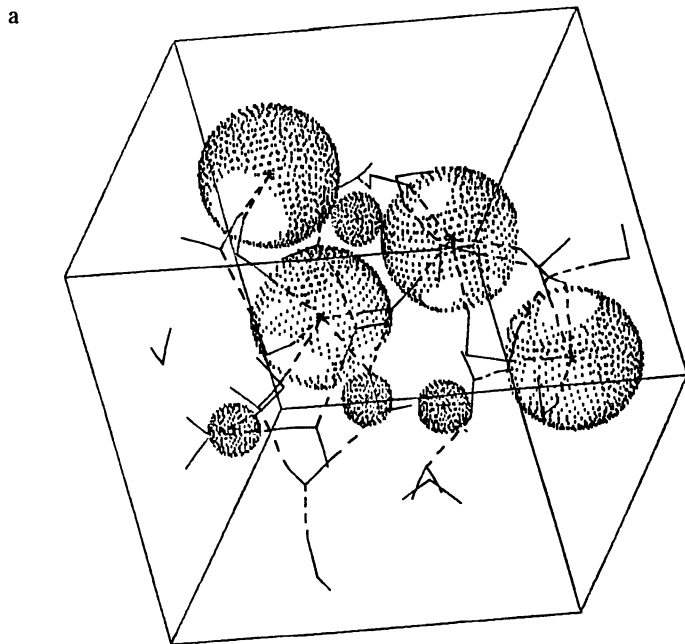


Figure 22a. Molecular dynamics calculation for $\text{LiCl} \cdot 8\text{H}_2\text{O}$.

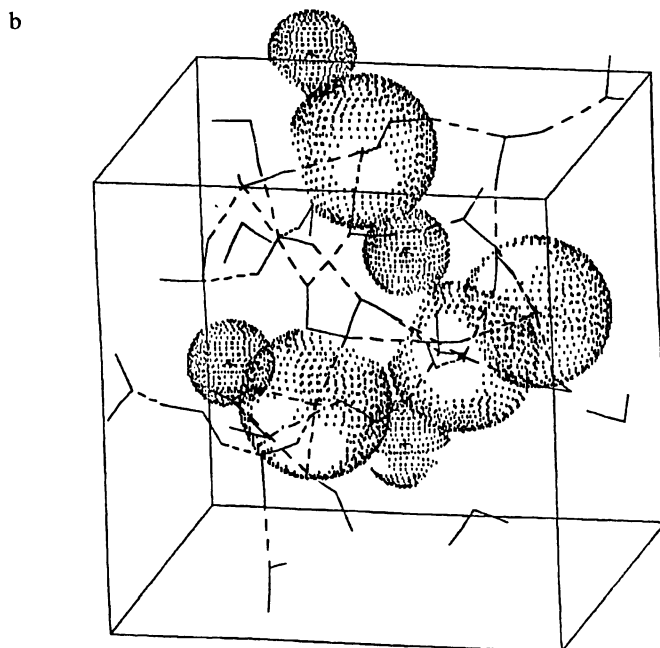


Figure 22b. Molecular dynamics calculation for $\text{NaCl} \cdot 8\text{H}_2\text{O}$ solutions with a force-field energy minimization model.

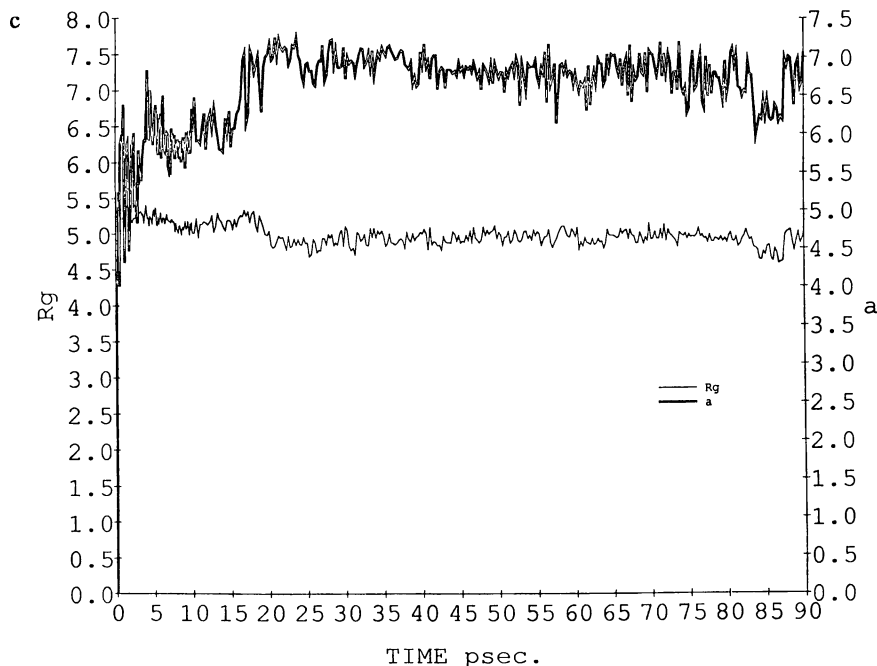
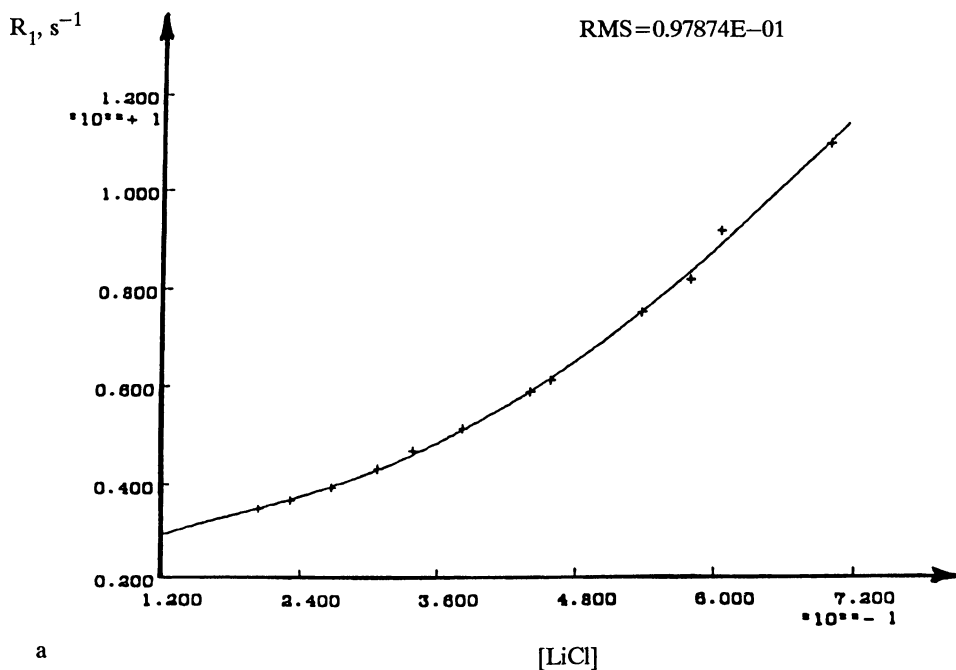


Figure 22c. Example of a computer run in the molecular dynamics calculation for $\text{LiCl} \cdot 8\text{H}_2\text{O}$ up to 90 ps illustrating the formation of stable hydrated clusters of radius of gyration, $R_g \approx 4.8 \text{ \AA}$ after about 20 ps; (also shown is the dynamics of the Debye inter-ion closest distance of approach for the Li^+ and Cl^- ions.)

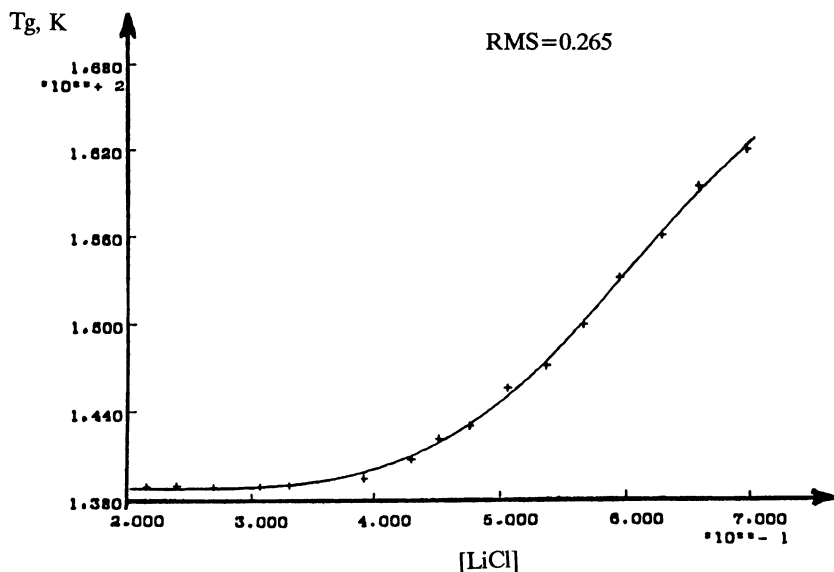
lithium cation make the principal contribution to the field gradient at the ^7Li nucleus. Since the residence time of water in the Li^+ hydration shell is about 30 ps at 298 K, and the ^{17}O reorientational correlation time is less than 24 ps for $4 < R < 12$, it would seem that water molecules undergo **fast hindered reorientation** in the hydration shell of Li^+ .

The ^{19}F nuclear magnetic relaxation of the F^- anion in CsF , KF and RbF solutions in D_2O (or H_2O) exhibited concentration dependences (Figure 24c) similar to that observed by us for $\text{NaCl} \cdot \text{RD}_2\text{O}$ solutions by ^{23}Na and ^{17}O NMR. The closest $\text{F}^- \cdots \text{F}^-$ separation was found to be $a = 3 \text{ \AA}$ (42), which is consistent with a model of two fluoride ions bridged through the same water molecule by breaking two of the hydrogen bonds of that water with its neighbor water molecules.

One could also analyze the ^{19}F NMR relaxation data with a thermodynamic linkage model and derive a more detailed structure of the $(\text{F}^- \cdots \text{HOH} \cdots \text{F}^-)$ clusters. It would also be interesting to compare the cluster sizes and the equilibria for CsF , KF and $\text{RbF} \cdot \text{RD}_2\text{O}$ with those already determined for $\text{LiCl} \cdot \text{RD}_2\text{O}$ solutions.



a



b

Figure 23. (a) Concentration dependence of the ^2H NMR longitudinal relaxation measurements for $\text{LiCl} \cdot n\text{D}_2\text{O}$ solution at 293 K. Fitted curve was obtained by nonlinear regression analysis of the ^2H NMR data with the ion-pair cluster model. (b) Nonlinear regression analysis of the concentration dependence of the glass transition temperature, T_g , for $\text{LiCl} \cdot n\text{H}_2\text{O}$ glasses with the ion-pair cluster model.

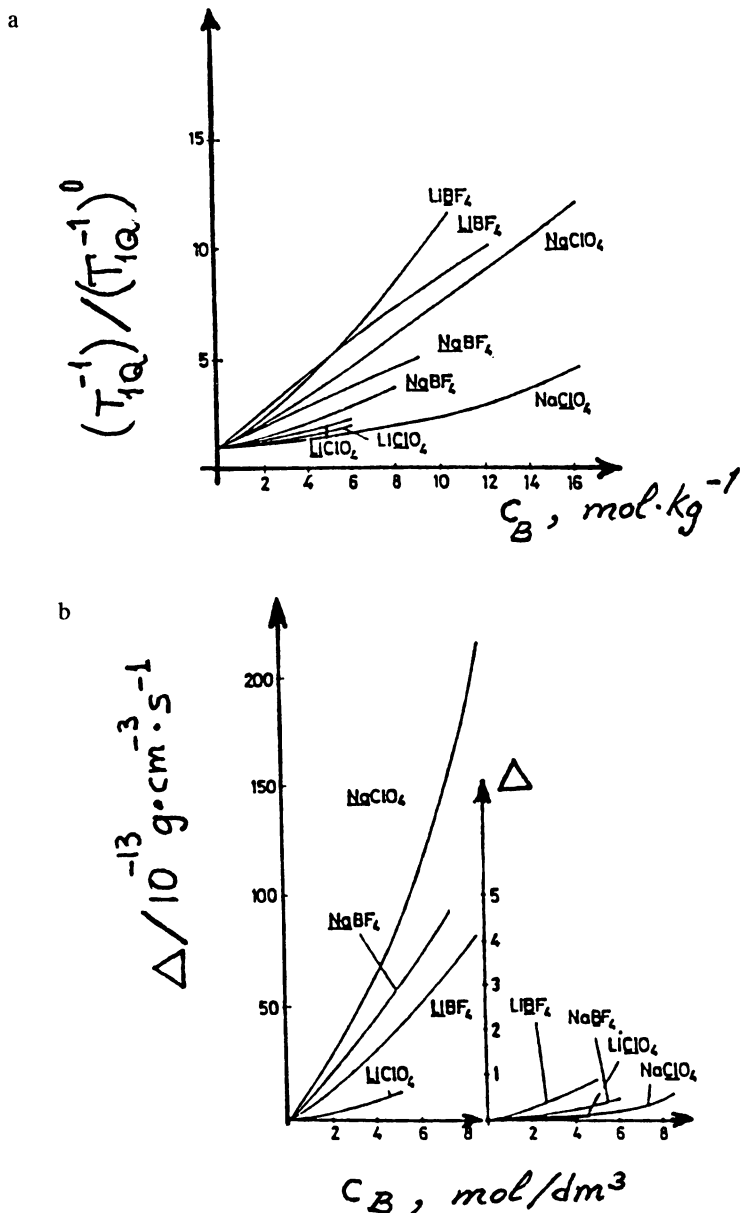


Figure 24. Concentration dependence of longitudinal magnetic relaxation rates for lithium and sodium perchlorates (a) and tetrafluoroborates (b) in aqueous solutions (modified from ref. 41). (c) experimental concentration dependences of longitudinal relaxation rates for KF, CsF and RbF solutions; (d) calculated concentration dependences of longitudinal relaxation rates for KF, CsF and RbF solutions. (Reproduced with permission from ref. 42. Copyright 1977, the Royal Society of Chemistry: Cambridge, UK).

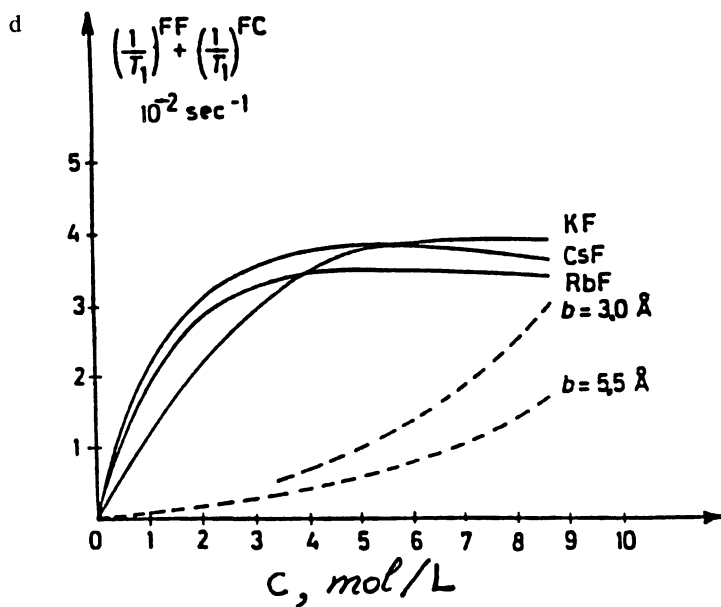
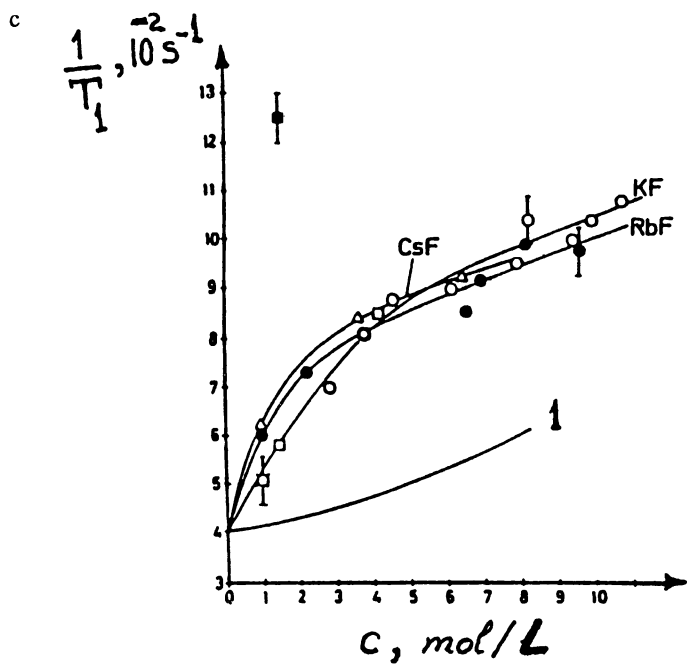


Figure 24. Continued.

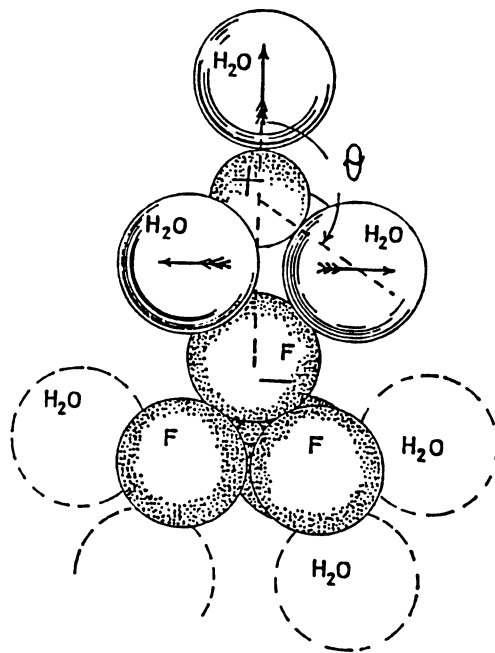


Figure 25. Molecular model of tetrafluoroborate hydration (according to ref. 42). (Reproduced with permission. Copyright 1977, the Royal Society of Chemistry: Cambridge, UK).

The variation of the longitudinal magnetic relaxation rates with concentration was reported also (43) for lithium and sodium perchlorates and tetrafluoroborates in aqueous solutions up to high concentrations ($R > 3$), when the solubility limit permitted (Figure 24a). Note that these concentration dependences were different from those observed for the $\text{LiCl} \cdot \text{RD}_2\text{O}$ or CsF , KF , $\text{RbF} \cdot \text{RD}_2\text{O}$ solutions (Figures 23a and 24c, respectively). It would seem that the large sizes of the perchlorate and tetrafluoroborate anions prevent, in these cases, the association with the cation into well-defined, hydrated n -mers; however, at high concentrations water molecules should still be bridged between the counterions (Figure 25). Therefore, MD simulations have not yet been reported for these complex anions in aqueous solutions.

Molecular Dynamics and Nuclear Spin Relaxation Studies of Glycine Hydration and Activity in Aqueous Solutions. This is the first report of extensive ^{17}O and ^2H NMR transverse relaxation studies that allowed us to quantitate glycine interactions, hydration and aggregation properties as a function of both pH and amino acid concentration.

In the presence of added LiCl to glycine solutions in ethanol-water mixtures (for $60\% \leq [\text{EtOH}] \leq 95\%$) at 25°C , it was found previously possible to screen out dipole-dipole interactions (45), as predicted by Kirkwood's theory for zwitterions in the absence of activity effects. On the other hand, for glycine in aqueous solutions it was predicted by Kirkwood (45) that interactions among glycine dipoles, as well as glycine-dipole interactions with the surrounding water molecules, would result in significant activity effects in these aqueous solutions. This theoretical prediction is

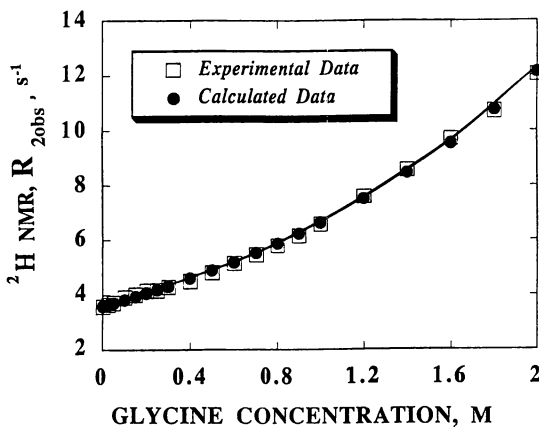


Figure 26. Nonlinear variation of ^{17}O NMR transverse relaxation rates of D_2O in glycine solutions **without** added salt.

borne out by our first ^{17}O NMR results presented in Figure 26 that show indeed a nonlinear concentration dependence of the water ^{17}O transverse relaxation rates, which is caused by the glycine activity in these aqueous solutions *without* added LiCl. Nonlinear regression analysis of the variation of water ^{17}O NMR transverse relaxation as a function of glycine concentration (Figure 26) with a simple activity model (see following paper in this book) yields the value of the second order virial coefficient, B_0 , for glycine in water.

Literature Cited

1. March, N. H.; Tosi, M. P. Chapter 5 in *Coulomb Liquids*, Academic Press: London, **1984**, pp.111-157.
2. Woodcock, L.V. *Nature (Lond)*, **Phys. Sci.** **1971**, *232*, 63.
3. Clementi, C.; Popkie, H. *J. Chem. Phys.* **1972**, *57*, 1077.
4. Enderby, J. E.; Neilson, G. W. *Phys. Bull.* **1978**, *29*, 360.
5. Baianu, I. C.; Boden, N.; Mortimer, M. *Chem. Phys. Letts.*, **1978**, *54*, 169.
6. Caminiti, R. et al. *Disc. Faraday Chem. Soc., Ion-Ion and Ion-Solvent Interactions*, Oxford **1977**, *64/4*, 1A.
7. Engström, B.; Johnson, B.; Jonsson, B. *J. Magn. Reson.* **1982**, *50*, 1.
8. Smith, I. C. P. In *NMR of Newly Accessible Nuclei.*; P. Laszlo, Ed.; Academic Press: New York, NY, **1982**, Vol. 2; pp.1-30.
9. Mantsch, H. H.; Saito, H.; Smith, I. C. P. *Intl. Rev. J. Progr. NMR Spectrosc.* **1977**, *11(4)*, 211.
10. Greenfield, M. S. et al. *J. Magn. Reson.* **1987**, *72*, 89-107.
11. Enderby, J. E. *Phys. Bull.* **1978**, *29*, 360-363.
12. Baianu, I. C.; Rubinson, K. A.; Patterson, J. *Physica Stat Solid., A* **1979**, *53*, k133-138.
13. Baianu, I.C. In *Rapidly Quenched Metals*, Cantor, B., Ed.; The Chameleon Press: London, England, **1978**, Vol. 2, pp.419, 425.
14. Warren, B. E.; Marel, G. *J. Sci. Instr.* **1905**, *36*, 196.
15. Neilson, G. W.; Enderby, J. E. *J. Phys. C: Solid State Phys. (UK)*, **1980**, *13*, L924-L926.

16. Neilson, G. W.; Howe, R. A.; Enderby, J. E. *Chem. Phys. Lett.* **1975**, *33*, 284.
17. Rotschild, W. G. *Dynamics of Molecular Liquids*; Wiley Publs: New York, **1984**.
18. Licheri, G.; Piccaluga, G.; Pinna, G. *J. Chem. Phys.*, **1976**, *64*, 2437.
19. Caminiti, R. et al. *Anali di Chim.* **1975**, *65*, 695-710.
20. Leadbetter, A. J. In *Neutron Inelastic Scattering Proc. Symp. Neutrons, Grenoble, France, IAEA Publ.: Vienna, 1972*.
21. Narten, A. H.; Vaslow, F.; Levy, H. A. *J. Chem. Phys.* **1973**, *58*, 5017.
22. Dzidic, I.; Kebarle, P. *J. Phys. Chem.* **1970**, *74*, 1466.
23. Arshadi, M.; Yamadagni, R.; Kebarle, P. *J. Phys. Chem.* **1970**, *74*, 1475.
24. Cieplak, P.; Lybrand, T. P.; Kollman, P. A. *J. Chem. Phys.* **1987**, *86*, 6393-6403.
25. Cieplak, P.; Kollman, P. *J. Chem. Phys.* **1990**, *92*, 6761-6767.
26. Impey, R. W.; Madden, P. A. *Mol. Physics* **1982**, *46*, 513-539.
27. Halle, B. et al. *J. Amer. Chem. Soc.* **1981**, *103*, 500-508.
28. Baianu, I. C. et al. In *Water Relations in Foods*. Levine, H.; Slade, L., Eds.; Plenum Press: New York and London, **1991**.
29. Nash, C. P.; Donnelly, T. C.; Rock, P. A. *Solution Chem.* **1977**, *6*, 663.
30. Okazaki, S.; Ohtori, N.; Okada, I. *J. Chem. Phys.* **1990**, *92*, 7505-7515.
31. Geiger, A.; Hertz, H. G. *J. Soln. Chem.* **1976**, *5*, 365-387.
32. Boden, N.; Mortimer, M. *J. Chem. Soc. Faraday Trans. II* **1978**, *74(2)*, 353-366.
33. Chiba, T. *J. Chem. Phys.* **1964**, *41*, 1352.
34. Jones, J.; DeFries, T.; Wilbur, D. J. *J. Chem. Phys.* **1976**, *65*, 582.
35. Peterson, S. W.; Levy, H. A. *Acta Crystallogr.* **1957**, *10*, 70.
36. Pauling, L. *J. Amer. Chem. Soc.* **1935**, *57*, 2680.
37. Wyman, J., Jr. *Adv. Protein Chem.*, **1964**, *19*, 223-311.
38. Kumosinski, T. F. *J. Agric. Food Chem.* **1988**, *35*, 669-672.
39. Angell, C. A.; Sare, E. J. *J. Chem. Phys.* **1970**, *52*, 1058.
40. Tokuhito, T. *J. Magn. Reson.* **1988**, *70*, 22-29.
41. Hertz, H. G.; Rädle, C. *Ber. Bunsenges. Physik. Chem. (Z.f. Elektrochem.)* **1974**, *78*, 509-514.
42. Contreras, M.; Hertz, H. G. *Faraday Chem. Soc. Discussion, Solvent Interactions* **1977**, *64*, 33-47.
43. Woodcock, L. V. In *Advances in Molten Salt Chemistry*. Braunstun, Mamantor, G.; Smith, G. P., Eds.; Plenum Press: New York, NY, **1975**, *Vol. 3*, pp.1-74.
44. *Physical Chemistry of Food Processes*; Baianu, I. C.; Pessen, H.; Kumosinski, T. F., Eds.; Van Nostrand Reinhold: New York, NY, **1993**, *Vol. 2*.
45. Kirkwood, J. G. *J. Chem. Phys.* **1934**, *2*, 351-361.
46. Howe, R. A.; Howells, W. S.; Enderby, J. E. *J. Phys. C: Solid State Phys.* **1974**, *7*, L111.
47. Madden, P. A.; Impey, R. W. *Ann. N. Y. Acad. Sci.* **1985**, *482*, 91-114.
48. Enderby, J. E.; Howells, W. S.; Howe, R. A. *Chem. Phys. Lett.* **1973**, *21*, 109-112.

RECEIVED July 8, 1994

Chapter 18

Molecular Dynamics and Multinuclear Magnetic Resonance Studies of Zwitterions and Proteins in Concentrated Solutions

Ion C. Baianu¹, E. M. Ozu¹, T. C. Wei¹, and Thomas F. Kumosinski²

¹Department of Food Science, Agricultural and Food Chemistry—Nuclear Magnetic Resonance Facility, University of Illinois at Urbana, 580 Bevier Hall, 905 South Goodwin Avenue, Urbana, IL 61801

²Eastern Regional Research Center, Agricultural Research Service, U.S. Department of Agriculture, 600 East Mermaid Lane, Philadelphia, PA 19118

Multinuclear spin relaxation observations were carried out for concentrated protein solutions, such as lysozyme, myosin or soy globulins. Ion-specific effects were observed and analyzed with thermodynamic linkage models involving reversible binding of hydrated ions. Molecular dynamics and multinuclear spin relaxation measurements were carried out for several proteins in order to elucidate the mechanisms for the cationic and anionic interactions with the protein binding sites. Specific interactions of the anions with the positively charged side chain groups of lysine, arginine and histidine, as well as nonspecific binding to amide groups, were found. Both anion- and cation- specific, as well as cooperative, interactions with charged side chain groups were found for myosin, tropomyosin, and for 7S and 11S globulins from soy. The comparison between the ¹⁷O and ²³Na NMR results strongly suggests that water is exchanged as the hydrated ion species between the myofibrillar and/or myosin protein binding sites and the bulk, aqueous solution of electrolytes.

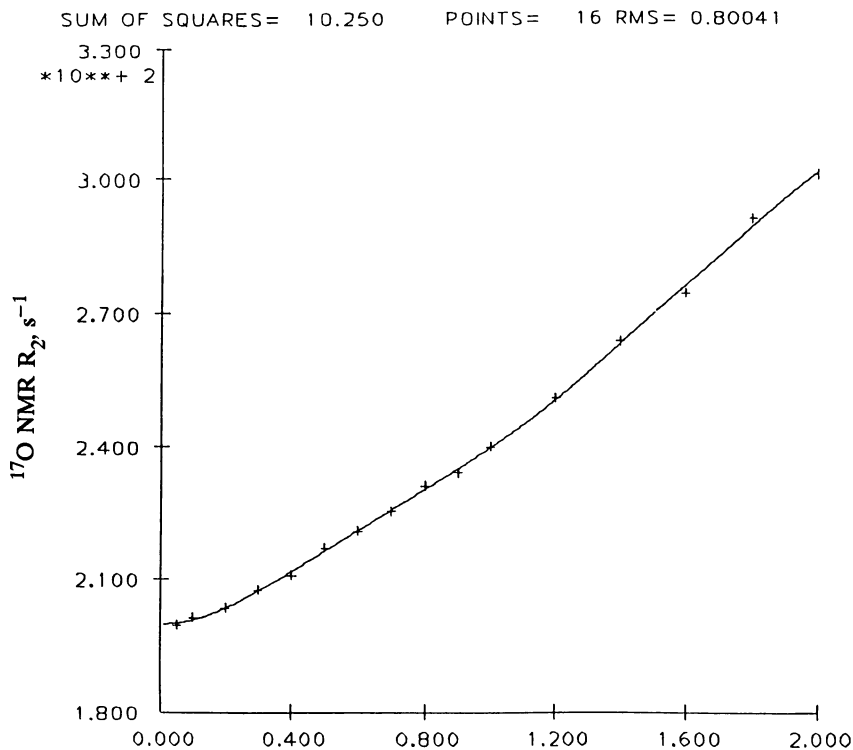
Molecular Dynamics and Nuclear Spin Relaxation Studies of Amino/Imino Acid Hydration and Activity in Aqueous Solutions

This is the first report of extensive ¹⁷O and ²H NMR transverse relaxation studies on concentrated imino acid and amino acid solutions in water that allowed us to quantitate amino acid interactions, hydration and aggregation properties as a function of both pH and amino/imino acid concentration. We have found that proline forms octahedral clusters at high concentrations in aqueous solutions in the pD range of 2.3 to 7.4 that reduce significantly water activity, (Figures 1, 2 and Color Plate 17.).

Figure 1 illustrates the variation of ¹⁷O NMR transverse relaxation of water in solutions of proline, as well as the calculated concentration dependence (continuous line). All initial structures were built using the Sybyl-Mendel molecular modeling program, and were then minimized utilizing Kollman and Tripos force-field computations. The NMR relaxation data analysis was carried out by nonlinear regression with a thermodynamic linkage model, according to Wyman's theory of

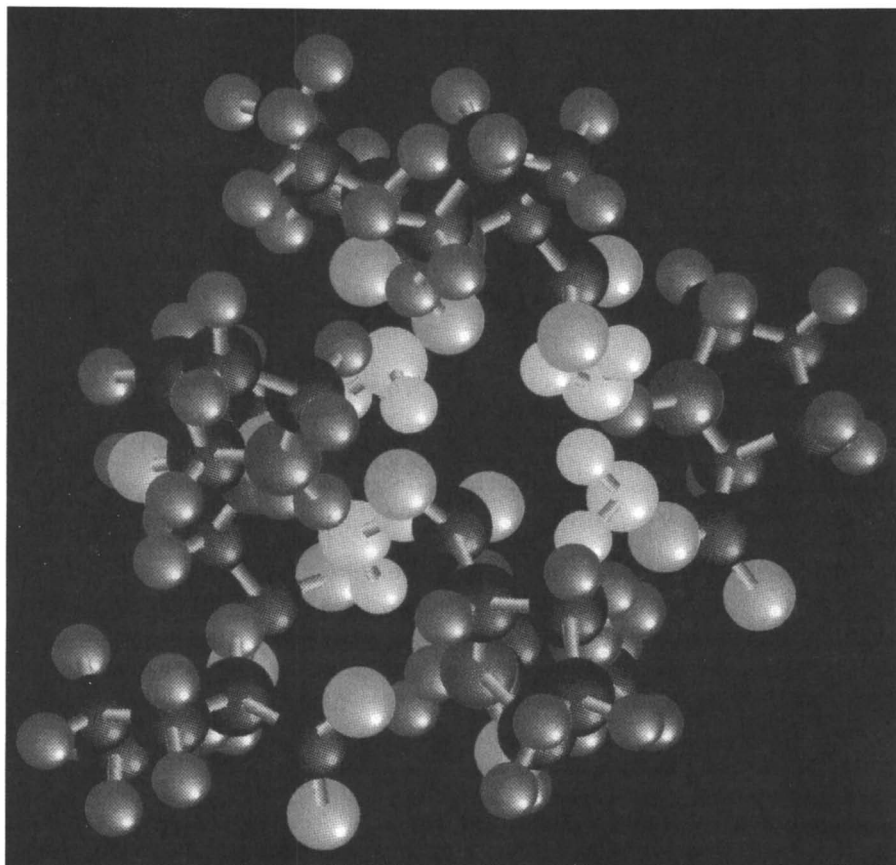
NOTE: The color plates can be found in a color section in the center of this volume.

0097-6156/94/0576-0325\$08.00/0
© 1994 American Chemical Society



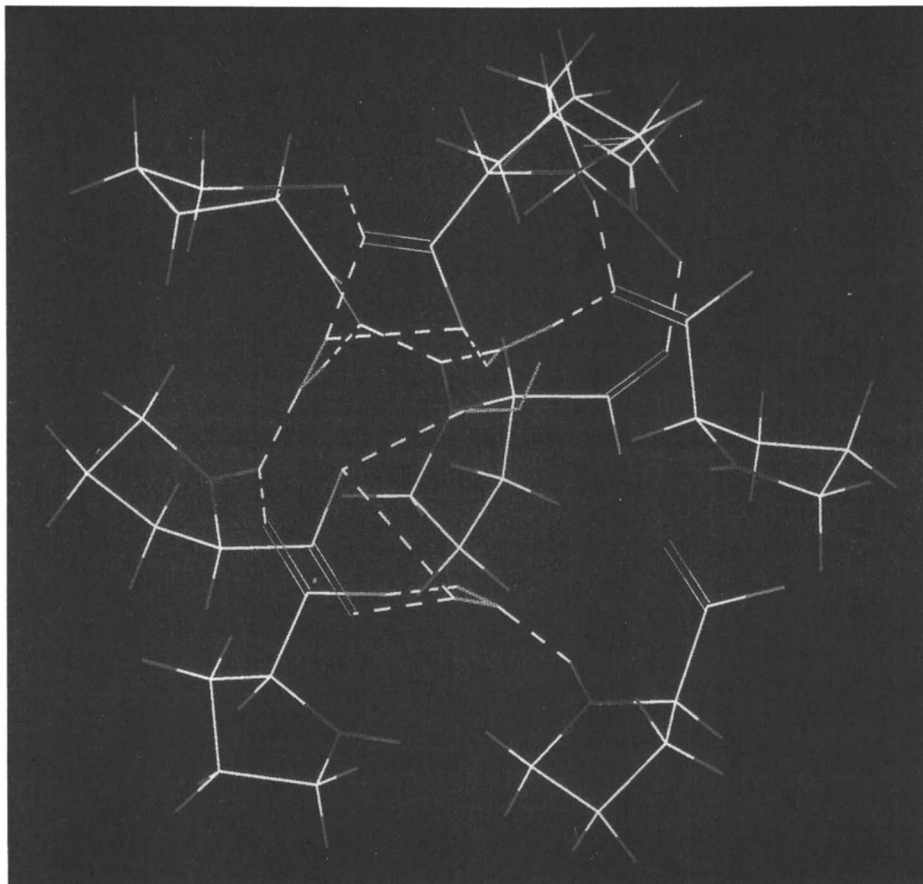
Proline

Figure 1. Variation of ^{17}O NMR transverse relaxation rates with proline concentration in D_2O solutions at pD 2.8 and 20°C . The line drawn between the experimental points represents the result of nonlinear regression analysis of NMR data with the amino acid activity model specified in the text.



A: Water 4, Proline 8; $t = 0$ ps

Figure 2. Single-frame illustrations of molecular dynamics of proline in water at 20 °C: A. The beginning of the simulation.



B: Proline 8, Water 4; $t = 50$ ps

Figure 2. Single-frame illustrations of molecular dynamics of proline in water at 20 °C: B. Wire model of Color Plate 17, at 50 ps.

linked functions, and allowed the determination of the degree of cooperativity and apparent binding constants for ion binding to myosin and soy globulins. Figure 2 and Color Plate 17 present a result of related molecular dynamics simulations for a proline cluster that is formed with 4 trapped water molecules. Charge-charge and dipole-dipole interaction models were employed to fit the ^{17}O NMR relaxation data in conjunction with nonlinear regression computer analysis; at concentrations above 1 M a thermodynamic linkage model was employed to analyze the aggregation of proline molecules. Lysine and arginine behavior in aqueous (D_2O) solutions, on the other hand, is dominated by repulsive charge-charge interactions characterized by a second order virial coefficient (B_0). The values of the coefficient B_0 (which were determined by nonlinear regression analysis of the ^{17}O NMR transverse relaxation dependences on amino/imino acid concentration) allowed us for the first time to quantitate and compare the effects of these small molecules on water activity.

Molecular Dynamics Computations for Hydrated Proteins

Significant progress was made in the last five years with the computer modeling of protein conformations and molecular motions of water surrounding proteins. Ahleström et al. (1) showed that the results of the molecular dynamics computations were distinct for proteins in vacuum and in the presence of water molecules; Lennard-Jones (6/12) potentials were assumed for such calculations. Figures 3 and 4 show representative results of such work: the calculated radial distribution functions of the protein oxygens are shown around the calcium ions at two different sites (CD in Figure 3 and EF in Figure 4). Notably, the radial distribution is "sharper" in vacuum than in the aqueous solution. The approach also yielded minimum distances between amino acid groups and water, time correlation functions for certain amino acid residues such as Phe (Figures 5a and 5b), and time constants for the residues forming calcium binding sites (Table I).

Table I. Time Constant, τ_1 , at the Calcium Binding Sites

Residue	Atoms	APO	VAC	AQ
<u>CD site</u>				
Asp 51	$\text{C}_\beta\text{-C}$	1.9	1.7	0.8
Asp 53	$\text{C}_\beta\text{-C}$	1.2	1.9	0.3
Glu 59	$\text{C}_\beta\text{-C}_\delta$	0.1	> 4.0	0.5
Glu 62	$\text{C}_\beta\text{-C}_\delta$	1.3	0.2	0.4
<u>EF site</u>				
Asp 90	$\text{C}_\beta\text{-C}$	0.1	0.8	2.3
Asp 92	$\text{C}_\beta\text{-C}$	0.3	2.4	0.3
Asp 94	$\text{C}_\beta\text{-C}$	0.1	3.2	0.2

SOURCE: Reprinted with permission from ref. 1. Copyright 1982.

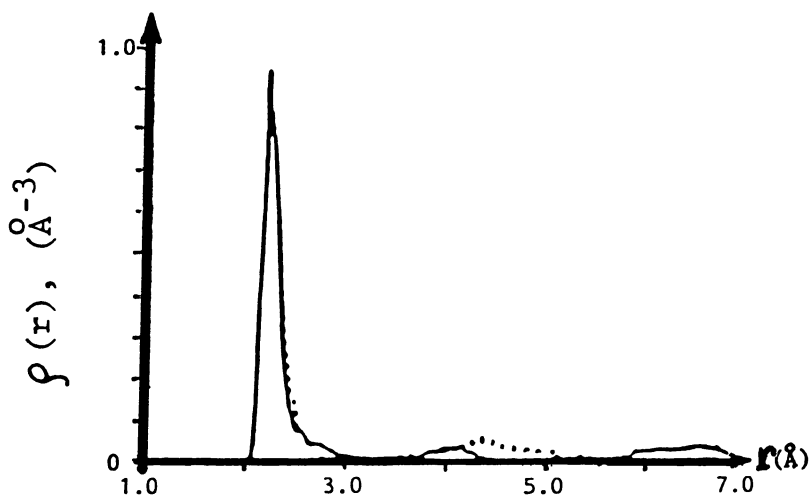


Figure 3. Radial distribution of protein oxygens around the calcium ions for the CD site. Solid lines refer to the vacuum simulation and dashed lines to the simulation in water (AQ). (Modified from ref. 44).

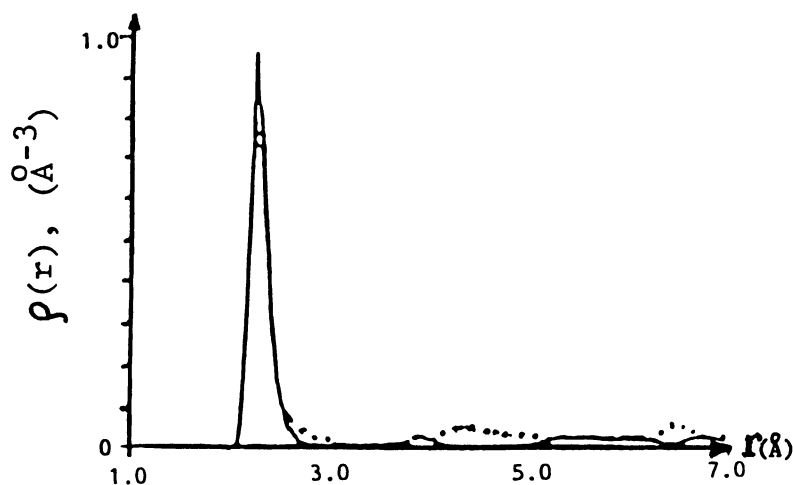


Figure 4. Radial distribution of protein oxygens around the calcium ions for the EF sites. Solid lines refer to the VAC simulation and dashed lines to the AQ simulation. (Modified from ref. 45).

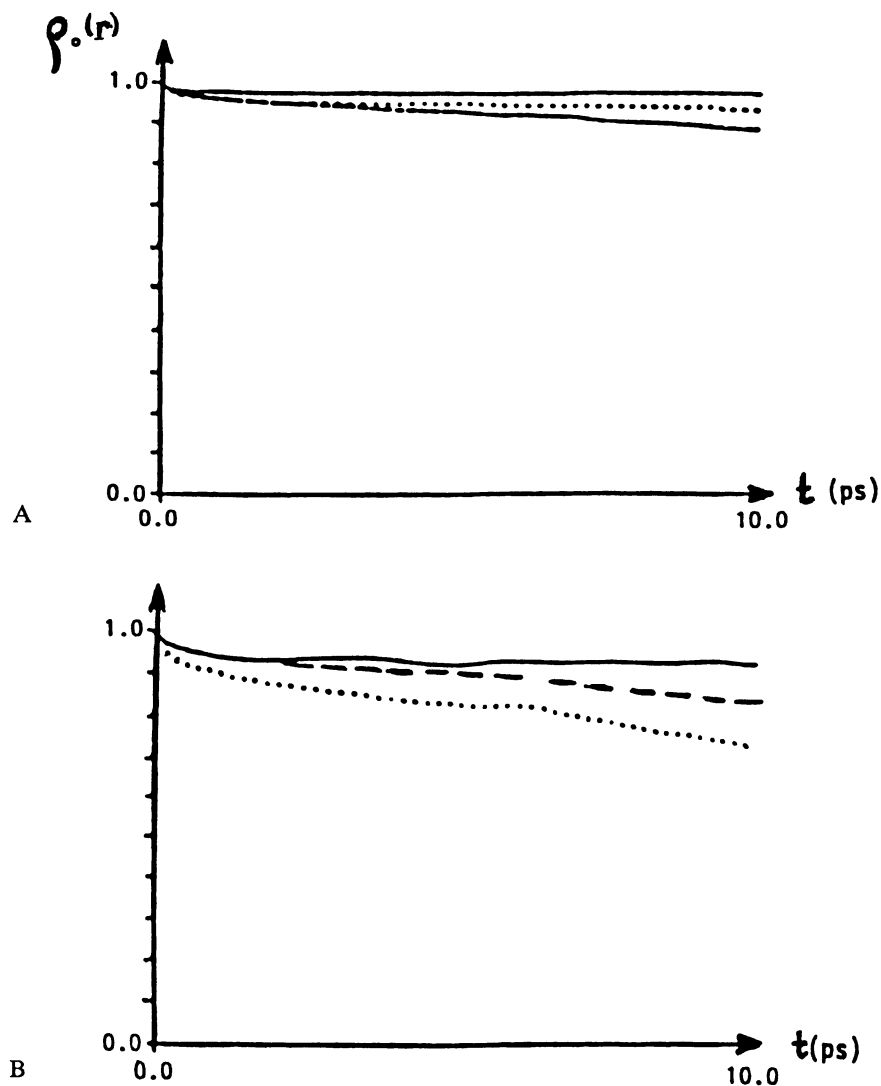


Figure 5. Time correlation functions ($C_1(t)$) for vectors in Phe 66 for : (A) DD vectors, and (B) GZ vectors. Solid lines refer to the VAC, dashed lines to the AQ, and dotted lines to the APO simulation. (Modified from ref. 45).

Proteins and Water Relaxation

If we mix proteins with water and then carry out nuclear spin relaxation experiments, the relaxation rates (both R_1 and R_2) tend to increase as a function of the added protein concentration (2-5). There are several reasons for this increase and the appropriate theory was previously tested and reviewed in detail (2,5). Briefly, the explanation is as follows:

1. - water "bound" to the protein has a different rate of relaxation than "free" water; the "bound" water moves essentially with the relatively slowly rotating protein (e.g., on a scale of a few nanoseconds for lysozyme and β -lactoglobulin).

2. - "free" and "bound" water exchange rapidly, so that r.f. excited "free" water molecules can bind to a protein, give up nuclear spin energy (relax), and rapidly return to the solution, or vice versa.

In this paper, we will focus on a simple two-"state" model for the analysis of NMR relaxation measurements which seems to be appropriate for ^{17}O NMR relaxation, and to a limited extent also to ^2H NMR relaxation.

Protein Activity Model

For a two-"state" model ("bound" and "free" water), Kumosinski and Pessen (4) have shown that for the change in R_{obs} , the observed longitudinal or transverse relaxation rate of water deuterons in the presence of a varying protein concentration, c , one has:

$$R_{\text{obs}} - R_f = (R_b - R_f) n_w a_p / W \quad (1)$$

where R_f is the appropriate relaxation rate of "free" water (R_1 or R_2), R_b is the corresponding relaxation rate of "bound" water, W is the total concentration of water, and a_p is the activity of the protein. n_w is the degree of hydration (i.e. basically, the average number of molecules of water "bound" per molecule of dry protein or, in units consistent with the concentration units employed, the number of grams of "bound" water per gram of dry protein). For other ligands, in general, n_w differs from N , the number of available binding sites per substrate molecules, the difference being a function of association constant and ligand concentration. In the case of water, however, (which is a ligand present in such vast excess that the substrate is saturated with it), the distinction between n_w and N disappears. In the following, we will, for simplicity and convenience, use the expression "hydration" for short to indicate the quantity n_w in units of g/g. Furthermore,

$$a_p = c \exp(2B_0 c + \dots) \quad (2)$$

where B_0 is the second virial coefficient of the protein, activity, a_p .

Values for R_{1b} or R_{2b} , n_w , and τ_c were obtained by simultaneous solution of the Kubo-Tomita-Solomon equations (6-7),

$$R_{1b} = 2K\tau_c[(1 + \omega_0^2\tau_c^2)^{-1} + 4(1 + 4\omega_0^2\tau_c^2)^{-1}] \quad (3)$$

and

$$R_{2b} = K\tau_c[3 + 5(1 + \omega_0^2\tau_c^2)^{-1} + 2(1 + 4\omega_0^2\tau_c^2)^{-1}] \quad (4)$$

where R_{1b} and R_{2b} are the longitudinal and transverse relaxation rates, respectively, and τ_c is the correlation time of the "bound" water; ν_0 (or $\omega_0 = 2\pi\nu_0$) is the nuclear angular precession frequency (Larmor frequency) in Hz or in radians per second, respectively, K is a measure of the strength of the nuclear interaction, i.e.

$$K = (3/80)(e^2qQ/\hbar)^2(\eta^2/3 + 1)S^2 \quad (5)$$

where, e is the electronic charge, 1.6022×10^{-19} coulomb, q is the electric field gradient, Q is the nuclear electric quadrupole moment, \hbar is Planck's constant divided by 2π , 1.056×10^{-27} erg \cdot s, η is a dimensionless parameter measuring the deviation from axial symmetry (8), and S is the order parameter for intermediate asymmetry of the motion of the "bound" water (9). Hence, this thermodynamic theory can be used whether isotropic ($S = 1$) or anisotropic motion ($S < 1$) of bound water is hypothesized. In the latter case, the "bound water" should be thought of in the sense of "hydrodynamically influenced layers" or "surface-induced probability distribution of water molecules." For these experiments, η is assumed to be zero, $\nu_0 = 9.17$ MHz and $e^2qQ/\hbar = 215.6$ kHz (10).

Lysozyme Solutions and Hydrated Powders

Lysozyme in aqueous solutions has become a test-system for theories of protein hydration based on nuclear spin relaxation studies. As in the cases of polysaccharides (11) and muscle (12), the interpretation of ^1H NMR relaxation measurements for lysozyme in aqueous solutions is complicated by cross-relaxation between water protons and protein protons (13), as well as by proton chemical exchange (14). In a recent report (14), it was shown that ^2H NMR relaxation of D_2O in lysozyme solutions is also affected by chemical exchange, and that the contribution from deuterium chemical exchange increases markedly with increasing lysozyme concentration (Figure 2 of ref. (14)) and increasing pH (Figure 3 of ref. (14)). These findings disagree with a previous claim that scaled water relaxation rates are the same for ^2H and ^{17}O NMR of water in lysozyme solutions (15).

The magnetic field, pH, and concentration dependence of the ^2H and ^{17}O NMR relaxation rates for lysozyme solutions suggest the presence of anisotropic motions of water "bound" to lysozyme, and are consistent with fast exchange of water between the "bound" and "free" water populations. However, two kinds of motions, **slow** and **fast**, must be postulated in order to explain the difference between R_2 and R_1 values at high field (in the extreme narrowing limit, $\omega_0 \cdot \tau_c > 1.0$), as shown in Figures 4 and 5 of ref. (14). The correlation time for the slow motions of water "bound" to lysozyme was determined from the field dependence of the ^{17}O NMR relaxation rates to be of the order of 4.7 to 7.4 ns (depending on salt content), in agreement with recent results by frequency-domain fluorescence. The fast motions of water "bound" to lysozyme had a correlation time value of the order of 30 ps, in excellent agreement with the results reported previously (16).

The NMR relaxation data for hydrated lysozyme appear to be sensitive to both the

state of protein aggregation and the ionization of the protein side chains. The molecular dynamics of lysozyme hydration was shown to be, therefore, best described by a dual-motion anisotropic model with fast exchange of water between the "bound" and "free" populations.

Hydration of Myoglobin Microcrystals

Myoglobin microcrystals provide an interesting case for protein hydration studies, because the microcrystals can be oriented in a high magnetic field and the $^1\text{HO}^2\text{H}$ signal can be readily separated from the resolved quadrupole splittings of selectively deuterated, C^2H_3 -labelled methionines-55 and -131 (17). The microcrystalline samples have to be maintained, however, in a 90% saturated $(\text{NH}_4)_2\text{SO}_4$ solution. As shown in Figure 6, a deuterium NMR spin-echo (E^{XX}) technique yields inverted, but resolved quadrupole spectra for the labelled methionines, whereas the single Lorentzian peak of $^1\text{HO}^2\text{H}$ is not inverted and has substantially slower relaxation rates than those of the methionines. It is also interesting that one of the two labelled methionines, the one exposed to the solvent, has a significantly slower relaxation than the buried methionine group. Both groups are, however, relaxing much faster than water ($^1\text{HO}^2\text{H}$). The difference in the ^2H NMR spin-echo (E^{XX}) responses of the labelled methionines, in comparison with water (Figure 6), is caused by the presence of significant dipolar interactions both between the methyl group deuterons and with the surrounding lattice of protons in the myoglobin microcrystals. Such dipolar interactions are averaged to zero in liquid water by the extremely fast motions of the water molecules. In agreement with other reports on hydrated lysozyme powders (3) and solutions (14), there are no observable, resolved deuterium quadrupole splittings for "bound" water in myoglobin microcrystals. The ^2H NMR spin-echo (E^{XX}) experiments illustrated in Figure 6 show, in fact, that "bound" water in oriented myoglobin microcrystals has only "liquid-like" character and experiences no significant dipolar interactions, as would be the case for groups, or molecules, that have slow motions, such as the labelled methionines. One has to conclude, therefore, that "bound" water in myoglobin microcrystals is not "solid-like" in any sense, and that its dynamic characteristics are close to that of bulk water. The "bound" water motions also appear to be essentially isotropic in oriented myoglobin microcrystals, since no orientation dependence of the water ($^1\text{HO}^2\text{H}$) peak was observed.

Myosin in Electrolyte Solutions

Hydrated myosin provides an interesting model system for hydration studies of large proteins. Unlike lysozyme and myoglobin, which are globular proteins, myosin is a rod-like protein of ~500,000 molecular weight. Furthermore, myosin solubility depends markedly on the presence of added electrolytes. In our recent studies of myosin in electrolyte solutions by NMR techniques (18), it was shown that the analysis of the salt or protein concentration dependences of the NMR relaxation rates requires the use of a thermodynamic linkage approach (19) to fit the relaxation data. The salt-variation curves for myosin (Figures 7A and 7B) and myofibrillar proteins (Figures 7C and 7D) are dominated by the ion binding to myosin and by the fast chemical exchange of both ions and water between "bound" and "free" populations.

Ion-specific effects were observed (18) by comparing the myosin behavior in NaCl solutions with that in KCl solutions. Ion and water "binding" were also strongly pH-dependent (data not shown), as one could predict from the known pH titration behavior of myosin (20).

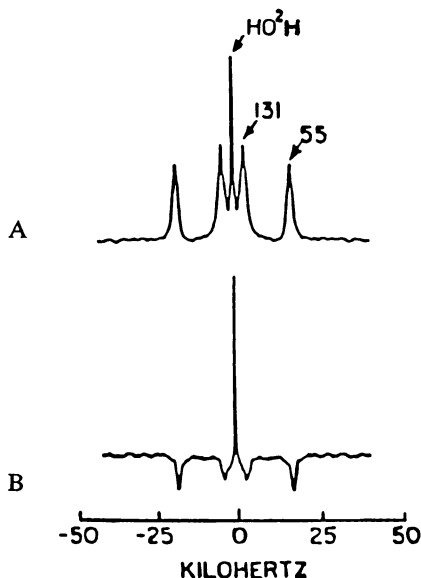


Figure 6. Deuterium NMR spin-echo spectra in polycrystalline S-methyl- $^2\text{H}_3$ methionine at 293 K for two pulse sequences: (A) spin-echo following an XY sequence with $\beta = 90^\circ$; (B) spin-echo following an XX sequence with $\beta = 55^\circ$ (β = pulse width or rotation angle); data were shifted to the echo maxima for XY echoes, or minima for XX echoes, and no instrumental phase corrections were made). Data were recorded at 55.3 MHz (corresponding to a magnetic field strength of 8.45 T). The 90° pulse widths were 2.0 to 2.5 μs , $2\tau = 100 \mu\text{s}$, recycle time 0.5 s, 2 MHz data acquisition rate, 4096 data points per spectrum, no line broadening, and 100 scans per spectrum. Spectra were symmetrized about zero frequency, because single phase-detection was employed. (Modified from ref. 17).

Myosin solubility and self-association were also found to be thermodynamically linked to ion-binding. Hydrophobic interactions between the tail parts of the myosin molecules are presumably the major cause of myosin self-association in the absence of salt. In addition, ionic interactions, mainly between carboxyl and ammonium or imidazolium groups, are also likely to be involved in myosin dimerization; dimer formation decreases myosin hydration, as observed in the ^{17}O NMR experiments. The heterogeneity of myosin **B** is reflected in its large, second virial coefficient and appears to be correlated with larger aggregate sizes and higher self-association rates than those observed for myosin **A** without salt.

In spite of the relatively large molecular weight of myosin ($M_w \sim 500\text{k}$) it has not yet been possible to determine if there is also a slow (ns) component of "bound" water motions in myosin solutions. However, it has been possible to determine, from a thermodynamic linkage analysis, the details of the cooperative self-association / dissociation processes of myosin in the presence of varying salt concentration (18). Such details could not be previously resolved by analyzing light-scattering data for myosin. Further studies of the magnetic field dependence of the NMR relaxation

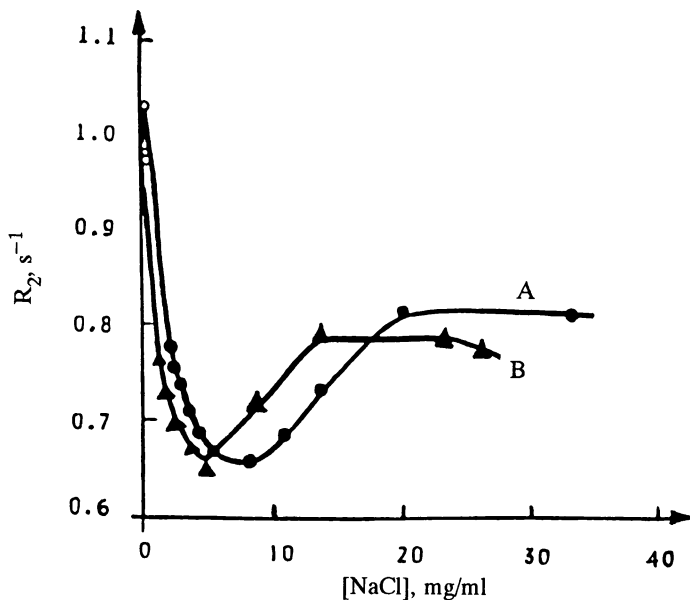


Figure 7A&B. Dependence of the ^1H NMR transverse relaxation rates, R_2 (s^{-1}), on NaCl concentration in aqueous solutions for (A) myosin A and (B) myosin B, at a fixed protein concentration of 8.0 and 8.5 % (w/v), respectively.

rates of water in myosin solutions with added electrolytes should provide additional information about the molecular dynamics of myosin hydration, and would allow one to evaluate the possibility of slower "bound" water motions related to the tumbling rate of myosin in solution.

Myofibrillar Proteins in Solutions with Electrolytes

Myofibrillar proteins are a major group of muscle proteins that are important both technologically and from a biomedical standpoint. Unlike purified myosin from this group, which has been extensively studied, the physicochemical properties of the mixture of myofibrillar proteins have been much less studied.

In a recent report, multinuclear spin relaxation measurements of myofibrillar proteins in solutions with electrolytes were presented (21), and the suggestion was made that their solubility and self-association properties depend on ion binding and pH in a manner broadly similar to soy protein isolates (22). Therefore, as explained above, a thermodynamic linkage analysis of myofibrillar protein aggregation and ion binding is appropriate. Nonlinear regression yielded the average aggregate sizes and 'apparent' ion binding constants. In the absence of added salt, the analysis of the ^{17}O NMR relaxation measurements suggested the formation of predominantly myosin dimers at low myofibrillar protein concentrations, followed by the cooperative aggregation into larger n-mers at the higher protein concentrations. The average, fast correlation time of water "bound" to myofibrillar proteins was about 17 ps in the presence of 0.5 M NaCl; the slow correlation time could not be estimated from NMR transverse relaxation time data, because of the very large size ($> 500,000 M_w$) and heterogeneity of the myofibrillar protein mixture. Sorption isotherms were determined in parallel experiments to the NMR studies and strongly indicated the aggregation of the myofibrillar proteins at higher concentrations, with or without

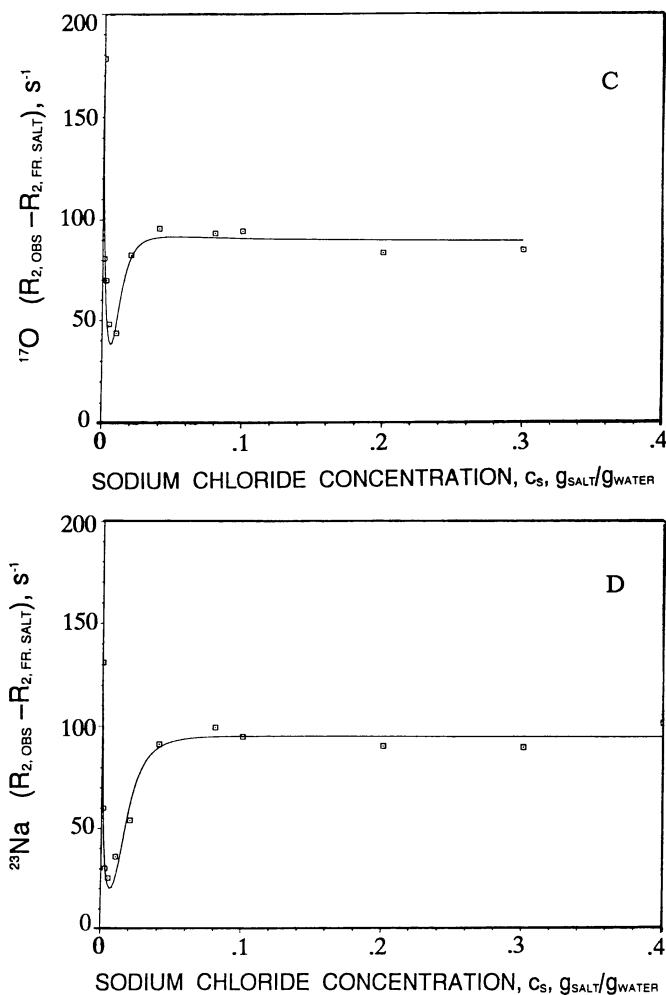


Figure 7C&D. Comparison of water ^{17}O NMR and ^{23}Na NMR transverse relaxation data (open squares) with cooperative ion-binding, n-mer models: (C) nonlinear regression analysis of ^{17}O NMR data with the $n = 4$ model; (D) nonlinear regression analysis of ^{23}Na NMR data with the $n = 4$ model.

NaCl. The salt binding to myofibrillar proteins resulted in a non-additivity of the “water activities” (i.e. relative vapor pressures) for the salt solution and the protein component (Figure 4 in ref. (21)). In the range of “water activities” above 0.90, the onset of non-equilibrium thermodynamic processes cannot be excluded, and therefore, the analysis of the results for the hydrated powders may require further detailed studies.

The molecular dynamics of water and ions in myofibrillar solutions occurs on a ten-picosecond time scale, similar to that of other proteins that were studied by NMR; the exchange of water and ions between the “bound” and “free” populations is determined to be fast by NMR.

Wheat Protein Hydration and Ion-Binding Properties

Wheat storage proteins are the major source of protein (in terms of bulk quantities) in the Western hemisphere, and certain other areas of the globe. Technological developments in the use of wheat proteins in foods are being rapidly engineered, although such developments are often by trial-and-error, rather than being based on a systematic, physicochemical approach. During the last ten years, however, major advances have also been made in the field of physical chemistry of wheat storage proteins. Among the techniques employed were high-resolution NMR, nuclear spin relaxation, X-ray scattering, dielectric relaxation, and photoacoustic spectroscopy (23-24). Several resolved ^1H NMR peaks were observed for wheat gluten proteins at high-field (Figures 2 and 3 in ref. (21) and Figure 2 in ref. (24)). The ^{17}O NMR peak of water in the hydrated wheat gluten, or wheat flour, is much broader than that of liquid $^1\text{HO}^2\text{H}$, and appears to follow the predictions of a two-"state", fast-exchange model, up to at least 15 to 20% solids. The effects of protein aggregation on water mobility and protein hydration in these systems, however, were not investigated in detail, partially because of the very heterogeneous nature of these systems. The molecular motions of several amino acid side chains in hydrated wheat grains, gluten, glutenins, and wheat doughs were found to be quite fast in the presence of 40% $^2\text{H}_2\text{O}$ (24), but were slow in the dehydrated powders (for less than 10% moisture contents). It was possible, however, to study in greater detail the hydration and ion-binding properties of purified wheat gliadin at pD 3.4 by employing a Mn^{2+} , paramagnetic ion probe (25). The range of salt concentrations that could be studied (0.2 to 4 mM MnCl_2) was limited by the solubility of wheat gliadins in the presence of Mn^{2+} ions. The average number of available binding sites for wheat gliadin was found to be $n = 7 \pm 1$, and the average binding constant was found to be $k_f = 12.9 \text{ M}^{-1}$. The two-"state", fast-exchange model could be employed to analyze the NMR relaxation results, up to about 15% wheat gliadin concentration in solution at pD 3.4. The interactions of wheat gliadin with fructose and sucrose could be followed by NMR relaxation up to high carbohydrate concentrations ($\sim 0.55 \text{ mol/L}$); the ^1H NMR transverse relaxation rate was found to decrease linearly with increased carbohydrate concentration. The observed NMR relaxation behavior was interpreted in terms of weak interactions of the wheat gliadin with fructose and sucrose, and *preferential hydration* of the wheat gliadins (25). Such results may also be technologically relevant to wheat protein stabilization, or to the use of carbohydrate cryoprotectants with wheat storage proteins.

Molecular Dynamics of Soy Protein Hydration

The solubility of soybean protein isolates exhibited a salt-dependent variation that has been investigated in detail, for both native and heat-denatured proteins (22). An attempt was made previously to interpret such solubility data in terms of the theory of Melander and Horvath (26). This theory predicts that at low salt concentration, the solubility of a protein should increase because of the electrostatic contributions to the free energy of the system; at high salt concentrations, the salting-out free energy is expected from such theories to predominate because of the increase of the surface tension of the electrolyte solution and the exposed hydrophobic regions of the protein. In the case of soy protein isolates, Shen (27) found the opposite behavior to occur, for

both the native and heat-denatured proteins. The protein solubility decreased with added salt to a minimum value, beyond which it increased to a constant limiting value, at about 1 M added NaCl. The shapes of the native and heat-induced solubility profiles of the soy protein isolates were similar for a series of salts: NaCl, NH₄Cl, NH₄Br, NH₄NO₃, and NaI. The limiting values of solubility at high salt concentrations followed the lyotropic, or Höffmeister, series. These observations were initially explained by an increase of the solvent-exposed hydrophobic area of soy proteins, as a result of the salt-association of the proteins upon salt addition; this process would be followed by a salting-in process, induced by the increased dipole moments of the soy proteins resulting from a non-specific solvation effect at the higher salt concentrations. Ion binding to soy proteins was neglected in this initial interpretation of the solubility data (27). In a subsequent, recent report (22), it was pointed out that the latter interpretation is likely to be incorrect; in fact, sodium and ammonium ion binding to soy proteins dominates the solubility properties of these proteins. The solubility profiles of soy protein isolates (27) were quantitatively analyzed by nonlinear regression, using equations derived on the basis of Wyman's theory of thermodynamic linkage (19). The latter approach demonstrated that the soy protein solubilities are linked to the free energy of ion binding to the proteins. All solubility profiles, for all added salts, were then fitted quantitatively, by employing ion-binding "constants" treated as parameters.

The apparent ion-binding constant for salting-out, k_1 , always exceeded the ion-binding constant for salting-in, k_2 , but the solubility, S_1 , for the salting-in processes had no apparent trend with the type of salt added. The solubility for the salting-out process, S_2 , for the denatured soy proteins correlated very well with the lyotropic series.

It was concluded from such studies that the solubility of soy proteins is **thermodynamically linked** to their salt-, or ion-, binding capacity (22), in terms of an **isoelectric binding model**. The salt cations bind to negative sites on the soy protein surface, with an average apparent constant k_1 , and produce a species of zero net charge with the corresponding solubility, S_1 . Further salting-in of the protein was predicted to occur, either as a result of only cation binding, k_2 , to the exposed protein sites (yielding proteins with a net positive charge), or because of the binding of both cations and anions (from the added salt) to the oppositely charged protein sites (yielding protein species with a zero or negative charge). For the heat-denatured soy isolate, the salting-in process is most likely caused by both salt anion and cation binding to the corresponding positive and negative sites on the protein, yielding an ion-protein complex with a net zero charge. The salting-out process of both native and heat-denatured soy protein isolates is caused by cation binding to proteins with a net negative charge, leading to **isoelectric precipitation**. Such results showed that the binding of salt cations and/or anions to soy proteins has a major influence on protein solubility, even if the values of the binding constants are small. Small binding constants (weak binding) would also suggest the presence of fast exchange between the bound ions and the electrolyte solution.

The ion binding to soy proteins was recently shown to strongly influence their hydration properties, as determined by multinuclear spin relaxation measurements. Both ¹⁷O and ¹H NMR relaxation could be interpreted with a fast-exchange, two-"state" model for water, if the soy protein activity was also taken into consideration. The NMR transverse relaxation rates of soy proteins in solution showed a marked dependence on both pH and salt variation. The interpretation of these NMR

relaxation data considered the effects of the ionized groups of the protein, the state of protein aggregation, and ion binding to the soy proteins. Generally, electrostatic repulsions between soy proteins increased with increasing pH, and the addition of salt tended to suppress such repulsions. When added at neutral pH, however, the salt ions enhanced protein intermolecular repulsions, because of the increase of the protein net charge upon binding of the ions.

Measurements of ^{13}C NMR longitudinal relaxation times for specific amino acid residues of soy glycinin revealed the presence of fast local reorientation of such groups, with pseudo-isotropic rotational correlation times ranging from 20 ps to 0.16 ns (28). Amongst such groups were Glu CH_2 , Lys ϵCH_2 , Arg δCH_2 , and The βCH . Not all groups, however, have such short correlation times; a large number of the amino acid residues, located in the structured domains of the soy glycinines, are likely to have much longer correlation times. The molecular dynamics of the amino acid residues (whose ^{13}C NMR peaks were resolved) exhibited a marked dependence on both pH and added salt. The ^{17}O NMR transverse and longitudinal relaxation time measurements were analyzed with fast exchange, two-"state", anisotropic dual-motion model (14) discussed earlier for lysozyme solutions. The value of the fast correlation time of water "bound" to soy proteins was about 32 ps, whereas the value of the slow, average correlation time was found to be - 14 ns. The latter value is much too small in comparison with the slow tumbling rate of a soy protein aggregate, such as, for example, an 11S soy glycinin hexamer; the value of 14 ns does, however, match rather well the tumbling rate of a 44,000 M_w subunit of the 11S fraction. Recently, ^{17}O NMR measurements on purified 7S and 11S soy globulin fractions yielded values of 62 and 125 ns, respectively, at pD 7.4, 20°C, with 0.5 M NaCl, (Wei and Baianu, unpublished results).

Conclusions

The mechanisms for the interactions of anions and cations with proteins in aqueous solutions were investigated by nuclear magnetic resonance over a wide range of salt concentrations. Markedly nonlinear dependences of ^{17}O and ^{23}Na NMR transverse relaxation rates on salt concentration were analyzed with a thermodynamic linkage model of salt-dependent solubility and hydration (**ligand-induced association model**), according to Wyman's theory of linked functions. Nonlinear regression analysis of both ^{17}O and ^{23}Na data suggested cooperative, reversible binding of hydrated ions to myofibrillar proteins. Both ions and water were found to exchange fast, on the NMR timescale, between the binding sites of the myofibrillar proteins and the aqueous solution. At sodium chloride concentration higher than about 0.1 grams salt/gram water, ion activities had marked effects upon the NMR relaxation rates of both ions and water. A salt activity model allowed quantitative fitting of the NMR data at high salt concentrations.

Literature Cited

1. Ahlström, P., et al. *J. Amer. Chem. Soc.* **1987**, *109*, 1541.
2. Kumosinski, T. F.; Pessen H. *Arch. Biochem. Biophys.* **1982**, *218*, 186-292.
3. Lioutas, T. S.; Baianu, I. C.; Steinberg, M. P. *Arch. Biochem. Biophys.* **1986**, *247*, 68.
4. Lioutas, T. S.; Baianu, I. C.; Steinberg, M. P. *J. Agric. Food Chem.*, **1987**, *35*, 133.

5. Kumosinski, T. F.; Pessen H. In *NMR in Agriculture*; Pfeffer, P. E.; Gerasimowicz, W. V., Eds.; CRC Press: Boca Raton, FL, 1989.
6. Kubo, R.; Tomita, K. *J. Phys. Soc. Jpn.*, **1954**, *9*, 888.
7. Solomon, I. *Phys. Rev.* **1955**, *99*, 559.
8. Abragam, A. *The Principles of Nuclear Magnetism*; Oxford Univ. Press (Clarendon): London, UK and New York, NY, 1961.
9. Smith, I. C. P. In *NMR of Newly Accessible Nuclei*; Laszlo, P., Ed.; Academic Press: New York, NY, 1983, Vol. 1, pp.1-30. .
10. Waldstein, P.; Rabideau, S. W.; Jackson, J. A. *J. Chem. Phys.* **1964**, *41*, 3407.
11. Mora, A.; Baianu, I. C. *J. Agric. Food Chem.* **1989**, *37*, 1459.
12. Edzes, H. T.; Samulski, E. T. *J. Magn. Reson.* **1978**, *31*, 207.
13. Kalk, A.; Berendsen, H. J. C. *J. Magn. Reson.* **1976**, *24*, 343.
14. Kakalis, L.; Baianu, I. C. *Arch. Biochem. Biophys.*, **1988**, *267*, 829.
15. Koenig, S. H.; Hallenga, K.; Shporer, M. *Proc. Natl. Acad. Sci. USA* **1975**, *72*, 2667.
16. Halle, B.; Andersson, T.; Forsén, S.; Lindman, B. *J. Am. Chem. Soc.* **1981**, *103*, 500.
17. Baianu, I. C.; Gutowsky, H. S.; and Oldfield, E. *Biochem.* **1984**, *23*, 3105.
18. Baianu, I. C. et al. In *NMR in Agriculture and Food Chemistry*; Finley, J., Ed., Plenum Press: New York, NY, 1990.
19. Wyman, J., Jr. *Adv. Protein Chem.* **1964**, *19*, 223.
20. Lewis, M. S.; Saroff, H. A. *J. Amer. Chem. Soc.* **1957**, *79*, 2112.
21. Lioutas, T. A.; Baianu, I. C.; Bechtel, P. J.; Steinberg, M. P. *J. Agric. Food Chem.* **1988**, *36*, 437.
22. Kumosinski, T. F. *J. Agric. Food Chem.* **1988**, *36*, 110.
23. Baianu, I. C.; Förster, H. *J. Appl. Biochem.* **1980**, *2*, 347.
24. Baianu, I. C.; Johnson, L. F.; Waddell, D. K. *J. Sci. Food Agric.* **1982**, *33*, 373.
25. Mora, A. *M.S. Thesis, University of Illinois*, **1989**.
26. Melander W.; Horvath, C. *Arch. Biochem. Biophys.* **1977**, *183*, 200.
27. Shen, J. L. In *Protein Functionality in Foods*; Cherry, J. P., Ed., ACS Symposium Series 147, American Chemical Society, Washington, D. C., 1981.
28. Kakalis, L.; Baianu, I. C. *J. Agric. Food Chem.* **1989**, *37*, 1222.

RECEIVED July 14, 1994

Chapter 19

NMR and Molecular Modeling Evidence for Entrapment of Water in a Simple Carbohydrate Complex

P. Irwin, Gregory King, Thomas F. Kumosinski, P. Pieffer, J. Klein¹, and L. Doner

Eastern Regional Research Center, Agricultural Research Service,
U.S. Department of Agriculture, 600 East Mermaid Lane,
Philadelphia, PA 19118

During structural investigations of a glucuronic acid derivative dissolved in DMSO we recognized a water activity-dependency between the NOESY cross-peaks of H₂O and the carbohydrate's hydroxyl protons. The -OH↔H₂O first order exchange rate constant increased from 0.32 to 11.14 s⁻¹ as the molar ratio of H₂O:sugar increased from only ca. 4 to 5. The latter finding indicated that the -OH↔H₂O proton exchange process, which is proportional to the translational diffusion of water, diminished as H₂O approached the concentration which exists in the crystalline structure and was, presumably, entrapped by our glucuronic acid derivative forming a stable complex. Supporting this, a significant upfield shift in the resonance frequencies of the hydroxyl (-OH $\Delta\delta_{\text{ave}} = 86.33$ Hz) protons was observed (CH $\Delta\delta_{\text{ave}} = 0.25$ Hz) when water was removed by reaction with 2,2-dimethoxypropane. Molecular dynamics calculations (100 ps) on the energy-minimized carbohydrate-water complex confirm the presence of 2-3 near neighbor H₂O molecules associated with the polar functional groups. In fact, the computationally-derived weighted average distance of all water molecules adjacent to the -OH groups was found to be inversely proportional to the individual -OH $\Delta\delta$ s.

Knowledge about the interactions between carbohydrates and water is of some consequence because important chemical and physical properties are imparted by the way these compounds coexist. DMSO is a good solvent for understanding these interactions because carbohydrates retain much of their H₂O-induced conformation (*I*) in DMSO and one can specifically observe, assign and study a carbohydrate's hydroxyl exchange with small quantities of H₂O because the -OH resonance frequencies are dissimilar.

In this chapter we present spin-lattice relaxation, 2D NMR, chemical shift and molecular dynamics evidence that *N*-phenyl (*N*-phenyl- β -D-glucopyranosylamine)-

¹Current address: Department of Field Crops, Volcani Center, Bet Dagan, Israel

uronamide's (*N*-phenyl uronamide; Figure 1) waters of crystallization are tightly bound to the polar functional groups of the sugar moiety even after extreme dilution in DMSO.

Materials and Methods

Sample Preparation. *N*-phenyl uronamide was synthesized and purified as described previously (2,3). D-glucopyranuronic acid was dissolved in H₂O (1.5 g/25 mL). Aniline (2 mL) was dropped slowly into the stirring mixture and the pH adjusted to 4.75 on a Radiometer (reference to brand or firm name does not constitute endorsement by the U. S. Department of Agriculture over others of a similar nature not mentioned) pH stat. Approximately 3 g of 1-ethyl-3-[3-(dimethylamino)propyl]carbodiimide (EDC) was added to the solution and the pH stat activated causing 0.1N HCl to be delivered to the reaction mixture to maintain the pH at ca. 4.75 (4-7). When no more titrant was needed to maintain a constant pH the reaction was complete. At this point an insoluble off-white precipitate had formed and was subsequently washed with H₂O to remove unreacted aniline or EDC. Excess water was removed by washing the precipitate with chilled EtOH. The acid sugar derivative was then dissolved in hot EtOH and 2-4 mm needle-like crystals formed overnight at room temperature. For production of the anhydrous form, the above procedure was repeated except that a small amount of 2,2-dimethoxypropane was added to the EtOH to react with unwanted water (e.g., H₂O + 2,2-dimethoxypropane → 2MeOH + acetone).

NMR Spectroscopy. Samples for NMR were prepared in a dry box. *N*-phenyl uronamide crystals were dissolved in DMSO-*d*₆ (≥ 99.5 atom % ²H) which had been stored several days with molecular sieve pellets under dry N₂ (the DMSO contained, except when specified, ca. 30 mM H₂O even in the presence of "dry" molecular sieves). Several DMSO-*d*₆ washed molecular sieves were kept in the 5 mm NMR tubes to maintain the sample in a relatively dry state; the NMR tubes were closed and wrapped with parafilm or sealed under vacuum to assist in the exclusion of extraneous H₂O vapor. The samples were stored at 3°C and underwent no obvious degradative process, such as pyranose ring opening and associated Amadori rearrangement (2) or loss of the C₁ amine functionality (Figure 2). Evidence for water activity dependent hydrolysis (2) is provided in Figure 3. Using reverse phase HPLC, one can see that increasing the water concentration from ca. 0 to 50% (v/v) in MeOH increases the first order rate constant by a factor of about 9. It is noteworthy that there was a significant degree of hydrolysis (*t*_{1/2} ~ 83 min) in absolute MeOH. For these kinetic experiments 3 mg of *N*-phenyl uronamide were dissolved in 10 mL of either 50% MeOH:H₂O or abs. MeOH and maintained at 40°C. At various times 100 μL of each solution was injected into an HP 1090 HPLC system equipped with a supelco LC-18 reverse phase (15 cm; 5 μm particle size) column; 50 % MeOH was used as the mobile phase (0.2 mL min⁻¹). The various peaks were checked against

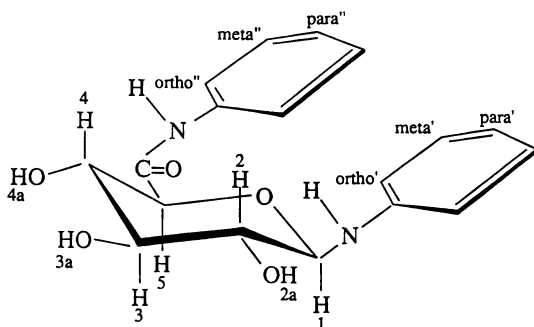


Figure 1. Structure and conformation *N*-phenyl uronamide ^1H position labels.

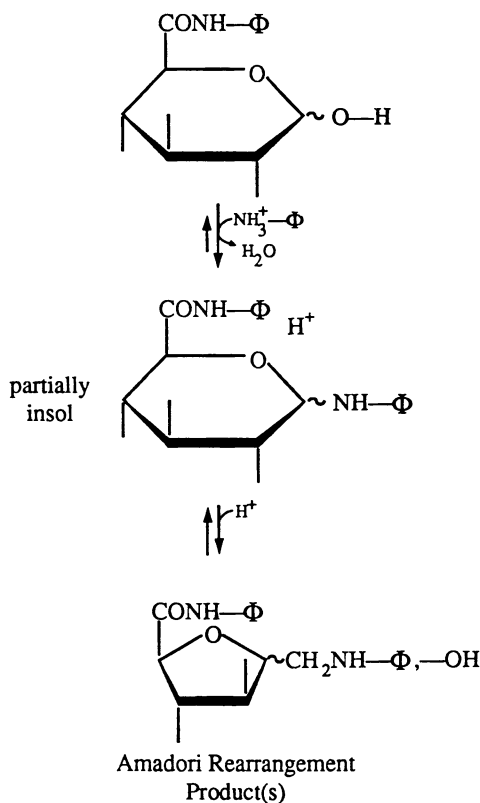


Figure 2. Reaction scheme proposed for the formation of *N*-phenyl uronamide and its Amadori rearrangement product(s). Reproduced with permission from Ref. 2. Copyright 1990, Journal of Carbohydrate Chemistry.

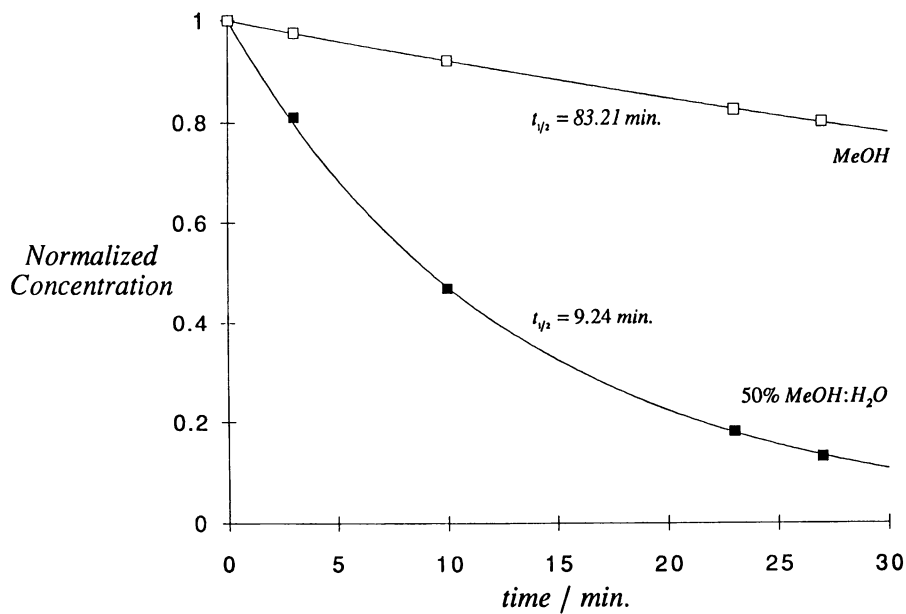


Figure 3. Change in the relative concentration of *N*-phenyl uronamide over time in 100% MeOH (open squares) and 50% MeOH:H₂O (closed squares) at 40°C.

standards of aniline and *N*-phenyl-D-glucopyranuronamide (e.g., *N*-phenyluronamide without the C₁ amine functional group).

Before NMR experiments, the 90° pulse was determined for each condition, such as variable temperature or concentration, utilizing standard methods (8). All NOESY spectra were collected on a JEOL GX-400 NMR spectrometer system operated at ca. 400 MHz (9.40 T) using 5 mm probes (3). Computer line broadening was selected to be approximately equal to the digital resolution. These experiments were acquired using a matrix of 128 x 1024 (t₁ x t₂), 256 x 2048 after zero-filling, complex data points which represented a spectral width of 953.1 Hz for either dimension. For each t₁ spectrum collected, 16 transients were acquired. A sine-bell apodization function was used to process these data. All quantitative 2D Overhauser enhancement matrices were processed without symmetrization. All ROESY (2) data were collected using a JEOL GSX-400 NMR spectrometer with a proton full-power 90° pulse of 10.5 μs. Acquisition data sets consisted of 2048 complex points for t₂ and 64 acquisitions for each t₁ data set. A spin-lock field of 3 kHz, 1 kHz off-resonance from the average chemical shifts of the residual H₂O protons and the -OHs, was used for mixing times (τ_m) of 0.075, 0.2, 0.4 and 0.6 s. The data sets were zero-filled to 4096 t₂ points and 2048 for t₁. A phase-shifted sine-bell algorithm was used as the window function. All the -OH resonances (H_{2a} → 4_a) were integrated and fitted to an exponential function (equation 1)

$$I = I_o \left\{ 1 - e^{-x\tau_m} \right\} \quad (1)$$

$$I_o = \lim_{\tau \rightarrow \infty} I \quad (2)$$

using a modified Gauss-Newton procedure developed in this laboratory by Dr. William Damert.

Proton T₁ inversion recovery experiments were performed on JEOL NMR spectrometers operated at either 400 or 270 MHz (9.40 or 6.34 T). Each τ value was signal averaged for 64 acquisitions with 16 dummy scans. T_{1sat} experiments (9-11) were performed identically to the above except that the H₂O resonance was irradiated 721.67 Hz upfield from the C₄-OH (H_{4a}) resonance. All peak intensity data were fit to an exponential function (equation 3) utilizing the aforementioned curve-fitting procedure.

$$I_i = I_o \left[1 - 2e^{-\frac{-(\tau_i - \tau_o)}{T_1}} \right] \quad (3)$$

The T_{1sat}-associated pseudo first-order rate constant (κ_{sat}) calculation was

accomplished as shown in equation 4

$$\kappa_{sat} = \frac{1}{T_{1sat}} - \frac{I^+ / I^\theta}{T_{1sat}} \quad (4)$$

where I^+ / I^θ is the ratio of hydroxyl resonance integrals with irradiation on the H₂O resonance and 721.67 Hz downfield, respectively. T_{1sat} is the normal T_1 measurement but with spin saturation of H₂O.

The correlation time (τ_c) for the *N*-phenyl uronamide·H₂O complex and individual resonance T_{1i} s (T_{1i}) were estimated using equation 5

$$\frac{1}{T_{1i}} = \frac{3}{10} \gamma^4 \left[\frac{h}{2\pi} \right]^2 \sum_j \frac{1}{r_{ij}^6} \left\{ \frac{\tau_c}{1 + (\nu_o \tau_c)^2} + \frac{4\tau_c}{1 + (2\nu_o \tau_c)^2} \right\} \quad (5)$$

where ν_o was either 270 or 400 MHz and the interproton distance parameter, r_{ij} , was assumed to be 2 Å since the interproton distance parameter has only a minor effect on the calculated T_{1i} s.

Modeling Studies. Initially, the SYBYL software package was used to construct the *N*-phenyl uronamide molecule and to prepare an initial dimer configuration (see Results and Discussion section). Certain dynamics simulations (Table III, Figure 9) were performed with the program SCHIZO which is a generalized version of the SCAAS model (12). In this model the solute is situated within a spherical droplet of solvent molecules and constraints are placed upon the solvent molecules near the surface of the sphere to prevent them from “evaporating” from the droplet. Other simulations (Figures 10 and 11) used the Sybyl software package alone as a comparison. Force-field parameters for bonding terms and nonbonded van der Waals interactions were from Clark and co-workers (13). Atomic charges for the carbohydrate and DMSO were obtained with the electronegativity equalization algorithm of Gasteiger and Marsili (14) except that the DMSO charges were subsequently scaled to make the dipole moment agree with the experimental value. For the SCHIZO model, all water force-field parameters were taken from King and Warshel (12); molecular dynamics trajectories were propagated using a 2 fs time step and “temperatures” were maintained at the desired values by utilizing gentle velocity scaling.

Results and Discussion

N-phenyl uronamide is an unusual by-product of the activation of D-glucuronic acid's carboxyl group with a carbodiimide reagent (4-7) in the presence of the nucleophile aniline. Originally (2), the H₂O resonance was misassigned as the CH₃ of DMSO since (Figures 2 and 3) the carbohydrate was found to be unstable

in the presence of free H₂O and considerable effort (see Experimental) was taken to eliminate water from the samples. Other studies were performed because we supposed the observed cross peaks between the -OH groups and the "solvent" were due to magnetization transfer via spin diffusion or 2nd order Overhauser effects. However, upon treating *N*-phenyl uronamide with 2,2-dimethoxypropane (spectra shown in Figure 4) we discovered that this "solvent" peak disappeared and, therefore, was in fact a small amount of H₂O which co-crystallized with the solute and/or which was absorbed from the head-space above the solvent (ca. 30 mM). When *N*-phenyl uronamide was recrystallized, as described previously, from hot EtOH without 2,2-dimethoxypropane and examined via ¹H NMR, using "100%" DMSO-*d*₆, the H₂O's of crystallization were found to exist in a ca. 4:1 molar ratio to the acid sugar derivative; upon vacuum drying at ca. 100°C the level of hydration could be reduced to 1 H₂O:*N*-phenyl uronamide (C₁₈H₂₀O₅N₂·H₂O; 3). A similar inaccuracy (*I*) may have been made on a similar-sized carbohydrate, cellobiose, inasmuch as comparable "DMSO"/-OH interactions (1st or 2nd order Overhauser effects), via NOESY NMR, have been hypothesized.

Exchange and Exchange-like Phenomena. In 2D Overhauser enhancement spectroscopic experiments, cross peaks not associated with scalar coupling result from either direct cross relaxation (15,16) or exchange phenomena (15-18). Direct cross relaxation (e.g., a "1st order Overhauser effect", 19) is a through-space dipolar spin-spin interaction proportional to the inverse 6th power of the distance between the interacting spins. Exchange-like phenomena (15-23) can be simplified as follows (3):

1. chemical exchange processes:
 - a. chemical exchange (15-18)
 - b. stereochemical exchange (17)
 - c. relayed exchange (18)
2. magnetization exchange processes:
 - a. "2nd order Overhauser effects" ($\nu_0\tau_c \sim 1-10$; 19,21-23)
 - b. "spatial" and "spectral" spin diffusion ($\nu_0\tau_c \gg 10$; 20)

The 2nd order Overhauser effects and other spin diffusion-like processes can only occur in solutions of molecules with relatively long τ_c 's at high magnetic fields. Cellobiose, whose molecular weight (mw = 342.29) is similar to the title compound (mw = 344.36), is an unlikely candidate for any of the magnetization equilibrium processes listed above.

We became interested in the exchange behavior of the H₂O:*N*-phenyl uronamide system since the water that was there was not acting as a catalyst for either the hydrolysis or Amadori rearrangement of our compound. Previous studies (Figure 3; 2) indicated that even minute quantities of water in MeOH induced a breakdown of about 36% per h. Thus, in the presence of up to 4-5 moles of water per mole of solute no breakdown was observed in DMSO over periods of several days indicating that the water was not available to interact under

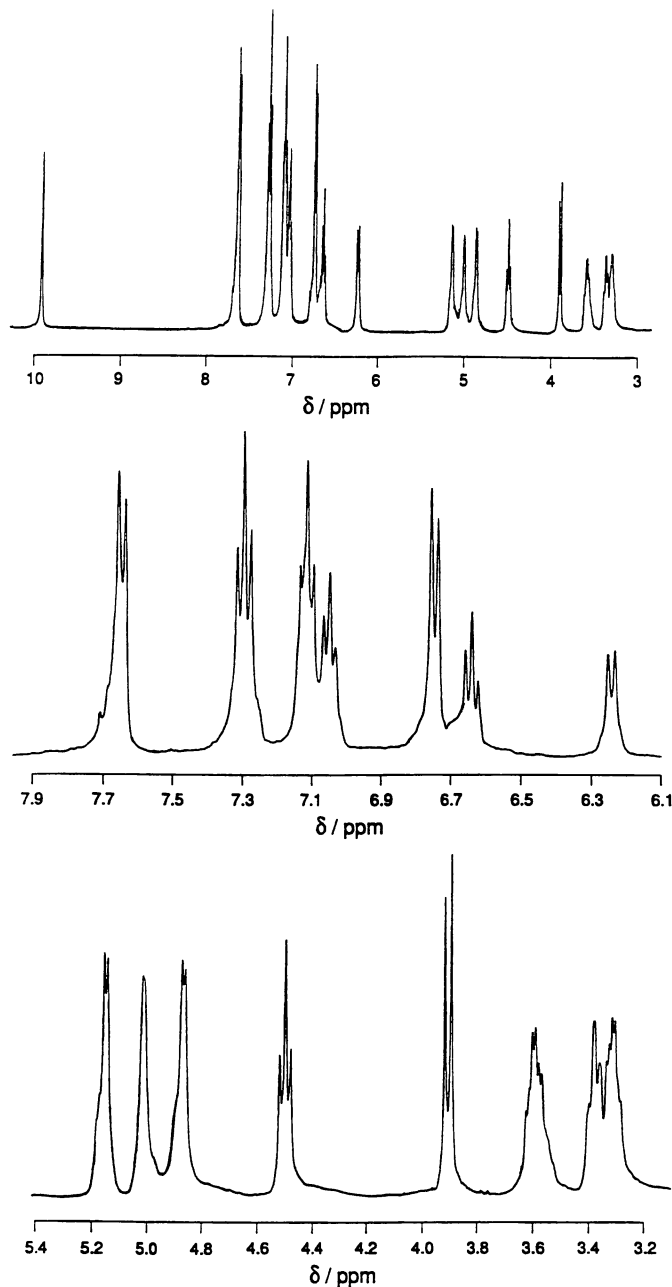


Figure 4. ^1H NMR spectra of *N*-phenyl uronamide in the anhydrous state. For the hydrated form, a water resonance would appear, depending on the temperature, at ca. 3.34 ppm. Reproduced with permission from Ref. 3. Copyright 1993, Journal of Carbohydrate Chemistry.

these conditions. The rate of chemical exchange was measured from the mathematical behavior of off-diagonal $-\text{OH} \leftrightarrow \text{H}_2\text{O}$ NOESY resonances as the mixing time (τ_m) was varied. For instance, during very slow chemical exchange (e.g., an amide-H, $\tau_c = 0.1\text{--}5$ ns; 18) NOESY cross peak integrals (I) should increase with mixing time (τ_m) as a typical 1st order rate process (equation 1) where I increases to a maximum, I_0 , as a function of τ_m and eventually levels off; in this relation κ is a 1st order rate constant in units of reciprocal time. For ^1H donor species, similarly-sized to the above and displaying a moderate exchange rate, I increases to I_0 in a similar fashion to the above example but rapidly declines thereafter (18). NOESY data (Figure 5; all data points resulted from the integration of all off-diagonal $-\text{OH} \leftrightarrow \text{H}_2\text{O}$ resonances) indicate that the $\text{H}_2\text{O}:\text{N}$ -phenyl uronamide complex underwent very slow exchange since $I_{-\text{OH}/\text{H}_2\text{O}}$ stabilized as τ_m approached 1 s. Also unusual was the fact that κ changed as a function of $[\text{H}_2\text{O}]:[\text{N-phenyl uronamide}]$. Rotating frame 2D Overhauser (ROESY; 3,15,16,24) enhancement $\text{H}_1 \leftrightarrow \text{H}_5$ cross peaks (Figure 6, upper spectrum) were negatively phased relative to the $\text{C}_{2,3}$ or $4\text{-OH}/\text{H}_2\text{O}$ cross peaks (Figure 6, lower spectrum) thereby eliminating 1st order Overhauser effects as a possible explanation of our data. With regard to N -phenyl uronamide $\cdot 4\text{H}_2\text{O}$, we found (Table I; 3) that τ_c was ca. 0.54 ns via spin-lattice relaxation measurements at two fields. The calculated $T_{1\text{Hs}}$ were observed to diverge from the experimental an average of only 0.48%. Clearly, based upon the τ_c calculation and ROESY experiments the observed cross peaks between the solvating species, H_2O , and the title compound's hydroxyl ^1Hs were due to chemical exchange. Further support for a slow exchange mechanism is presented in Figure 7 (3) whereupon inversion recovery experiments were performed with simultaneous irradiation at the H_2O 's resonance frequency ($T_{1\text{sat}}$) and at an equivalent frequency off-set downfield (T_1) from the observed hydroxyl proton ($\text{C}_4\text{-OH}$). Based upon $T_{1\text{sat}}$ and I^+/I^0 a κ_{sat} was found (equation 4) to be approximately 0.3 s^{-1} . The process of longitudinal relaxation without exchange effects would be most nearly represented by the $T_{1\text{sat}}$ curve (open diamonds). The differences between these two treatments demonstrates the profound effect of exchange on the relaxation behavior of the $-\text{OH}$ groups. Of course, magnetization equilibrium processes can be eliminated *a priori* as the basis of our observations since these processes occur only when τ_{cs} for molecular reorientation are much longer (19,21-23).

As mentioned previously, we noted that the $-\text{OH} \leftrightarrow \text{H}_2\text{O}$ exchange rate constant increased from 0.32 to 11.14 s^{-1} as the molar ratio of $[\text{H}_2\text{O}]:[\text{N-phenyl uronamide}]$ increased only from ca. 4.5 to 5.2 (Figure 8). This latter finding indicated that κ , which is proportional to the translational diffusion of H_2O (18), diminished as the water concentration approached that which was inherently complexed in the crystalline structure ($\text{C}_{18}\text{H}_{20}\text{O}_5\text{N}_2\cdot 4\text{H}_2\text{O}$) and was, presumably, tightly hydrogen bound to the $-\text{OH}$ and $-\text{NH}$ functional groups when put into solution in DMSO. To further support this contention (Table II), the anhydrous

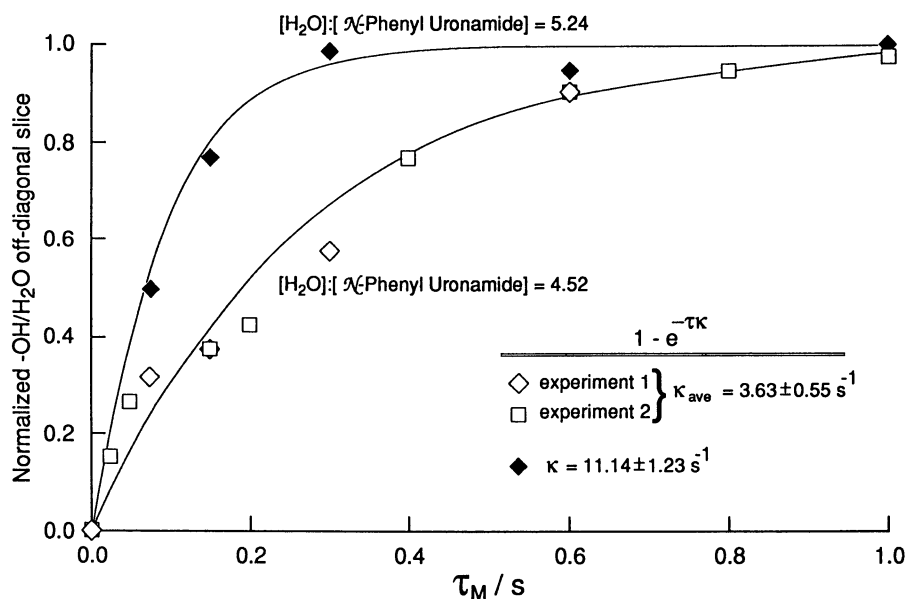


Figure 5. -OH \leftrightarrow H₂O cross peak areas plotted as a function of mixing time, τ_m , at 40°C. Curves resulted from best fits to a first order exponential rate expression (equation 1). Reproduced with permission from Ref. 3. Copyright 1993, Journal of Carbohydrate Chemistry.

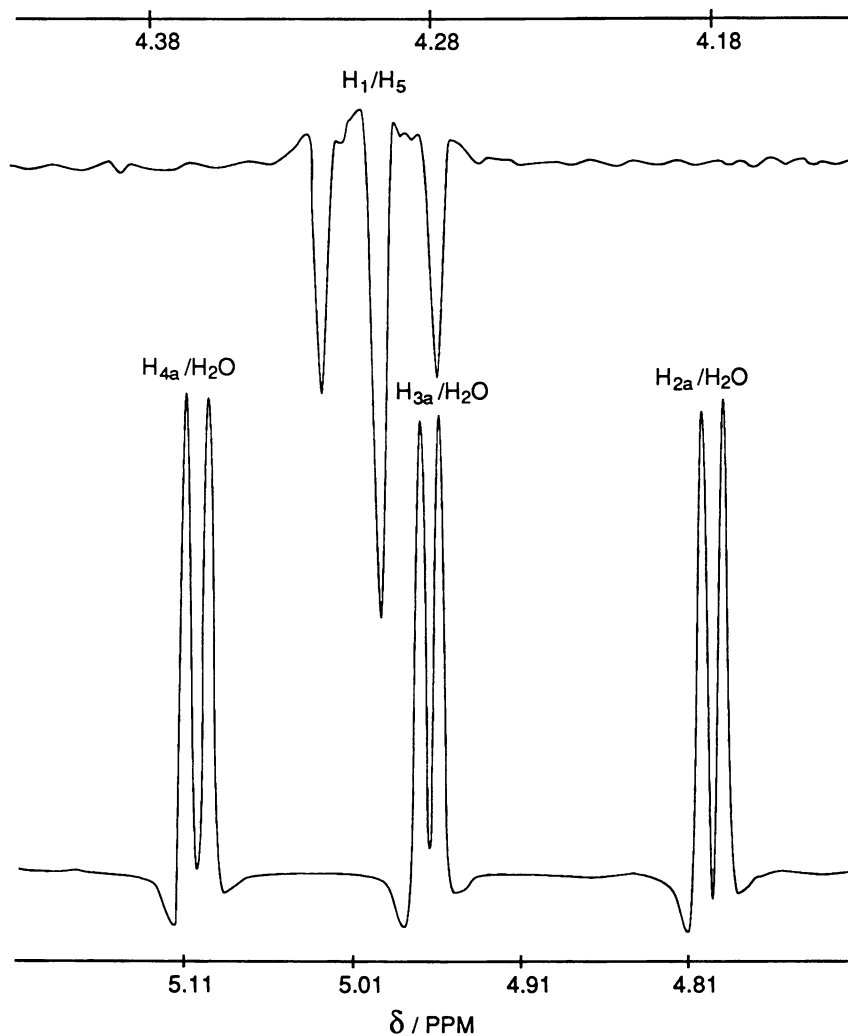


Figure 6. $H_1 \leftrightarrow H_5$ and $-OH \leftrightarrow H_2O$ (40°C) cross peaks from a ROESY experiment. Reproduced with permission from Ref. 3. Copyright 1993, Journal of Carbohydrate Chemistry.

version of the title compound was compared to *N*-phenyl uronamide·4H₂O with respect to chemical shift changes ($\Delta\delta$) upon dehydration. When no H₂O was present in the DMSO/*N*-phenyl uronamide solution significant upfield

Table I. Proton NMR spectral assignments and spin-lattice relaxation times (observed, calculated and difference) at 2 fields. The correlation time ($\tau_c = 5.4 \times 10^{-10}$ s) was based upon the static field dependence as shown in equation 5.

$\delta\{^1H\}$	$T_{1H}/s^a \{\tau_c = 0.54ns\}$					
	270 MHz			400 MHz		
	obs. ^b	calc.	Δ	obs.	calc.	Δ
4.5 {C ₁ -H}	0.25	0.32	0.07	0.61	0.57	-0.04
6.4 {C ₁ -NH}	0.33	0.32	-0.01	0.56	0.57	0.01
5.34 {C ₂ -OH}	0.84	0.83	-0.01	1.46	1.47	0.01
5.31 {C ₃ -OH}	0.84	0.81	-0.03	1.42	1.44	0.02
5.01 {C ₄ -OH}	0.85	0.79	-0.06	1.37	1.40	0.03
3.89 {C ₅ -H}	0.31	0.41	0.10	0.77	0.72	-0.05
7.66 {Ortho"}	1.09	1.05	-0.04	1.84	1.86	0.02
7.29 {Meta"}	1.11	1.19	0.08	2.15	2.10	-0.05
7.04 {Para"}	1.59	1.46	-0.13	2.50	2.57	0.07
6.76 {Ortho'}	0.66	0.67	0.01	1.20	1.19	-0.01
7.11 {Meta'}	1.02	1.04	0.02	1.86	1.86	0.00
6.64 {Para'}	1.05	1.14	0.09	2.08	2.08	0.00

^a42 mM solution at 40°C; see equation 5 for τ_c - T_{1H} calculation.

^bAll T_{1H} calculations had $\leq 5\%$ error.

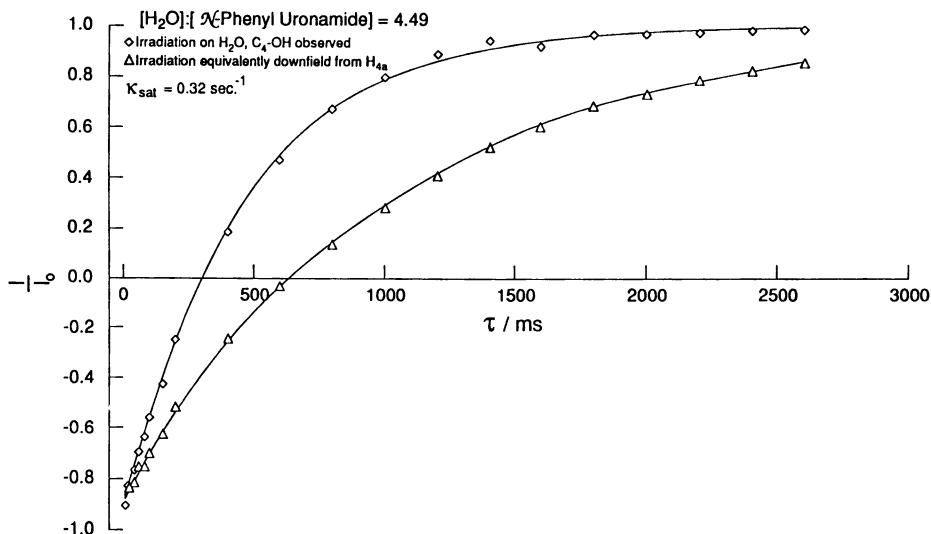


Figure 7. Inversion recovery experiments on *N*-phenyl uronamide (40°C; 400 mM; C_4 -OH observed) with irradiation 721.67 Hz upfield (e.g., on the H_2O resonance) from C_4 -OH (diamonds) and with the same treatment 721.67 Hz downfield from C_4 -OH (triangles). Reproduced with permission from Ref. 3. Copyright 1993, Journal of Carbohydrate Chemistry.

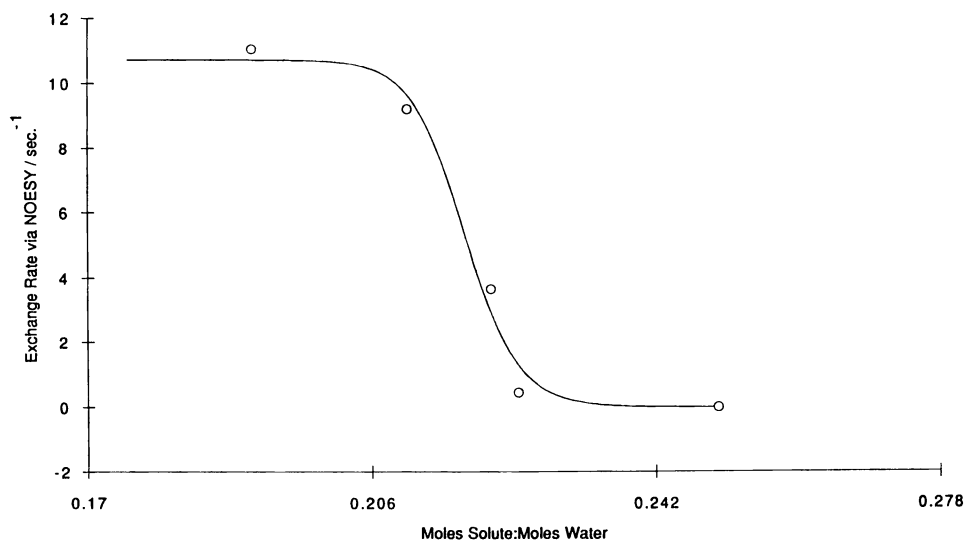


Figure 8. Plot of exchange rate constants as a function of [*N*-phenyl uronamide]: $[H_2O]$ indicating extreme concentration of water dependence.

Table II. Proton NMR spectral assignments for *N*-phenyl uronamide in the hydrated and dehydrated forms.

Assignment	δ / ppm		$\Delta\delta / \text{Hz}$
	[water] : [N-Phenyl Uronamide]		
	0:1	4:1	
Amide N-H ^a	9.91 ^b	10.09 ^c	72.00
Amine N-H	6.24	6.40	63.68
C ₄ -O-H	5.14	5.34	78.40
C ₃ -O-H	5.01	5.31	121.34
C ₂ -O-H	4.86	5.01	59.24
			x = 78.93
Ortho"	7.64	7.66	6.34
Meta"	7.29	7.29	0.00
Para"	7.05	7.04	-2.28
Ortho'	6.75	6.76	5.42
Meta'	7.11	7.11	0.00
Para'	6.64	6.64	0.00
C ₁ -H	4.50	4.50	0.00
C ₂ -H	3.31	3.29	-7.64
C ₃ -H	3.37	3.38	2.28
C ₄ -H	3.59	3.60	4.32
C ₅ -H	3.90	3.89	-5.74
			x = 0.25

^a300 mM solution at 50°C; made with "100%" DMSO-*d*₆.^bReacted with 2,2-dimethoxypropane prior to crystallization.^cca. 4.41 H₂O molecules per molecule of *N*-phenyl (*N*-phenyl-*b*-D-lucopyranosylamine)uronamide.

shifts on the resonance frequencies of all the polar (-OH $\Delta\delta=86.33$ Hz; -NH $\Delta\delta=67.84$ Hz) functional groups, relative to the methine protons (CH $\Delta\delta=0.25$ Hz), were observed.

All these data are evidence that the water molecules in this system experienced the slowest exchange as the molar ratio of [H₂O]:[*N*-phenyl uronamide] approached 4. Apparently, the translational diffusion of water, as measured by exchange, diminished dramatically as the molar ratio of H₂O:*N*-phenyl uronamide approached that level bound in the crystalline structure and argues that the complex had a relatively long lifetime in solution. At the above levels of hydration most of the H₂O might not be available for translational diffusion thereby causing κ to diminish. However, as H₂O activity increased beyond the capacity of *N*-phenyl uronamide to bind it, the average residence time would necessarily diminish resulting in an elevation of k as a function of increasing the water concentration. At H₂O levels well above 4:1 *N*-phenyl uronamide hydrolyses (2) to *N*-phenyl-D-glucopyranuronamide + aniline (Figures 2 and 3).

Modeling the Behavior of H₂O and *N*-Phenyl Uronamide. We were interested in determining if molecular dynamics simulations would agree with our experimental observation that about 3-4 water molecules bind to *N*-phenyl uronamide with a long residence time even when in dilute solution. For this purpose, an *N*-phenyl uronamide dimer was constructed, energy minimized and placed in a sphere of water molecules with a radius of 12 Å. The temperature of the system was gradually raised from 0 to 300 K over 300 ps. At this point water molecules > 6 Å from the carbohydrate's polar groups were removed. The dimer and 33 water molecules were then placed in a sphere of DMSO (radius of 18.5 Å). The temperature was gradually raised from 0 to 300 K over 300 ps. After 50 ps at 300 K, the two sugars had drifted apart. The 2 monomers, with their neighboring water molecules, were separated into two new systems (Table III). One sugar had 8 water molecules within 6 Å of polar groups while the other had 10. The two systems were placed in new spheres of DMSO as before and the temperature was raised progressively to 300 K. At this point in time both of the sugars had 7 near-neighbor water molecules. After an additional 100 ps at 300 K both systems had 2-3 water molecules ≤ 10 Å of the polar functional groups, a number somewhat lower than expected from the NMR data (e.g., ca. 4). Assuming that the Interaction energy change (ΔU_{HB}) necessary to form one H-bond is ca. 5 kcal/mol then the ratio of all interaction Internal energy changes ($\Delta U_{\text{I}}/\Delta U_{\text{HB}}$) calculated for our complex as a function of time should provide a measure of the number of interactions per water molecule. Table III shows that the near neighbor water molecules have a little over 1 H-bond with the polar functional groups of our carbohydrate model. However, showing better agreement with the NMR data was the observation that the weighted average distance (d) of all water molecules the -OH groups was found to be inversely proportional (equation 6) to the individual -OH $\Delta\delta$ s (Figure 9).

$$d = \sum_{j=1}^3 \frac{1}{r_j^3} \quad (7)$$

In this relationship the same 3 H₂O molecules were used for the calculation depicted in Figure 9. A similar relationship was noted for the N-H functional groups. Other modeling studies using the SYBYL software package exclusively indicated (Figures 10 and 11) that the spatial variation of water molecules close to the polar functional groups was quite low (2.5-3Å) after 50 ps of simulated time.

Table III. $\Delta U_{\text{Interaction}}$ and $\Delta U_{\text{Interaction}}/\Delta U_{\text{Hydrogen Bond}}$ calculations on the H₂O:*N*-phenyl uronamide complex as a function of time at 300K.

<i>System</i>	<i>Time (ps) after warmup to 300K</i>	$\Delta U_I/\text{kcal mol}^{-1}$	$\frac{\Delta U_I}{\Delta U_{HB}}$	$H_2O \leq 10\text{\AA}$
Monomer 1	0	-32.5	6.5	7
	50	-29.4	5.9	4
	100	-25.3	5.2	3
Monomer 2	0	-36.3	7.3	7
	50	-12.4	2.5	3
	100	-9.9	1.9	2

Conclusions

NMR experiments indicate that a stable complex of approximately 4:1 H₂O:*N*-phenyl uronamide exists even with extreme dilution in DMSO. Evidence for this was: a) the carbohydrate underwent no obvious hydrolysis in the presence of these hydration waters; b) the rate of chemical exchange was slower than one would expect for a -OH \leftrightarrow H₂O interaction; c) there was a concentration of water dependence for the exchange rate constant; d) upon dehydration of the complex there was a large chemical shift change in the polar functional groups. The molecular modeling studies were found to agree with the NMR results since the weighted average distance of all water molecules the -OH groups was found to be inversely proportional to the individual -OH $\Delta\delta$ s.

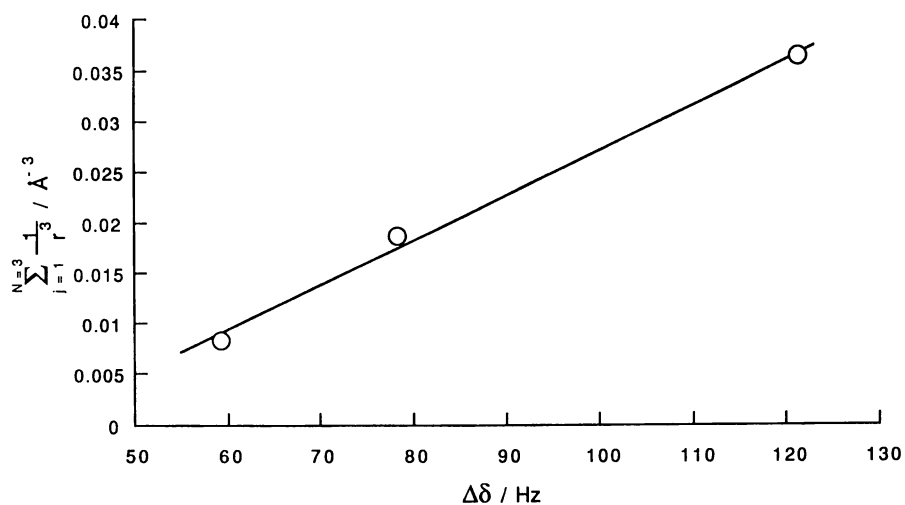


Figure 9. Plot of reciprocal weighted average distance of all water molecules the -OH groups, from dynamic simulations, as a function of individual -OH $\Delta\delta$ s.

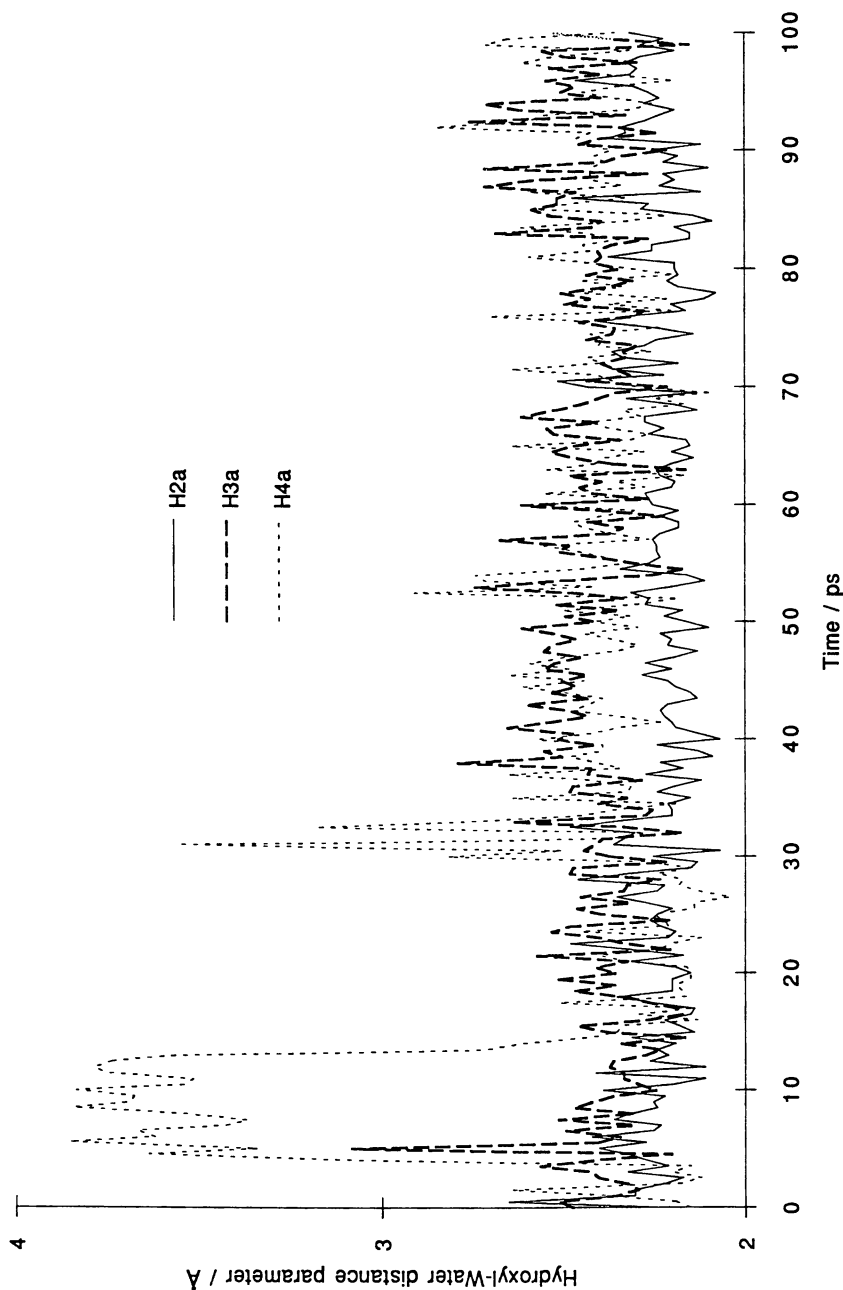


Figure 10. Variation of *N*-phenyl uronamide -OH \leftrightarrow H₂O distances as a function of simulated time in DMSO.

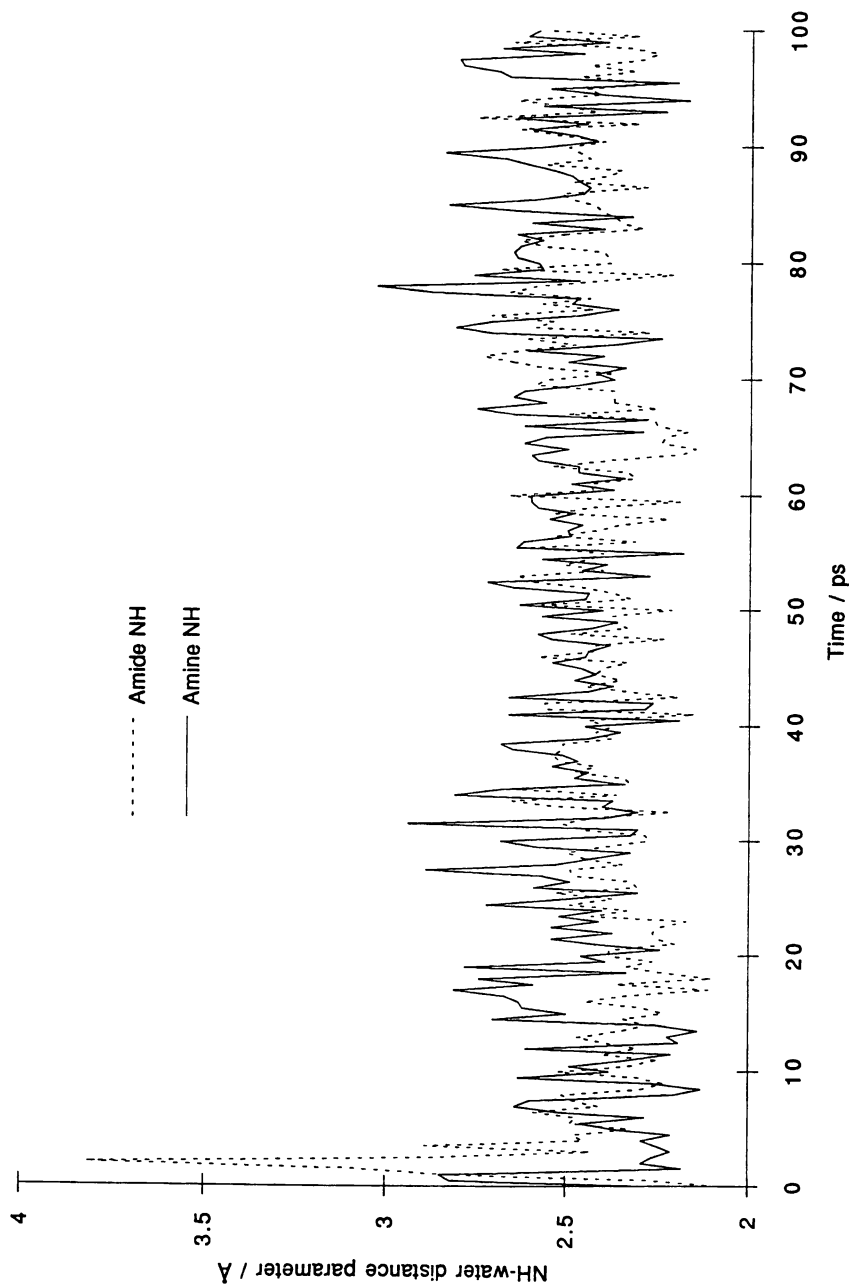


Figure 11. Variation of *N*-phenyl uronamide -NH \leftrightarrow H₂O distances as a function of simulated time in DMSO.

Acknowledgements

Some of this research was funded by a grant from the U.S.-Israel Binational Agricultural Research and Development Fund (BARD).

Reference to a brand or firm name does not constitute endorsement by the U.S. Department of Agriculture over others of a similar nature not mentioned.

Literature Cited

1. Laschi, F.; Pogliani, L. *Spectrosc. Lett.*, **1989**, *22*, 191.
2. Irwin, P.; Pfeffer, P.; Doner, L.; Lillie, T.; Frey, M. *J. Carbohydr. Chem.*, **1990**, *9*, 269.
3. Irwin, P.; Pfeffer, P.; Doner, L.; Ferretti, J. *J. Carbohydr. Chem.*, **1993**, *12*, 63.
4. DeTar, D.; Silverstein, R.; Rogers, F. *J. Am. Chem. Soc.*, **1966**, *88*, 1024.
5. Hoare, D.; Koshland, D. *J. Biol. Chem.*, **1967**, *242*, 2447.
6. R. L. Taylor and H. E. Conrad, *Biochem.*, **1972**, *11*, 1383.
7. Irwin, P.; Sevilla, M.; Osman, S. *Macromol.*, **1987**, *20*, 1222.
8. Haupt, E. *J. Magn. Resonance*, **1982**, *49*, 358.
9. Forsén, S.; Hoffman, R. *J. Chem. Phys.*, **1963**, *39*, 2892.
10. Shoubridge, A.; Briggs, R.; Radda, G. *FEBS Lett.*, **1982**, *140*, 288.
11. Balaban, R.; Kantor, H.; Ferretti, J. *J. Biol. Chem.*, **1983**, *258*, 12787.
12. King, G.; Warshel, A. *J. Chem. Phys.*, **1989**, *91*, 3647.
13. Clark, M.; Cramer, R.; van Opdenbosch, N. *J. Comput. Chem.*, **1989**, *10*, 982.
14. Gasteiger, J.; Marsili, M. *Tetrahedron*, **1980**, *36*, 3219.
15. Otting, G.; Wüthrich, K. *J. Am. Chem. Soc.*, **1989**, *111*, 1871.
16. Bothner-By, A.; Stephens, R.; Lee, J.; Warren, C.; Jeanloz, R. *J. Am. Chem. Soc.*, **1984**, *106*, 811.
17. Willem, R. *Prog. NMR Spectrosc.*, **1987**, *20*, 1.
18. van de Ven, F.; Janssen, H.; Gräslund, A.; Hilbers, C. *J. Magn. Resonance*, **1988**, *79*, 221.
19. Macura, S.; Ernst, R. *Mol. Phys.*, **1980**, *41*, 95.
20. Suter, D.; Ernst, R. *Phys. Rev. B*, **1985**, *32*, 5608.
21. Kumar, A.; Wagner, G.; Ernst, R.; Wüthrich, K. *J. Am. Chem. Soc.*, **1981**, *103*, 3654.
22. Kalk, A.; Berendsen, H. *J. Magn. Resonance*, **1976**, *24*, 343.
23. Hull, W.; Sykes, B. *J. Chem. Phys.*, **1975**, *63*, 867.
24. Bax, A.; Davis, D. *J. Magn. Resonance*, **1985**, *63*, 207.

RECEIVED June 14, 1994

Chapter 20

Structure—Serologic Relationships of the Immunodominant Site of Foot-and-Mouth Disease Virus

F. Brown¹, P. G. Piatti¹, I. Toth², and J. F. E. Newman¹

¹North Atlantic Area, Plum Island Animal Disease Center, Agricultural Research Service, U.S. Department of Agriculture, P.O. Box 848, Greenport, NY 11944-0848

²Department of Pharmaceutical Chemistry, The School of Pharmacy, Brunswick Square, London, United Kingdom WC1N 1AX

The extensive antigenic variation of foot-and-mouth disease virus and the localization of the dominant immunogenic site on a short loop region on the surface of the virus particle make it a valuable model for studying the structural basis of this variation. Synthetic peptides corresponding to the loop region elicit high levels of specific protective neutralizing antibody. The results of serologic studies have led to the conclusion that a major influence on the antigenicity of the loop is the presence of a Pro residue at position 153. Equally important is the sensitivity of position 148 to the presence of either Leu or Phe.

The extensive antigenic variation of foot-and-mouth disease virus (FMDV) presents problems in the control of the disease by vaccination. Seven serotypes of the virus, A, O, C, SAT1, SAT2, SAT3 and Asia1, are known and vaccines prepared from viruses of an individual serotype do not afford any protection against infection with viruses of the other six serotypes. In addition there is also considerable variation within the individual serotypes so that a vaccine prepared from one isolate may not afford protection against other viruses of the same serotype.

The current vaccines are prepared by chemical inactivation of virus grown in tissue culture cells. During the past 30 years the composition and structure of the virus-specific particles and proteins in the virus harvests have been determined and their role in providing protective immunity established(1). The major immunogen in the harvests is the intact virus particle. This is an icosahedral particle consisting of one molecule of single-stranded RNA surrounded by a shell consisting of 60 molecules of each of four proteins VP1-VP4. The mol. wt. of VP1-VP3 is c 24×10^3 and that of VP4 is c 8×10^3 . The latter protein is located internally. Stepwise dissection has revealed that the immunodominant site is located in VP1 within an exposed and highly flexible loop region comprising 28 amino acids. Because of its high

0097-6156/94/0576-0362\$08.00/0
© 1994 American Chemical Society

flexibility, the structure of the loop (residues 135-156 of VP1) has not been solved by X-ray crystallography except in the case of one isolate of serotype O under reducing conditions(2). The residues 132-134 and 157-159 are situated at the base of the loop (Color plate 18). These two segments are c 8 Å apart and protrude from the capsid surface suggesting that the loop is essentially a separate structural domain.

Peptides corresponding to this region in all seven serotypes of the virus are immunogenic and can provide protective immunity in experimental animals. Moreover, this region is highly variable between isolates in both sequence and antigenicity, thus providing the opportunity for studies on the structural basis for antigenic variation.

In this chapter we focus on one example, a virus of serotype A, isolated from an outbreak of the disease in England in 1932. Passage in baby hamster kidney cells of virus isolated from tongue epithelium of an infected animal showed that the isolate contained at least three related species which could be differentiated by serologic tests (3). Crucially, these studies showed that the viruses differed in only one amino acid in the loop region, the remainder of the four capsid proteins being identical. This chapter describes the results obtained in an extension of these studies and is complementary to the chapter by Lee France and her colleagues on the structural analysis of this loop region by circular dichroism spectroscopy and molecular modeling.

Materials and Methods

Viruses. The viruses were isolated from tongue epithelium of an infected steer. This isolate had been grown in baby hamster kidney cells and individual species were obtained by plaque picking. The sequences of the nucleic acid coding for the loop region were determined by conventional methods(3).

Preparation of anti-peptide antisera. Peptides corresponding to the 141-160 region of the loop region were linked to activated bovine serum albumin via an added Cys residue at the C terminus of each peptide, mixed with an oil adjuvant, and injected subcutaneously into guinea pigs. Blood samples were collected at intervals and the sera separated for subsequent analysis.

Enzyme linked immunosorbent assay (ELISA). The uncoupled peptides (0.1 µg well in 15 mM carbonate/35 mM bicarbonate buffer, pH 9.6) were adsorbed on to 96-well microtiter plates overnight at 4°C. The plates were then washed with 40 mM Na phosphate buffer, 3% gelatin, 1% Tween 20 for 1 hour at 37°C. Anti-peptide antiserum raised in guinea pigs was added to the plates in two-fold dilutions in 40 mM Na phosphate buffer, 1% gelatin, 1% Tween 20 and incubated for 1 hour at 37°C. After washing the plates, alkaline phosphatase-conjugated goat anti-guinea pig IgG in PBS, 1% gelatin was added for 1 hour at 37°C. Excess of conjugate was removed by washing as before. The reaction was quantified by adding p-nitrophenyl disodium phosphate for 30 min. at 37°C and another 30 min. at room temperature. The

NOTE: The color plates can be found in a color section in the center of this volume.

reaction was stopped with 3N NaOH and the color read at 405 nm with an ELISA reader.

Radioimmunoprecipitation (RIP). Radiolabelled virus particles were prepared by growing them in the presence of ^{35}S -methionine and purifying them as described by Brown and Cartwright(4). Twenty microliters of serum dilution were mixed with the virus and the volume was brought to 250 μl with TNEN (50 mM Tris-HCl, pH 7.5, 0.2 mM NaCl, 5 mM EDTA, 0.05% Nonidet-P40) containing 0.05% normal guinea pig serum and incubated for two hours at 37°C. Fifty microliters of a 10% suspension of *S. aureus* ghosts, washed three times in TNEN, were added to each sample and the mixtures incubated for one hour at 4°C. The pellets were washed three times in TNEN containing 0.05% normal guinea pig serum, dissolved in a solubiliser solution and counted by liquid scintillation.

Neutralization tests. Ten-fold dilutions of each virus were mixed with an equal volume of serum dilution and the mixtures added to wells containing monolayers of baby hamster kidney in microtiter plates. The cells were incubated at 37°C for 3 days and then stained with methylene blue-formalin. The cells in those wells in which the virus dilution had been neutralized remained alive and stained blue. In those wells where the virus had not been neutralized the cell sheet was destroyed and no staining was obtained. The neutralizing activity of the serum dilution is expressed as the difference between the titers of the virus alone and the virus-antiserum mixture.

Results and Discussion

Nine species have been isolated from the original virus (Table I). The results obtained with five of these species are given in Tables II-IV. These analyses show that the viruses fall into two distinct groups, determined by the presence of a Pro residue at position 153. However, Phe and Leu residues at position 148 also have a marked influence on the serologic cross-reactivity because the relationship between the viruses containing Phe Pro and Leu Pro at positions 148 and 153 is less close than exists between the three viruses in the other group.

Our main reason for investigating the structural basis of antigenic variation in FMDV was to design a structure simulating the dominant immunogen of the virus which would induce antibodies that cross-reacted with several antigenic variants. It has been shown for FMDV that antipeptide sera show greater cross-reactivity than the corresponding antiviron sera, and it has been proposed that this arises from the greater diversity of conformational states accessible to a peptide in solution than to the analogous sequence on the viral protein(5,6). This suggests that use of a peptide vaccine designed to induce antibodies which cross-reacts with all the viruses rather than the current inactivated vaccine might overcome the problem of antigenic variation.

The data presented in this paper indicate that such an objective can be achieved.

Table I. Amino acid changes at positions 148 and 153 of the GH loop region of FMDV, serotype A12

Virus	140	148	153	160
A	Ser Gly Ser Gly Val Arg Gly Asp	Ser	Leu	Arg Val Ala Arg Gln Leu Pro
B		Leu	Pro	
C		Ser	Ser	
USA		Phe	Pro	
LPb		Phe	Leu	
Leu Leu		Leu	Leu	
Ser Pro		Ser	Pro	
6FF3		Phe	Ser	
7SF3		Phe	Gln	

Table II. ELISA of peptides from the loop region of capsid protein VP1 of FMDV with the corresponding anti-peptide antiserum

Antiserum		Peptide					
		148 153	Leu Pro	Phe Pro	Phe Leu	Phe Ser	Phe Gln
148	153						
Leu	Pro		100	8	10	12	17
Phe	Pro		125	100	15	17	34
Phe	Leu		25	11	100	74	166
Phe	Ser		7	13	38	100	155
Phe	Gln		3	9	32	39	100

The results are expressed as a percentage of the homologous reaction.

Table III. Radioimmunoprecipitation of ³⁵S-methionine labelled viruses with anti-peptide antisera

Antiserum		Virus					
		148 153	Leu Pro	Phe Pro	Phe Leu	Phe Ser	Phe Gln
148	153						
Leu	Pro		100	50	2	2	2
Phe	Pro		11	100	7	5	5
Phe	Leu		8	3	100	45	63
Phe	Ser		4	≤ 1	32	100	100
Phe	Gln		3	≤ 1	89	35	100

The results are expressed as a percentage of the homologous reaction.

Table IV. Neutralization of viruses with anti-peptide antisera

Antiserum		Virus					
		148 153	Leu Pro	Phe Pro	Phe Leu	Phe Ser	Phe Gln
148	153						
Leu	Pro		100	10	≤ 1	≤ 1	≤ 1
Phe	Pro		32	100	3	3	≤ 1
Phe	Leu		≤ 1	3	100	10	10
Phe	Ser		≤ 1	10	32	100	32
Phe	Gln		≤ 1	1	10	10	100

The results are expressed as a percentage of the homologous reaction

Literature Cited

1. Brown, F. *Proc. R. Soc. London B.* **1986**, *229*, 215-226.
2. Logan, D.; Abu-Ghazaleh, R.; Blakemore, W.; Curry, S.; Jackson, T.; King, A.; Lea, S.; Lewis, R.; Newman, J.; Parry, N.; Rowlands, D.; Stuart, D.; Fry, E. *Nature* **1993**, *362*, 566-568.
3. Rowlands, D. J.; Clarke, B. E.; Carroll, A. R.; Brown, F.; Nicholson, B. H.; Bittle, J. L.; Houghten, R. A.; Lerner, R. A. *Nature* **1983**, *306*, 694-697.
4. Brown, F.; Cartwright, B. *Nature* **1963**, *199*, 1168-1170.
5. Ouldridge, E. J.; Parry, N. R.; Barnett, P. V.; Bolwell, C.; Rowlands, D. J.; Brown, F.; Bittle, J. L.; Houghten, R. A.; Lerner, R. A. In *New Approaches to Immunization*; Brown, F.; Chanock, R. M.; Lerner, R. A. Eds.; Cold Spring Harbor Press, Cold Spring Harbor, NY, 1986; pp. 45-49.
6. Geysen, H. M.; Barteling, S. J.; Melen, R. H. *Proc. Natl. Acad. Sci. USA* **1985**, *82*, 178-182.

RECEIVED January 13, 1994

Chapter 21

Predicted Energy-Minimized α_{s1} -Casein Working Model

Thomas F. Kumosinski, Eleanor M. Brown, and Harold M. Farrell, Jr.

Eastern Regional Research Center, Agricultural Research Service,
U.S. Department of Agriculture, 600 East Mermaid Lane,
Philadelphia, PA 19118

A previously reported three dimensional model of α_{s1} -casein (J. Dairy Sci 74: 2889, 1991) was constructed using sequence based prediction algorithms in conjunction with global experimental secondary structural information obtained from Raman Spectroscopy. This model is now energy minimized using primarily, the Kollman force field. Both the original and energy minimized structures contain a hydrophilic domain and a hydrophobic domain which are connected by a segment of α -helix. However, within the hydrophobic domain, three anti-parallel hydrophobic sheets are in different spatial orientations for the two structures. To mimic the self-association properties of the α_{s1} -casein B at low ionic strength, a tetramer model was constructed using the largest hydrophobic antiparallel sheets and the hydrophobic ion-pair, which occurs within the deletion peptide of the A variant, as interaction sites. Two tetramers were self associated via hydrophobic sites to model the octamer of α_{s1} -casein B which occurs at high ionic strengths. All energy minimized working models are in agreement with many of the biochemical, chemical and physico-chemical properties of α_{s1} -casein A, B and C.

In a previous report (16), a predicted three dimensional model of α_{s1} -casein was presented. This structure was built using sequence based secondary structure prediction algorithms, global secondary structural information from Raman Spectroscopy and molecular modeling techniques which, in part, minimized bad van der Waals contacts. The model consisted of a hydrophilic domain which contained most of the serine phosphates and a hydrophobic domain which contain two (one large and one small) stranded anti-parallel β -sheet structures with hydrophobic side chains. Both domains were connected via an extended α -helix, and another sheet structure. In addition, the model accounted for a variety of functional properties, derived biochemical, chemical and physico-

This chapter not subject to U.S. copyright
Published 1994 American Chemical Society

chemical results for α_{s1} -casein A, B and C. However, this structure was not energy minimized for interactions resulting from van der Waals energies, electrostatics and intra-molecular hydrogen bonding. In fact, a destabilizing energy of over two million kcal/mole could be calculated for this structure using a Kollman force field (31).

In this paper, we will present a structure refined by energy minimization techniques to further improve the original α_{s1} -casein model. This model will be compared with the original structure and discussed with respect to structural motifs present in α_{s1} -caseins of other species, biochemical cleavage information via renin, chemical modification results, and solution physico-chemical experiments. Also, to mimic the self-association of α_{s1} -casein, interaction sites on the monomer will be utilized to build energy minimized tetramer and octamer structures.

Materials and Methods

Predictions of Secondary Structures. Selection of appropriate conformational states for the individual amino acid residues was accomplished by comparing the results of sequence-based predictive techniques, primarily those of Chou and Fasman (5), Garnier et al. (12) and Cohen et al. (6,7). Assignments of secondary structure (α -helix, β -sheet or β -turn) for the amino acid sequences were made when either predicted by more than one method, or strongly predicted by one and not predicted against by the others. Such methods have previously been applied to the caseins (9,13,19). In addition, because of the large number of proline residues in the caseins, proline-based turn predictions were made using the data of Benedetti et al. (2) and Ananthanarayanan et al. (1).

Energy Minimization - Molecular Force Field. The concept, equation for and a full description of a molecular force field was given in a previous communication (17). In these calculations, a combination of the Tripos (24) and Kollman force fields (31) were employed. Both force fields used electrostatic interactions calculated from partial charges given by Kollman (31). A united atom approach with only essential hydrogens for reasonable calculation time on the computer was also used. A cutoff value of 8 Å was employed for all non-bonded interactions for both force fields. Both the BFGS (Boyden, Fletcher, Goldfarb and Shanno) and conjugate gradient techniques were employed as minimization algorithms when applicable. The Tripos force field contains fewer parameters than the Kollman force field, i.e. no 6-12 potential for specific hydrogen bonding, and is used in this study to overcome large energy barriers.

Molecular Modeling. The three dimensional structure for α_{s1} -casein was approximated using molecular modeling methods with an Evans and Sutherland

PS390¹ interactive computer graphics display driven by Sybyl (Tripos, St. Louis, MO) software on a Silicon Graphics (Mountainview, CA) DW-4D35 processor. The original structure was built with assigned ϕ and ψ angles characteristic of the respective predicted structures. All ω angles were assigned the conventional trans configuration. In addition, aperiodic structures are in the extended rather than totally random configuration. The Sybyl subroutine "SCAN" was used, on the side chains only, to adjust torsional angles and relieve bad van der Waals contacts. The individual pieces were then joined together to produce the total polypeptide model. This initial structure has been presented in a previous paper (16).

The original 3D model, which was constructed using a combination of sequence based secondary structure prediction algorithms constrained to be compliance with experimental global secondary structure determined from vibrational spectroscopy, was divided into three parts. The cleavage point was chosen to be in the middle of an extended structure to allow joining for reassembly of the total structure after minimization with known ϕ and ψ angles. Since secondary structure sequence based prediction algorithms are not informative with the respect to the type of turns, each of the three portions of the α_{s1} -casein B structure was reconstructed with differing turns and energy minimized. Each structure were tested by this method, and the model with the lowest energy was chosen for reassembly with the other pieces. Thus, by this method a larger sampling of conformational space was performed. However, this methodology does not in any way allow one to conclude that the global energy minimization structure was achieved.

The energies of the three pieces of the initial structure were calculated using a Tripos and a Kollman force field. Because of the large destabilizing van der Waals, bond stretching, due to the high proline (see Figure 1) and turn content, as well as the positive hydrogen bonding energy, each piece was first minimized with respect to total energy using a Tripos force field without electrostatics. The BFGS technique (24), which requires a large block of computer memory, was employed as the minimizing algorithm. This technique is more useful when a large number of energy barriers are suspected. The pieces were then subjected to a Tripos force field with electrostatics and finally a Kollman force field. The conjugate gradient technique was used only when the Kollman force field was employed.

The three energy minimized portions of α_{s1} -casein B were then joined together using ϕ , ψ values for an extended structure. The total structure was then energy minimized to an energy of ± 0.05 kcal/mole using a conjugate gradient algorithm and a Kollman force field (31).

The above procedure is an extension of the method of Cohen and Kuntz (38) where sequence based secondary structure prediction methods are used as a

¹Mention of brand or firm name does not constitute an endorsement by USDA over products of a similar nature not mentioned.

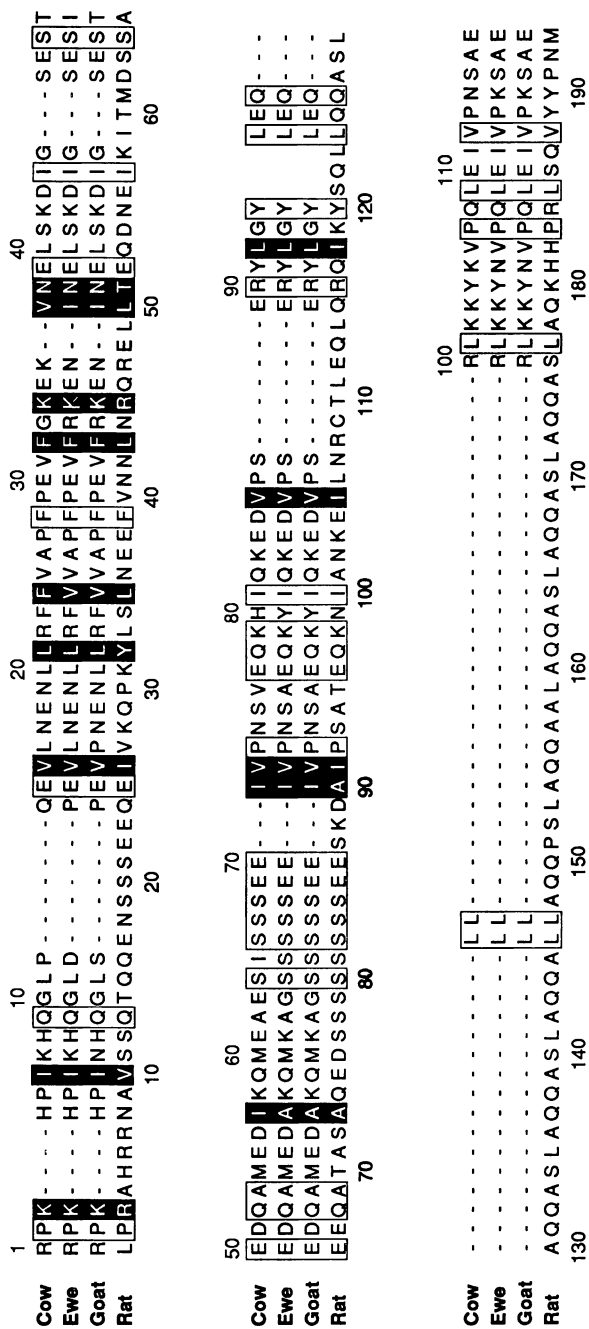
starting point for tertiary structure prediction. Since the choice of the sequence based algorithm is arbitrary (dubious) an added constraint of agreement between the individual or consensus algorithm and experimental global secondary results in conjunction with energy minimization techniques was utilized for sampling more of the possible conformational space.

This method may be tested with a protein whose global secondary structure and X-ray crystallographic structure is known. In addition, the protein, like casein should contain no disulfide bonds. Avian pancreatic polypeptide with 63 residues was chosen. Here, the global secondary structure results were obtained from the circular dichroism (CD) experiments of Noelken et al. (22) which showed a minimum of 80% helix under a variety of environmental conditions. The consensus prediction algorithm which agreed with the CD results showed a polyproline helix for residues 1-9, α helix in residues 13-32 and turns at residues 10 and 11. All other residues were given an extended conformation. After several energy minimization processes during which the turns were changed, the structure with the model with lowest energy, i.e. -599.4 kcal/mole, was chosen and is shown as a backbone ribboned structure in the lower portion of Figure 2. The upper portion shows the backbone of the X-ray crystal structure (1PPT) in the representation. Comparison of the backbone atoms between the predicted model X-ray structure yielded a root-mean square deviation of 3.37Å using the algorithm fit of the Sybyl program. Hence, a reasonable low resolution structure can be obtained by this methodology. It should be stressed that all structures built by these methods should be viewed as low resolution working models and not perfect structures for which the global energy minimum has been attained.

Assembly of Aggregate Structures. Aggregates were constructed using a docking procedure on the modeling system previously described. The docking procedure of this system allowed for individually manipulating the orientation of up to four molecular entities relative to one another. The desired orientations could then be frozen in space and merged into one entity for further energy minimization calculation utilizing a molecular force field. The criterion for acceptance of reasonable structures was determined by a combination of experimentally determined information and the calculation of the lowest energy for that structure. At least ten possible docking orientations were attempted, each structure was then energy minimized and assessed for the lowest energy in order to provide a reasonable sampling of conformational space.

Results and Discussion

Generation of Energy Minimized Three Dimensional Models. The caseins in general represent a unique class of proteins which are neither globular nor fibrous. They are characterized by a high content of the amino acid proline. In our initial attempts at undertaking structural motifs for casein (15,16) we were



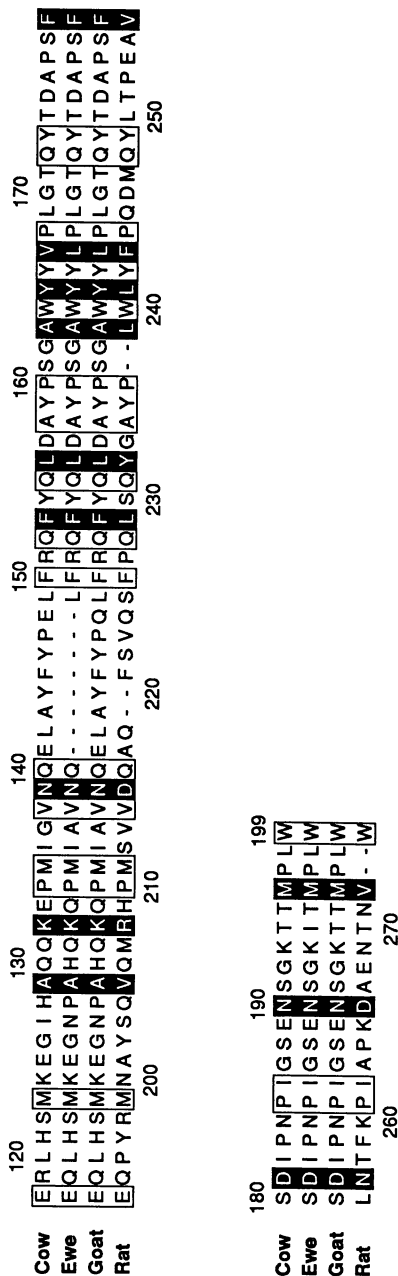


Figure 1. Sequence of bovine α_{s1} -casein C compared with those of ewe, goat and rat. The alignment follows reference (13) except that the goat has been added here. α_{s1} -Casein B represents replacement of glycine (G) 192 with glutamic acid in bovine protein.

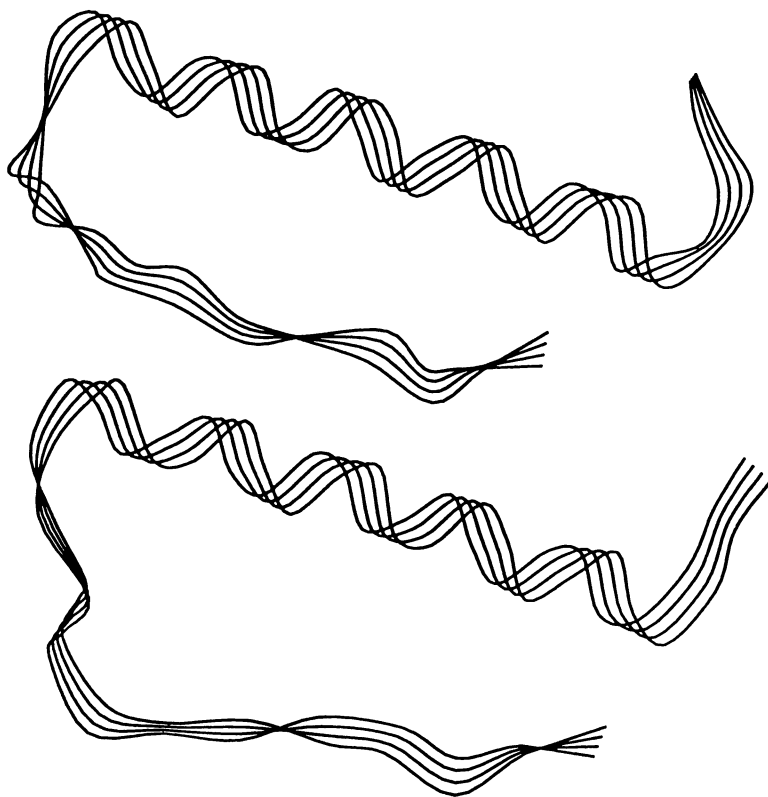


Figure 2. Ribboned backbone models of avian pancreatic polypeptide, upper: X-ray structure (1PPT); lower: predicted model.

struck by the fact that data from Raman spectroscopy predicted a high degree of turns and that there appeared to be a correlation between the percent proline and the percent of predicted turns. As pointed out by others, proline disrupts regular structure but excels at making turns (1,2,7,27). In globular proteins turns are almost always on the surface and are hydrophilic in nature. Caseins, however, have a propensity to self-associate (10,13,28) and most data point toward hydrophobic cores in the associated complexes (10,28). Thus proline-driven turns in hydrophobic areas may serve to facilitate the formation of interaction sites for hydrophobic core formation. Therefore, in the construction of casein monomers, proline turns in hydrophobic areas can appear on the monomer surface because it is anticipated these will be buried in subsequent polymer formation. In point of fact, the absence of a hydrophobic modeling term and the lack of water in these in vacuo calculations may actually be a positive element in these studies. The monomer constructed here actually could represent the monomer-within-a-polymer for the individual caseins. Exposed hydrophobic sites help to explain the nature of casein self-associations. In fact, these hydrophobic proline-based turns could well be the signature of an α_{s1} - or κ -casein, the structural motif which sets these proteins apart from both globular and fibrous proteins and even to some extent β -casein (17).

In the work that follows, it must be understood that we begin with secondary structural predictions. These predictions do not have a high degree of certitude, but they are based upon bond angles which occur in aqueous crystals for proteins. The beginning structures, like the previously proposed linear model (9,13) are modified to account for global circular dichroic and Raman data for α_s -casein, so that the assigned structures are not altogether arbitrary nor without precedent (9,13). Energy minimization of these structures could trap a less than favorable energetic state, but we have used algorithms which could avoid this problem. However, given the task at hand, de novo calculations from randomized structures would also not guarantee a native structure for a molecule the size of α_{s1} -casein. The structure presented herein thus represents a computational short-cut, and therefore is a working model, subject to change and refinement, and not a final exact structure. It is presented because it is unlikely that exact crystal structures will ever be available, and it is meant to stimulate research and discussion. In that light, the following steps were used to assemble the α_{s1} -casein model.

The refined, energy minimized model of α_{s1} -casein B was built using the computer generated structure presented in a preceding publication (16). In this latter paper, three segments were individually built, residues 1-84, 85-160 and 161-199, and then joined with appropriate ϕ , ψ angles to produce the polypeptide chain. Here, we individually energy minimized the three pieces by the following procedure. For the N-terminal piece, a destabilizing energy of over 900,000 kcal was calculated for the initial model using a Kollman force field as seen in Table I. A large portion of this energy resulted from improper van der Waals and hydrogen bonding interactions, with a smaller contribution

attributed to bond stretching energies. This large net energy is most likely due to the conformation of the six proline residues which occur in the first half of this segment of α_{s1} -casein (Figure 1). Because of these high energy values, energy minimization was performed using a Tripos force field first without electrostatic and then with electrostatic interactions. For these calculations, a BFGS algorithm was used for the minimization algorithm in order to avoid high energy wells with local minima. This combination of methods proved successful in reducing the calculated energy; results of the Tripos energy minimization with electrostatics are presented in column two of Table I. Here, the van der Waals energies are considerably lower than the initial values presented in column one of the same Table and H-bond energies are omitted by the program. Finally, this structure was refined using a Kollman force field in conjunction with a faster conjugate gradient minimization algorithm which takes H-bonding into account. The results are presented in column three of Table I, which shows a favorable total energy of -682 kcal/mole.

TABLE I. Energy calculations for the N-terminal segment of α_{s1} -casein (residues 1-84)

Energies k cal/mole	Initial	Tripos	Kollman
Bond Stretching	4223	365	42
Angle Bending	683	437	196
Torsional	440	207	246
Out of Plane Bending	53	53	21
1-4 van der Waals	206	-24	118
van der Waals	590463	171	-346
1-4 Electrostatic	780	155	905
Electrostatic	-2474	-2612	-1844
H-Bond	319688	---	-20
Total	914062	156	-682

The same procedure was applied to the other two pieces of the initial structure. The three minimized structures were then joined with the appropriate ϕ , ψ angles determined from the secondary structure sequence based prediction results presented in the preceding paper (16). This final total polypeptide chain was further energy minimized using a Kollman force field with a conjugate gradient minimizer. The result of this calculation as, seen in Table II, yield a

stabilizing energy of -2002 kcal/mole or ≈ 10 kcal/mole/residue. Such values are consistent with results obtained from the use of force fields to energy minimize structures derived from X-ray crystallography.

Refined Three Dimensional Structure of α_{s1} -Casein. The energy minimized structure, generated for α_{s1} -casein B as described above is shown in Color Plate 19 where it is displayed from carboxyl- to amino-terminal (left to right). Analysis of this structure shows the molecule to be composed (right to left) of a short hydrophilic amino-terminal portion, a segment of rather hydrophobic β -sheet, the phosphopeptide region, a short portion of α -helix connects this N-terminal portion to the very hydrophobic carboxyl-terminal domain containing extended antiparallel β -strands. For clarity the backbone without side chains is shown in Figure 3A with prolines (P) indicated, and an accompanying α -carbon chain trace stereo view (Figure 3B) is given.

TABLE II. Energy calculation for the refined α_{s1} -casein B structure

Bond Stretching Energy :	34	
Angle Bending Energy :	425	
Torsional Energy :	427	
Out of Plane Bending Energy :	16	
1-4 van der Waals Energy :	306	
van der Waals Energy :	-872	
1-4 Electrostatic Energy :	2135	
Electrostatic Energy :	-4426	
H-Bond Energy :	-47	
Total Energy :	-2002	kcal/mol

From the overall shape of the α_{s1} - model (Color Plate 19 and Figure 3), it is apparent that neither a prolate nor an oblate ellipsoid of revolution can be used to approximate its structure, as was done in the case of the β -casein refined structure (17). Indeed, a rather large degree of asymmetry of structure is observed. As noted above the hydrophilic and hydrophobic domains are joined by extended structures whose central feature is an α -helix with its pitch perpendicular to the two domains. It is speculated that this α -helix would be important for preserving the integrity of the two domains when a dynamic calculation is performed.

In the bovine casein this segment of α -helix occurs from residues 92 to 100 (Color Plate 19 and Figure 3). In the protein from all three ruminant milks, starting with residue 90 there are no amino acid substitutions in this region. For rat α_{s1} -casein a near exact or functional homology occurs for the beginning

NOTE: The color plates can be found in a color section in the center of this volume.

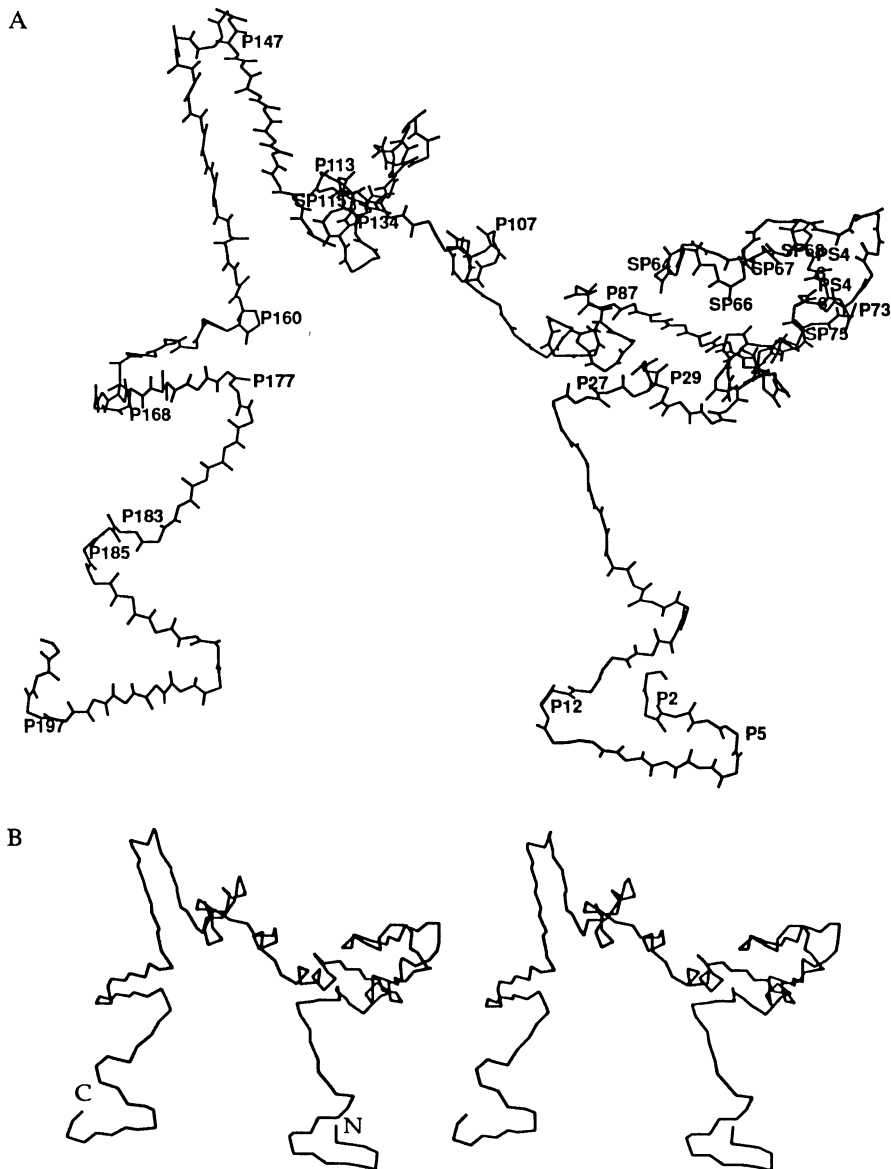


Figure 3. A) Backbone of the refined model of α_{s1} -casein, without side chains, prolines (P) indicated. B) Stereo view of the refined three dimensional molecular model of α_{s1} -casein; the N- and C-terminal ends of the molecule are labelled. C) Initial backbone structure of α_{s1} -casein B (Kumosinski et al., 16). D) Stereo chain trace of initial structure of a comparison with refined energy minimized structure. The ticked line represents the suggested stereo center, where the center of the stereo viewer should be placed.

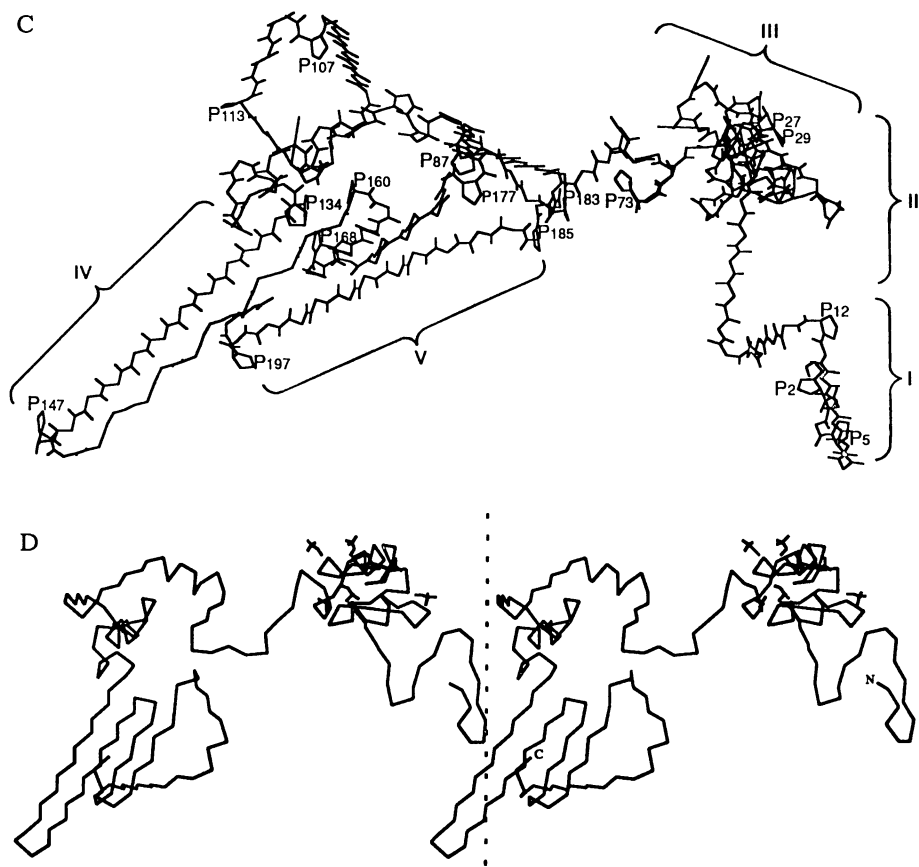


Figure 3. Continued.

Downloaded by NORTH CAROLINA STATE UNIV on October 26, 2012 | http://pubs.acs.org
 Publication Date: December 14, 1994 | doi: 10.1021/bk-1994-0576.ch021

of the α -helical region. In the rodent α_{s1} - however, there is then a large segment of repeating structure inserted at this point (Figure 1). It is interesting to speculate that this segment, which is based on the DNA-sequence, may also be helical in nature. Holt and Sawyer (13) calculate β -sheet for this extended region but in either case the rodent molecule has an even greater space between the hydrophilic and hydrophobic domains of the protein.

For bovine α_{s1} -casein, a radius of gyration (R_g) of 31 Å with a dipole moment of 3002 Debye units and a net charge of -26 can be calculated from its structure. In contrast, the refined β -casein structure yields an R_g of only 21 Å and a dipole moment of an 1825 D (17). These values demonstrate the large global differences between the α_{s1} -casein and β -casein refined structures, which may account for their differing functional properties, especially with respect to submicelle and micelle formation. The backbone structure without side chains as well as the stereo chain trace structure for the unrefined initial structure (15) is presented in Figure 3C and D, respectively. Comparison of Figures 3A and B with C and D, show major and minor differences between the refined energy minimized structure and the initial model. Although the initial and refined structures both consist of a hydrophilic and hydrophobic domain, the spatial orientation in the hydrophobic domain is quite different. A major change in the structure centering about proline 134 opens the structure dramatically. Thus the long 20 residue hydrophobic stranded antiparallel β -sheet (centered on proline 147) is opened up as is the short 8 residue piece (centered on proline 168). These are the major differences between the initial and refined structures Figure 3A, B and Figure 3C, D, respectively. This orientational difference resulted from the large portion of van der Waals energies from the proline based turns of the initial structure. In addition, the van Waals energies in the hydrophilic region due to proline and serine phosphates in the initial structure are reduced because these regions are partially changed to loops in the refined structure.

To test the accuracy of the refined structure of α_{s1} -casein B, comparisons between this model and experimental data, derived from solution studies such as Raman spectroscopy, chemical and biochemical experiments and solution physico-chemical studies on the hydrophobically induced, ionic strength dependent self-association of α_{s1} -casein, are presented.

Secondary Structure Analysis. The Ramachandran plot of ψ , ϕ , backbone dihedral angles (open circles) calculated from the refined α_{s1} -casein B structure, using the Tripos' Sybyl molecular modeling software is shown in Figure 4. Also, bounded areas of acceptable ψ , ϕ angles for β -sheet, α -helix and β - and γ -turn regions are presented with appropriate labels. Hence, by summing the number of points that are present within the limits of the particular periodic or turn structure, the global amount of α -helix, β -sheet or turn can be obtained. However, this type of analysis does not allow for the residue length of the periodic structure or its placement within the sequence structure. Hence, a visual inspection of the secondary structure by use of a chain trace or ribbon as

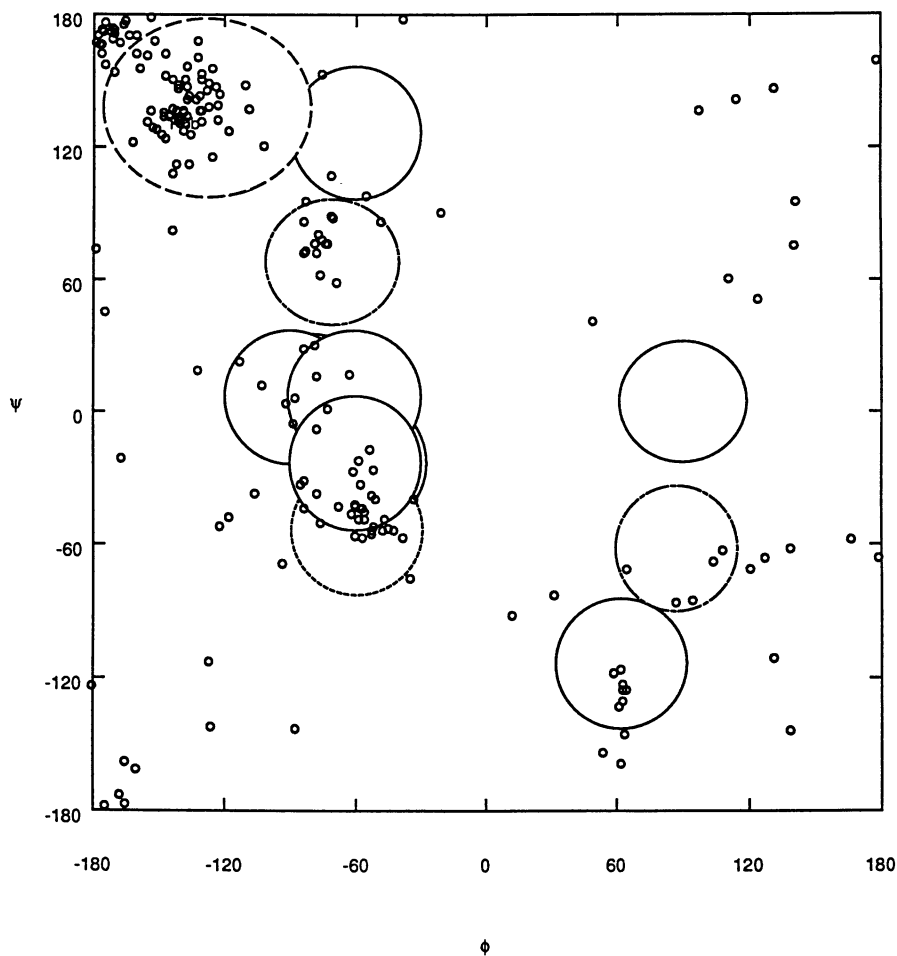


Figure 4. Ramachandran plot of ψ , ϕ angles calculated from energy minimized structure of α_{s1} -casein B. Dashed line (—) indicates area for β -sheet structure, dotted line (···) for α -helix, solid lines for β -turns, and dashed-dot line (— · — ·) γ -turns.

in Color Plate 19 should also be employed when calculating the global secondary structure of any model.

The above procedure was employed on the α_{s1} -casein B initial and refined structure and, as seen in Table III, the global secondary structural results are in good agreement with those obtained via Raman spectroscopy in D_2O . In fact, the refined structure yields values which agree more closely with the Raman results than the initial structure.

Holt and Sawyer (13) have recently used other secondary structure prediction algorithms and have theorized that α_{s1} -casein could contain up to 10% α -helix, 10% turns, 50% β -structure and 30% unordered structure. Their predictions are by their nature biased toward β -structure and so relative to the experimental Raman data (Table III) β -sheet is over predicted while the % of turns are underpredicted. However the N-terminal half of their predicted structure is in relatively good agreement with the 3D model presented in this work. In fact the major α -helical residues between residues 120-130 while somewhat frame shifted by one or two are also in good agreement. The major difference resides in our preselection of proline residues to be in turns as noted above. In the structures predicted by Holt and Sawyer (12) expansion of the % turns to include prolines and their adjacent 2 to 3 residues would increase their turn content primarily at the expense of β -sheet and bring their predictions closer to the global Raman data. Interestingly, our hydrophobic based turns, which depend upon prolines 147, 160, 168, 177 and 185 are in areas predicted by Holt and Sawyer to contain at least one residue in a β -turn conformation. It would appear then that there is overall good but not exact agreement between the algorithms used here and those of Holt and Sawyer (13). One major difference is for the rodent protein with its large insertion (Figure 2) where our algorithms favor α -helix while that used by Holt and Sawyer (13) favors β -sheet.

TABLE III. Comparison of the initial and the final secondary structures of α_{s1} -casein with spectroscopic data

Sample		% Helix	% β -Structure	% Turns	% unspec
α_{s1} -Casein	Raman ¹	8 - 13	18 - 20	29 - 35	33 - 40
	Initial	15	22	45	18
	Refined	8	18	34	40

¹ Reference 4.

Chemistry of α_{s1} -Casein and the Refined Model

Sites of Phosphorylation in α_{s1} -casein. α_{s1} -Casein B is a single polypeptide chain of 199 amino acid residues with a molecular weight of 23,619 (20). The α_{s1} -B molecule contains eight phosphate residues, all in the form of serine monophosphate. Seven of these phosphoserine residues are clustered in an acidic portion of the molecule bounded by residues 43 and 84 (the second fifth of the molecule from the amino-terminal end). This highly acidic segment contains 12 carboxylic acid groups as well as seven of the phosphoserines. The model shows all seven of the phosphate residues in this cluster to be located on β -turns which is compatible with known phosphorylated residues in crystallized proteins (27). The folding of the chain brings these phosphate residues into close proximity. This cluster forms a highly hydrophilic domain on the right shoulder of the molecule (Figure 3) and is bounded by prolines 29 and 87.

The α_{s1} -A Deletion and Chymosin Cleavage Sites. The rare α_{s1} -casein A genetic variant exhibits self association reactions which are highly temperature dependent (10,11,13). The A variant is the result of the sequential deletion of 13 amino acid residues between residues 13 and 27; the majority of these deleted amino acids are apolar (11) but Glu 14, Glu 18 and Arg 22 are also deleted (11). The deleted segment encompasses a region predicted to be in a β -sheet. This sheet provides a spacer-arm between the hydrophilic amino-terminal region (five o'clock in Figure 3A) and the phosphopeptide region. Deletion of this spacer-arm brings the hydrophilic N-terminal section (Figure 3A) closer to the phosphate rich shoulder. In addition, the Phe-Phe bond (residues 23 and 24) is deleted in the A variant; this represents a major chymosin cleavage site (15,19) and its absence may account for the poor quality products prepared from α_{s1} -casein A milks (28). Additional chymosin cleavage sites are located between residues 32-33, 149-150, 169-170 and 179-180 (19,21). The site between residues 32 and 33 is relatively exposed thus facilitating enzymatic attack. The last three sites, 149-150, 169-170 and 179-180 appear to be less exposed and could be involved in hydrophobic sheet-sheet interactions.

Hydrophobic Interactions. For α_{s1} -casein, the high degree of hydrophobicity exhibited by the carboxyl terminal half of the molecule (residues 100 to 199) may be responsible, in part, for the pronounced self-association of the α_{s1} -casein monomer in aqueous solution (28,29). This self-association approaches a limiting size (\sim tetramer) under conditions of low ionic strength; the highly charged phosphopeptide region can readily account for this phenomenon through charge repulsions. At elevated ionic strength ($>0.1M$) the polymer size increases to an octamer, and at ionic strengths >0.5 , α_{s1} -casein is salted out of solution at 37°C. The hydrophobic C-terminal domain contains two segments of extended β -sheets (residues 134 to 160 and 163 to 178) and one smaller extended hydrophobic segment from 180 to 188, found to the left in Figure 3A. These segments are directed by prolines 134,

147, 160, 168, 177 and 185. The crystal structure for insulin dimer (24), each monomer contains a similar pair of extended β -strands at the monomer-monomer interface. These extended β -strands may lend stability to casein polymers through sheet-sheet interactions. Of the six noted proline residues 134, 160, 168 and 185 are conserved across the four species shown in Figure 1. In addition proline 147 and 177 are compensated for in the rat protein by proline residues at $n+4$ and $n-1$ positions respectively, so that a formal structural homology for hydrophobic interactions occurs in all four species, the differences being in the length of the three hydrophobic segments. For bovine α_{s1} -caseins tyrosine residues play an important role in interactions with κ -casein; nitration of tyrosines of α_{s1} -casein leads to decreased stability in reconstituted micellar structures (32). It is interesting to note the tyrosines at 159, 166, 173 are conserved across species as are the two tryptophan residues. Residues 150 through 185 contain 13 hydrophobic residues which are exactly or functionally conserved (Figure 1). Thus the hydrophobic nature of these proline directed turns appear to be functionally preserved in most species. Such extended sheets are also predicted for κ -casein (15) so that sheet-sheet interactions may play an important role in casein micelle formation.

In α_{s1} -casein C, Glu 192 of the B variant is replaced by Gly through a point mutation (17). The association properties of α_{s1} -C are changed relative to α_{s1} -B by this replacement (21); the C variant has stronger associative properties. The segment of the bovine α_{s1} - molecule from residues 188 to 192 represents a hydrophilic turn centered on a functionally conserved amide (N/D) at positions 190 (Figure 1). This hydrophilic turn is followed, in ruminants, by a lysine residue at 193. The conversion of 192 to a glutamic residue in the B variant may alter the charge distribution here perhaps leading to weaker associative properties.

Correlation with Physico-Chemical Studies. To date, no indepth small angle X-ray or neutron scattering studies have been reported on any of the variants of α_{s1} -casein, especially, at low temperatures and ionic strengths where α_{s1} -caseins disaggregate. Hence, no correlation between an experimental radius of gyration and a value calculated from the refined structure is possible. However, there are light scattering studies on the B and C variants of α_{s1} -casein from which a stoichiometry of the α_{s1} -casein self-association maybe obtained under a variety of temperatures and ionic strengths. From the results, it was concluded that α_{s1} -casein undergoes a concentration dependent reversible association from monomer to dimer then tetramer, hexamer, octamer and even higher if the ionic strength is increased (28,29). To test the refined three dimensional structure presented in this paper, we attempted to construct energy minimized dimer, tetramer and octamer structures using primarily hydrophobic sites.

The first step was to create a dimer using the large antiparallel stranded β -sheet i.e., residues 136-158 as the interaction site. The side chains are predominately hydrophobic and the hydrogen bonding of the antiparallel sheet

secondary structure lends rigidity to this site. After energy minimization, a dimer could easily be formed if two of these stranded sheets are docked in an antiparallel fashion (Figure 5A). Such an asymmetric arrangement could minimize the dipole-dipole interactions of the backbones while allowing the hydrophobic side chains to freely interact. In fact, this minimized structure (Figure 5A) maintains a net stabilizing energy of -520 kcal/mole, i.e., $-520 = E_2 - 2 \cdot E_m$ where E_2 is the energy of the dimer and E_m is the monomer energy. Here an interesting prediction on the self association of the ovine α_{s1} -protein can be made; residues 141 through 148 are deleted. This would considerably shorten the length and symmetry of this extended sheet as turn centered at proline 147 is replaced by proline 160 while prolines 168, 177 and 185 are conserved. Thus ovine α_{s1} -casein should have weaker self associations than goat or cow. For the rodent α_{s1} -casein a proline follows 5 residues later (rat proline 230) making a better homology with cow and goat for this segment.

Another possible interaction site for hydrophobic dimerization resides in the deletion peptide of α_{s1} -casein A (i.e. the peptide, residues 14 to 26 inclusive, which is deleted from α_{s1} -casein B to form α_{s1} -casein A). Closer inspection of the residues of this peptide show a β -sheet secondary structure with hydrophobic as well as acidic and basic side chains. Thus, by docking two molecules in an antiparallel fashion, a hydrophobically stabilized inter-molecular ion pair between arginine 22, of one chain and glutamic 18 of another chain can be formed upon the construction of a dimer. The dimer was then energy minimized and is shown in Figure 5B. The formation of this hydrophobic ion pair has been used by several investigators to explain the difference in the calcium-induced solubility and colloidal stability between α_{s1} -caseins B and A (11,14). Small-angle X-ray scattering of micelles reconstituted from whole caseins containing α_{s1} -B and A show large differences with respect to submicellar packing density within the micelle structure i.e. 3 to 1 verses 6 to 1 for B and A, respectively. This difference in packing density may be a result of destructuring interference in the scattered intensity due to an asymmetric structure with a center of inversion (23). It was speculated that this asymmetric structure was due to the formation of a hydrophobically stabilized intermolecular ion pair in α_{s1} -B at this deletion peptide site (21), and that the α_{s1} -A molecule, lacking this segment behaves mole like a β -casein in its properties.

A tetramer of α_{s1} -casein can be modeled starting with the dimer formed by the intermolecular hydrophobic ion pair (IPr in, Figure 5B). A molecule of α_{s1} -casein is then added to each side of this structure via the hydrophobic (Hb) sheet-sheet interaction shown in Figure 5A. Such an energy minimized tetrameric structure is presented in Figure 6A. This tetramer structure is highly asymmetric and also contains two possible hydrophobic ion pair sites at each end of the molecule. Such sites at either end of the tetramer could lead to further aggregation resulting in a rod with a large axial ratio and dipole moment. It is noteworthy that the hydrophobic antiparallel stranded sheets of residues 163-174 could not be docked with the same site one from another

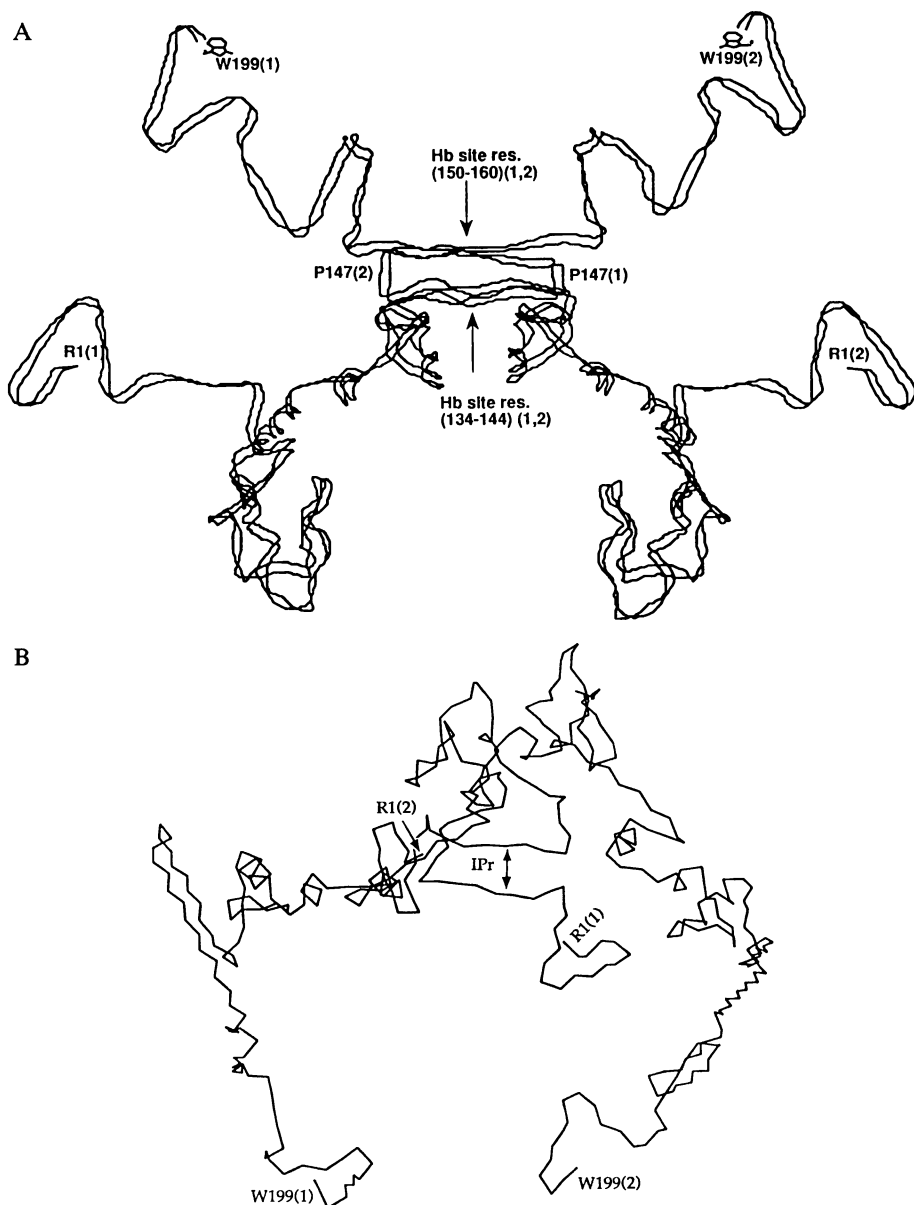


Figure 5. A) α -Carbon chain trace of backbone without side chains of hydrophobic (Hb) stabilized antiparallel sheet dimer; the large sheets centering on proline 147 are docked. B) Backbone structure of hydrophobic ion-pair dimer (IPr) from α_{s1} -casein B; this area contains the α_{s1} -A deletion peptide and centers about residues 14 to 25 in each molecule. Key for labels: R1(1), W 199(1) represent N and C terminals of molecule 1 (arginine 1 and tryptophan 199 of molecule 1); the 2 in parenthesis refers to the same residues of molecule 2. Both dimeric structures energy minimized to -10 kcal/mole/residue.

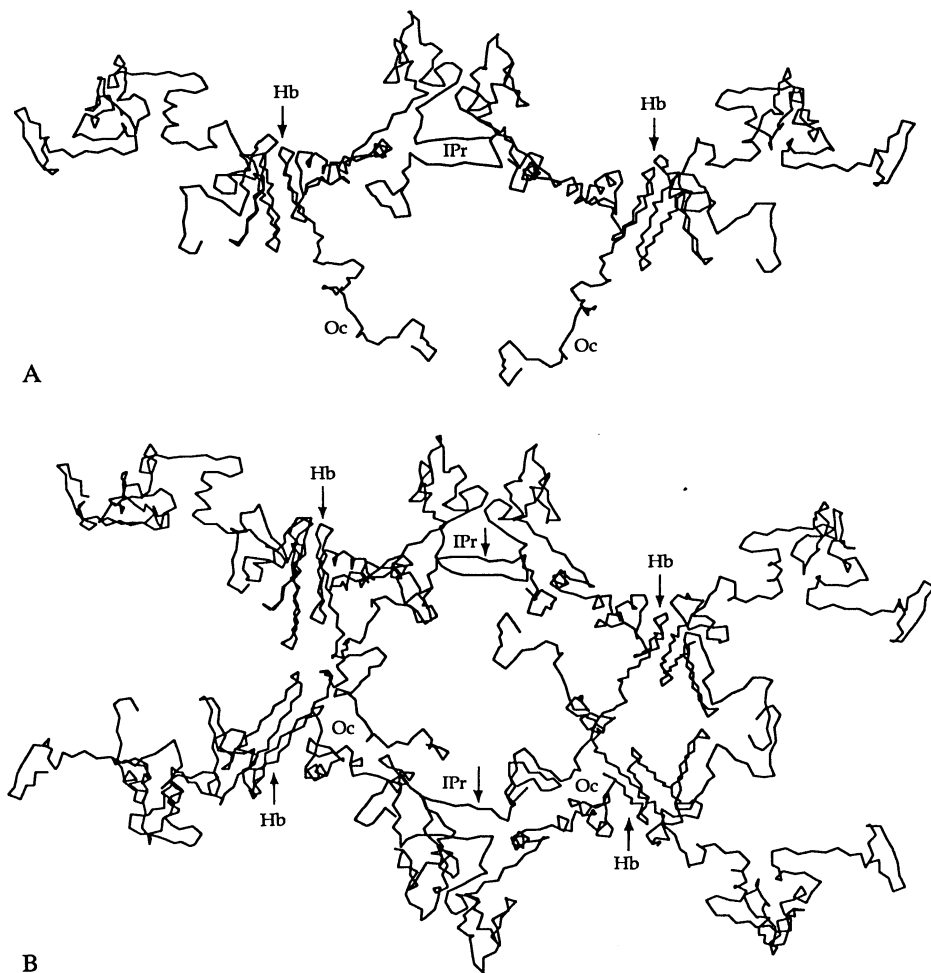


Figure 6. A) α -Carbon chain trace of α_{s1} -casein B tetramer resulting from the docking of two molecules (left and right sides) of α_{s1} -through the hydrophobic (Hb) scheme shown in 5A with the IPr (ion pair) dimer scheme given in Figure 5B. The Oc refers to potential sites for octamer formation. B) α -Carbon chain trace of octamer showing Hb, IPr and Oc interaction sites. All structures were energy minimized to -10 kcal/mole/residue.

tetramer structure due to large steric factors. Such a structure could not possibly be made without major changes in the α_{s1} -casein B model. However, a more plausible site for octamer formation via simple hydrophobic interactions can be constructed from use of the hydrophobic side chains centered between prolines 177 and 185 which are located on the lower side of the asymmetric tetramer structure (Figure 6A, Oc) and are solvent accessible. With this in mind, two tetramers were docked in an antiparallel fashion with a center of inversion using these hydrophobic side chains as interaction sites. The energy minimized octamer structure is shown in Figure 6 B and C. This model yielded a favorable energy of approximately -10 kcal/mole/residue. Such an octamer structure would still allow for water to flow through part of the polypeptide chain yielding an apparently high hydrodynamic hydration value. It would also be stable in solution since the hydrophobic side chains are predominately in the center of the model and all eight hydrophilic domains are solvent accessible, i.e., two on the upper and lower center part and two at either end of the structure (see Figure 6 B and C).

A very interesting feature of all the α_{s1} -caseins is the preservation of the C-terminal tryptophan. Ribabeau-Dumas and Garnier (25) showed that carboxypeptidase A could quantitatively remove the C-terminal tryptophan of α_{s1} -casein alone, and in native and reconstituted micelles. This was interpreted as a demonstration of the open network of the casein micelles which allowed penetration of the protease into the micelle. This is in accord with the model shown in Figure 3 A, B, C, where the C-terminal tryptophan is extended in space at the left side of the model. Thus although residues 134 to 185 participate in hydrophobic interactions a hydrophilic turn then intervenes and the C-terminal tryptophan can still be exposed in monomeric and polymeric structures. The tryptophan is thus available to digestion with carboxypeptidase A.

Finally, two of the small hydrophobic antiparallel stranded sheets of residues 163-174 are located on either end of the octamer structure between each hydrophilic domain and are therefore solvent accessible. The residues have the possibility of interaction with one or both κ -casein hydrophobic antiparallel stranded sheets even in the presence of α_{s1} -self-associations. As discussed above, tyrosines 159 and 166, proline 160 and tryptophan 164 are conserved, perhaps for relatively specific chain-chain interactions with κ -casein. Such interactions may be important in all species for micelle formation.

It is hoped that with the presentation of these working models, investigators will be inspired to perform detailed small angle scattering and other physical and biochemical experiments to ascertain the validity of these aggregate structures.

Concluding Remarks

In this paper, we have presented an energy minimized predicted three dimensional structure of α_{s1} -casein using a combination of secondary structure

sequence based prediction algorithms, global secondary structural results from Raman spectroscopy and molecular modeling techniques for energy minimization. This structure is in agreement with biochemical cleavage results using carboxy peptidase and chymosin on α_{s1} -casein B. It is also in agreement with other experimentally derived results from solution physico-chemical experiments and provides a molecular basis for the self-associations of α_{s1} -casein. However, this structure should be viewed as a working model with the ability to be changed as more precise experiments are performed to ascertain the validity and predictability of this three dimensional structure. In future studies, molecular dynamics calculations will be performed on this and the aggregate structure to test its stability when a kinetic energy equivalent to a bulk temperature is applied. In addition, it may be possible in the future to ascertain how the α_{s1} -casein B molecule specifically interacts with κ -casein to produce a synthetic submicelle structure of four α_{s1} -casein B molecules to one κ -casein molecule, the low weight-ratio complexes observed experimentally (9,13,26) in reconstitution experiments from purified caseins.

Literature Cited

- 1 Ananthanarayanan, V.S.; Brahmachari, S.K.; Pattabiraman, N. *Arch. Biochem. Biophys.* 1984, 232, 482.
- 2 Benedetti, E.; Bavoso, A.; Blasio, B.D.; Pavone, V.; Pedone, C.; Toniolo, C.; Bonora, G.M. *Biopolymers* 1983, 22, 305.
- 3 Byler, D.M.; Farrell, H.M., Jr. *J. Dairy Sci.* 1989, 72, 1719.
- 4 Byler, D.M.; Farrell, H.M., Jr.; Susi, H. *J. Dairy Sci.* 1988, 71, 2622.
- 5 Chou, P.Y.; Fasman, G.D. *Adv. Enzymology* 1978, 41, 45.
- 6 Cohen, F.E.; Abarbanel, R.M.; Kuntz, I.D.; Fletterick, R.J. *Biochemistry* 1983, 22, 4894.
- 7 Cohen, F.E.; Abarbanel, R.M.; Kuntz, I.D.; Fletterick, R.J. *Biochemistry* 1986, 25, 266.
- 8 Cohen, F.E.; Kuntz, I.D. *In Prediction of Protein Structure and the Principles of Protein Conformation*, Fasman, J.D., ed., Plenum Press, New York, 1989, p 617.
- 9 Creamer, L.K.; Richardson, T.; Parry, D.A.D.; *Arch. Biochem. Biophys.* 1981, 211, 689.
- 10 Farrell, H.M., Jr. *In Fundamentals of Dairy Chemistry*, N. Wong, Ed.; Physical Equilibria: Proteins, 3rd Edition. Van Norstrand Reinhold, New York, NY, 1988, p 461.
- 11 Farrell, H.M., Jr.; Kumosinski, T.F.; Pulaski, P.; Thompson, M.P. *Arch. Biochem. Biophys.* 1988, 265, 146.
- 12 Garnier, J.; Osguthorpe, D.J.; Robson, B.; *J. Mol. Biol.* 1978, 120, 97.
- 13 Holt, C.; Sawyer, L. *Protein Engineering* 1993, (in press).
- 14 Kumosinski, T.F.; Farrell, H.M. Jr. *J. Protein Chem.* 1991, 10, 3.

- 15 Kumosinski, T.F.; Brown, E.M.; Farrell, H.M. Jr. *J. Dairy Sci.* **1991**, *74*, 2879.
- 16 Kumosinski, T.F.; Brown, E.M.; Farrell, H.M. Jr. *J. Dairy Sci.* **1991**, *74*, 2889.
- 17 Kumosinski, T.F.; Brown, E.M.; Farrell, H.M. Jr. *J. Dairy Sci.* **1993**, *76*, 931.
- 18 Lawrence, R.C.; Creamer, L.K.; Giles, J. *J. Dairy Sci.* **1987**, *70*, 1748.
- 19 Loucheux-Lefebvre, M.H.; Aubert, J.P.; Jolles, P.; *Biophys. J.* **1978**, *23*, 323.
- 20 Mercier, J.C.; Grosclaude, F.; Ribadeau-Dumas, B. *Eur. J. Biochem.* **1971**, *23*, 41.
- 21 Mulvihill, D.M.; Fox, P.F.; *J. Dairy Research* **1979**, *46*, 641.
- 22 Noelkin, M.E.; Change, P.J.; Kimmel, J.R. *Biochemistry*, **1980**, *19*, 838.
- 23 Pessen, H.; Kumosinski, T.F.; Farrell, H.M. Jr.; Brumberger, H.; *Arch. Biochem. Biophys.* **1991**, *284*, 133.
- 24 Press, W.H.; Flannery, B.P.; Teukolsky, S.A.; Vetterling, W.T.; Numerical Recipes, The Art of Scientific Computing, Cambridge University Press, 1988, pp 301-327.
- 25 Ribadeau-Dumas, B.; Garnier, J. *J. Dairy Res.* **1970**, *37*, 269.
- 26 Richardson, J.S. *Advances Protein Chem.* **1981**, *34*, 167.
- 27 Rose, G.D.; Gierasch, L.M.; Smith, J.A. *Adv. Prot. Chem.* **1985**, *37*, 1.
- 28 Schmidt, D.G. *In* Developments in Dairy Chemistry -1., P.F. Fox, Ed., Applied Science Pub. Ltd., Essex, England, 1982, p 61.
- 29 Swaisgood, H.E.; Timasheff, S.N. *Arch. Biochem. Biophys.* **1968**, *125*, 344.
- 30 Thompson, M.P.; Gordon, W.G.; Boswell, R.T.; Farrell, H.M. Jr. *J. Dairy Sci.* **1969**, *52*, 1166.
- 31 Weiner, S.J.; Kollman, P.A.; Nguyen, D.T.; Cose, D.A. *J. Comput. Chem.* **1986**, *7*, 230.
- 32 Woychik, J.H.; Wondolowski, M.V. *J. Dairy Sci.* **1975**, *52*, 1669.

RECEIVED July 7, 1994

Chapter 22

Three-Dimensional Molecular Modeling of Bovine Caseins

Energy-Minimized Submicelle Structure Compared with Small-Angle X-ray Scattering Data

Harold M. Farrell, Jr., Thomas F. Kumosinski, and Gregory King

Eastern Regional Research Center, Agricultural Research Service,
U.S. Department of Agriculture, 600 East Mermaid Lane,
Philadelphia, PA 19118

To develop a molecular basis for structure-function relationships of the complex milk protein system, an energy minimized three dimensional structure of a casein submicelle was constructed consisting of one κ -casein, four α_{s1} -casein and four β -casein molecules. The models for the individual caseins were from previously reported refined, three dimensional structures. Docking of one κ -casein and four α_{s1} -casein molecules produced a framework structure through the interaction of two hydrophobic antiparallel sheets of κ -casein with two small hydrophobic antiparallel sheets (residue 163-174) of two preformed α_{s1} -casein dimers. The resulting structure is approximately spherically symmetric, with a loose packing density; its external portion is composed of the hydrophilic domains of the four α_{s1} -caseins, while the central portion contains two hydrophobic cavities on either side of the κ -casein central structure. Symmetric and asymmetric preformed dimers of β -casein formed from the interactions of C terminal β -spiral regions as a hinge point, could easily be docked into each of the two central cavities of the α - κ -framework. This yielded two energy minimized three dimensional structures for submicellar casein, one with two symmetric β -casein dimers and one with two asymmetric dimers. These refined submicellar structures were tested by generating theoretical small-angle X-ray scattering (SAXS) curves and comparing them with experimental data. Agreement between experimental and theoretical curves was best when 120 bound water molecules were included. Comparison of SAXS data with the theoretical curves generated from X-ray data for bovine pancreatic trypsin inhibitor, by the same programs, gave similar results.

This chapter not subject to U.S. copyright
Published 1994 American Chemical Society

The caseins occur in bovine milk as colloidal complexes of protein and salts, commonly called casein micelles. Removal of calcium is thought to result in the dissociation of this micellar structure into noncolloidal protein complexes called submicelles (12). These submicelles consist of four proteins, α_{s1} -, α_{s2} -, β -, and κ -casein, in the ratios of 4:1:4:1 (8). All are phosphorylated to various extents, have an average monomer molecular weight of 23,300, and were considered to have few specific secondary structural features, such as sheets or helices (12). Recent infra-red and Raman spectroscopic data, however, have demonstrated the existence of turns and more β -sheet than expected in casein monomers and polymers (4, 5). The isolated fractions exhibit varying degrees and mechanisms of self-association, that are thought to be mostly hydrophobically driven (12, 18, 36). However, less work has been done on the tertiary and quaternary structure of these proteins in mixed associations in their native state. There is hydrodynamic evidence that, in the absence of calcium, whole casein associates to form aggregates with an apparent upper limit of 94 Å for the Stokes radius with a molecular weight of 220,000 (submicellar form) (30, 31), and this is in general agreement with the type of protein particles formed upon dissociation of casein micelles (12, 18, 36).

It has long been hypothesized that, upon the addition of calcium, these primarily hydrophobically stabilized, self-associated casein submicelles further aggregate via calcium-protein side chain salt bridges to the colloidal micelles, with a size distribution centering upon 1500 Å diameter (12, 18, 36). However, the exact supramolecular structure of the casein micelle remains unknown. Models presented have ranged from those having discrete submicelles to those having the structure of a loose porous gel (12), and to a newer model of a homogeneous sphere with a "hairy" outer layer (41). For a recent up-to-date review of micelle structure, the reader is referred to Holt (18).

To better understand the nature of the protein-protein interactions involved in micelle formation, small angle X-ray scattering experiments have been performed on whole casein in the presence and absence of calcium to mimic the micelle and submicelle structures, respectively (15, 32). It was found that micellar structures are indeed composed of submicellar particles whose structure may be approximated by an inhomogeneous spherical aggregate of two concentric electron dense regions. The inner high electron density core of the submicelle is still seven times lower than the electron density of a globular protein and has a radius of 53 Å. From these values, it was speculated that this core predominately contains hydrophobic groups. The outer loose spherical region probably contains hydrophilic groups with very low packing density. The overall radius of the spherical structure would be approximately 103 Å. From the low electron density, it was also concluded that large amounts of water could easily flow through the polypeptide chains within this structure (15, 32).

Recently, three dimensional models refined via energy minimization techniques were constructed for κ -casein (22), α_{s1} -casein (23) and β -casein (21). These predicted structures were built from secondary structure sequence-based prediction algorithms in conjunction with global secondary structure results obtained from vibrational spectroscopy experiments (4, 5). All energy minimized structures were also in agreement with these global secondary structure determinations (21-23).

Several energy minimized aggregate structures were also presented to mimic the self-association processes for each of these caseins. In addition, qualitative speculation was presented for the interaction sites for κ -casein with α_{s1} -casein, but none were immediately obvious for κ -casein interaction sites with β -casein (21, 22).

In this paper, we will attempt to build an energy minimized submicelle structure composed of one κ -casein with four α_{s1} -caseins and four β -caseins via plausible docking sites consistent with solution physical chemical, biochemical and chemical experimental information. The energetics of all structures will be presented to ascertain if the formation of a synthetic submicelle structure, i.e., one κ -casein with four α_s -casein molecules, is the predominant intermediate or framework structure for submicelle formation. Finally, this refined structure will be compared with the geometric parameters calculated from the small angle X-ray scattering results for the submicelle particle using a variation of the procedure for computer generated models developed by Lattman (26).

Methods

Construction of Aggregate Structures

All complex aggregate structures employed the various casein monomer structures previously refined via energy minimization (21-23). Aggregates were constructed using a docking procedure on an Evans and Sutherland (St. Louis, MO) PS390 interactive computer graphics display driven by Sybyl molecular modeling software (Tripos, St. Louis, MO) on a Silicon Graphics (Mountainview, CA) W-4D35 processor. The docking procedure of this system allowed for individually manipulating the orientation of up to four molecular entities relative to one another. The desired orientations could then be frozen in space and merged into one entity for further energy minimization calculation utilizing a molecular force field. The criterion for acceptance of reasonable structures was determined by a combination of experimentally determined information and the calculation of the lowest energy for that structure.

Molecular Force Field Energy Minimization. Studies concerned with the structures and/or energetics of molecules at the atomic level require a detailed knowledge of the potential energy surface (i.e., the potential energy as a function of the atomic coordinates). For proteins, molecular mechanics methods have been used. The applications of these techniques to casein monomers have been detailed elsewhere (21-23).

Briefly in this study, the AMBER force field (44, 45) in Tripos' Sybyl software package uses electrostatic calculations which include atomic partial charges (q_i) obtained by the Kollman group (44, 45) and a united atom approach with only essential hydrogens. All molecular structures were refined with an energy minimization procedure using a conjugate gradient algorithm, in which the positions of the atoms are adjusted iteratively so as to achieve a minimum potential energy value. Energy minimization calculations were terminated when the energy difference between the current and previous iterations was less than 1 kcal/mol of protein. A nonbonded cutoff (the distance beyond which hydrogen bonding is not

considered) of 5 Å was used initially to save computer time, and then an 8 Å cutoff was used as the structures became more refined. A stabilization energy of at least -10 kcal/residue of protein was achieved for all structures, which is consistent with values obtained for energy minimized structures determined by X-ray crystallography.

Construction of Hydrated Structures. Low hydrated structures of the refined, energy-minimized casein submicelle models were constructed using a docking procedure on an Evans and Sutherland (St. Louis, MO) PS390 interactive computer graphics display driven by the Tripos Sybyl (Tripos, St. Louis, MO) molecular modeling software on a Silicon Graphics (Mountain View, CA) W-4D35 processor. The docking procedure allowed for individually manipulating the orientation of 120 energy minimized water molecules in up to four molecular display areas relative to one another. The desired orientation could then be frozen in space and merged into one molecular display area for energy minimization calculation using a molecular force field. The criterion for acceptance of reasonably hydrated structures was determined by a combination of experimentally determined information, i.e., DNMR relaxation results (13, 15) in combination with the calculation of the lowest energy for that structure. All water molecules with unacceptable van der Waals interactions were eliminated.

For hydrated structures with large amounts of water, the Tripos (St. Louis, MO) "Droplet" algorithm was employed. This procedure creates a monomolecular layer of water around an entire structure in an objective manner. In these calculations, a structure with a low hydration value (120 water molecules) was created using the above docking procedure, then the high hydration model was generated using the "Droplet" algorithm. Thus, a total of 2723 water molecules could be objectively added to the low hydrated structure yielding a total hydration value of 0.244 g water/g protein.

Calculation of SAXS Profiles. All small-angle X-ray scattering profiles were calculated for the unhydrated and hydrated structures using a computer program based on an algorithm developed by Lattman (26). This methodology not only allows for rapid calculation of SAXS profiles, i.e., at least ten times faster than other procedures, but also allows for optimization of the residual between calculated and experimental SAXS profiles using adjustable temperature factors for protein, bound water and solvent water. The effects of solvent have been modeled by subtracting from each protein atom a properly weighted solvent water molecule. Protein hydrogen atoms are implicitly accounted for using the strategy of Gelin and Karplus (6).

The scattering profile is given by

$$I(R) = \frac{1}{2} \sum_{n=0}^x \sum_{m=0}^n \epsilon_m N_{m,n} |G_{m,n}(R)|^2 \quad (1)$$

where

$$G_{m,n}(R) \sum_j F_j Y_{m,n}(\theta_j, \phi_j) j_n(2\pi r_j R) \quad (2)$$

I is the scattering intensity, and $R = 2 \sin \theta / \lambda$ where θ is the scattering angle and λ is the wavelength of incident radiation. N is the number of atoms and ϵ is a constant related to the order (n or m) of the Legendre polynomial used. $Y_{m,n}$ are complex spherical harmonics, j_n are the spherical Bessel functions. The index j runs over all atoms, i.e., protein, protein-bound water and solvent water, and r , θ and ϕ are their corresponding spherical atomic coordinates. The expanded structure factor, F_j is given by the following:

$$F_j = (\alpha_P f_P + \alpha_B f_B - \alpha_W f_W) \exp(2\pi i r_j R) \quad (3)$$

where α_P , α_B and α_W are the occupancies of the protein atoms, bound water and solvent water, respectively. The temperature factors, B , are related to the structure factors by

$$f_P = f_P^0 e^{-B_P} \quad (4a)$$

$$f_{BW} = f_{BN}^0 e^{-B_B} \quad (4b)$$

and

$$f_W = f_W^0 e^{-B_W} \quad (4c)$$

where the f^0 are the structure factors in the absence of thermally induced vibrational motion and the B factor compensates for temperature induced changes (a Debye-Waller constant). All subscripts of P , B and W represent the atoms due to the protein, bound water and solvent water. The units of B are \AA^2 .

All calculations using the Lattman program were performed on a VAX 8350 (Digital Equipment, Waterbury, MA) computer. All BPTI calculations took 20 min. to complete whereas submicelle structures required at least 22 hrs.

Results and Discussion

Refined Casein Complex Structures

Synthetic Submicelle: Framework. Following the discovery by Waugh and von Hippel (42) of κ -casein, the stabilizing factor of casein micelles, many studies aimed at understanding the nature of the protein-protein interactions involved were conducted. For the bovine system, these studies focused upon reconstituting micelles with mixtures of α_{s1} - and κ -caseins (28, 35). The reasons for this selection included:

1. Synthetic micelles roughly resembling those of parent micelles could be formed from these two fractions alone.
2. Historically, β -casein was readily separable from the other fractions by mild procedures, so that it was not considered a primary reactant.
3. Separation of κ -casein from the α -complex had been a relatively difficult task, indicating a high degree of interaction.

All of these factors pointed toward the importance of α_{s1} - κ -casein interactions in the bovine casein system (20, 28, 35).

Initial micelle reconstitution experiments (42) suggested that maximum stability of reformed micelles occurred at a ratio of 4:1, α_{s1} - κ -casein. Later, Noble and Waugh (28) suggested a ratio of 10:1 overall but with stronger 1:1 complexes as nucleating sites. In consideration of these studies three factors are important: first, these were whole κ -casein fractions and Groves et al. (17) have recently shown that these preparations contain polymers ranging up to octamers and above as well as some monomers, depending upon the degree of disulfide bonding; secondly, the ratio of 4:1:4:1 for α_{s1} - α_{s2} - β - κ -casein, is about 9:1 in terms of phosphorylated calcium sensitive caseins to κ -casein; finally, the redox potential of the bovine mammary gland lies far toward the reducing end of the scale as the ratio of NADP to NADPH is 4×10^{-5} (2). All of these factors considered, along with the potential reactivity of κ -casein as a monomer (46), a ratio of 4:1 for the interaction of α_{s1} -casein with a reduced κ -casein monomer appeared to be a logical starting point for the construction of a theoretical submicelle.

Figures 1(a), (b) and (c) show the backbone structures for the energy minimized models of κ -casein B, α_{s1} -casein B and β -casein A², respectively. The κ -casein structure which has been colloquially referred to as a "horse and rider" model (Figure 1(a)) contains two sets of "dog-leg" structures. These so-called "dog-leg" structures are the result of two sets of antiparallel sheet structures each connected via a proline residue in a γ -turn configuration, i.e., prolines 27 and 47. It may be noted that for κ -casein, these prolines and the preceding and following sequences appear to be functionally preserved across a variety of species whose primary structures are known (19, 23). Hydrophobic groups, notably tyrosine and valine which are evolutionarily conserved, are the predominant side chains located on both the smaller "dog-leg" (residues 20-34) and the larger one, (residues 39-55). In addition, each "dog-leg" contains a lysine side chain near the pivotal proline residue; there is conservation of these positive charges in almost all κ -caseins (19, 23). This positive charge could conceivably form hydrophobically stabilized ion pairs with another "dog-leg" structure from another casein containing an acidic group in a similar position. Such "dog-leg" structures could easily be docked in an antiparallel fashion to maximize attractive dipole-dipole interactions and yield an acceptable stabilization energy.

Inspection of the β -casein A² structure of Figure 1(c) shows no such "dog-leg" structures. Only two distorted arm structures are observed; but, these arm structures contain hydrophilic side chains, and are unlikely candidates for hydrophobic interaction with κ -casein. The hydrophobic domain, left side of

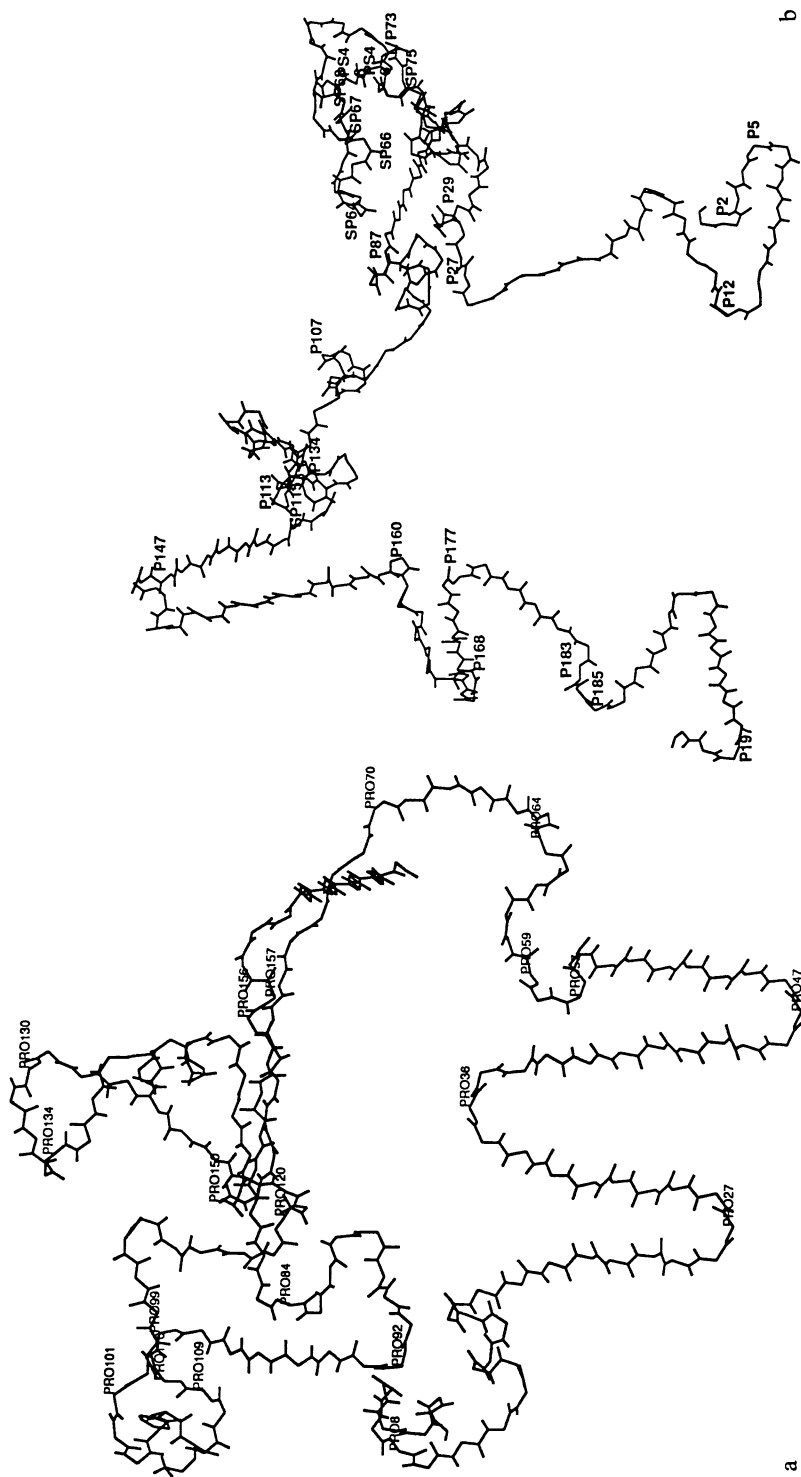
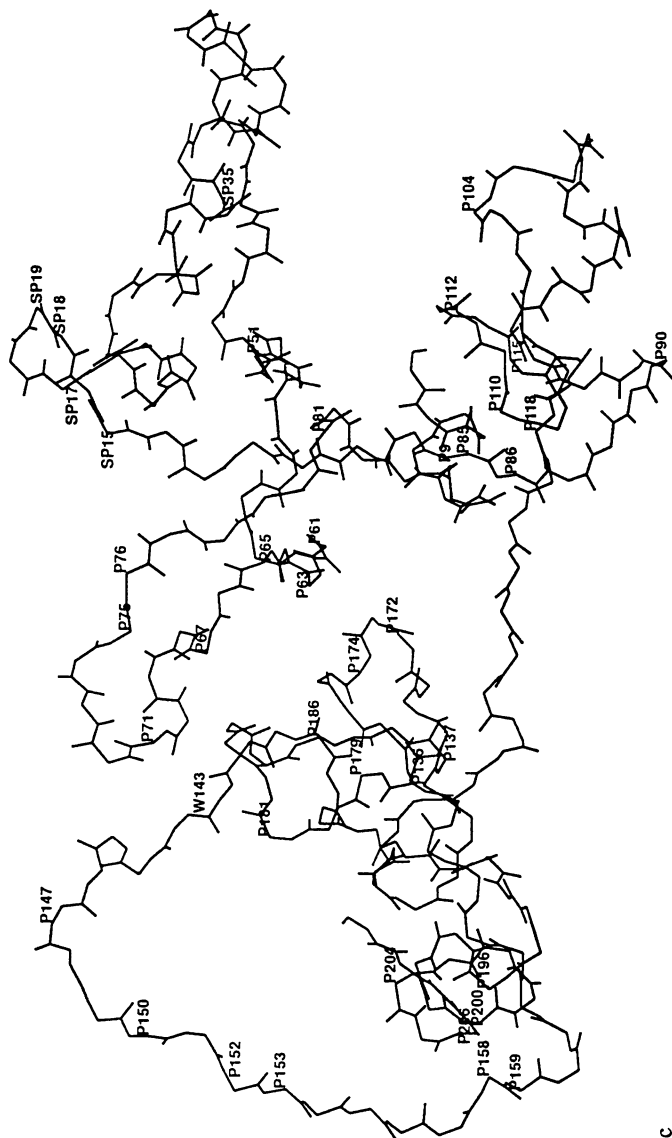


Figure 1a and 1b. (a) Backbone structure with labeled prolines (PRO) of α_s1 -casein B. (b) Backbone structure with labeled prolines (P) and phosphoserines (SP) of α_s1 -casein B. (Figure 1a reproduced with permission from reference 22. Copyright 1993 American Dairy Science Association.)



c

Figure 1c. Backbone structure with labeled prolines (P) and phosphoserines (SP) of β -casein A². (Reproduced with permission from reference 22. Copyright 1993 American Dairy Science Association.)

Figure 1(b), of the α_{s1} -casein B structure however has two such "dog-leg" structures, i.e., a large one, residues 136-159, and a smaller one, residues 162-175, whose side chains are predominantly hydrophobic. Both of these "dog-leg" structures contain proline residues, i.e., residues 147 and 168, as pivotal points for the stranded antiparallel sheet structures. Unlike the γ -turn structures in κ -casein, these prolines are in the 2 position of a β -turn configuration, allowing for greater intra-chain hydrogen bonding. The larger structure, residues 136 to 159 with a pivotal proline at residue 147, appears to have been deleted in ovine α_{s1} -casein, but is functionally conserved in rat at n+3 residues from bovine (19, 23). The larger "dog-leg" thus has some variance in charge and size across species. In contrast, the smaller "dog-leg" centering on proline 168 is functionally preserved in all species of α_{s1} -casein molecules examined, as is the proline at 160 which begins this structure (19, 23). Additionally, tryptophan 164 and tyrosine 166 are invariant and may potentiate hydrophobic interactions. There are no positive charges preceding the pivotal proline but a negatively charged aspartic 175 is conserved following this residue (19, 23).

In a previous report, it was shown that the large "dog-leg" structure, residues 136-159, of α_{s1} -casein B was an excellent site of dimerization of α_{s1} -casein, yielding an interaction energy of -505 kcal/mole, as calculated from the resulting difference between the energy of the dimer (Col. 2 of Table 1) and two times the energy of the monomer (Col. 1 of Table 1). As noted above, variance across species in this region has led to the prediction that ovine α_{s1} -casein may have altered self-association properties relative to its caprine and bovine counterparts (23). Dimer formation (see Figure 2(a)) at the larger "dog-leg" permits the easy docking of one of the small α_{s1} -"dog-leg" structures (residues 162-175) in an antiparallel fashion to a κ -casein "dog-leg" structure. In fact, the last residue of the small α_{s1} -casein "dog-leg", aspartic 175, would easily interact with either lysine residue (24 or 46) on each of the "dog-leg" structures of κ -casein, resulting in a hydrophobically stabilized ion pair formation. The choice of α_{s1} -casein interacting as a dimer is supported by the studies of van de Vroot et al. (39) who examined the interactions of whole κ -casein and α_{s1} -casein by sedimentation equilibrium. They suggested dimer formation by α_{s1} - could precede α_{s1} - κ -interactions. Furthermore, they demonstrated that κ - κ -polymers were larger in size than the resultant α_{s1} - κ -complexes, indicating a more energetically favorable state for the complexes. Pepper (30) and Pepper and Farrell (31) showed similar changes by gel chromatography. Slattery and Evard (37) observed complex formation between reduced κ -casein and α_{s1} -casein by sedimentation velocity studies. Association constants at 20°C for α_{s1} - and κ -casein complexes range from 2 to $8 \times 10^4 \text{ M}^{-1}$ ($K_D = 12$ to $50 \mu\text{M}$) depending upon the method of measurement (39). For the polymerization of reduced κ -casein to its "micellar" complex an association constant of $4.5 \times 10^4 \text{ M}^{-1}$ ($K_D = 22 \mu\text{M}$) can be calculated from Vreeman et al. (40). Association constants for α_{s1} -casein polymerization range from 8 to $11 \times 10^4 \text{ M}^{-1}$ ($K_D = 9$ to $12 \mu\text{M}$) as calculated from Schmidt and Payens (35) at 21°C and ionic strength equal to .1. Thus, considering the evidence for complex formation and the similarity of the association constants, there is a very high probability that α_{s1} - κ -caseins could form these postulated mixed complexes.

Table 1. Calculated energy of α_{s1} -casein and β -casein dimers

Structure	α_{s1} -Monomer	α_{s1} -Dimer	β -Casein Dimer
Bond Stretching	34.3	75.7	82.5
Angle Bending	425.5	892.9	1314.5
Torsional	427.3	840.9	44.6
Out of Plane Bending	15.6	36.6	602.6
1-4 van der Waals	305.9	616.3	1091.6
van der Waals	-872.4	-1805.0	-2094.2
1-4 Electrostatic	2135.3	4268.2	4400.1
Electrostatic	-4426.3	-9337.6	-10230.0
H-Bond	-46.7	-96.4	-123.8
Total	-2001.6	-4508.5	-4912.1

Docking two α_{s1} -casein dimer structures via their small "dog-leg" structures in an antiparallel fashion with the two "dog-leg" structures of κ -casein, one interaction in front of the κ -casein (hydrophilic ends up) and one behind the other "dog-leg" structure (hydrophilic ends down), yielded a rather spherically symmetric structure. Energy minimization of this model composed of one κ - with four α_{s1} -casein monomers yielded an excellent energy of -11811.1 kcal/mol as seen in Col. 1 of Table 2. The architecture of this refined synthetic submicelle structure is presented in Figure 3 as a ribboned backbone structure.

Here, all the hydrophilic domains of α_{s1} -casein which contain the serine phosphates (pictured as ball and stick models) and acidic groups, are located on the outside of the structure for easy access by water and calcium as potential sites for calcium binding and cross-linking which leads to micelle formation. The dimeric phosphate clusters are however diagonal (left to right) from each other decreasing charge repulsions. The internal portion of this structure is divided into four open sectors. The two parallel to the κ -casein central structure (top view of Figure 3) are largely hydrophobic in which fats and other hydrophobic solutes could bind. The other two quarters, which are perpendicular to the κ -casein (left and right of 3), are hydrophilic and are easily water and enzyme accessible. The massive hydrophobic surface area produced by the two central hydrophobic quadrants (Figure 3) can also be potential interaction sites for four β -casein structures via hydrophobic interactions provided the area is large enough not to cause poor van der Waals contacts. It should also be noted at this time, that the docking of four α_{s1} -casein structures to the κ -casein structures could not be accomplished by use of any of their larger "dog-leg" structures without producing large positive

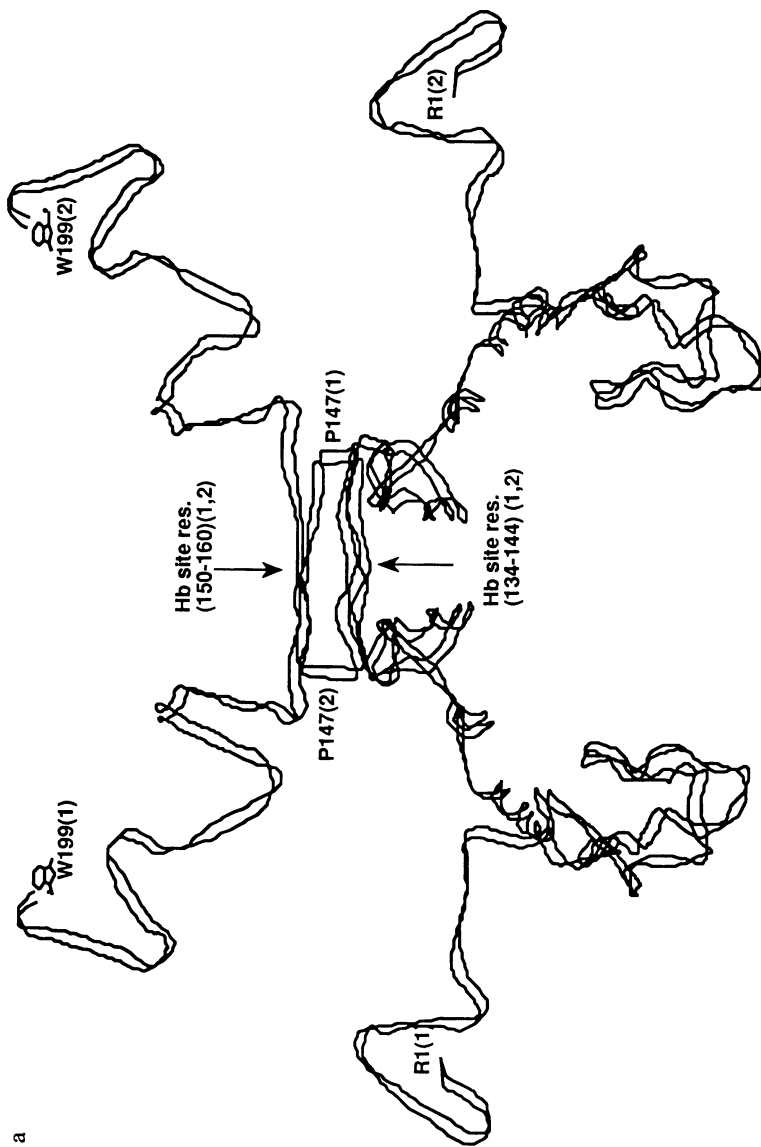


Figure 2. (a) Double ribbon structure of α_{31} -casein B dimer constructed by docking two large hydrophobic sheets in an antiparallel fashion interactions sites noted (1,2) refer to molecules 1 and 2). Hb = hydrophobic sitem, W199 = C-terminal tryptophan, R1 = N-terminal arginine, P147 = proline 147.

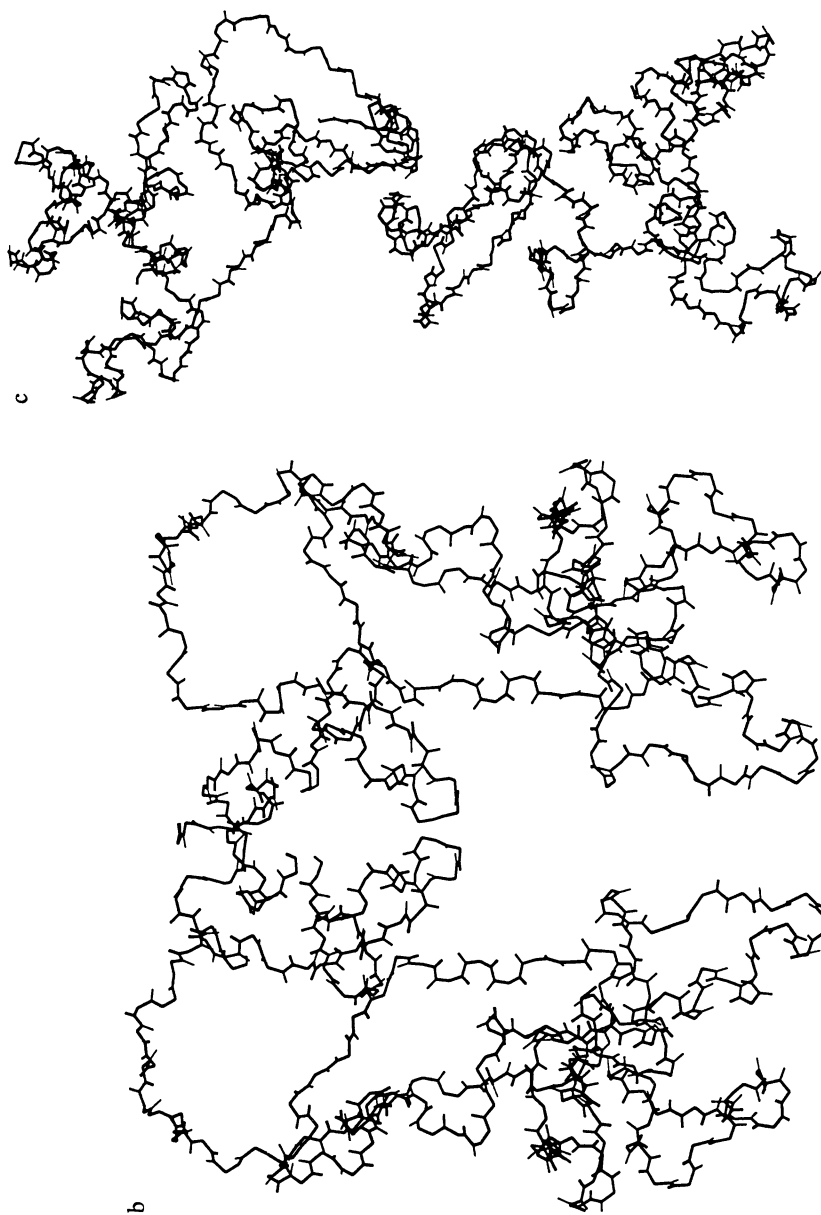


Figure 2. Continued. (b) Backbone structure for β -casein A^2 asymmetric dimer.
(c) Backbone structure for β -casein A^2 symmetric dimer.

Table 2. Energy for refined casein submicelle structures

Structure	Syn. Sub. ¹	Submicelle ²	
		Asym.	Sym.
Bond Stretching Energy :	177.7	208.7	363.2
Angle Bending Energy :	2280.1	4909.0	3430.1
Torsional Energy :	1996.4	2888.4	3272.6
Out of Plane Bending Energy :	93.1	200.5	496.1
1-4 van der Waals Energy :	1311.5	2697.3	2939.1
van der Waals Energy :	-4003.7	-8379.8	-8200.5
1-4 Electrostatic Energy :	12046.8	19000.7	19292.8
Electrostatic Energy :	-25454.9	-43813.0	-45120.4
H-Bond Energy :	-258.1	-513.9	-519.0
Total Energy :	-11811.1	-21802.2	-24047.0

¹ Syn. Sub = synthetic micelle framework of one κ -casein and two α_{s1} -casein dimers.

² Submicelle = theoretical submicelle, consisting of the one κ -casein and two α_{s1} -casein dimer framework with two β -casein dimers added in asymmetric or symmetric fashion.

destabilization energies, caused initially by poor van der Waals contacts. This (Figure 3) is the best structure yielding the lowest energy as determined by up to 10 to 15 docking combinations between two α_{s1} -dimers and one κ -casein monomer.

From the above rationale, it appears on the basis of experimental evidence and from modeling considerations that before the complete casein submicelle can form, a synthetic submicelle framework consisting of κ -casein with α_{s1} -casein structures could form as an energetically favorable intermediate. After this structure is formed, β -casein can then interact to form a hypothetical casein submicelle.

Submicelle Structure. As the interaction of β -casein with the micelle is primarily hydrophobic (18, 34, 43) it seemed plausible to dock two β -caseins within each of the two larger hydrophobic quadrants of the refined synthetic submicelle structure which occur to the right and left of the "rider" (see Figure 3). However, whether this should be performed in a random fashion was not immediately evident.

In a previous report, the energy minimized structure of β -casein monomer and aggregates were presented (21). A radius of gyration of 23 Å could be calculated for the monomer model which can be approximated by a prolate ellipsoid of revolution of radii 42 Å by 21 Å. It resembles a detergent molecule in as much as one end of the molecule contains two hydrophilic arms while the other end and central portion of the structure contains predominantly hydrophobic side chains. To comply with chymosin cleavage experiments of Creamer (7), an asymmetric dimer was constructed as a precursor to β -casein polymeric structures (21). (The terms symmetric and asymmetric will be used here to describe structures with and

without a center of inversion, respectively.) The asymmetric dimer was constructed in an antiparallel fashion with a β -spiral region of residues 190-206 used as a hinge point; the importance of this region in β -casein interactions has been established by Beery and Creamer (3). In the resulting dimer, all the hydrophilic groups remain on one side of the structure and the hydrophobic groups are located on the other side (Figure 2(b)). After energy minimization this dimeric structure yields a total energy of -4912.1 kcal/mole (Col. 3 of Table 1). Two of these β -casein A² dimers were then docked in an asymmetric fashion within the two hydrophobic quadrants of the synthetic submicelle framework structure of Figure 3, with their hydrophilic arms pointed outward from the central cavity. This resulting structure (Color Plate 20) was next energy minimized and yielded an acceptable energy of over -10 kcal/mole/residue (Table 2). Although no stabilization energy is observed, such a structure is highly likely since the interaction between the β -casein dimers and the synthetic submicelle structure is for the most part hydrophobic, which is supported by the dissociation of β -casein from submicelles and micelles at 4°C and below (1, 9, 10, 34).

Another submicellar structure, which will be referred to as a symmetric model, can be built by the symmetric docking of two β -casein A² symmetric dimers into the two hydrophobic cavities of the synthetic submicelle structure. The β -casein A² symmetric dimer (Figure 2(c)), contains two hydrophilic sites at either end of the structure and a central hydrophobic region. The symmetric dimer has no loss in stabilization energy and after minimization gives a total energy of -5484.4 kcal/mole, which is different from the asymmetric dimer and will be discussed in detail in the energetics section of this manuscript. In this submicelle structure, the symmetric dimers must be docked with their central hydrophobic portion in contact with the hydrophobic cavity of the synthetic submicelle structure so that their hydrophilic areas are actually perpendicular to corresponding β -casein dimeric structures docked within the asymmetric submicellar model (Color Plate 20). The resulting structure was energy minimized (Table 2) and is presented in Plate 20. It should be noted that one of the β -casein hydrophilic portions of the symmetric dimer partially covers the view of the GMP of the κ -casein within this structure, but it in no way interferes with access of the chymosin to the phenylalanine-methionine cleavage site of κ -casein, nor does it hinder access to glycosylation sites on the κ -casein.

Several other approaches to docking β -casein dimers were attempted. Most of these resulted in extreme loss of stabilization energy. The dominating factor in docking the four β -caseins is the proximity of the four hydrophilic ends to each other and to the phosphate rich portions of α_{s1} -casein. Charge repulsions in these areas prevent many hypothetical approaches. To test the above energy minimized structures, comparisons of the models will be made with experimental evidence derived from Raman spectroscopy and small-angle X-ray scattering.

Secondary Structural Analysis. Global secondary structural analysis was initiated on the energy minimized three dimensional submicelle structures. The results were compared with the reported global secondary structure calculated from Raman spectroscopy (Table 3). No significant changes in the ϕ , ψ angles of the backbone peptide bonds between the individually refined κ -, α_{s1} - and β -casein structures and

NOTE: The color plates can be found in a color section in the center of this volume.



Figure 3. Refined structure of casein synthetic submicelle framework, i.e., one κ -casein B and four α_{s1} -casein B monomers. Ribbon backbone; for α_{s1} -casein monomers (Figure 1b) side chains of serine phosphates only are shown as ball and stick models. The view is from the top in which κ -casein (see Figure 1a) can be seen as essential nucleation point for the formation of this framework structure.

Table 3. Comparison of initial structures with spectroscopic data

Sample		% Helix	% β -Structure	% Turns	% unspec
Submicelle (Lyophilized)	Raman ¹	8 - 18	24 - 30	36 - 39	16 - 32
κ -casein	Refined	16	27	30	26
α_{s1} -casein	Refined	8	18	34	40
β -casein	Refined	10	20	34	36
Submicelle ²	Calculated	10	20	34	36

¹References 4, 5.

²Constructed for asymmetric and symmetric models (see text).

those within the two submicelle structures was observed. Table 3 also contains the secondary structure analysis calculated in previous communications for the individual casein structures. Hence, an average of the molecular fractions of caseins in the submicelle could be used to calculate the global secondary structure of the submicelle structure. The results of this calculation are given in Table 3 as the Calculated Submicelle row and is in reasonable agreement with the experimentally determined values from Raman spectroscopy (Row one of Table 3), even though the Raman spectroscopy experiments were performed on a lyophilized powder of whole sodium caseinate. It is hoped that in the future, precise global secondary experiments will be performed in solution using Raman or FTIR spectroscopy because this same Raman study (5) showed that conformational changes can occur in β -casein A² during lyophilization. While this type of analysis does not prove the refined submicelle structure, it does add further validation to the possibility of such a refined model. It should be stressed, however, that molecular dynamic calculations should be performed in the future, for it is the dynamic structure which ultimately should be correlated with solution physical chemical properties.

Comparison with Solution Structural Results. The energy minimized three dimensional models of the casein asymmetric and symmetric submicelles, shown in (Color Plate 20), can all easily be approximated by a spherical particle of two packing densities. The distance from one end of the β -casein through the α -casein structure and to the end of the opposite β -casein molecule is 100 Å in both models. The packing density of this region would be considered higher than any other within the overall structure. The longest distance measured from one α_{s1} -casein hydrophilic domain to an opposite one is about 200 Å. These values, symmetry and packing densities, agree qualitatively with recent small angle X-ray scattering results (15, 32). Here, the data were modelled as an inhomogeneous sphere of two electron densities of diameter 106 Å by 203 Å with the same center of symmetry. In addition, the low electron density is consistent with the low packing density observed in the refined structure. Such a low electron density could be interpreted as a high hydration value or a particle in which water can easily flow throughout the polypeptide chain or both.

From our present structural study it would be reasonable to assume that the hydrophilic domains of α_{s1} -casein B within the synthetic submicelle structure (Figure 3(b)) are potential interaction sites for colloid formation by self-association of the synthetic submicelle structure via calcium salt bridges. Such a colloidal matrix structure could easily allow for the temperature induced hydrophobic association or disassociation of β -casein from that colloidal matrix structure. In fact the remnant after cold dissociation would mimic the micelle framework postulated by Lin et al. (27). It should be noted that the above is not a conclusive proof for the mechanism of submicelle and micelle formation. It is a working hypothesis, as is the structure, and is in need of a large amount of quantitative experiments to disprove, prove or further refine this structure and mechanism. As with all such models it is not an end in itself but a spring-board to further research.

Many studies of several investigators (12, 18, 35) show that casein submicelles have a significantly higher hydration value than for globular proteins, i.e., 3-6 g

water/g protein. Quantitative comparison of this submicelle structure with small angle X-ray scattering results required the development of a methodology suitable for assessing differences in the models as well as the effects of added water.

Bovine Pancreatic Trypsin Inhibitor. To test the programs developed for comparison of energy minimized structures with SAXS profiles for hydrated and unhydrated casein submicellar structures, and to more fully understand the parameters calculated from the Lattman program, it was important to first use the procedure on a protein molecule with a known hydrated three dimensional X-ray crystallographic structure. We followed the lead of the Lattman paper (26) and used the X-ray and neutron crystallographic structure of BPTI, i.e., the 5BPTI file in the Brookhaven protein databank. The structure obtained in the protein data bank consists of all protein hydrogens as well as 63 water molecules with hydrogens and one bound anion. To mimic the Lattman calculation to be used on the casein models, all hydrogens were removed, the anion was eliminated as well as the three waters associated with this anion. This structure was presented to the Lattman program using the scattering data provided in the program files which were originally determined by Pickover and Engelman (33). To be consistent with the calculation of submicellar casein, only 20 equally spaced data points were used. The resulting calculated B values for the protein, B_P , the bound water, B_B , and the solvent water, B_W , as well as the residual (the variance of the calculated data from experimental data) as a measure of the goodness-of-fit, are presented in the first row of Table 4. The overall profile of the calculated and experimental SAXS results, as filled triangles with a connecting line, are shown in Figure 4(a). As seen in Figure 4(a), the fit of the structure to the SAXS data is acceptable; even the residual value of 0.220 (Table 4) represents an error of only 3 percent. The B values for the bound water and solvent are in reasonable agreement with those calculated by Lattman (26), i.e., 59 \AA^2 and 72 \AA^2 , respectively; however, the B_P of 125 \AA^2 is much lower than that presented in his paper, i.e., 284 \AA^2 . Presumably, this discrepancy may be caused by the use of a lower number of experimental data points. Since the casein submicelle models, developed above, initially contained no water and were energy minimized, it was decided to first test the effect of energy minimization on the goodness-of-fit with experimental SAXS profiles using the Lattman procedure. Energy minimization of the BPTI structure with 60 water molecules from X-ray crystallography, resulted in an energy of -1712.8 kcal , which is well below the energy criterion of -10 kcal per residue or water molecule, i.e., -118 kcal/mole , that we have previously chosen to impose as an acceptable criterion for improvement in the energy value (21-23). The resulting energy minimized structure was then presented without hydrogen to the Lattman program and the resulting B and R values are given in row 2 of Table 4. The residual, R, value of 0.160 for the refined structure is somewhat lower than that of the unrefined model while their corresponding B_P and B_B values are much higher. $B_P = 224$ actually approaches the value of 284 found by Lattman (26). The reasons for the better fit of the energy minimized over the initial structure with the experimental SAXS profiles are not clear at this time. However, since the unrefined structure was determined in the crystal state and the SAXS profiles are obtained in solution where dynamic processes occur, a hypothesis concerning the

Table 4. Temperature factors for bovine pancreatic trypsin inhibitor (BPTI) from SAXS*

Refined	Waters	B_p ¹	B_w ²	B_B ³	R ⁴
No	60	125	55	89	0.220
Yes	60	224	58	200	0.160
No	0	-48	-45	—	0.186
Yes	0	-18	-60	—	0.133
No	4	-12	-47	102	0.0833
Yes	4	-3	-41	122	0.0731
Yes	202	137	120	35	0.148

*SAXS—Small Angle X-ray Scattering.

¹ B_p —Temperature factor for protein, Å².

² B_w —Temperature factor for free water, Å².

³ B_B —Temperature factor for bound water, Å².

⁴R—Residual: deviation of calculated from experimental SAXS data.

necessity of a protein adapting to a lower possible energy structure in solution may be offered.

This hypothesis can be tested at this time by using, the Lattman program, together with theoretical unhydrated and hydrated forms of BPTI in their original and energy minimized structures. Here, the refined and original structures of BPTI with all waters eliminated and with only 4 internal waters retained can be used to calculate theoretical SAXS profiles (Table 4). In addition, calculated energies for the refined structures of BPTI containing no waters and four internal waters are shown in Table 5. The results of the Lattman program for the initial and refined structures with 0 and 4 waters are listed in rows three to six of Table 4. All temperature factors for the protein, B_p , are negative (Table 4). Thus, no physical interpretation regarding the vibrational motion of atoms of the protein can be made. It appears that energy minimization of structures always improves the goodness-of-fit between theoretical and experimentally determined SAXS profiles when using the Lattman methodology and the Tripos software. Comparison can be made for the variation of hydration levels versus goodness of fit to SAXS (R values) for only the energy minimized structures of the BPTI. The results of the correlation of experimental and theoretical SAXS profiles for the refined unhydrated and various hydrated structures of BPTI are given in rows 2, 4, 6 and 7 of Table 4. The lowest value of the residual, R, between the experimental and theoretical SAXS profiles is obtained for the refined structure of BPTI with four bound internal waters. The fit of the theoretical SAXS profile for the structure is presented in Figure 4(a) as open circles; this structure is depicted in Figure 4(b). Comparison

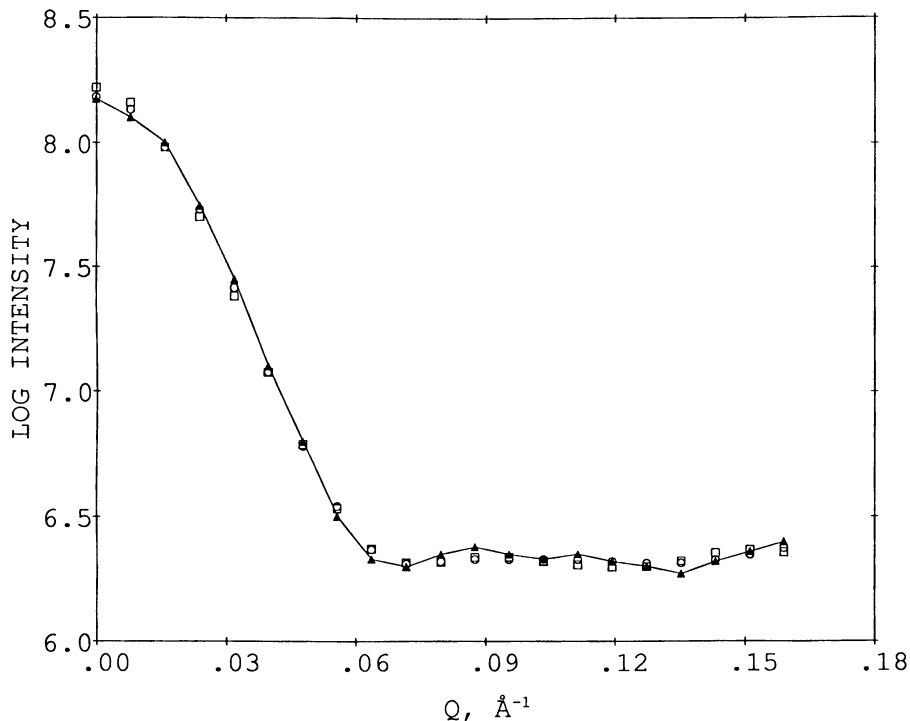


Figure 4. (a) SAXS profiles of Bovine Pancreatic Trypsin Inhibitor, BPTI: filled triangle with connecting lines, experimental data (33); squares, optimized theoretical curve from X-ray crystallographic structure with 60 waters determined via neutron diffraction; circles, optimized theoretical curve from energy minimized structure with 4 internal waters.

with the 60 water BPTI structure, depicted as squares in Figure 4(a) with the four water structure (open circles) and the experimental data (filled triangles) show only slight differences at high Q values but deviate mostly at the very low Q values where the radius of gyration and molecular weight are calculated. Overall, even the 60 water structure profile shows a satisfactory fit between theoretical and experimental SAXS profiles even though its R value is three times larger than the refined internally bound water BPTI structure (Table 4). However, since the B value of the protein is negative for the four internally bound water BPTI structure, we would conclude that some more bound waters are necessary for proper simulation of the data. Using the 202 water structure yields a more acceptable value of 137 \AA^2 for the B_p value as well as a positive B_w value, but with an increase in the R value by a factor of two. Here the actual number of water molecules that should be added to the protein surface appears to be somewhere between 0 and 198. Lattman (26), as well as other investigators (29) using NMR

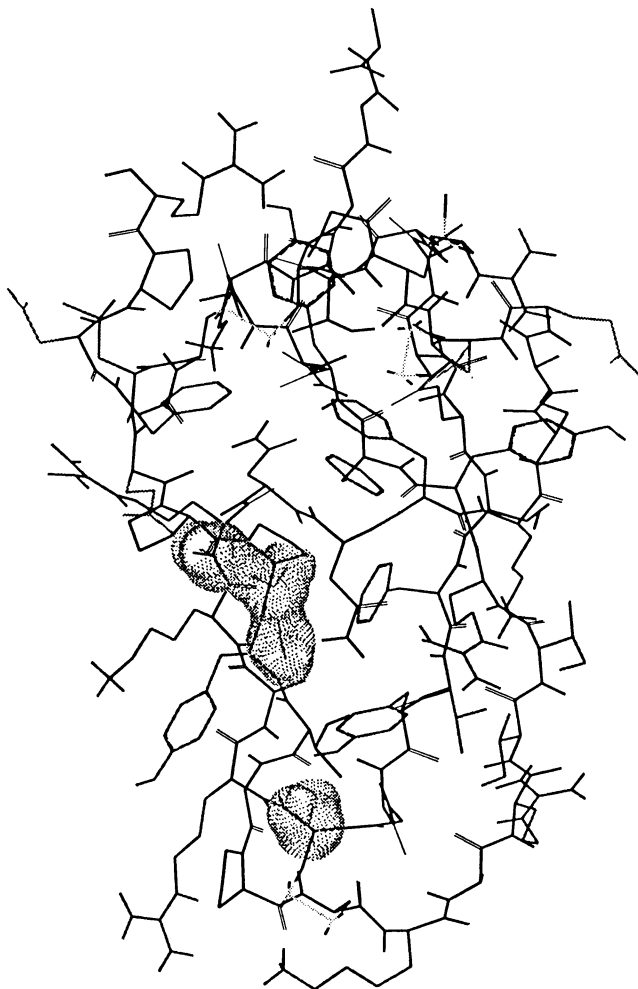


Figure 4. (b) Structure of BPTI with 4 internal waters. The structure shows both backbone and all side chains, as well as van der Waals dot surface (gray) of 4 internal waters.

experiments, also concluded that at least four but no more than ten water molecules, six by NMR, are bound to BPTI. The obvious problem of the number of water molecules, four to ten, bound to BPTI and the exact location of the other two to six surface waters does not allow us to further test their methodology using the SAXS profiles of BPTI. However, the results so far indicate that the method of Lattman is still useful for determining whether or not an energy minimized three dimensional structure can be tested by generation of theoretical SAXS profile results and comparing these with experimental data to determine if structures contain bound waters, externally or internally, when they exist in solution.

Bovine Casein Submicelle Structure. We now turn our attention to attempt to test the energy-minimized casein submicelle structures using the Lattman methodology with published experimental SAXS profiles following the methodology developed for BPTI. We will also attempt to ascertain the need for and location of bound waters within the structures. A search of the literature showed two papers (24, 38) in which the SAXS data were in graphical form for computer aided digitalization, and the appropriate conditions were used to insure submicellar (maximally aggregated) casein solution structure. However, one of the papers (38) only contained SANS data in D₂O and, even though contrast variation experiments were performed at several H₂O/D₂O mixtures, no molecular weights were calculated. Only one report exists whereby precise SAXS experimental profiles were obtained in H₂O and not in D₂O where hydrophobic protein self-association would be increased (24). Also in this study, molecular weights were given with statistical errors calculated to insure that submicellar casein was present. No variation in molecular weight was observed with respect to protein concentration insuring the elimination of particle polydispersity. However, whole casein was used for this SAXS study which contains 10 percent of α_{s2} -casein (8) and as noted above no α_{s2} -casein structure exists within this predicted three dimensional submicellar model. Nevertheless, since only 10 percent of the submicelle casein particle contains α_{s2} -casein, it is thought that comparison of the theoretical curves from the submicelle structure and the experimental SAXS profiles would still lead to fruitful results.

Without energy minimization, both the asymmetric and symmetric unhydrated submicelle structures were subjected to the Lattman procedure for comparison with experimental SAXS profiles. The comparisons with experimental results are presented in Figure 5(a). Here, the experimental data are shown as filled triangles with connecting lines while the theoretical curves, using the asymmetric and symmetric structures, are represented by circles and squares, respectively. As can be seen, both structures yield rather unfavorable SAX profiles. While agreement with experimental data is moderate at large Q values, disagreement is unacceptable at low values which would yield erroneous calculated molecular weights and radii of gyration. The Lattman temperature parameters (B values) from these calculations are given in Table 5. The B values for the protein are positive with close values of 33 and 35 Å², respectively, but, the large R values of 16.9 and 16.4 for both the asymmetric and symmetric models reflect the poor agreement between theoretical and experimental SAXS profiles. Such lack of agreement between theoretical and experimental curves, especially at low Q values, Figure 5(a), may reflect the absence of internally bound water molecules within these submicellar models just as four internally bound water molecules were necessary to obtain the best R value of 0.0731 for the BPTI structure.

With this in mind, 120 water molecules were energy minimized and docked within the κ -casein cavity for both the asymmetric and symmetric submicellar structures. This cavity within the κ -casein molecule would in reality either contain either bound or free water. Bound water was chosen since 120 bound waters would mimic the amount of bound water determined via DNMR Relaxation results (5, 6), i.e., 0.007 g water/g protein, assuming all bound water fell within this cavity.

Subjecting these low hydrated structures (120 water molecules) to the Lattman method yielded good results with more acceptable R values of 0.682 and 0.611 for the asymmetric and symmetric models, respectively (Table 5). Also, the corresponding B_p values are both positive, i.e., 43 \AA^2 whereas the B_w are both negative, -540 \AA^2 . The negative B_w values could be an artifact of the approximation by Lattman for positioning of the free water molecules at the same position as the protein and bound water atoms.

To follow the methodology developed with BPTI study, we now add a larger surface of water, i.e., 2703 water molecules to each of the low bound water submicelle structures using the droplet algorithm. This amount of water, 2823 moles water/mole protein, would mimic an accepted hydration value of 0.244 g water/g protein for globular proteins. The normal hydrodynamic hydration value for casein submicelles is on the order of 2 to 3 g water/g protein (13, 24), while the gravimetric hydration of isoelectric casein is about 0.7 g water/g protein (13, 15). However, due to the large number of water molecules involved to achieve these numbers as well as the length of the calculation, it would seem prudent to first attempt to solve the 2823 water molecule structure to ascertain if any improvement is observed in the R values. Using the Lattman procedure on the two droplet structures (i.e., asymmetric and symmetric + 2823 water molecules), which we shall define as high hydration structures, R values were obtained that were twice as large as those found when using the low hydration (120 water molecules) structures (see Table 5). Both B_p values for the asymmetric and symmetric high hydration structures were 36 \AA^2 and in close agreement with the value of 43 \AA^2 obtained using the low hydration structure. However, as in the BPTI study, the B_w values were now positive and the B_B values were more realistic, i.e., on the order of 100 \AA^2 . Hence, it would appear that the true hydration value for the submicellar casein structure lies somewhere between 120 and 2823 molecules of water per molecule of protein. The exact amount and location of these new bound waters can not be determined at this time. Further studies may, in time, resolve this problem.

We next energy minimized all four hydrated complexes, i.e., the low and high hydrated asymmetric and symmetric models, to further mimic the BPTI study, and subjected all four hydrated energy minimized structures to the Lattman procedure. Interestingly little improvement in fit occurred following the minimization of the protein water complexes (Table 5). The energy minimized theoretical and experimental profiles for the low and high hydrated structures are shown in Figures 5(b) and 5(c), respectively for comparison. In both figures the experimental data are given as filled triangles with connecting lines while the SAXS profiles that form the asymmetric and symmetric models are circles and squares, respectively. Figure 5(c) shows clearly that a less acceptable fit to the experimental data is obtained with both high hydrated structures, especially at low Q values. The zero Q value from which the molecular weight is calculated yields an acceptable value with the experimental value but all other Q values deviate significantly in a positive then a negative manner to a Q value of 0.05 \AA^{-1} . Such a non-monotonic curve with a maximum at low Q values is indicative of high virial or ordering effects. Since this ordering is not observed in the experimental data, it appears that the droplet

Table 5. Temperature factors for submicelle structures from SAXS

Structure	Water	B_p	B_w	B_B	R
Asymmetric	None	33	-181	—	16.9
Asymmetric	Low	43	-540	413	0.682
Asymmetric	High	36	53	98	1.17
Symmetric	None	35	25	—	16.4
Symmetric	Low	43	-540	503	0.611
Symmetric	High	36	57	130	1.26
Asymmetric*	Low	46	-550	436	0.684
Asymmetric*	High	30	85	156	1.81
Symmetric*	Low	45	-550	500	0.612
Symmetric*	High	31	27	100	1.29

*Denotes energy minimized submicelle-water complex.

Note: All other parameters defined in Table 4.

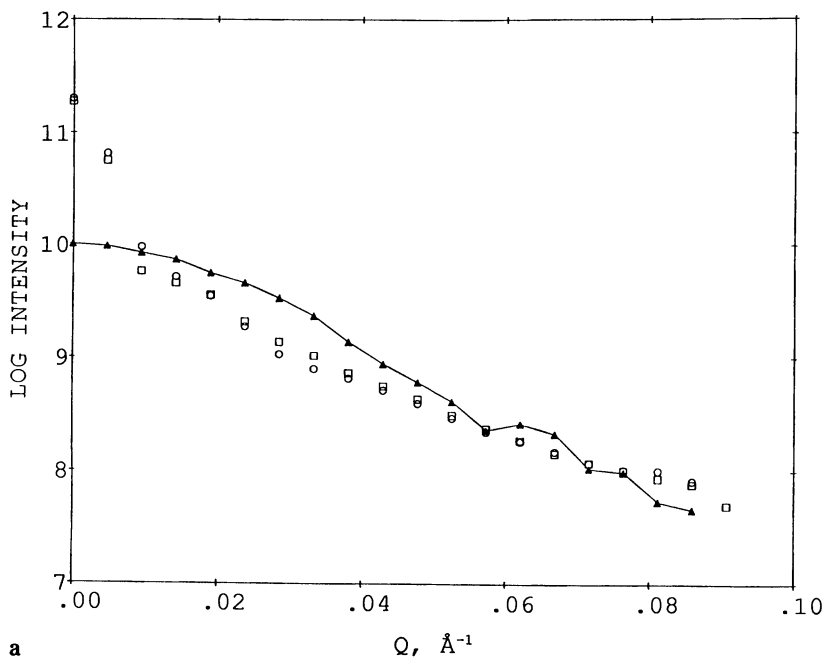
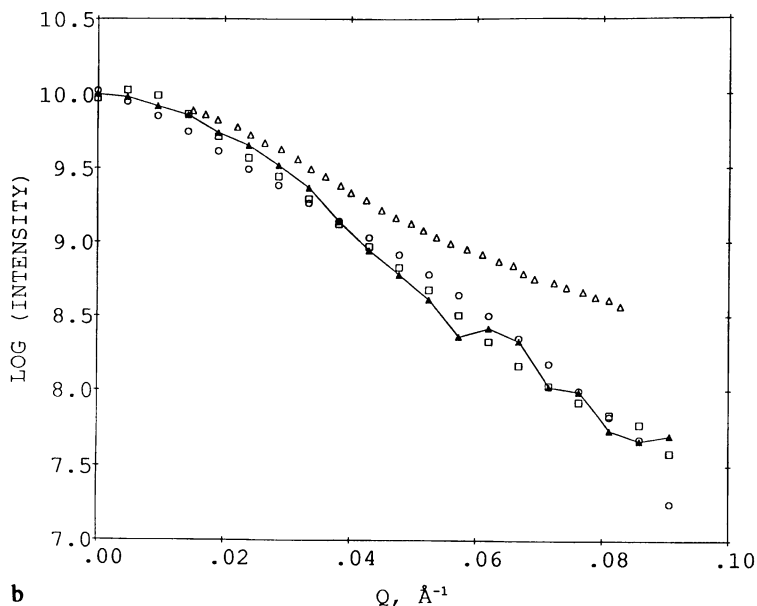
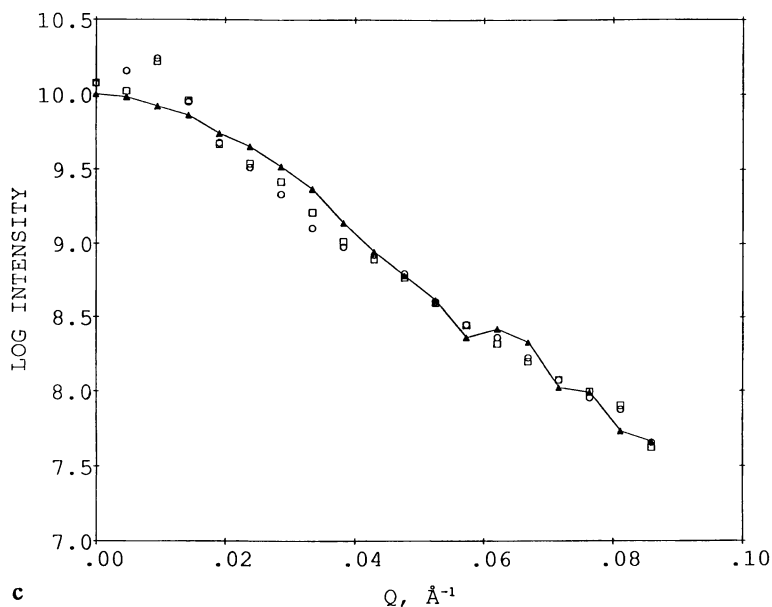


Figure 5. SAXS profiles of submicellar casein; solid triangles (▲) with connected lines represent experimental data (24). (a) Theoretical optimized curves from unhydrated asymmetric (circles) and symmetric (squares) energy minimized structures.



b



c

Figure 5. Continued. SAXS profiles of submicellar casein; solid triangles (\blacktriangle) with connected lines represent experimental data (24). (b) Theoretical optimized curves of energy minimized low hydration (120 waters) asymmetric (circles) and symmetric (squares) structures; open triangles (Δ) are experimental small-angle neutron scattering in D_2O (38). (c) Theoretical optimized curves for energy minimized high hydration (2823 waters) asymmetric (circles) and symmetric (squares) structures.

water added to the low hydrated structures is not tightly bound water but are most likely free waters and easily exchanged with the bulk solvent molecules.

Figure 5(b) shows good agreement with theoretical and experimental SAXS profiles. Here, it can readily be seen that the circles are closer to the filled triangles than the squares which represent the SAXS profile calculated from the low hydrated asymmetric structure at almost all Q values. This fit is further reflected in Table 5 by a slightly lower R value of 0.612 for the low hydrated symmetric structure than the corresponding asymmetric model, 0.684. However, these results should indicate only that the low hydrated symmetric structure has only a slightly higher probability of existence in solution. These calculations and experiments do not justify the elimination of a symmetric form. What can be also concluded from this study is that the casein submicelle is structurally more rigid than globular proteins but overall more flexible, in that submicelle formation may occur under a variety of conditions with varying combinations of monomers (12). This is seen by the fact that the B_p values for all submicellar casein models yield values on the order of 35 to 40 \AA^2 (Table 5), while the BPTI values were much higher, i.e., 135 to 200 \AA^2 (see Table 4). These consistently lower B_p values for casein may be ascribed to the existence of proline residues throughout the polypeptide chain which would yield an open but more rigid structure. Conversely the movement of β -casein in and out of the complex with temperature is most likely not microscopically reversible (12).

In addition, the reported neutron scattering data of casein in D_2O (38) is shown in Figure 5(b) as open triangles. Here, the neutron data were normalized at zero Q value to be compared with the SAXS profiles. As can easily be observed, large differences exist between experimental SAXS profiles and neutron scattering data in D_2O . Whether this difference is a direct result of structural changes induced by the addition of D_2O is not clear at this time. However, these results do suggest to the casein investigator that care must be taken to avoid the addition of D_2O in casein solutions.

Finally, because of the rigid structure of all submicellar structures, it would be prudent to ascertain if the protein structure had an influence on the structure of the water molecules within the various hydrated forms. Figure 6(a) shows these 120 added water molecules within the α -casein cavity for the energy minimized low hydrated submicelle structure. Only the water, shown as "wire-frame" structure and the ribboned backbone of α -casein are displayed. The dashed lines indicate the presence of all hydrogen bonds. It can be easily seen that a worm-like structure of the 120 water molecules is present within the energy minimized α -casein cavity. This super structure of waters is obviously due to the influence of protein electrostatic interactions and resembles a solid distorted cylinder as seen by the space-fill model of these waters shown in Figure 6(b). Previous studies have shown that the α -casein molecule exhibits a dipolar character (22). The protein dipolar character is apparently superimposed on the internal water molecules based upon their stability within the cavity following minimization (Figure 6a,b). Here, then, is a clear representation of the influence of protein structure and energetics on internal bound water structures. That the protein structures did not change when water was added and the complexes were energy minimized, further suggests

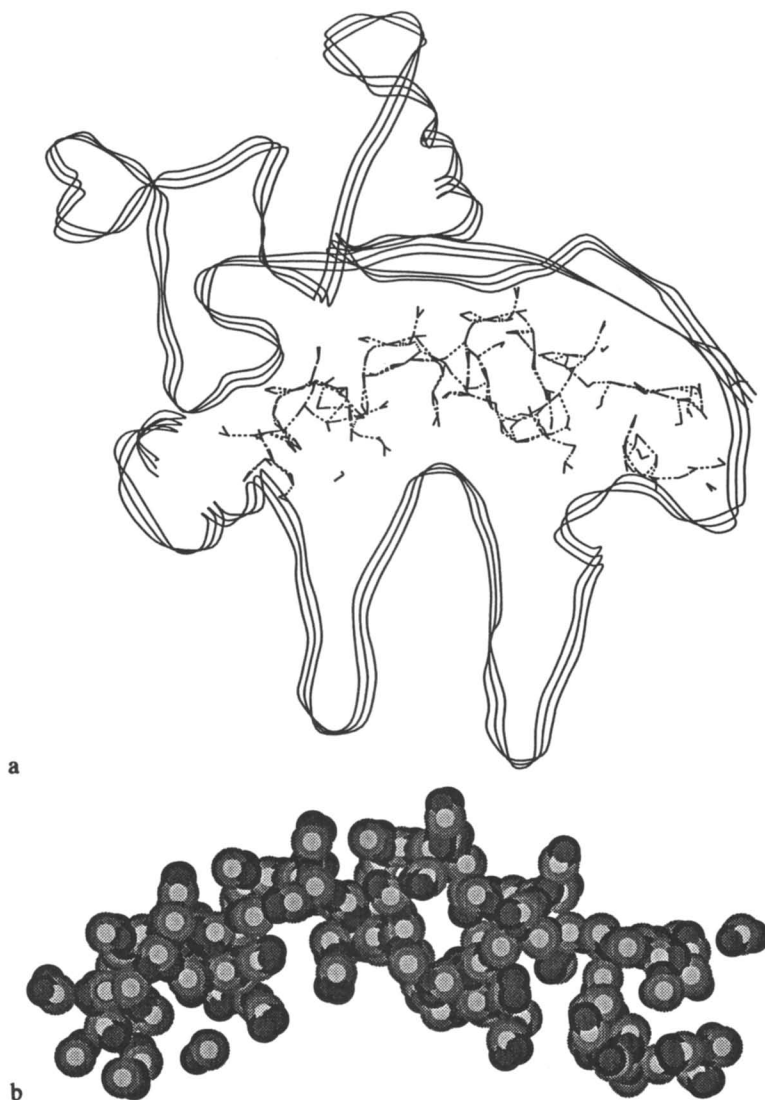


Figure 6. (a) Energy minimized docked low hydration waters (120 waters) displayed with ribboned α -casein backbone structure; water in black v-shape, with dashed lines representing hydrogen bonding. (b) Space-filled energy minimized model of low hydration waters colored by atom types; oxygen in light to medium shading and hydrogen in dark to black shading.

the hypothesis that α -casein, the backbone structure is somewhat rigid, while side chains have a great deal of mobility.

As emphasized in previous papers on the monomeric caseins (21-23), it must be kept in mind that these structures represent working models. They are not the final native structures but are presented to stimulate discussion and to be modified

as future research unravels the nature of these non-crystallizable proteins. Inspection of a recent drawing of the casein micelle by Holt (18) demonstrates how structures such as those presented here could be further aggregated into the casein micelle. Continued dialogue and research in this area may bring together the new concepts necessary to finally produce an accurate model. It is hoped that this work is a start in that direction.

Note

The mention of brand or firm names does not constitute an endorsement by the U.S. Department of Agriculture over others of a similar nature not mentioned.

Literature Cited

1. Ali, A. E.; A. T. Andrews; G. C. Cheeseman. *J. Dairy Sci.*, **1980**, *47*, 371-382.
2. Baldwin, R. L.; Yang, Y. T. In *Lactation: A Comprehensive Treatise*, Larson, B. L.; Smith, V. R. Eds. Academic Press Inc., New York, NY, 1974, p. 349-355
3. Berry, G. P.; Creamer, L. K. *Biochem.*, **1975**, *14*, 3542-3535.
4. Byler, D. M.; H. M. Farrell, Jr. *J. Dairy Sci.*, **1989**, *72*, 1719-1721.
5. Byler, D. M.; H. M. Farrell, Jr.; H. Susi. *J. Dairy Sci.*, **1988**, *71*, 2622-2629.
6. Clore, G. M.; Bax, A.; Wingfield, P. T.; Gronenborn A. M. *Biochem.*, **1990**, *29*, 5671-5676.
7. Creamer, L. K. *N.Z. J. Dairy Sci. and Technol.*, **1976**, *11*, 30-39.
8. Davies, D. T.; Law, A. J. R. *J. Dairy Res.*, **1980**, *47*, 83-90.
9. Davies, D. T.; Law, A. J. R. *J. Dairy Res.*, **1983**, *50*, 67-75.
10. Downey, W. K.; Murphy, R. F. *J. Dairy Res.*, **1970**, *37*, 361-372.
11. Eigel, W. N.; Butler, J. E.; Ernstrom, C. A.; Farrell, H. M., Jr., Harwalker, V. R.; Jenness, R.; McL. Whitney; R. *J. Dairy Sci.*, **1984**, *167*, 1599-1631.
12. Farrell, H. M., Jr. In *Fundamentals of Dairy Chemistry, 3rd Edition*. N. Wong, ed., Van Nostrand Reinhold, New York, NY, 1988, p. 461-510.
13. Farrell, H. M., Jr.; H. Pessen; Kumosinski, T. F. *J. Dairy Sci.*, **1989**, *72*, 562-574.
14. Farrell, H. M., Jr.; Kumosinski, T. F.; Pulaski, P.; Thompson, M. P. *Arch. Biochem. Biophys.*, **1988**, *265*, 146-158.
15. Farrell, H. M., Jr.; Pessen, H.; Brown, E. M.; Kumosinski, T. F. *J. Dairy Sci.*, **1990**, *73*:3592-3601.
16. Gelin, B. R.; Karplus, M. *Biochem.* **1979**, *18*, 1256-1259.
17. Groves, M. L.; Dower, H. J.; Farrell, H. M.; Jr., *J. Protein Chem.*, **1992**, *11*, 21-28.
18. Holt, C. *Adv. in Prot. Chem.*, **1992**, *43*, 63-113.
19. Holt, C.; Sawyer, L. *J. Chem. Soc.*, **1993**, *89*, 2683-2690.
20. Kumosinski, T. F.; Farrell, H. M., Jr. Calcium-induced associations of the caseins: thermodynamic linkage of calcium binding to colloidal stability of casein micelles. *J. Protein Chem.* **1991**, *10*, 3-11.
21. Kumosinski, T. F.; Brown, E. M.; Farrell, H. M., Jr. *J. Dairy Sci.*, **1993**, *76*, 931-945.

22. Kumosinski, T. F.; Brown, E. M.; Farrell, H. M., Jr. *J. Dairy Sci.*, **1993**, *76*, 2507-2520.
23. Kumosinski, T. F.; Brown, E. M.; Farrell, H. M., Jr. In *Molecular Modeling. ACS Monograph Series*, Kumosinski, T. F.; Liebman, M. N., eds. Denver, CO, 1994.
24. Kumosinski, T. F.; Pessen, H.; Farrell, H. M., Jr.; Brumberger, H. *Arch. Biochem. Biophys.*, **1983**, *266*, 548-561.
25. Kyte, J.; Doolittle, R. F. *J. Mol. Biol.*, **1982**, *157*, 105-132.
26. Lattman, E. E. *Prot. Stru. Funct. and Gene.* **1989**, *5*, 149-158.
27. Lin, S. H. C.; Leong, S. L.; Dewan, R. K.; Bloomfield, V. A.; Morr, C. V. *Biochem.*, **1972**, *11*, 1818-1821.
28. Noble, R. W., Jr.; Waugh, D. F. *JACS*, **1965**, *87*, 2236-2242.
29. Otting, G.; K. Wüthrich. 1989. Studies of protein hydration in aqueous solution by direct NMR observation of individual protein-bound water molecules. *J. Am. Chem. Soc.* *111*, 1871-1875.
30. Pepper, L., *Biochim. Biophys. Acta* **1972**, *278*: 147-155.
31. Pepper, L.; Farrell, and H. M., Jr. Interactions leading to formation of casein submicelles. *J. Dairy Sci.*, **1982**, *65*, 2259-2266.
32. Pessen, H.; Kumosinski, T. F.; Farrell, H. M., Jr.; Brumberger, H. *Arch. Biochem. Biophys.*, **1991**, *284*, 133-142.
33. Pickover, C. A.; Engleman, D. M. On the interpretation and prediction of X-ray scattering profiles of macromolecules in solution. *Biopolymers* *21*:817-828.
34. Rose, D. *J. Dairy Sci.*, **1968**, *51*, 1897-1902.
35. Schmidt, D. G.; Payens, T. A. J. In *Colloidal aspects of casein in surface and colloid science, Vol. 9*. E. Matijevic (Ed.). Wiley and Sons, Incorp. New York, NY. 1976.
36. Schmidt, D. G. In *Developments in Dairy Chemistry -1*, P. F. Fox (Ed.). Applied Science Publish Limited, Essex, England, 1982, p. 61-93.
37. Slattery, C. W.; Evard, R. *Biochim. Biophys. Acta.*, **1973**, *317*, 529-538.
38. Stothart, P. H. *J. Mol. Biol.* **1989**, *208*, 635-639.
39. van de Vroot, F. R.; Ma, C.-Y.; Nakai, S. *Archives Biochem. Biophys.* **1979**, *195*, 596-603.
40. Vreeman, H. J.; Brinkhuis, J. A.; Van Der Spek, C. A. *Biophys. Chem.* **1981**, *14*, 185-193.
41. Walstra, P. *J. Dairy Sci.* **1990**, *73*, 1965-1979.
42. Waugh, D. F.; Von Hippel, P. H. *J. Am. Chem. Soc.* **1956**, *78*, 4576-4582.
43. Waugh, D. F.; Creamer, L. K.; Slattery, C. W.; Dresdner, G. W. *Biochem.*, **1970**, *9*, 786-795.
44. Weiner, S. J.; Kollman, P. A.; Case, D. A.; Singh, U. C.; Ghio, C.; Alagona, G.; Profeta, S.; Weiner, P. *J. Am. Chem. Soc.*, **1984**, *106*, 765-784.
45. Weiner, S. J.; Kollman, P. A.; Nguyen, P. T.; Case, D. A. *J. Comput. Chem.* **1986**, *7*, 230-252.
46. Woychik, J. H.; Kalan, E. B.; Noelken, M. E. *Biochem.* **1966**, *5*, 2276-2288.

RECEIVED June 30, 1994

Chapter 23

Molecular Dynamics of Salt Interactions with Peptides, Fibrous Proteins, and Casein

Thomas F. Kumosinski and Joseph J. Unruh

Eastern Regional Research Center, Agricultural Research Service,
U.S. Department of Agriculture, 600 East Mermaid Lane,
Philadelphia, PA 19118

Controversy exists concerning the molecular basis for salt induced solubility (salting-in) of proteins. Specifically, the mechanism for the anionic or cationic interactions with either backbone or side chain groups on the protein is still speculative. We initiated molecular dynamics, MD, calculations to clarify this dilemma. MD results for oxytocin in CaCl_2 indicate anions H-bond to the peptide N-H backbone causing a conformational change (from turn and extended structure to loop); which is in agreement with the FTIR spectroscopy of oxytocin. MD calculations on predicted and energy minimized structures of a tropocollagen template molecule in aqueous CaCl_2 suggest that Ca^{2+} or Cl^- bind to the tropocollagen N-H backbone bonds, but unlike oxytocin stabilizes the structure. Finally, MD calculations on the N-terminal half of native and dephosphorylated α_{s1} -casein A showed ion binding and hydration energies in agreement with experimental data.

Biotechnology holds the promise of new product development through the use of new designer-type materials possessing tailor-made functionalities obtained via genetic engineering of proteins and/or creation of new protein functionality by controlling co-solutes (15). However, the problem of developing quantitative measures for protein structure-function relationships still remains unresolved. Without knowledge of these relationships, new material developments are of limited value due to the low probability of success.

One important functionality is the salt-induced resolubilization of proteins in solutions of mono and divalent salts at concentrations in excess of 0.5 M. Over the decades, many investigators (Steinhardt and Reynolds) (19) have performed experimental studies to provide a mechanistic basis for the salt-induced resolubilization of proteins. In a study on the influence of salt to the conformations of proteins, Von Hippel (23) defined salts according to their ability to become potential structure formers or structure breakers with respect

**This chapter not subject to U.S. copyright
Published 1994 American Chemical Society**

to thermal denaturation. Structure formers increase the temperature for protein thermal denaturation whereas structure breakers decrease this temperature. In these traditional studies, only the helix and disordered conformations were considered.

Timasheff's group (2,21), on the other hand, has studied the salt-induced resolubilization of proteins as a function of their ability to increase the preferential binding of salt over water in the vicinity of the protein surface. Remarkably, this classical thermodynamic study has provided a quantitative prediction of the resolubilization of protein at high salt concentrations for both mono and divalent salts. In addition, Robinson and Jenks (17), through their binding studies of a model peptide compound (acetyl tetraglycine ethyl ester) which contains no formal charge, speculate that salts at high concentrations bind to the high dipole moment regions of the peptide bonds.

In recent years, the emergence of molecular modeling as a technique for refining existing three dimensional molecular structures or building newly predicted structures has yielded a technology with the capability of studying the molecular basis for structure-function relationships (8,9). Not only proteins, but also salts and their interactions with proteins may be studied for their potential effect on structure-function relationships.

We now attempt to define and model this simple functionality relationship for oxytocin, tropocollagen and the hydrophilic domain of α_{s1} -casein A, using computer-generated three dimensional structures.

Using molecular modeling techniques (such as energy minimization and molecular dynamics, MD) the protein-salt-water interactions were simulated for each protein in the presence of enough CaCl_2 to mimic a greater than 1M solution. This process should establish the most probable protein-salt binding site, but can only mimic the total amount of salt bound to a protein. Whether this controls the salting-in or salting-out of a protein or has any effect on protein solubility can not yet be established.

Theory

Molecular Modeling. All complex structures employed monomer structures previously refined via energy minimization (10,11). They were constructed using a docking procedure on an Evans and Sutherland PS390 interactive computer graphics display driven by the Tripos Sybyl (St. Louis, MO) molecular modeling software on a Silicon Graphics 4200 Unix-based computer. The docking procedure allowed for individual manipulation of the orientation of up to four molecular entities relative to one another. The desired orientations could then be frozen in space and merged into one entity for further energy minimization calculations based on molecular force fields i.e. either the Kollman or Tripos generated force fields. The criterion for acceptance of reasonable structures was determined by a combination of experimentally

determined information and the calculation of the lowest energy for that structure.

Force Field Calculation. Studies concerned with the structures and/or energetics of molecules at the atomic level require a detailed knowledge of the potential energy surface (i.e., the potential energy as a function of the atomic coordinates). For systems with a small number of atoms, quantum mechanical methods may be used, but these methods become computationally intractable for larger systems (e.g., most systems of biological interest) because of the large number of atoms that must be considered. For these larger systems, molecular mechanics methods are used. Molecular mechanics is based on the assumption that the true potential energy surface can be approximated with an empirical potential surface consisting of simple analytical functions of the atomic coordinates. The empirical potential energy model treats the atoms as a collection of point masses that are coupled to one another through covalent (bonded) and noncovalent (nonbonded) interactions. The potential energy function (6,19) generally has the form:

$$\begin{aligned}
 E_{total} = & \sum_{bonds} K_r(r-r_{eq})^2 + \sum_{angles} K_\theta(\theta-\theta_{eq})^2 \\
 & + \sum_{dihedrals} \frac{1}{2}K[1 + \cos(n\phi - \gamma)] \\
 & + \sum_{i < j} \frac{B_{ij}}{R_{ij}^{12}} - \frac{A_{ij}}{R_{ij}^6} + \frac{q_i q_j}{\epsilon R_{ij}}
 \end{aligned} \tag{1}$$

The first three terms are due to covalent interactions and represent the difference in energy between the geometry of the actual structure and a geometry in which the bond lengths, bond angles, and dihedral angles all have ideal values. The remaining terms represent nonbonded van der Waals and electrostatic interactions. In Equation 1, r , θ , ϕ , and R_{ij} are variables, determined by the atomic coordinates. All other entities are constant parameters chosen to reproduce experimental observations as closely as possible. Although empirical potential energy functions such as Equation 1 are relatively crude, they have been applied successfully to the study of hydrocarbons, oligonucleotides, peptides and amino acids, as well as systems containing a large number of small molecules such as water. The Tripos force field in Tripos' Sybyl software package uses the above functional equation, and also includes a bump factor which allows atoms to approach each other within a fraction of the van der Waals radius to compensate for H-bonding if so chosen by the user. The parameters used for electrostatic calculations include atomic partial charges (q_i) calculated by the Kollman group (7,22) using a united atom approach with only essential hydrogens. All molecular structures were refined with an energy minimization procedure using a conjugate gradient algorithm, in

which the positions of the atoms are adjusted iteratively so as to achieve a minimum potential energy value. Energy minimization calculations were terminated when the energy difference between the current and previous iterations was less than 1 kcal/mol. A nonbonded cutoff of 5 Å (i.e. all non covalently bonded interactions were not calculated for distances greater than 5 Å) was used to save computer time, and is an appropriate value to use for a function that varies with distance. A stabilization energy of at least -10 kcal/mol/residue was achieved for all structures, which is consistent with values obtained for energy minimized structures determined by X-ray crystallography.

Molecular Dynamics. In the previous section we considered only static structures. However, the dynamic motion of molecules in solution contributes to their functionality. The molecular dynamics approach is a method of studying motion and molecular configuration as a function of time (t). All atoms in the molecule are assigned a kinetic energy through a velocity term which can be related to the local temperature as well as to the average temperature of the system. These calculations can be performed in vacuum or in the presence of a desired number of solvent molecules such as water. Environmental effects of constant temperature and volume (using a periodic boundary condition to confine the calculation within a prescribed volume), can also be handled in the calculations (18). For these calculations a force field describing the potential energy is combined with Newton's second law of motion

$$F_i = m_i a_i(t) = m_i \frac{dv_i(t)}{dt} = m_i \frac{d^2 x_i(t)}{dt^2} = -\nabla_i E \quad (2)$$

where F_i is the force on atom i which has mass (m_i), velocity (v_i), acceleration (a_i), and position (x_i). ∇_i is the gradient or the derivative with respect to position, t is the time displacement and E is the potential energy of the molecule described by the chosen force field. Equation (2) is integrated at various time intervals for the desired molecule using the chosen force field via a prescribed numerical integration method. The time interval chosen must be small in comparison with the period associated with highest frequency of motion within the molecule. This is usually stretching of a bond associated with a hydrogen atom, i.e. one femtosecond. Numerical integration of Equation (2) over one femtosecond intervals to 100 psec for a protein molecule of 2000 atoms or more necessitates a fast computer with a large memory capacity. The results of these calculations can mimic the motions of molecules in solution and also yield time dependent geometric parameters. For example the distance from the center of moment for a set of atoms may be related to correlation times derived from NMR, EPR and fluorescence experiments.

Avian Pancreatic Polypeptide

Three dimensional X-ray crystallographic and circular dichromatic (CD) data exist for avian pancreatic polypeptide (1PPT) (Brookhaven Protein Data Bank, 23). Therefore, 1PPT was chosen to ascertain whether the usual molecular force field and dynamics parameters need to be modified to mimic experimental physical-chemical solution studies. The three dimensional structure of 1PPT consists of 45 residues containing a large α -helix, two gentle turns and a proline helix motif. The two helices are anti-parallel with hydrophobic groups existing within their interface. The molecular force field model must match the X-ray structure. The amount of helix in the model agrees with the X-ray crystallographic structure. The CD results of Noelken et al. (16) indicate a large amount of helix conformation (i.e., over 80%) under both dimeric and monomeric conditions. This data agrees with the amount of helix calculated using the X-ray crystallographic structure. Thus, the molecular dynamics (MD) calculations in water must maintain the integrity of both helix conformations to mimic the experimental CD results.

Figure 1 shows the ribbon backbone structure at 50 psec (light ribbon) in the presence of water, with a dielectric constant of unity and a cutoff of 8Å for all non covalently bonded interactions, versus the three dimensional structure (dark ribbon). As can easily be seen, at 50 psec the proline helix structure starts to unwind from the C-terminal end. Such a destabilization in the proline helix is unacceptable when compared with the solution structural CD results. In addition, the time dependence of the radius of gyration was not constant. The use of the large non-interaction cutoff distance of 8 Å with a dielectric constant of unity assumes that this system maintains a uniform dielectric region. While this assumption is appropriate for water, it is well known that proteins do not possess a uniform dielectric constant throughout their structure (20). Thus, this approximation predicts a destabilized secondary structure as seen in Figure 1.

Figure 2 shows the same comparison at 50 psec but with a dielectric constant which varies with distance and a cutoff of 5Å for all non-bonded interactions. Here, all helix and turn structures are preserved (50 psec as light ribbon and original X-ray structure as dark ribbon). This holds true even when a Tripos force field (either with all the hydrogen atoms or with only the essential hydrogen bonding atoms) is used instead of the Kollman force field. In addition, the radius of gyration and root-mean-square fluctuation of all atoms becomes constant at 20-25 psec for all calculations. Also, using a larger cutoff value for the non-bonded interactions with additional water molecules did not change the structure of 1PPT. Hence, for the sake of computer time and speed, the lower cutoff value of 5Å was employed for all calculations. Additionally, the appropriate number of water molecules to maintain a density of unity and a periodic boundary was included in all calculations.

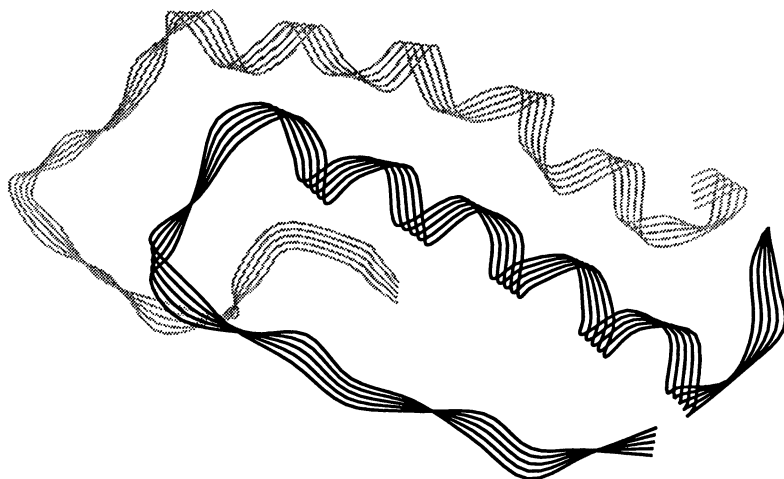


Figure 1. Ribboned backbone dynamic structure of avian pancreatic polypeptide. Crystallographic structure: dark ribbon. Molecular dynamic structure at 50 psec and 300 °K in water with dielectric constant of unity: light ribbon.

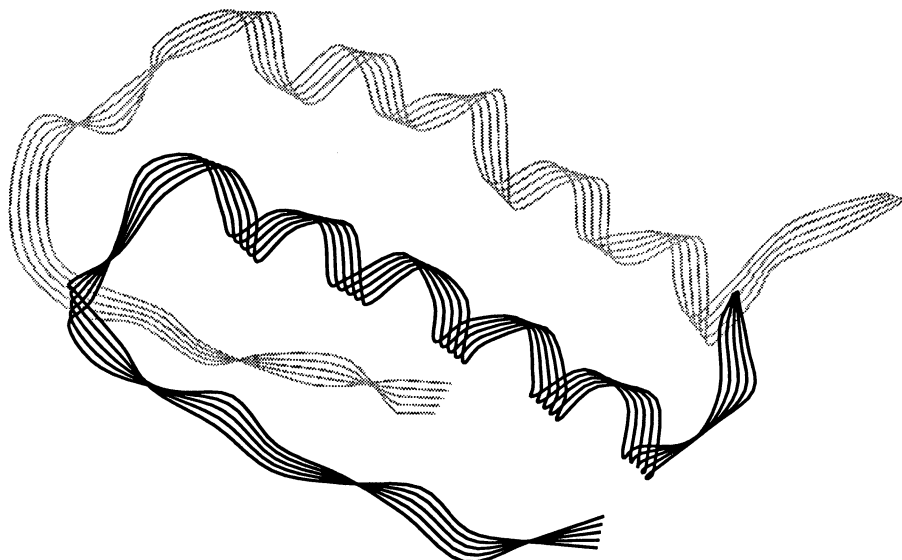


Figure 2. Ribboned backbone dynamic structure of avian pancreatic polypeptide. Crystallographic structure: dark ribbon. Molecular dynamic structure at 50 psec and 300 °K in water and utilizing a dielectric constant that varies with distance i.e. $D=R$: light ribbon.

Oxytocin

To test the molecular dynamics calculations for modeling the salt induced conformational change of a protein, we will use the X-ray crystallographic structure and results from the Fourier Transform Infra Red Spectroscopy of oxytocin. Oxytocin is a peptide of nine residues with one disulfide bond and two β -turns. It is available in pure form as an acetate salt. The biological role of oxytocin is to induce smooth muscle contraction and to initiate lactation in all mammals.

Fourier Transform Infra Red Spectroscopy (FTIR). Figure 3A shows the FTIR spectrum of oxytocin in water at 25°C. The amide II envelope, which ranges from 1520 to 1580 cm^{-1} results from N-H and N-C deformations of the backbone peptides groups. The amide I envelope which ranges from 1620 to 1700 cm^{-1} results from the carbonyl stretching of the backbone peptide groups. Both the amide I and II envelopes are sensitive to the conformational state of the peptide bonds in proteins. In this figure, three large absorbances at 1676, 1615 and 1556 cm^{-1} are observed. A smaller absorbance at 1510 cm^{-1} caused by the tyrosine side chain O-H deformation also appears. Analysis of the envelopes into component Gaussian bands was not considered necessary for the interpretation of these results. The 1676 cm^{-1} envelope has been assigned by Krim and Bandakar (8) as a turn conformation. The 1610 cm^{-1} envelope arises from the carbonyl group stretching of the acetate anion within the sample, while the 1555 cm^{-1} is assigned to the amide II peptide deformation for a turn conformation (8).

When 0.05M CaCl_2 is added, the amide II absorptivity is almost eliminated and only one absorbance at 1645 cm^{-1} in the amide I envelope is observed (Figure 3B). The 1645 cm^{-1} frequency band has been classified by Byler and Susi (3) as an irregular or disordered structure. Hence, when 0.05M CaCl_2 is added to oxytocin a conformation change is induced from a turn to an irregular or disordered non-hydrogen bonded conformation. In addition, since the amide II absorptivity is virtually eliminated, the conformational change must be caused by salt binding of at least the N-H backbone group. Such a conformational change caused by salt binding to N-H groups is in accord with the results of Robinson and Jencks (17) and should be possible to model using MD calculations.

Molecular Dynamics (MD) Simulations. To simulate the above salt-induced conformational changes in oxytocin we performed MD calculations until equilibrium was established (as measured by the time dependence of the geometric parameters of oxytocin) in the absence and presence of CaCl_2 atoms. Here the 3d structure from the Brookhaven Protein Data Bank was in actuality β -mercaptopyrroline oxytocin (wet form) which was modified to cysteine for compatibility with the experimental protein results. The modified structure was

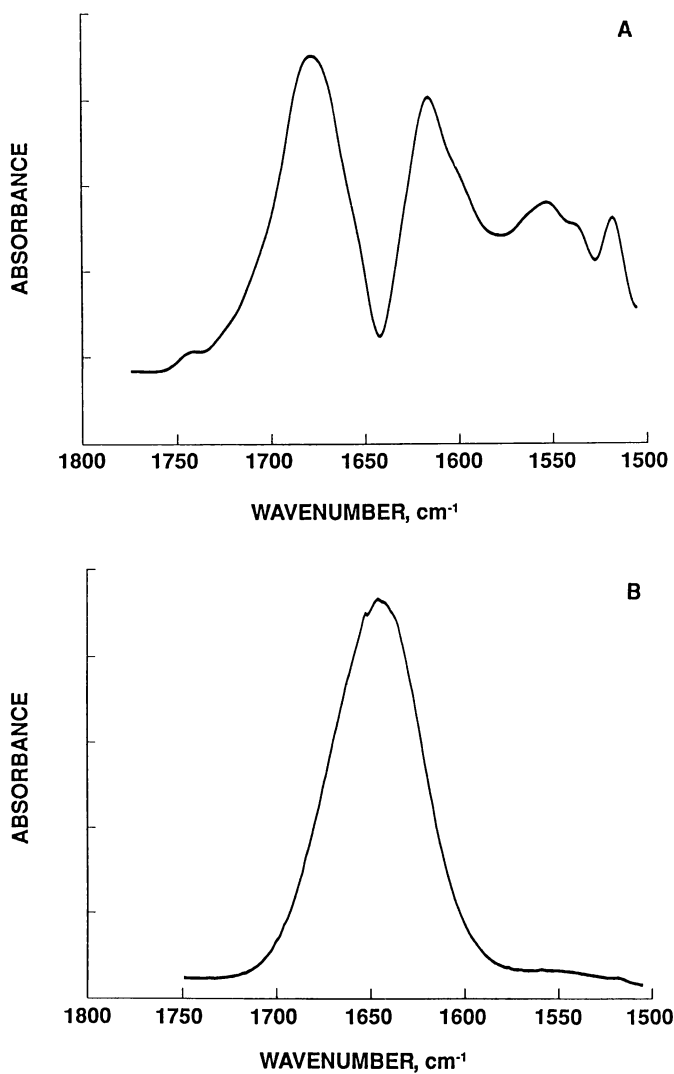


Figure 3. Fourier transform infra red spectra of oxytocin A: in water; and B: in 0.05M CaCl₂.

then energy minimized before performing MD calculations. It should be noted that no change in the two β -turns (of the original Brookhaven structure) resulted from this theoretical modification. In addition, in order to mimic salt binding to the backbone N-H group, we allowed the chlorine atoms to become hydrogen bond acceptors. The Tripos force field was used since it contains a parameter which allows a hydrogen atom to come within a specific fraction of the van der Waals radius of the chlorine atom. This fraction is called a bump factor. For these calculations, as well as all other MD calculations in this paper, a default value of 0.7 was maintained for the bump factor. The number of water molecules for all MD calculations was 400. When present in any MD calculations, CaCl_2 atoms were docked into the models as Ca and Cl atoms with their van der Waals radii and charges adjusted to mimic Ca^{2+} and Cl^- ions.

Figure 4A shows the variation of oxytocin with time during the MD calculation at 300 °K. As can be seen, the system become equilibrated within 70 psec by the invariance with time of the radius of gyration, r , and root-mean-square fluctuations, σ , of all atoms of oxytocin. However, throughout this simulation the tow turn conformation of oxytocin remains constant as exhibited by the constancy of the internal hydrogen bonds shown in Figure 4B. Figure 4B shows the dynamic structures of oxytocin at 70 psec in a wireframe structure with water (water not shown) and internal hydrogen bonds represented by dashed lines. Here, the backbone structure remains essentially the same as the backbone of the X-ray crystallographic structure. Only the orientation of the side chains have changed due to the MD calculations. Also, shown in Figure 4B is the ribbon backbone for the oxytocin molecule.

Next, 11 Ca^{2+} and 22 Cl^- ions were docked into the equilibrated oxytocin-water system of Figure 4B and energy minimized. The resulting oxytocin with a Cl^- Ca^{2+} and Cl^- distribution is shown in Figure 4C. The oxytocin is represented by a ribboned backbone and the Ca^{2+} and Cl^- ions by balls with radii 0.15 times that of their van der Waals radii. The smaller radii simplified the cumbersome nature of the system. It also should be noted that the Ca^{2+} and Cl^- ions display a different distribution than seen in Figure 4D after MD calculations. The latter distribution is closer to the ribbon oxytocin molecule and generally more compact.

In Figure 4D, 11 Ca^{2+} and 22 Cl^- ions were docked into the oxytocin-water system with a periodic boundary, energy minimized and subjected to MD calculations at 300 °K for 70 psec where equilibrium was easily established. The results show that the chloride ions have hydrogen bonded to the NH backbone atoms as well as the OH group of the tyrosine side chain. Since only a bump factor of 0.7 was utilized, electrostatics appear to be the interaction responsible for this phenomenon. The chloride ion also binds through electrostatics to the Ca^{2+} ion which in turn binds to another Cl^- ion until a network of salt bridges is created with the end Cl^- hydrogen bonded to the backbone N-H groups. In fact all N-H groups are hydrogen bonded to a salt bridge network which destroys both turn conformations and creates a loop-type or irregular structure. This change in conformation is illustrated in Figure 4E,

where the turn conformation is shown as a backbone ribboned structure and the loop conformation is a wireframe representation with Cl^{-1} bonded to N-H groups as dashed lines. It is apparent that the inertia of the CaCl_2 salt-bridge network changed the $-\phi$, ψ angles of the oxytocin to create a more open and disordered structure. It is also surprising that the water molecules did not destroy the salt-bridge network since the Cl^{-1} ions could have hydrogen bonded to the water molecules. The reason for this dilemma may arise from the difference in inertia between a water molecule, which has fast motion about a substantial dipole movement, and the Cl^{-1} ion which is spherically symmetric thus preferring to be associated with a group possessing a lower kinetic energy. It also should be noted that MD calculations of the oxytocin and CaCl_2 system when water was eliminated yielded the same results.

The above represents experimental and theoretical studies which remarkably agree with one another. To date, the authors are now aware of any other study that correlates MD calculations with FTIR results. However, for the remainder of this paper only MD calculations will be utilized, since no FTIR observations of CaCl_2 interaction with fibrous proteins or the hydrophilic half of α_{s1} -casein A exist.

Tropocollagen

For a fibrous protein model we use the tropocollagen structure. Since no X-ray crystallographic structure for tropocollagen exists in the Brookhaven Protein Data Banks, it was necessary to construct such a structure from ϕ , ψ angles reported in the literature. Here, we utilized the reported ϕ , ψ angles of Miller and Scheraga (14), who also constructed a lowest energy tropocollagen with closest packing. A template repetitive sequence of gly-ala-Hpro for a total of 180 residues per strand was constructed using appropriate ϕ , ψ angles. Three such strands were docked together in a closest packing manner with the appropriate translation adjustments to create a super helix motif with a pitch of 32 to 33 residues. This structure was energy minimized and the resulting structure containing the super helix is shown in Color Plate 21. In Color Plate 21 the three chains are colored magenta, green and red-orange respectively, and the atoms are represented by spacefilling for easy perception of the super helix motif. While a gly-ala-Hpro template is utilized in this tropocollagen structure a gly-pro-Hpro template could also be so constructed and would yield the same super-helix structure.

Figure 5A shows the intermolecular hydrogen bonding network as dashed lines holding the three individual strands in a stable tropocollagen structure. It is apparent that any disruption of these intermolecular hydrogen bonds between backbone C-O and NH groups would cause dissociation of the three stranded super helix structure.

MD calculations were performed on a smaller portion of the tropocollagen structure (i.e. 45 residues per helical strand for a total of 135 residues) to

NOTE: The color plates can be found in a color section in the center of this volume.

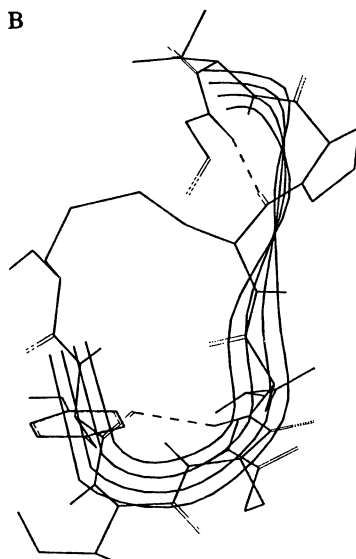
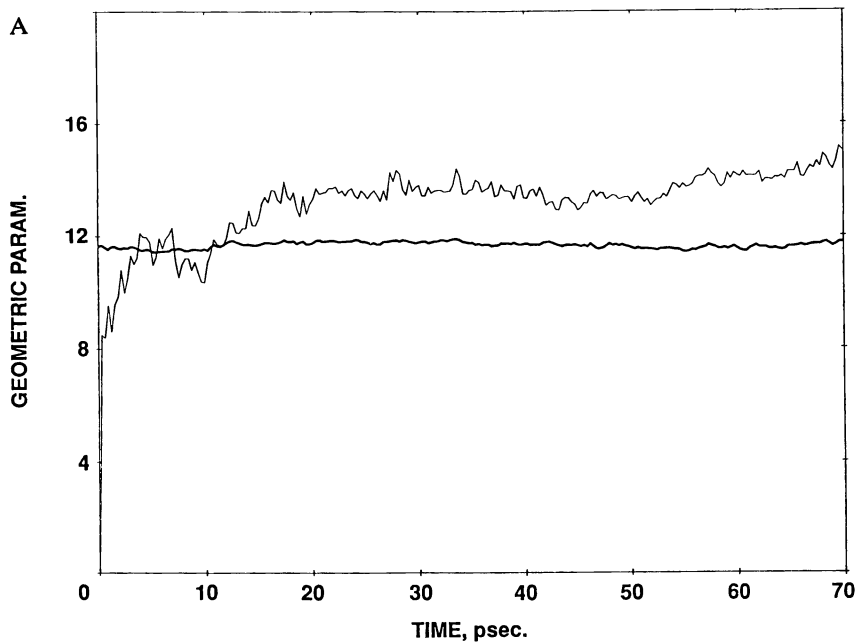


Figure 4. Molecular dynamics of oxytocin at 300 °K. A: Time dependency of geometric parameters of oxytocin: solid line, radius of gyration of all oxytocin atoms, r ; double line: root-mean-square fluctuations of all atoms, a , of oxytocin. B: Wireframe structure with dashed lines as internal hydrogen bonds and ribboned backbone of oxytocin in water at 70 psec.

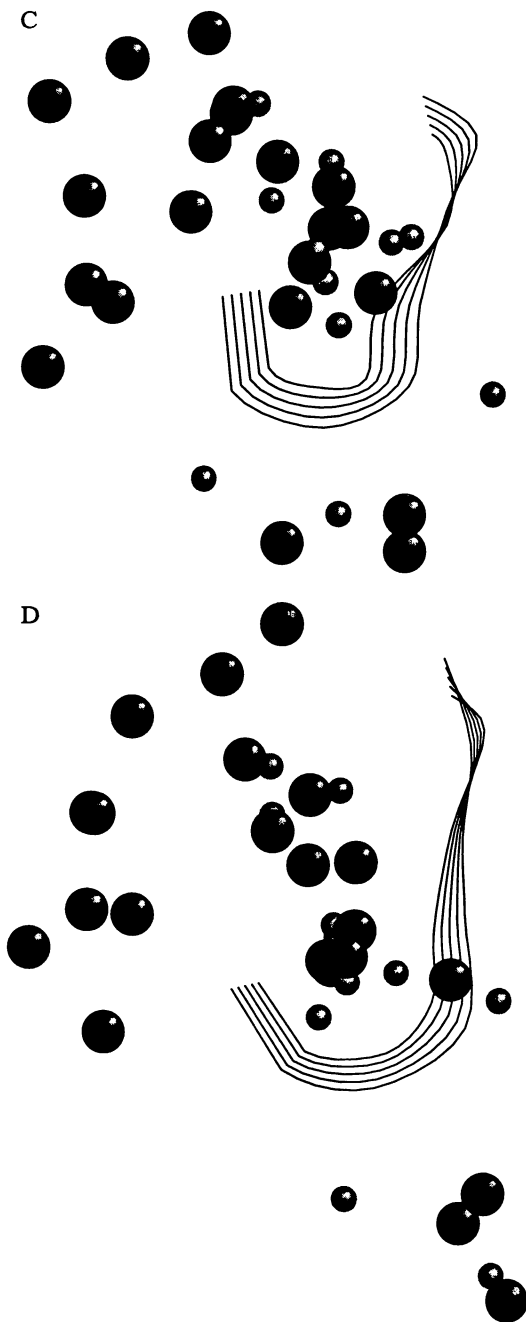


Figure 4. Continued. C: Ball pattern of fluctuations of CaCl_2 atoms before molecular dynamics at 0 psec (Ball at 0.15 van der Waals radius). Oxytocin exhibited as a ribboned backbone structure. D: Same as C but after dynamics at 70 psec. *Continued on next page.*

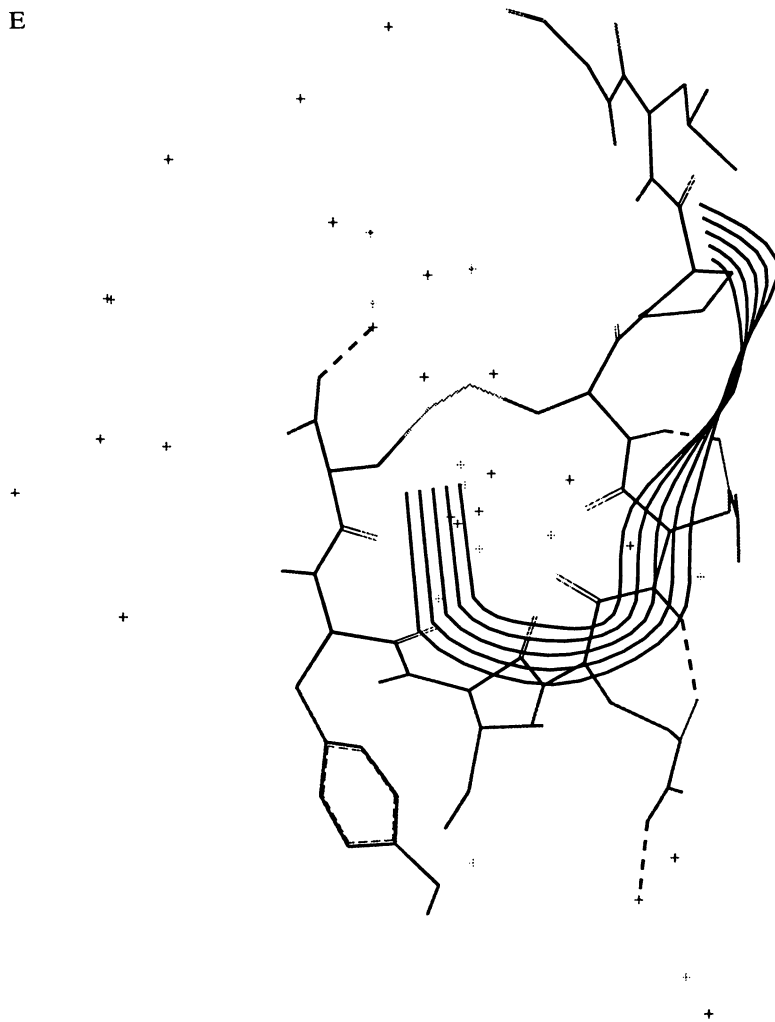


Figure 4. Continued. E: Ribboned backbone of oxytocin at 0 psec wireframe structure with dotted H bond lines at 70 psec.

decrease the time length of the calculation. Four hundred water molecules were added to insure a density of one, along with a non-bonded cutoff of 5Å and a period boundary condition. This small piece would insure at least one turn of the super-helix of the tropocollagen structure. After 50 psec at 300° K, equilibrium was well established and no disruption in the tropocollagen structure (ribboned structure of Figure 5B) was observed. Thus, this small portion of the tropocollagen structure is dynamically stable at 300 °K and 50 psec. Figure 5B also shows the added Ca^{2+} and Cl^{-1} ions after docking into equilibrated tropocollagen and water system.

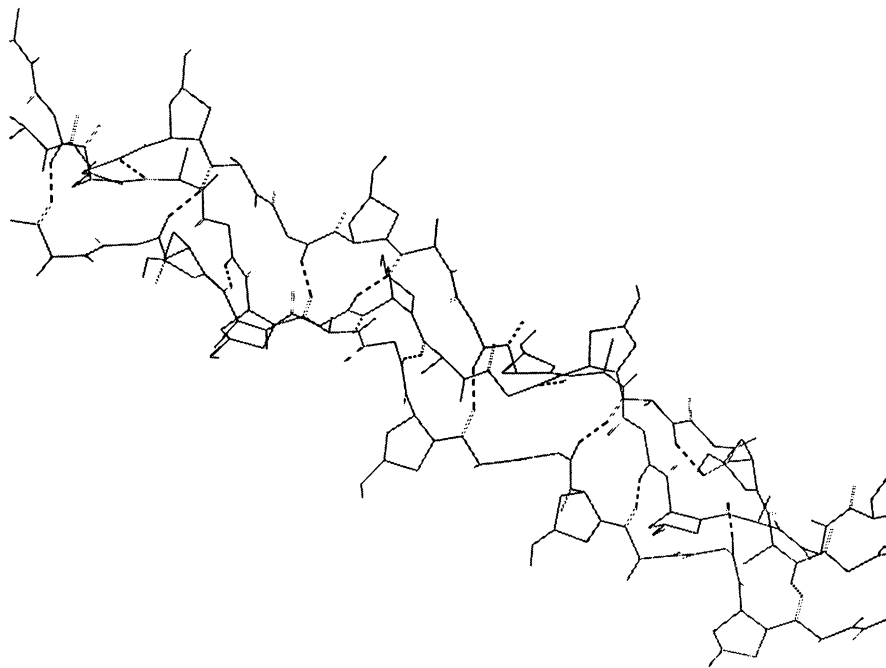


Figure 5. Three-dimensional structure of Tropocollagen. A: Wireframe structure showing dashed lines as hydrogen bonding. *Continued on next page.*

Addition of CaCl_2 to mimic a 3M solution (i.e. 22 Ca^{2+} and 44 Cl^{-1} atoms) was then achieved using the docking and merge procedure in Sybyl. The system was energy minimized and submitted to MD calculations at 300 °K for 50 psec where equilibrium was well established. Figure 5C shows a ribboned backbone of tropocollagen with the Ca^{2+} and Cl^{-1} represented as balls (as in Figure 5B) after 50 psec with added Ca^{2+} and Cl^{-1} ions. Here, it is seen that no disruption in the tropocollagen's structure occurs throughout any of the dynamic calculations. The Cl^{-1} of a stable salt-bridge network of CaCl_2 hydrogen bond to the solvent-exposed backbone N-H groups. However, unlike in the oxytocin case, the CaCl_2 salt bridge network tends to stabilize the tropocollagen structure and may even cause a less dynamically flexible structure. It also should be noted that the distribution of the Ca^{2+} and Cl^{-1} are more compact and closer to the ribboned tropocollagen structure after MD calculations for 50 psec (see Figure 5C) than prior to MD calculations (see Figure 5B). The salt-bridge network can easily be seen in Figure 5C where the Ca^{2+} and Cl^{-1} ions are represented as dotted van der Waals surfaces and tropocollagen as a ribboned backbone structure as in Figure 5B.

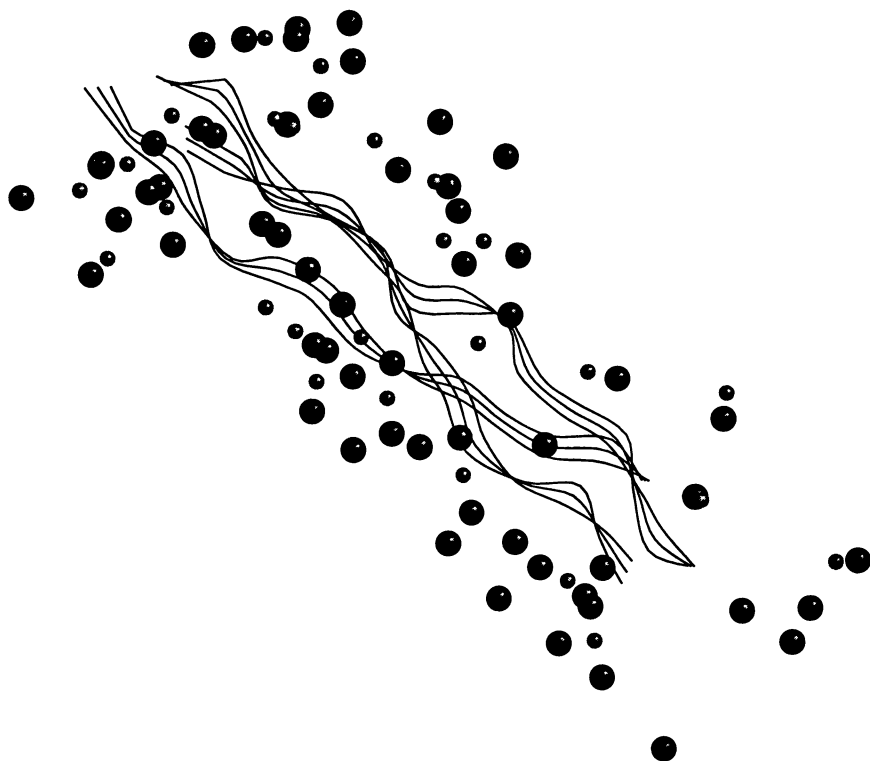


Figure 5. Continued. Three-dimensional structure of Tropocollagen. B: Dynamic structure at 300 °K and 0 psec for a 135 residues tropocollagen ribboned backbone structure in water with added CaCl_2 representing balls and no water shown.

It should be stressed that, although MD calculations show a large salt-bridge network in all structures, the calculation does not take into account the translational diffusion of the salt ions. In reality such a large network most probably would not exist due to the translational diffusion of the salt.

Hydrophilic domain of α_{s1} -casein A

Molecular Dynamics of CaCl_2 and MgCl_2 in Water. In this section we shall attempt to form a structural basis for the thermodynamics of the salting-in process which was characterized using thermodynamic linkage and non-linear regression analysis of the calcium-induced solubility profiles of α_{s1} -casein (6,13,13). We have chosen the hydrophilic N-terminal domain of the monomeric α_{s1} -casein A structure to model this solubility profile. This model was built from the α_{s1} -casein B model (12,13) by excising residues 14 through 26 and then deleting residues 100 to 199 of α_{s1} -B. The resulting hydrophilic domain, i.e., residues 1 through 99 of the α_{s1} -casein B minus 14 to 26 (a total of 86 residues), was then energy minimized. Eleven hundred water molecules

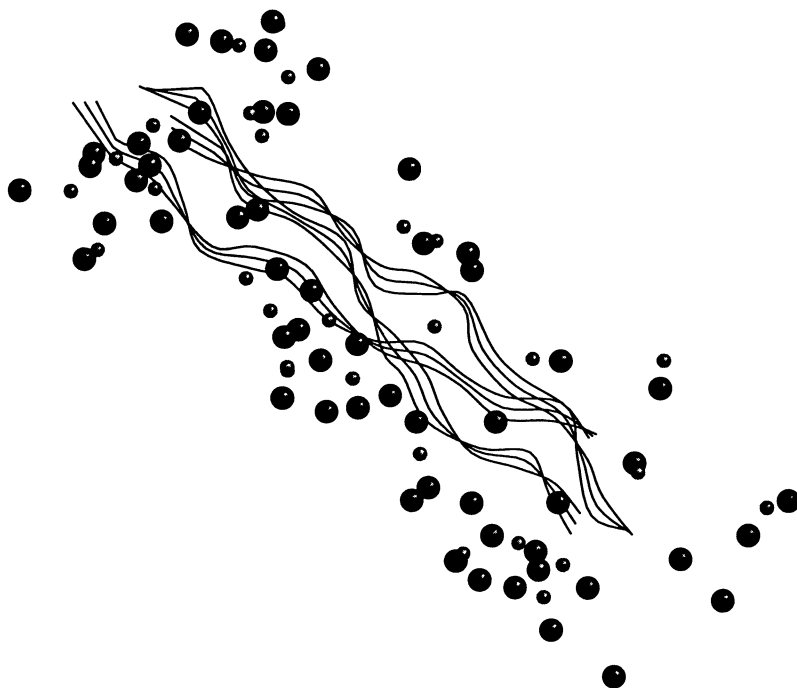


Figure 5. Continued. Three-dimensional structure of Tropocollagen. C: Dynamic structure of B at 300 ° K and 50 psec.

were added and the resulting structure in solution was again energy minimized with a cutoff for all non-bonded interactions of 5 Å, while maintaining a periodic boundary condition. The resulting structure with water was subjected to molecular dynamics (MD) calculation for 20 psec above equilibrium conditions which were determined by the stabilization of potential energy, radius of gyration of the protein backbone, root-mean-square fluctuations of the backbone atoms, and change in second moment. Such an equilibrated dynamic structure should approximate the structure, energetics and dynamic motion of this protein domain in solution.

To mimic the salt binding mechanism, 22 molecules of calcium or magnesium and 44 molecules of chloride with appropriate ionic charges were added to the above system in a pseudo random fashion. The system was energy minimized and subject to MD for a full 40 psec. Equilibrium was easily established once again at 15 to 20 psec. The resulting structure for the native half of α_{s1} -A in CaCl_2 is shown in Figure 6 A and B. The amount of salt added was chosen to comply with a condition which would result in saturation of the Ca^{2+} binding sites. Using a literature value of 380 M^{-1} for K_A (6,9) with 8 binding sites, 22 molecules of CaCl_2 per 1100 water molecules per molecule of protein is equivalent to greater than 99% occupancy of these calcium binding sites and greater than 80% occupancy with 8 additional putative salting-in sites derived using k_2 from Farrell, et al. (6). A similar structure for the dephosphorylated half of α_{s1} -A was also studied and is shown in Figure 7. In total, seven MD calculations were performed on the hydrophilic domain (H) of α_{s1} -A (residues 1-99) in the presence of 1100 water molecules to 40 psec: two in the absence of salt for native (H) and dephosphorylated (HO-P); two in the presence of CaCl_2 for H and HO-P; one with added MgCl_2 for H; and one each for MgCl_2 and CaCl_2 with no protein. Each calculation utilized a cutoff of 5 Å for non-bonded interactions, a Tripos force field, and a "bump" factor of 0.7 for simulating hydrogen bond formation. MD calculations require a running time on the Silicon Graphics Unix computer system of at least 3 days. It should be noted at this time that the energetics and geometric parameters estimated by the MD calculation reflect the total salt binding to the protein, i.e., the sum of the free energies of salt binding for the protein salting-out, as well as the salting-in process. More detailed analysis of these MD calculations, which are beyond the scope of this paper, must be performed for separation of the protein precipitation and resolubilization processes.

The average results of several calculated geometric parameters with their corresponding errors are presented in Table 1. The subscript 1 denotes the salt component i.e. both Ca^{2+} and Cl^{-1} , while the subscript 2 describes the protein. R is the calculated dynamic radius of gyration, x is the average center of mass of the component atoms and a is the root-mean-square fluctuation of component atoms from the center of mass. Here, a can be thought of as the dynamic Stokes radius of the chosen ions, and x is the spherical center of mass.

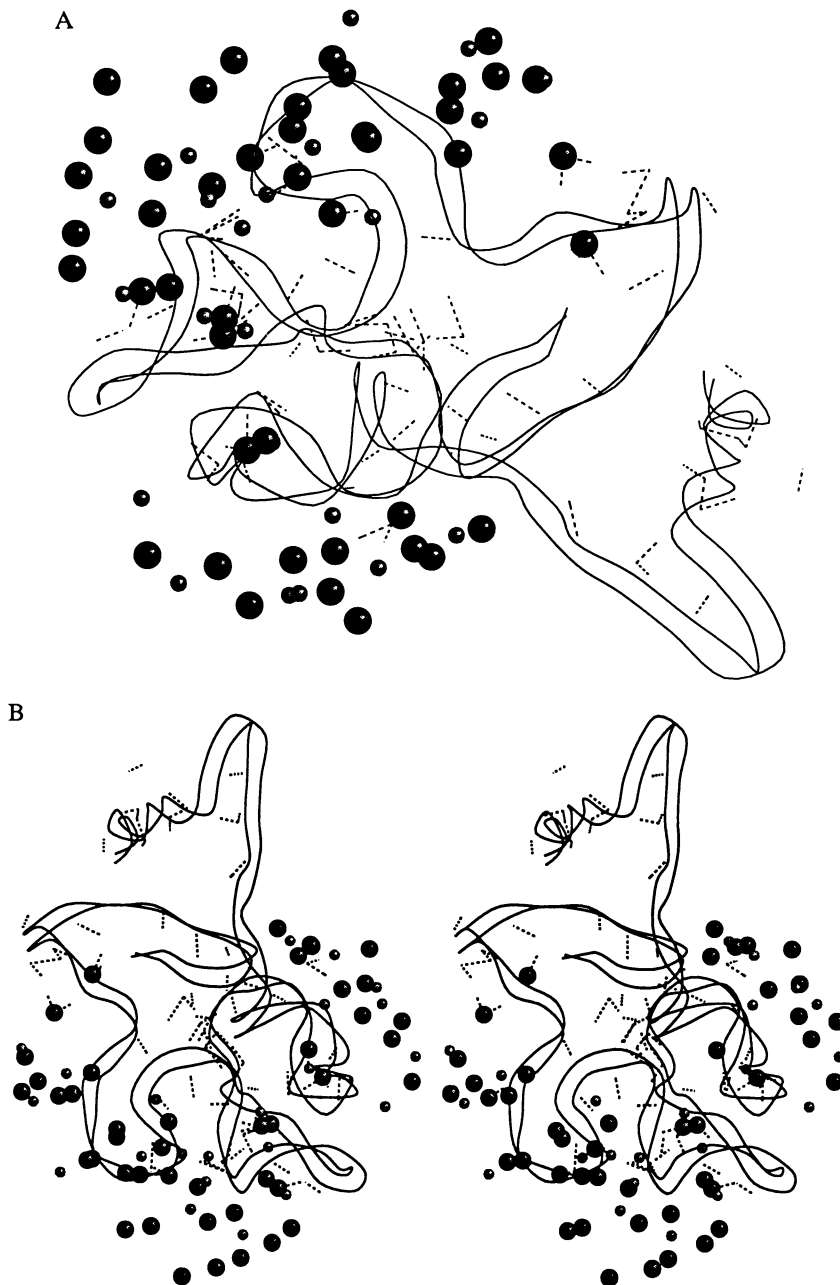


Figure 6. A: Backbone ribbon structure of the hydrophilic half of α_{s1} -casein A (residue 1 through 86) after molecular dynamics at 40 psec with 22 molecules of CaCl_2 in the presence of 1100 water molecules. Ca and Cl atoms are shown as ball models of radii equal to 0.15 times their known van der Waals radius. Dashed lines represent hydrogen bond formation. B: Stereo view (relaxed view) of A.

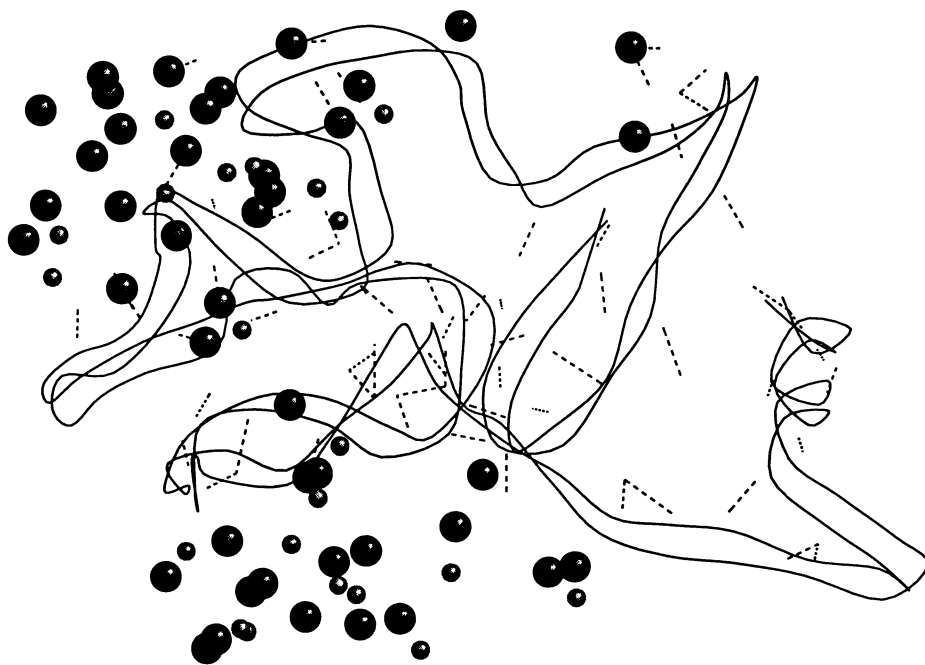


Figure 7. Same as in Figure 6A but for dephosphorylated α_{sI} -A. All serine phosphates mutated to serine with appropriate partial charges.

Table I. Molecular Dynamics of Hydrophilic Half (H) of α_{s1} -casein A in Water with Salt: Geometric Parameters

Protein	Salt	$a_1, \text{\AA}^2$	$x_1, \text{\AA}$	$R_2, \text{\AA}$	$a_2, \text{\AA}^2$	$x_2, \text{\AA}$
H	—	—	—	14.6 ± 0.2	8.2 ± 0.5	0.33 ± 0.04
HO-P	—	—	—	14.2 ± 0.04	7.0 ± 0.2	0.33 ± 0.03
H	CaCl ₂	11.5 ± 0.7	0.8 ± 0.10	14.6 ± 0.2	8.7 ± 0.4	0.38 ± 0.04
HO-P	CaCl ₂	9.38 ± 0.67	0.69 ± 0.06	15.5 ± 0.2	10.1 ± 0.4	0.51 ± 0.03
H	MgCl ₂	7.8 ± 0.6	0.67 ± 0.07	14.6 ± 0.2	7.7 ± 0.5	0.37 ± 0.04
—	CaCl ₂	9.1 ± 0.9	2.2 ± 0.1	—	—	—
—	MgCl ₂	6.8 ± 1.1	1.5 ± 0.4	—	—	—

R is radius of gyration

a is the RMS fluctuation of all atoms from center of mass; a dynamic Stokes radius.

x is the calculated spherical center of mass.

subscript 1 denotes salt while 2 denotes protein atoms.

Using Table 1 we can attempt to describe the distribution of the salt ions at the end of the MD calculations. To do so, we inspect the a and x values in Table 1 for both CaCl_2 and MgCl_2 alone and in the presence of both the hydrophilic half of α_{s1} -A (H) and its dephosphorylated form HO-P. No significant changes are seen in the a_1 values, since the mass of the proteins is far greater than that of the salts. The x_1 values are the best descriptors of the binding of ions to the proteins in these MD calculations. In all cases the x_1 values have decreased in the presence of either protein component. Such a decrease in the average spherical center of mass of the ions is a clear indication of protein-salt interactions, since the center of mass for salt ions alone would be larger as they move randomly about. However, x_1 would be smaller for salt bound to protein where movement is restricted. A difference between x_1 for the native H and the O-P form exists. It can easily be observed (Figures 6 and 7) that the salt ions associated with the H and HO-P forms have a different respective overall distribution. This difference needs to be verified but it is in agreement with the decreased x_1 in the presence of the protein fragments. It is assumed that the x_1 of the MgCl_2 was smaller due to the smaller value of the van der Waals radius for Mg^{2+} (0.66 Å) vs. Ca^{2+} (0.99 Å).

To observe the effect of salt binding on the dynamic structure of the hydrophilic half of the protein, we have calculated for the protein (in the presence or absence of salt ions) its radius of gyration R_2 , the a_2 and the x_2 values. These values are presented as columns 5, 6 and 7 of Table 1. Virtually no changes within the calculated error are observed for the R_2 , a_2 or x_2 values for the native H form either in the absence or presence of CaCl_2 or MgCl_2 . The HO-P form is not dramatically different either. However large increases in these descriptors are observed in the dephosphorylated HO-P form when CaCl_2 is added in the MD calculations. This change may reflect a general expansion of the HO-P structure when CaCl_2 is added which can also be observed by inspection of Figure 8A. Here the two structures are compared by representations of the protein backbones, with no side chains displayed. The H form is represented by a ribbon trace of the backbone, while the HO-P form is represented by a backbone-atomic stick model. The backbone model is much more expanded than the ribbon model. For easier observation of this conclusion a stereo view is shown in Figure 8B. The reason for this phenomenon is most likely due to the hydrogen bonding of chloride ions to the serine side chains as well as to N-H atoms of the backbone. Such anion-protein hydrogen bonding can impose important dynamic structural changes on the protein component. In the H form these interactions may be "screened-out" by the negatively charged phosphate groups. Whether this interaction causes increased solubility of the protein (i.e. salting-in) cannot be established at this time, because the potential role of the deleted residues is not considered. More MD calculations in conjunction with physical chemical experiments (e.g. FTIR of proteins in solutions) must be performed in the future to test this hypothesis.

To further correlate the binding free energies calculated from the literature for thermodynamic linkage analysis of the salting-out and salting-in

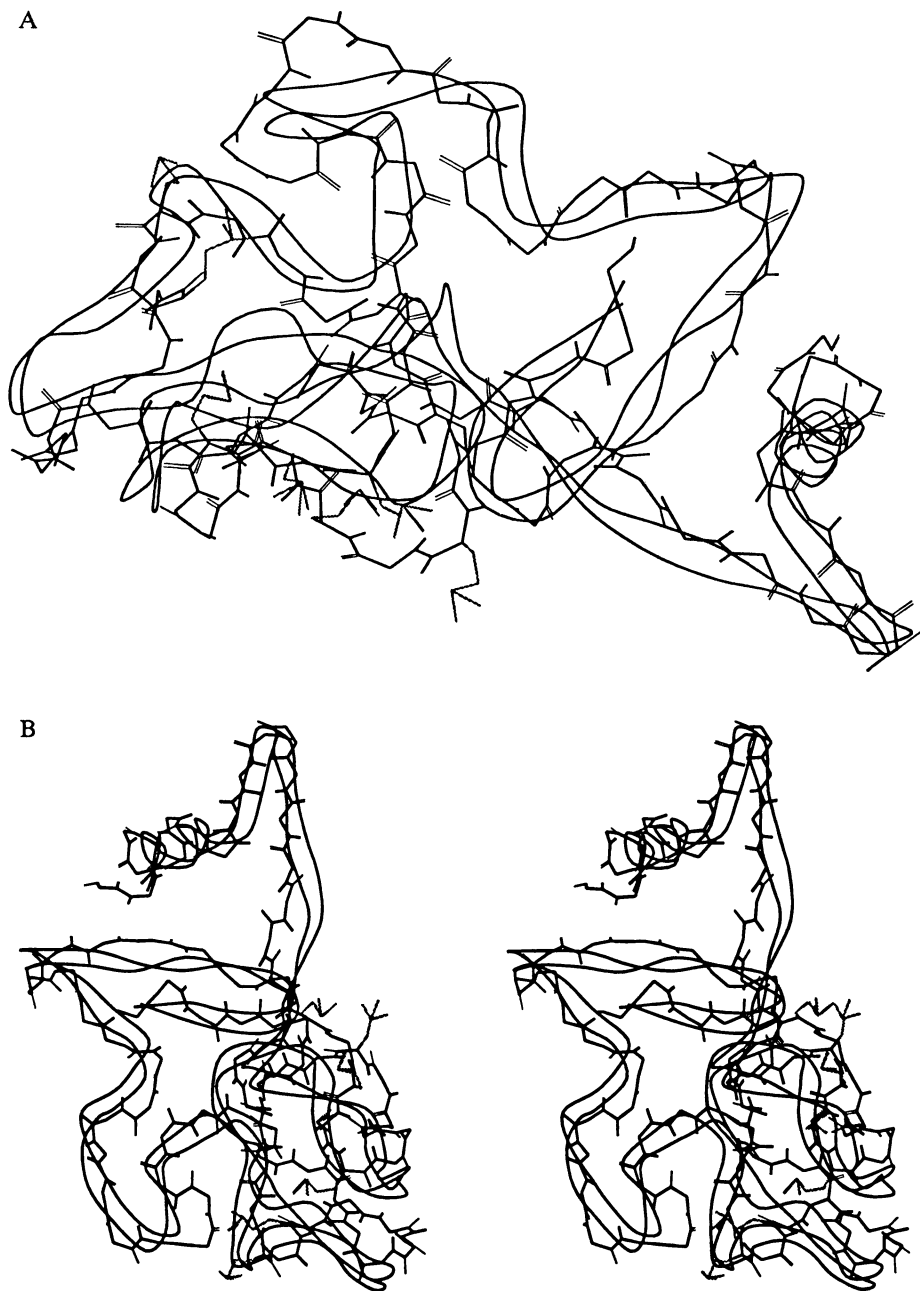


Figure 8. A. Comparison of hydrophilic domain of α_{s1} -A (shown with ribboned backbone) with dephosphorylated α_{s1} -A after molecular dynamics with 22 molecules of Ca and Cl₂ at 40 psec in the presence of 1100 water molecules dashed lines represent hydrogen bond formations. B. Stereo view (relaxed) of A.

experiments, ΔF_B , (6), the energetics of the seven MD calculations for the modeled H and HO-P were calculated. These values are presented in Table 2 along with the experimental total binding free energies, ΔF_B calculated from Farrell et al. (6), for native α_{s1} -A in CaCl_2 and MgCl_2 as well as for dephosphorylated α_{s1} -A in CaCl_2 . In this table, the calculated descriptors of energy are: E_T , for the average total potential energy of the system; E_W , for the water; E_P , for the energy of the protein; E_{PW} , protein-water interaction energy; E_{SW} , for the salt-water interaction energy; and E_{INT} , for the potential energy for the salt-protein interaction. It should be stressed that the energies estimated from MD calculations are the internal energy at constant volume and not the Gibbs or the Helmholtz free energy derived from binding experiments. However using the approximation that the potential interaction energy is equal to the ΔF_B , a qualitative correlation between these two parameters can be utilized to describe variations in protein components or type of salt used. Inspection of columns 4 and 5 of Table 2 shows that for native α_{s1} -A and HO-P in CaCl_2 and native A in MgCl_2 , the experimental changes of ΔF_B correlate well with changes in ΔE_{INT} when these environmental factors are varied; the absolute values for the two parameters are, however, significantly different. The large differences in absolute values between ΔF_B and ΔE_{INT} are due to the bump factor of 0.7 used in these calculations to quantitate hydrogen bonding in MD. If this factor were optimized then perhaps better correlations could be achieved. However, for this study, the value of 0.7 was used since it is the default value used by Sybyl for the Tripos force field. Such a small bump factor most likely allows for inordinately high electrostatic energy terms. Future work will tend to optimize this parameter. No conclusions can be made at this time concerning the meaning of the other energetic descriptors of Table 2 but they are presented along with their error for inspection by the reader.

Summary Correlations and Conclusions

From all the above, it can be concluded that molecular modeling techniques such as energy minimization and molecular dynamics can provide a powerful multifaceted approach for defining structure-function relationships such as the salt-protein interactions. Here, the dynamic changes in the protein structure responsible for the salt binding processes can be established using energy minimization and molecular dynamics calculations of the protein in water, in the presence and absence of salt ions. In particular, predictions concerning the type of protein primary structure groups responsible for binding and conformational change which occurs can be utilized to increase the desired protein functionality in a rational way through chemical or genetic modification. (15).

The molecular dynamics calculations could not at this present time distinguish between the salting-out and salting-in binding free energy. Only the total binding could be correlated in the MD results. In fact, significant

Table II. Molecular Dynamics of Hydrophilic Half (H) of α_{s1} -casein A in Water with Salt: Energetics and Comparison with Experimental Data

Protein	Salt	$-E_T$	$-E_{int}$	ΔF_B	$-E_{PW}$	$-E_W$	$-E_{SW}$
H	—	20,000 ± 152	—	—	20,000 ± 152	13,000 ± 64	—
HO-P	—	19,000 ± 54	—	—	19,000 ± 54	13,000 ± 101	—
H	CaCl ₂	31,000 ± 134	8,746 ± 162	33.2	13,000 ± 154	11,000 ± 142	20,000 ± 128
HO-P	CaCl ₂	28,000 ± 132	4,612 ± 198	14.3	14,000 ± 26	11,000 ± 91	22,000 ± 155
H	MgCl ₂	34,000 ± 92	11,000 ± 175	36.6	12,000 ± 133	11,000 ± 142	22,000 ± 155
—	CaCl ₂	27,000 ± 266	—	—	—	13,000 ± 122	27,000 ± 266
—	MgCl ₂	29,000 ± 198	—	—	—	12,000 ± 198	29,000 ± 198

E denotes potential energy in kcal/mole i.e. internal energy at constant volume subscripts:

T denotes total atoms

P denotes protein atoms

W denotes water atoms

S denotes salt atoms and int denotes protein-salt interaction.

ΔF_B denotes the total co-operative salt binding free energy in kcal/mole calculated from Tables 5 and 6 in Ref. 6), for the solubility data for α_{s1} -casein A (NA) and its dephosphorylated form (NAO-P).

differences in absolute values occur when the total salt binding free energy calculated from thermodynamic linkage, ΔF_B of the hydrophilic half of α_{s1} -casein A, is compared with the interaction energy derived from MD, ΔE_{INT} (see Table 2). This discrepancy is most likely due to the bump factor value chosen for the MD calculations; a lower value should be chosen for future work. However, the changes in ΔF_B with the change in salt from Ca to Mg for the native systems as well as the change with calcium between native and the dephosphorylated form show good correlations with ΔE_{INT} . More importantly, the geometric parameters for the proton from MD calculations (i.e. R_2 , a_2 , and x_2) of the HO-P change dramatically over the H form in the presence of $CaCl_2$. This phenomenon was interpreted as an expansion of the dephosphorylated form with added $CaCl_2$ which is shown in Figure 8A and 8B. It is apparent that this structural change can also be labeled as a conformational change since many internal hydrogen bonds are disrupted. In the future, this phenomenon will be tested experimentally by FTIR.

Literature Cited

1. Andersen, H. C. *J. Chem. Phys.* **1980**, *72*:2384-2394.
2. Arakawa, T.; Timasheff, S. N. *Biochem.* **1984**, *23*:5912-5923.
3. Byler, D. M. and Susi, H. *Biopolymers.* **1986**, *25*:469-487.
4. Eigel, W. N.; Butler, J. E.; Ernstrom, C. A.; Farrell, H. M., Jr.; Harwalkar, V. R.; Jenness, R.; Whitney, R. McL. *J. Dairy Sci.* **1984**, *67*:1599-1631.
5. Farrell, H. H., Jr.; Thompson, M. P. The caseins of milk as calcium binding proteins, pp. 117-137. In "Calcium Binding Proteins," M. P. Thompson (Ed). CRC Press Boca Ratan, FL, 1987.
6. Farrell, H. M., Jr.; Kumosinski, T. F.; Pulaski, P.; Thompson, M. P. *Archives Biochem. Biophys.* **1988**, *265*:146-158.
7. Kollman, P. A. *Ann. Review Phys. Chem.* **1987**, *38*:303-333.
8. Krimm, S. and Bandekar, J. *Adv. Protein Chem.* **1986**, *38*:181-364.
9. Kumosinski, T. F.; Farrell, H. M., Jr. *J. Protein. Chem.* **1991**, *10*:3-16.
10. Kumosinski, T. F.; Brown, E. M.; Farrell, H. M., Jr. *Trends in Food Sci. Technol.* **1991a**, *2*:110-115.
11. Kumosinski, T. F.; Brown, E. M.; Farrell, H. M., Jr. *Trends in Food Sci. Technol.* **1991b**, *2*:190-195.
12. Kumosinski, T. F.; Brown, E. M.; Farrell, H. M., Jr. *J. Dairy Sci.* **1993a**, *76*:931-945.
13. Kumosinski, T. F.; Brown, E. M.; Farrell, H. M., Jr. *J. Dairy Sci.* **1993b**, *76*:2507-2520.
14. Miller, M. H. and Scheraga, H. A. *J. Polymer Sci. Symp.* **1976**, *54*:171-200.
15. Noble, R. W.; Waugh, D. F. *J. Amer. Chem. Soc.* **1965**, *87*:2236-2257.

16. Noelken, M. E.; Change, P. J. and Kimmel, J. R. *Biochemistry*, **1980**, 19:1838-1843.
17. Robinson, D. R.; Jencks, W. P. *JACS*, **1965**, 87:2470-2479.
18. Sillen, L. G.; Martell, A. E. *Stability Constants of Metal Ion Complexes*, Special Publication No. 25 of The Chemical Society of London, Alden Press, Oxford, U.K., **1971**.
19. Steinhardt, J. and Reynolds, J. A. In "Multiple Equilibria in Proteins", p. 325. Academic Press, New York, **1969**.
20. Tanford, C. *Physical Chemistry of Macromolecules*. John Wiley & Sons, New York, **1967**.
21. Timasheff, S. N. and Arakawa, T. In "Protein Structure and Function: A Practical Approach" (Creighton, T. E., Ed.) pp 331-345. IRL Press, Oxford, **1988**.
22. Weiner, S. J.; Kollman, P. A.; Nguyen, D. T.; Cose, D. A. *J. Comput. Chem.* **1986**, 7:230-252.
23. Von Hippel, P. H.; Schleich, T. In *Structure and Stability of Biological Macromolecules*. S. N. Timasheff and G. D. Fasman, eds. Marcel Dekker, P. 417, New York, NY, **1969**.

RECEIVED July 7, 1994

Chapter 24

Probabilistic Approach to NMR-Based Determination of Accurate Local Conformation and Three-Dimensional Structure of Proteins in Solution

Leela Kar, Simon A. Sherman¹, and Michael E. Johnson²

Center for Pharmaceutical Biotechnology and the Department of Medicinal Chemistry and Pharmacognosy, University of Illinois at Chicago, 833 South Wood Street, Chicago, IL 60612-7231

We have recently shown that a probabilistic approach to NMR-based protein structure determination can yield higher quality structures than the more commonly used deterministic approaches. In addition to NMR data, the probabilistic approach uses *a priori* information on empirical distribution functions for backbone conformations, generated from the high resolution X-ray structures in the Protein Data Bank. This approach leads to greater accuracy in the determination of local conformations, and hence to greater precision and accuracy in the spatial structure built on the basis of these local conformations. In this review, we describe the application of this approach to two specific types of structural problems: (i) comparison of protein structures at the level of local conformations, which is particularly useful for large proteins for which NMR assignments exist, but for which spatial structures in solution have not been determined; cytochrome *c* is used as an example; and (ii) determination of three dimensional structures of proteins in solution, using accurate local conformations based on NOE data, and a build-up strategy that works well even with sparse NOE data sets; the C-terminal tryptic fragment of human plasma fibronectin is used as an example.

Structural comparisons of proteins in solution are often required to examine structure-functional relationships, to study structural effects of sequence variations, or to distinguish between two forms of the same molecule under different conditions. The methods commonly used to obtain spatial structures from NMR data (1,2) generally determine an accurate protein fold, but local conformations, obtained as a by-product

¹Current address: Eppley Institute, University of Nebraska, 600 South 42nd Street, Omaha, NE 68198

²Corresponding author.

of the spatial structures, are not so accurate. However, the biological function of a protein is often dependent on subtle changes in conformation at specific residues only. In such cases a global comparison of two structures, in terms of root mean square deviations (rmsd) in either atomic or angular coordinates, may show no significant differences. An alternative approach, using a probabilistic method to determine accurate local conformations of proteins in solution from NOE data (3,4), provides a simple, yet accurate and useful way to compare two solution structures in terms of dihedral angles of individual amino acid residues. In contrast to X-ray diffraction, the experimental observations in NMR mostly reflect the local geometry of the molecule rather than its overall spatial structure. Hence, a comparison of local conformations (in angular space) estimated from the NMR data, appears more appropriate for NMR than the comparison of atomic coordinates.

The probabilistic approach mentioned above does not require any knowledge regarding the overall spatial structure. It is, therefore, especially useful for large proteins for which NMR data exists, but solution structures have not been determined. A good example is eukaryotic cytochrome *c*, an electron transfer protein found between the outer and inner membranes of mitochondria. We have applied the probabilistic method to determine accurate backbone conformations for ferri- and ferrocyanochrome *c* in solution (5). This enabled detailed comparisons between the conformations of the two redox states in solution, as well as between cytochrome *c* conformations in solution and in crystalline environments. The need for such studies, and the insights provided by them are discussed.

The local conformations estimated by the probabilistic approach may also be used as initial angular coordinates for the determination of spatial structure by an efficient build-up strategy (4). Conventional approaches to NMR-based solution structure determination use a global optimization strategy, entailing the simultaneous use of all NMR constraints, and must deal with the complications of multiple minima. In contrast, the build-up strategy applied here uses a hierarchical approach, pursuing a minimum energy pathway to a global optimum. It starts with the estimation of accurate local conformation for each residue, then builds those fragments which contain the largest number of NMR constraints, and finally connects these energy refined fragments in several steps to obtain the complete structure.

We have used this build-up strategy to determine the 3D-structure of the C-terminal tryptic fragment of human plasma fibronectin (Fn) (6). Fn is a large glycoprotein, with two nearly identical subunits of 230-250 kD each, connected at the C-terminal by two inter-chain disulfide bonds. It was not known whether the two monomers were linked in a parallel or antiparallel arrangement. This question was investigated by the NMR-based 3D-structure determination of the C-terminal 6 kD Fn fragment containing the two inter-chain disulfide bonds. It was shown that the parallel/antiparallel question could be resolved by the probabilistic approach, but not by the more conventional deterministic approach based on distance geometry (6). This review describes the use of the build-up strategy to determine solution structures, and uses the Fn fragment as an example to illustrate the effectiveness of the probabilistic method in practice, especially with sparse NOE data sets.

Methodology and results are presented first for cytochrome *c*, then for Fn.

Comparison of Solution Structures using Accurate Local Conformations

Methodology. Methodology related to the comparison of structures involves: (i) determination of accurate backbone conformations; (ii) interpretation of data used; (iii) choice of suitable statistical criteria and (iv) strategies to search for correlations between conformational differences and molecular functions.

Determination of Accurate Backbone Conformations. Details regarding the probabilistic approach, and its performance in comparison to other methods for the determination of accurate backbone conformations from NMR data, have been reviewed elsewhere (4). Briefly, in the probabilistic approach, the (ϕ, ψ) conformational space is divided into several regions, such that conformations in each region correspond to the same set of sequential d connectivities ($d_{\alpha N}$, d_{NN} and $d_{\beta N}$). Empirical distribution functions for ϕ and ψ for each of the 20 amino acids, determined from the crystal structures in the Protein Data Bank (PDB) (7), are used to calculate the mathematical expectations (and standard deviations) for ϕ and ψ in each region, and these are used as the best representative angular coordinates for that region. The set of d connectivities for each residue in a protein then corresponds to a particular region in (ϕ, ψ) space, and hence to a particular set of angular coordinates describing the most probable backbone conformation for that residue.

The method has been coded in a program called *fisinoe* (3). This program was used to estimate the most probable values of (ϕ, ψ) and the associated standard deviations for each residue, corresponding to a set of $d_{\alpha N}$, d_{NN} and $d_{\beta N}$ connectivities. Two general assumptions were made: (i) the upper limit of distance constraints is less than 3.3 Å for sequential d connectivities; and (ii) all dihedral angles lie within sterically allowed regions in conformational space.

Uncertainties in d connectivity information (e.g., due to bleaching of cross-peaks caused by water suppression using pre-saturation, lack of resolution or lack of assignment) were treated in a conservative manner, considering both the presence and the absence of the corresponding d connectivity to be equally likely. In such cases, the sequential NOE information may fit more than one region in (ϕ, ψ) space, and hence may correspond to more than one set of most probable (ϕ, ψ) values for a particular residue. If these regions were close in conformational space, an average of the most probable (ϕ, ψ) values was used; otherwise the residue was excluded from the analysis.

Conformations of glycines determined by the probabilistic approach also contain ambiguities. Each set of $d_{\alpha N}$ and d_{NN} connectivities (no $d_{\beta N}$ for glycines) corresponds to at least two distinct regions in conformational space (4). Only that conformation which corresponded to the highest probability was considered. Statistical analyses were performed on data sets including and excluding glycines, in order to check for any undue influence of uncertainties in glycine conformations on results.

Data used. Sequential d connectivities for horse ferro- and ferricytochrome *c* were obtained from published NOE connectivity diagrams (8,9). [Information regarding coupling constants, NOE intensities and long range NOE connectivities have

not been published.] Atomic coordinates for tuna ferro- and ferricytochrome *c* (10) were obtained from the PDB.

Statistical Criteria. The angular root mean square deviation (armsd), was used as a criterion to estimate *precision* (when comparing different estimates of the same structure) and *variability* (when comparing different structures). The precision, in terms of armsd, for ϕ and ψ estimated by the probabilistic method was about 25° . A deviation greater than 75° (three times the armsd) in ϕ or ψ between two solution structures being compared was considered to indicate statistically different main chain conformations.

The precision, in terms of armsd values, for the X-ray structures was estimated by comparing the (ϕ, ψ, χ_1) values for the two crystallographically independent molecules in the asymmetric unit of tuna ferricytochrome *c*. Assuming the cores of these two essentially identical molecules to be statistically indistinguishable, the *variability* in the core residues was used as a measure of *precision* in the crystal structure determination. The calculated precision values were 7.4° for ϕ and ψ , and 14.1° for χ_1 ; so that deviations greater than 22° were considered to be statistically significant for ϕ and ψ (three times the armsd of 7.4), and greater than 42° for χ_1 (three times the armsd of 14.1) for the crystal structures. Conformational differences were considered significant only when observed for *both* molecules in the asymmetric unit of tuna ferricytochrome *c*.

Strategies for Structural Comparison. The analysis was confined to backbone conformations, since NMR assignments for backbone protons are likely to be unambiguous. Conformational differences between the following pairs of structures were examined: (i) the two redox structures in solution, using the NMR data for horse cytochrome *c*; (ii) the two redox structures in the crystalline state, using the X-ray data for tuna cytochrome *c*; and (iii) the solution versus crystal structures separately for each oxidation state. When comparing solution versus crystal structures, the higher armsd for the solution structure was used as a baseline to check for significant conformational change. Secondary structures deduced from the estimated backbone dihedrals were also compared in each case.

For each case listed above, correlations were examined between the observed conformational difference (in terms of armsd) and the spatial location of the corresponding residue on the protein. For this purpose, residues were categorized into (a) *surface* and *core* residues, depending on the solvent accessibility of the main chain and C_β atoms (11); and (b) *near* and *far* residues, depending on distances of the main chain and C_β atoms from the heme iron. Since the spatial structure in solution for the horse proteins has not yet been determined, the tuna crystal coordinates were used in the calculations of solvent accessibility and distances from the heme iron.

The software package SAS (SAS Institute Inc., Cary, NC) was used on a microvax for all statistical analyses.

Results and Discussion

Comparison of backbone conformations were done in two ways: (i) residue to residue

comparison between two structures; and (ii) statistical comparisons in terms of arms values after sorting the residues into classes (surface/core; near/far from the heme iron) as described in the Methodology section. Conformational differences between the two oxidation states of cytochrome *c* are discussed first, contrasting the information from the solution structures for the horse proteins with that from the crystal structures of the tuna proteins. Each solution structure is then compared to the corresponding crystal form.

Conformational Differences between the Two Redox States. The sequential *d* connectivity data obtained from the published NOE connectivity diagrams for horse ferro- and ferricytochrome *c* (8,9) was used as input for the FISINOE program and the main chain solution conformation for each residue was determined as described in the Methodology section. The calculated ϕ , ψ values have been reported elsewhere (5). Dihedral angles ϕ , ψ and χ_1 for tuna ferro- and ferricytochrome *c* were calculated from the crystallographic coordinates in the PDB. Statistically significant differences in backbone conformation ($\Delta\phi$ and/or $\Delta\psi > 3\sigma$) between the two redox states in solution are observed for fourteen residues: Lys27, Thr28, Leu32, Gln42, Thr47, Tyr48, Thr49, Asn52, Glu69, Lys72, Met80, Phe82, Ile85 and Lys86 (Table I). In the tuna crystals, this number is only three: Lys27, Val28 (observed in solution also) and His26 (Table I); in addition, the side chain conformation of Asp50 shows a significant change with redox state ($\Delta\chi_1 > 3\sigma$). [The sequences of the horse (104 residues) and tuna (103 residues) proteins differ at the following nineteen positions: 4, 9, 22, 28, 33, 44, 46, 47, 54, 58, 60, 61, 62, 89, 92, 95, 100, 103, 104. Of these, only two (28 and 47) show redox state dependent changes in backbone conformation. Most of these residues (14 of 19) are either on the surface or far from the heme ($>12 \text{ \AA}$ from the heme iron) and hence do not critically affect the results discussed here.]

The conformational differences between the two redox states are highlighted in a pictorial representation in Color Plate 22, using the crystal structure for tuna ferrocycytochrome *c* [since coordinates for the horse protein are not available, and since the overall folding of the two proteins is the same (12)]. Residues that show significantly different backbone conformations between the two oxidation states have been color coded to differentiate between solution (NMR) and crystal structure (X-ray) information. Residues that show redox state dependent conformational changes in solution are primarily clustered in three regions around the heme: residues 42, 47, 48, 49 and 52 are at the bottom of the heme; 80 and 82 are on one face of the heme and 27, 28 and 32 are at the opposite face (close to His18). Structural changes in similar regions have been reported earlier, on the basis of proton chemical shift differences (13). A total of 29 residues were reported to show chemical shift differences, including 27, 42, 52, 72 and 81. Shifts in atomic positions were reported in these regions for the tuna crystal structures also (10). However, most of these changes were limited to the motion of side chain atoms relative to the heme, associated with the displacement of a water molecule hydrogen bonded to side chain atoms of residues 52, 67 and 78. No significant changes in main chain conformations were observed in these regions in the tuna crystal structures. Main chain conformations of the N- and C-terminal helices are observed to be independent of the oxidation state in both solution and crystal structures.

NOTE: The color plates can be found in a color section in the center of this volume.

Table I. Differences in local conformation

Residue ^a	Reduced versus Oxidized State			Solution versus crystalline State		
	Solution ($\Delta\phi, \Delta\psi$) ^b	Crystal ($\Delta\phi, \Delta\psi, \Delta\chi_1$) ^c	Crystal ($\Delta\phi, \Delta\psi, \Delta\chi_1$) ^d	Reduced ($\Delta\phi, \Delta\psi$) ^e	Oxidized ($\Delta\phi, \Delta\psi$) ^f	Oxidized ($\Delta\phi, \Delta\psi$) ^g
His26		15,42,1	8,34,7			
Lys27	45,165	41,69,100	34,72,69	34,140	45,127	52,124
Thr28 (Val)	65,145	69,12,3	78,9,3	29,151		
Leu32	10,115			3,128		
Gln42	10,115					
Thr47 (Ser)	10,115				12,132	23,122
Tyr48	65,145				5,78	4,82
Thr49	40,165			45,140		
Asp50		1,12,147	1,10,166	5,168		
Asn52	40,165				31,171	36,164
Glu69	20,165				31,173	34,163
Pro71				3,152	9,157	4,163
Lys72	25,145			28,154		
Thr78				5,155	9,259	10,159
Met80	40,140			21,114		
Phe82	10,115			64,108		
Ile85	40,140			41,111		
Lys86	25,145			22,165		
Lys87				13,104	13,134	3,137

^aResidue substitutions for tuna cytochrome *c* at positions 28 and 47 are given in parenthesis.

^bStatistically significant differences in backbone conformation (in degrees), between horse ferro- and ferricytochrome *c* in solution ($\Delta\phi$ or $\Delta\psi > 75^\circ$).

^cConformational differences between tuna ferro- and ferricytochrome *c* crystals, considering the 'inner' molecule in the asymmetric unit.

^dConformational differences between tuna ferro- and ferricytochrome *c* crystals considering the 'outer' molecule in the asymmetric unit. For statistical significance, ($\Delta\phi$ or $\Delta\psi$) $> 22^\circ$, or $\Delta\chi_1 > 42^\circ$ for both molecules in the asymmetric unit of ferricytochrome *c*.

^eStatistically significant differences in backbone conformation between horse (solution) and tuna (crystal) ferrocycytochrome *c* ($\Delta\phi$ or $\Delta\psi > 75^\circ$).

^fStatistically significant differences in backbone conformation between horse (solution) and tuna (crystal) ferricytochrome *c* ('inner' molecule in the asymmetric unit).

^gStatistically significant differences in backbone conformation between horse (solution) and tuna (crystal) ferricytochrome *c* ('outer' molecule in the asymmetric unit).

Functional Implications: Consideration of Highly Conserved Residues. Leu32, Tyr48, Met80 and Phe82 are evolutionarily invariant residues in cytochrome *c* (10). In the crystal structures, Leu32 is part of the hydrophobic heme pocket and Tyr48 stabilizes the heme. The sulfur of Met80 is liganded to the heme iron. Results of site specific mutation experiments have implicated a regulatory role for Phe82 in the electron transport process (14). The oxidation state dependent conformational changes indicated for these residues in solution, therefore, could be critical in defining their functional role and the underlying mechanism in electron transfer. In contrast to the results obtained here for the solution structures, the crystal structures of tuna cytochrome *c* show no significant redox state dependent changes in backbone conformation for these residues.

Glycines at positions 1, 6, 34, 41, 77 and 84 are also evolutionary invariant residues in cytochrome *c* and occupy critical positions in the spatial structure where there is no room for side chains (10). No differences in conformation between the two redox structures in solution are observed for 1, 6, 34, 41 and 77, considering the conformation with the highest probability for each glycine. No comparison can be made for Gly84, since its resonances were not assigned in ferricytochrome *c*.

Consideration of Effects on Dipole Moments. Redox state dependent conformational changes in charged residues like Lys27, Glu69, Lys72 and Lys86 are likely to have a significant effect on charge distribution, dipole moments and relative orientation of the heme group with respect to its redox partners, thus affecting the electron transfer process. For all fourteen residues showing redox state dependent conformational changes in solution, an extended backbone conformation changes to a twisted one (or vice versa), and this change is also expected to affect the local dipole moment. For residues 47, 48, 52, 69 and 85, the changes in (ϕ, ψ) between the reduced and oxidized forms of the protein in solution are such that $(\psi_i + \phi_{i+1})_{\text{ferro}} - (\psi_i + \phi_{i+1})_{\text{ferri}} \approx 180$. In other words, the orientation of the peptide bond, C_i-N_{i+1} , is rotated by 180° upon change in the oxidation state, but the peptide plane and the vectors $C\alpha_i-C\beta_i$ and $C\alpha_{i+1}-C\beta_{i+1}$, that position the side chains, are not significantly affected. Such conformational changes may have significant effects on molecular properties associated with the peptide bond, and hence on molecular function, while leaving the structure essentially unperturbed. Any change in the orientation of the peptide bond will affect the direction of the dipole moment associated with it. Therefore, such conformational changes may lead to the modulation of local dipole moments, and hence prove to be functionally significant for the electron transfer properties of cytochrome *c*. Most studies involving dipole moment calculations for cytochrome *c* have focussed on the magnitude and direction of the net dipole moment of the molecule, and its effect on the interaction of cytochrome *c* with its redox partners (15,16). In general, such calculations ignore contributions from all bond dipoles, assuming random orientations; and consider only the contributions from charged side chains, α -helices and the heme. However, calculations using a tetrapeptide model (5) show that peptide bond reorientation may lead to significant changes in both magnitude and direction of the tetrapeptide's dipole moment. For cytochrome *c*, it is plausible that such changes in local dipole moments close to the heme (residues 47, 48, 52, 69 and 85 are all within 7 to 12 Å of the heme iron) will affect its electron transfer function.

Correlation Between Redox Dependent Conformational Changes and the Location of Residues. The fourteen residues showing significant differences in backbone conformation between the two redox states in solution were carefully examined to check for correlations between conformational change and parameters like proximity to the heme, solvent accessibility, polarity and position in the sequence. Only the first parameter showed a direct correlation. About 50% of the residues in cytochrome *c* are within 12 Å of the heme iron in the crystal. Therefore, the residues were categorized as *near* and *far* according to whether they were located inside or outside a sphere of radius 12 Å, centered at the heme iron. Calculation of *armsd* between ferro- and ferricytochrome *c* in solution showed that the *armsd* for residues within 12 Å of the heme iron was much larger than that outside this region (Table II). Significant changes in backbone conformation between the redox states in solution, therefore, are confined to residues close to the heme. Similar calculations for the tuna crystal structures showed the same trend, but with much smaller conformational differences between the two states (Table II). Since the inherent flexibility of residues on the protein surface may lead to differences between the two forms of the protein, the residues were also categorized as *surface* or *core* residues, depending on their solvent accessibility. Calculation of *armsd* between ferro- and ferricytochrome *c* for these two categories gave the expected results for both solution and crystal structures: the *armsd* was slightly larger for surface residues than for residues in the core (Table II).

Secondary Structure of the Two Redox States. The large differences in ψ (about 115° to 165°) between the two redox states in solution indicate changes in secondary structure (for instance, helical to extended conformation or vice versa). To examine this feature, the backbone conformations derived from the NMR data were used to determine secondary structural elements (extended, helical and turn structures) for horse cytochrome *c* in solution. The results are shown schematically in Figure 1. Overall, the secondary structure elements are the same in the two redox states in solution, except for certain turn structures (turns at 21-24, 72-75 and 81-84 in the reduced protein, but not in the oxidized; turns at 30-33 and 39-42 in the oxidized protein, but not in the reduced). A short stretch of extended residues (24-33) and a helical region (48-55) in the reduced structure are replaced by three turns (25-29, 30-33 and 49-52) in the oxidized form.

In the tuna crystals (10), the two redox forms have identical secondary structures: five helical regions (2-14, 49-55, 61-69, 71-75, 87-102), five turns (21-24, 32-35, 35-38, 43-46, 75-78) and two short extended regions (38-40, 57-60). The helical and extended regions match fairly well with those observed in the crystal structure of horse ferricytochrome *c* (12), but three of the turns do not match (14-17, and 67-70 present in horse ferricytochrome *c*; 43-46 present in the two tuna structures).

The most obvious differences at the secondary structure level are between the solution and crystal structures: the extended region, 24-33, observed in horse ferrocycytochrome *c* in solution, is not observed in the crystal; while the helical region

Table II. Angular Root Mean Square Deviations

	<i>Residues Categorized by Solvent Accessibility</i>				
	<i>Reduced versus Oxidized Protein</i>		<i>Solution versus Crystal Structures</i>		
	<i>Solution (ϕ, ψ)</i>	<i>Crystal (ϕ, ψ)</i>	<i>Ferro (ϕ, ψ)</i>	<i>Ferri (ϕ, ψ)</i>	<i>BPTI^a (ϕ, ψ)</i>
Surface	20,66	15,13	22,63	23,44	46,72
Core	16,59	10,12	24,56	24,54	23,24
	<i>Residues Categorized by Distance from the Heme Iron</i>				
Near ^b	20,73	14,14	26,69	26,58	
Far ^c	13,32	9,9	19,41	20,36	

^aThe armsd values for BPTI are included to illustrate that, in general, when solution and crystal structures are compared, surface residues usually exhibit larger armsd than core residues, and that corresponding results for cytochrome *c* deviate markedly from this expected trend.

^bResidues with C, N, C α , and C β atoms within 12 Å of the heme iron.

^cResidues with C, N, C α , and C β atoms further than 12 Å from the heme iron.

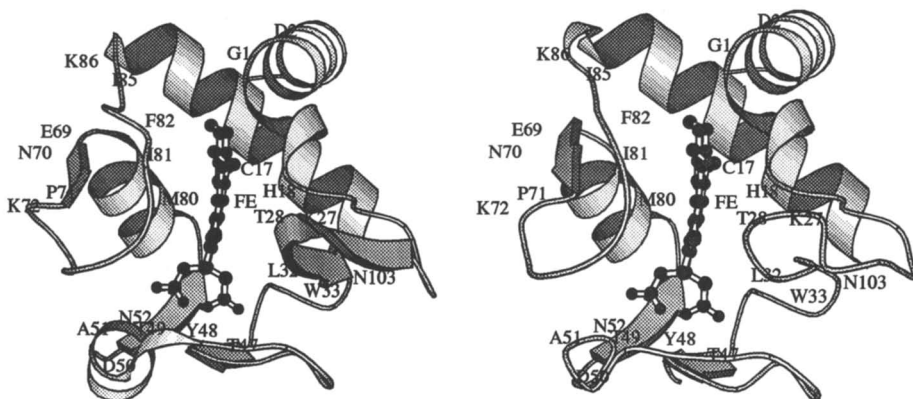


Figure 1. Schematic illustration of secondary structure in horse ferro- (left) and ferricytochrome *c* (right) in solution. The secondary structure was deduced from the backbone dihedrals (ϕ, ψ) determined by the probabilistic method, using sequential NOE data. The MolScript program (16) was used to create these schematic drawings, based on the atomic coordinates of tuna cytochrome *c* (9).

(71-75) in the crystals is absent in the solution structures. Implications of such structural differences are not clear without building the spatial structure in solution. However, determination of accurate backbone conformations facilitates the examination of such differences that are not obvious from a qualitative treatment of the NMR data or examination of 'secondary shifts' in proton chemical shifts (18,19).

Solution versus Crystal Structures. Our results show that the oxidation state dependent changes in backbone conformations are larger in solution than in the crystalline form (see Table I and Color Plate 22). This difference cannot be rationalized solely in terms of the amino acid substitutions between the horse and tuna proteins. While differences in the primary structure may lead to differences in local conformations, such conformational differences are likely to be confined to the side chain. Of the eighteen differences between the two sequences, only two (at positions 28 and 47) show oxidation state dependent conformational changes in solution. Redox state dependent conformational changes could be different if the solution and crystal structures are significantly different at the level of *local conformation*. A careful comparison shows that the backbone conformations of the NMR-based solution structures and the X-ray determined crystal structures are significantly different for 13 residues in ferrocycytochrome *c*: 27, 28, 32, 48, 49, 71, 72, 78, 80, 82, 85, 86 and 87; and for 8 residues in ferricytochrome *c*: 27, 42, 47, 52, 69, 71, 78 and 87 (Table I). Examination of the temperature factors for these residues in the tuna crystal structures do not suggest any disorder or poorly defined electron densities in the crystals at these locations. [It should be noted that the backbone conformation of Lys27 is in a sterically forbidden region in tuna ferricytochrome *c* (for both molecules in the asymmetric unit). Since the distribution functions used in the probabilistic approach are generated from the PDB, conformations corresponding to the highest probability generally fall within sterically allowed regions for all residues except Gly. Hence a sterically forbidden conformation is not estimated for Lys27 in the solution structures of the horse protein.]

Correlation between Differences in Solution and Crystal Forms and the Location of Residues. The armsd between the solution and crystal structures were examined after sorting the residues according to their solvent accessibility and distances from the heme iron. Comparison of both ferro- and ferricytochrome *c* structures in solution versus crystal form show a large armsd for residues within 12 Å of the heme iron (Table II), indicating that significant differences between the solution and crystal structures are confined to residues close to the heme.

In general, large conformational differences between solution and crystal structures are expected for surface residues that are likely to be affected by packing forces in the crystal. BPTI (bovine pancreatic trypsin inhibitor), for example, shows exactly such a correlation (3), with a high armsd corresponding to residues on the surface and a relatively lower armsd for residues within the globular core of the protein (Table II). In contrast, an anti-correlation is observed for both ferro- and ferricytochrome *c* when backbone conformations in the solution and crystal structures are compared, indicating that these structures are statistically different even at the core of the molecule. A plausible cause of such differences is the large difference in ionic

strengths between the solution and crystal environments. The highly charged surface of cytochrome *c* may make this protein unusually sensitive to the ionic strength of its surroundings. In such cases, the crystal structure may not provide a dependable model for molecular interactions *in vivo*.

In conclusion, it is demonstrated that the probabilistic approach provides a simple method for the determination of accurate backbone conformations using sequential NOE data. Consideration of other experimental data, like NOE intensities and coupling constants, would improve the precision to some extent and permit the estimation of side chain conformations required for determining the complete spatial structure; but these observations have not been published for cytochrome *c*. However, it is shown here that a careful analysis of accurate backbone conformations estimated from sequential NOE data alone may prove very useful in enhancing our understanding of structural and functional aspects of proteins, even when no spatial structure is known for the protein in solution.

The local conformations estimated by the probabilistic method may be used as initial angular coordinates for the determination of the complete 3D-structure by a build-up strategy described below.

NMR-Based 3D-Structure Determination

Experimental Methods: Sample Preparation. Fibronectin (Fn) was isolated from fresh-frozen human plasma and the carboxyl-terminal 6 kD fragment was purified as described previously (6). For NMR measurements, about 3 mg of the 6 kD fragment was lyophilized to dryness and redissolved in 0.5 mL of either D₂O, or H₂O containing 5% D₂O, giving a final peptide concentration of about 1 mM.

¹H NMR Spectroscopy. NMR data were accumulated on a General Electric GN500 spectrometer. Phase-sensitive two-dimensional (2D) COSY and NOESY data sets were collected in the hypercomplex mode (20), with standard pulse sequences and phase expressions (21,22). NOESY data were acquired with mixing times of 125 and 250 ms. Relayed COSY experiments in the absolute value mode (23) were used to help identify spin systems of side chains.

Computational Methods. Two different approaches were examined for the determination of spatial structure from NMR data: (i) the deterministic distance geometry (DG) approach (1), followed by energy refinement; and (ii) a build-up strategy (BUILD), using a probabilistic model of protein conformation (3,4).

The Distance Geometry Approach. The DSPACE software package (Hare Research, Inc., Woodinville, WA) was used in the DG approach, to generate structures consistent with covalency constraints and semi-quantitative estimates of inter-proton distance constraints, starting with random initial atomic coordinates (24). No anti-distance constraints (25) were applied. Since the NMR data indicate a symmetrical structure, symmetry was used as a restraint in the energy refinement process.

A set of structures, generated by DSPACE, and selected qualitatively for their conformational diversity by comparison of (ϕ, ψ) plots, was also used as initial

structures for restrained energy minimization (DGREM) and restrained molecular dynamics (DGRMD) calculations using the CHARMM software package (26).

The BUILD Approach. The BUILD procedure involves three steps: (a) estimation of an initial set of angular coordinates (local conformations) from the NMR data; (b) determination of the spatial structure by a gradual build-up process; (c) structure refinement by energy minimization on unrestrained structures.

The FISINOE program (3) was used as described above, to determine the backbone conformations (ϕ , ψ values) for each residue, using the sequential d connectivity information. Extension of the probabilistic method, and consideration of the intra-residual NOE data between amide- and C_{β} -protons was used to obtain the possible combinations of χ_1 and χ_2 angles for the side chain conformations (4). These (ϕ , ψ , χ_1 , χ_2) values were used as the starting set of angular coordinates for the BUILD procedure.

The BUILD strategy utilizes an optimality principle in which the fragment under construction at any stage has a minimum number of residues and a maximum number of restrictions. Long-range NOE requirements and the value of conformational energy are used as steering parameters to guide the build-up process. The interactive graphics package INSIGHT (Biosym Technologies, San Diego, CA) was used to construct the starting structures. All energy minimizations were performed using CHARMM. The force constant for the dihedral constraints was reduced in gradual steps from 50.0 ($\text{kcal}\cdot\text{mol}^{-1}\cdot\text{rad}^{-2}$) to 2.5 and finally to 0.0, decreasing it by about a factor of two following each cycle of 250 steps of conjugate gradient minimization. Symmetry was not used explicitly as a restraint in the energy minimization process; only the same initial angular coordinates were used for both chains. Distance constraints also were not included explicitly in the calculations. When checking for agreement between structures built and NOE constraints, it was assumed that the upper limit is $\leq 3.3 \text{ \AA}$ for sequential and intra-residue NOEs and $\leq 4.0 \text{ \AA}$ for all other NOEs. No semi-quantitatively estimated distance constraints were considered.

All calculations were performed on a Silicon Graphics work station.

Results and Discussion

NMR Assignments. The primary sequence of the 52 residue C-terminal dimer fragment is shown in Figure 2. The primary aim of this study was to use NMR to determine whether the inter-chain disulfide bonds linked the monomers in a parallel (Figure 2, top) or antiparallel manner (Figure 2, bottom). Details regarding assignments and chemical shift information have been reported elsewhere (6). Relevant NOE data are summarized in Figure 3. Several long-range NOEs, crucial for answering the parallel/antiparallel question, were observed in the 125 ms NOESY spectrum in D_2O : the ϵ - and ζ -protons of Phe12 showed NOEs to both β -methylene protons of Cys7 and Cys11; one of the β -methylene protons of Cys7 showed NOEs to both β -methylene protons of Cys11.

Data Interpretation. The NMR information indicated a symmetrical structure for the dimer, since chemical shifts were identical for the same residue in both chains. This

complicated NOE assignments, since no distinction could be made between *intra-chain* and *inter-chain* NOEs. However, a statistical analysis of short proton-proton distances in a data set containing high resolution protein crystal structures (27), shows that protein folding patterns in nature very rarely lead to short proton-proton distances relating main chain and C_{β} -protons of residues unless these involve immediate neighbors on a polypeptide chain. Therefore, all sequential and intra-residue NOEs were interpreted as *intra-chain*. This interpretation was supported by the results of a set of calculations using conformational analysis alone, without any consideration of the NOE data (6). No assumptions were made regarding the long-range NOEs, so that these constraints could be satisfied through either intra-chain or inter-chain (or both) connectivities in the 3D-structures derived, as long as all sequential NOE requirements were also satisfied.

Comparison of ^1H -chemical shifts (6) of the Thr1 to Arg25 segment with those of random coil structures (28) showed substantial differences for several resonances within the Thr3-Pro14 sequence, indicating a preferred conformation for this segment containing the inter-chain disulfide bonds. Also, inter-residue NOEs other than those showing sequential ($d_{\alpha\text{N}}$, d_{NN} , $d_{\beta\text{N}}$) connectivities were observed only within this segment. The NMR data indicates that the structure is more flexible further away from the inter-chain disulfide bonds. Therefore, spatial structure was calculated only for the Thr3-Pro14 segment. This proved to be more than adequate for demonstrating that the two inter-chain disulfide bonds link the Fn monomers in an antiparallel fashion in the 6 kD C-terminal dimer fragment.

Data Analysis: Results using DG. A set of 55 approximate inter-proton distance constraints per chain, derived from the observed NOE data, was used both with, and without the additional symmetry constraint, to calculate 3D-structures of the Thr3 to Pro14 segment. Several DG, DGREM and DGRMD structures, with reasonably small distance constraint violations, were obtained for both parallel and antiparallel models, showing that 'restrained' structures, roughly satisfying all NOE requirements, were possible for both models. The structures did not fall into any closely related sets, and showed large variations in the main chain conformation when compared in pairs. Also, all of the structures calculated in this way contained several dihedral angles well outside the sterically allowed regions. Therefore, comparison of the calculated conformational energies was not an adequate criterion, either for selecting a set of preferred structures from the many converged restrained structures, or for deciding whether the NMR data indicated a parallel or an antiparallel arrangement of the two chains in the dimer. We concluded that the number of inter-residue NOE constraints available (less than 5 per residue) was not sufficient for this strategy to work.

Results using BUILD. We then applied the BUILD procedure described above, using the d connectivities shown in Figure 3 ($d_{\alpha\text{N}}$, d_{NN} , $d_{\beta\text{N}}$) to estimate corresponding regions in ϕ, ψ space for each residue. The most probable ϕ, ψ values corresponding to each region were used as the starting set of ϕ and ψ angles for the 12 residue monomer segment, Thr3-Pro14. Extension of this method, and consideration of the intra-residual NOE data between amide- and C_{β} -protons was used to obtain the possible combinations of χ_1 and χ_2 angles for the side chain conformations (4). A four step

'build-up' procedure was followed to construct the final 3D-structure. (i) The Cys7-Cys11 segment, containing the inter-chain disulfide bonds was constructed first, because of the four long-range NOE constraints between the C_{β} -protons of Cys7 and Cys11 present in this segment. The Ile9 and Glu10 side chains were initially truncated to alanine, making the testing sequence Cys-Pro-Ala-Ala-Cys. (ii) Phe12 was added to this sequence, and, in two subsequent steps, alanines at positions 9 and 10 were replaced by Ile9 and Glu10. (iii) The Thr3-Asn6 segment was then added, using alanines at positions 4 and 6 in the initial calculations, before introducing side chain atoms for Asn4 and Asn6. (iv) Met13 and Pro14 were finally added to complete the segment. This procedure reduced the total number of calculations required from 4096x2 (for parallel and antiparallel structures) to only 90. Pro8 and Pro14 were modeled to be in the *trans* configuration, as indicated by the presence of strong $\alpha_i\text{-}\delta_i$ NOE cross-peaks.

Arrangement of Monomer Chains in the Dimer. The 'parallel/antiparallel' question was answered by constructing the Cys7-Cys11 segment in step (i). This segment, containing the local conformations determined by the probabilistic method, was built using INSIGHT. CHARMM was then used to 'patch' two such segments in either a parallel or an antiparallel fashion, via the two inter-chain disulfide bonds. Energy minimization and structure refinement were then performed as described in the Methodology section.

Of the four possible conformations with different combinations of χ_1 values for the pair of cystines in the monomer, none satisfied all four long-range NOE requirements between the C_{β} -protons of Cys7 and Cys11 for a parallel dimer structure. In the antiparallel structure, it was possible to select one of the four combinations of χ_1 conformers (with $\chi_1 \approx \bar{g} = -60^\circ$ for both Cys7 and Cys11), since only this conformation satisfied all sequential and long-range NOE requirements. Also, the symmetry requirement was satisfied in all four of the unrestrained antiparallel structures, but in none of the parallel structures.

Conformational Analysis. Since the results above were obtained assuming all $d_{\alpha N}$, d_{NN} , $d_{\beta N}$ and $d_{N\beta}$ connectivities to be intra-chain for the Cys7-Pro8-Ala9-Ala10-Cys11 sequence, conformational analysis was performed for this sequence to check whether other parallel or antiparallel structures (not predicted by the assumed intra-chain 'sequential' d connectivity patterns) were energetically favorable for this dimer fragment. Eight additional structures were found, with conformational energies comparable to the single conformation selected above. All eight were antiparallel structures, and all exhibited conformational symmetry. However, inter-proton distance calculations showed that none of these additional structures contained inter-chain d connectivities in place of missing intra-chain sequential d connectivities; *so that none of them satisfied all NOE requirements.* Therefore, even if the Cys7-Cys11 segment does exhibit multiple conformations in solution, it appears likely that a substantial population exists in the single conformation obtained assuming all sequential connectivities to be intra-chain. Only this structure was extended to build the 3D-structure of the Thr3-Pro14 segment.

It is important to note here that the observation of NOE contact between Cys7

and Cys11 did not, of itself, rule out a parallel structure; extensive additional conformational analysis was required to show that *only* an antiparallel structure satisfied *all* NOE constraints.

Structure Consistent with NMR Data. On addition of Phe12 to the Cys7-Cys11 segment, only one of two possible conformations, with $\chi_1 \approx -60^\circ$ (or *g*⁻) and $\chi_2 \approx 90^\circ$ (or *p*), for the Phe12 side chain was found to satisfy the long-range NOEs between the C_β-protons of Cys7 and the ring protons of Phe12 (considering both intra- and inter-chain connectivities). Since there were no strong long-range NOE constraints relating the Cys7-Phe12 segment to the rest of the structure, all subsequent calculations in the 'build-up' procedure used the following energy criteria to choose the set of most probable conformations at each step: for a set of χ_1 (or χ_2) rotamers, all conformations with energy greater than the lowest energy conformation by 4 kcal/mol (calculated using a dielectric constant of 10), were eliminated. On replacing alanine by Ile9 in the Cys-Pro-Ala-Ala-Cys-Phe segment, four energetically equivalent conformers were obtained. Replacing the second alanine by Glu10 led to 16 conformers, 8 of which were eliminated by energy criteria. Similarly, extending the sequence by the Thr3-Asn6 segment led to 32 possible conformers for the Thr3-Phe12 segment, 24 of which were eliminated by energy criteria. Addition of Met13-Pro14 resulted in 32 possible conformers, 16 of which were selected, using energy criteria, to be the final set of structures consistent with the NMR data. Although symmetry was not used explicitly as a constraint in the energy minimization process, the symmetry built into the backbone conformation at the start was largely retained during the build-up steps and in the final set of energy refined unrestrained structures. Similarly, long-range NOE requirements were also found to be satisfied in the final 16 structures, and were obtained as a by-product of energy refinement, without the use of distance constraints in the minimization process. In all 16 final structures, the long-range NOEs relating the C_β-protons of Cys7 and Cys11 were found to be *inter-chain*, while NOEs relating the C_β-protons of Cys7 and ring protons of Phe12 were found to be *intra-chain*.

Estimation of Precision. The backbone conformations of the set of 16 final structures were very similar, and were indistinguishable by χ^2 statistical criteria. The 16 structures consisted of combinations of side chain conformations for Ile9 (with the following rotamer conformations for χ_1 and χ_2 : *tt*, *tg*⁺, *g*⁻*g*⁻, *g*⁻*t*; where *g*⁺ = 60° and *t* = 180°), Glu10 (*g*⁻*t*, *g*⁻*g*⁻) and Met13 (*g*⁻*t*, *g*⁻*g*⁻) that are consistent with the NMR data, and are indistinguishable by energy criteria. The angular root mean square deviation (armsd) in backbone conformations for pairs of structures within this set of 16 was 6° for ϕ and 10° for ψ (average of armsd values for all pairs within the 16 structures). The armsd between FISINOE estimates and calculated values of backbone dihedrals was 25° for ϕ and 29° for ψ (average of armsd values for all 16 structures), and the estimated and calculated backbone structures are statistically indistinguishable, using χ^2 criteria. In terms of atomic coordinates, the average pair-wise rmsd for the 16 structures were: 0.63 ± 0.15 Å (all atoms); 0.44 ± 0.11 Å (heavy atoms only); 0.13 ± 0.04 Å (backbone atoms only). The list of estimated and final dihedral angles have been reported elsewhere (6). Color Plate 23 shows the fragments constructed

in the four successive steps of the build-up process. Superposition of the final 16 NMR-derived solution structures is shown in Color Plate 24. The backbone dihedrals for all 16 structures are shown in a Ramachandran plot in Figure 4.

Unlike the types I, II and III repeat units of Fn, which contain dominant structural features that are common to many proteins, the structure of the C-terminal dimer segment reported here is somewhat unusual. The backbone consists of two helical or twisted segments, Thr3-Asn6 and Ile9-Phe12, connected by an extended region at Cys7-Pro8. These twisted regions may serve as recognition sites for the monomers, and thus help to bring the two cystines in each monomer into close proximity, specifically in the antiparallel orientation, so that the required inter-chain disulfide bridges are formed in the correct manner in the dimer. The exposed surface of the two twisted regions contain several hydrophobic side chains (Val5, Ile9 and Phe12 followed by Met13 and Pro14). However, these may be covered by the rest of the monomer (Leu15-Glu26) folding back over itself. The aromatic ring of Phe12 in each monomer lies across the two disulfide bridges, and the hydrophobic interactions involving these aromatic side chains may help stabilize the disulfide bonds. The Cys7-Cys11 segment forms a loop that brings the two cystines in the same chain close enough to make intra-chain disulfide bonds also feasible. Why inter-chain disulfide bonds are observed in the 6 kD fragment of Fn remains an intriguing question.

A thorough understanding of the functions of Fn requires knowledge of the precise spatial arrangements of various binding domains and their interactions in the two similar subunits of the dimeric Fn molecule. Our NMR demonstration of the antiparallel arrangement of the inter-chain disulfide bridge near the C-termini of Fn suggests that similar binding domains, such as the gelatin and cell-binding domains in different subunits, may be arranged in a diagonal manner rather than in a mirror image. The present result is consistent with previous work which also suggested an antiparallel arrangement for the inter-chain disulfide bridge of Fn, based on HPLC patterns of peptides derived from the C-terminal 6 kD fragment (29).

In conclusion, a build-up strategy for obtaining spatial structures was described, using the FISINOE method for estimating local conformations from sequential *d* connectivities. Using this procedure, the solution structure of 24 residues of the dimer segment of Fn was determined from a set of NOE data that is too sparse for conventional DG methods. It was shown that the 2D-NMR data are consistent with only an antiparallel arrangement of the two monomers connected via the two inter-chain disulfide bridges in the 6 kD C-terminal fragment of human plasma fibronectin.

Summary

It is shown that the determination and careful statistical analysis of accurate local conformations can provide useful information regarding protein structure and function. Since the local conformation is not a by-product of the spatial structure, detailed comparisons between conformations of individual amino acid residues may be obtained without knowledge of the complete 3D-structure. Backbone conformations determined using the probabilistic approach are accurate enough to make such comparisons meaningful. The NOE information required is simply the presence/absence of sequential NOEs. Although use of properly scaled intensity data

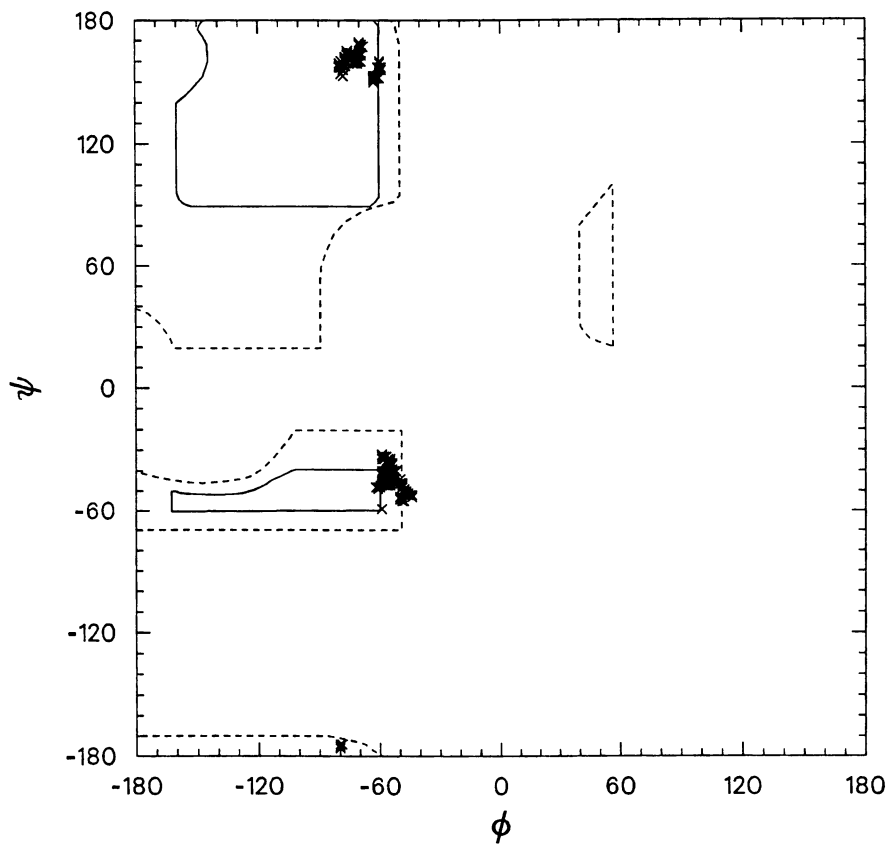


Figure 4. Ramachandran plot of the backbone dihedrals for all 16 energy refined final structures determined using the BUILD strategy and consistent with all NMR data.

and coupling constants would improve the precision of this method, and permit the estimation of side chain conformations, sequential *d* connectivity information is adequate for backbone conformations. This method, therefore, would be especially useful for large proteins for which partial or complete resonance assignments exist, but a 3D-structure has not been determined. Structural comparisons at the level of local conformations could provide important information regarding conformational effects of site specific mutations and structural changes at active site regions of enzymes in solution.

Local conformations determined from NOE data by the probabilistic approach may be used as initial angular coordinates for spatial structure determination. This strategy circumvents the multiple minima problem encountered by methods that attempt to build the overall spatial structure prior to detailed analysis of the local structure. An efficient 'build-up' strategy is presented that can construct an *unrestrained*, energy refined, 3D-structural model consistent with all NMR data even when the NOE data is sparse. Use of accurately determined local conformations as initial coordinates in the 'build-up' steps leads to enhanced precision for the family of structures satisfying all NMR data.

Acknowledgments

This research was supported in part by grant HL45977 from the National Institutes of Health.

Literature Cited

1. Wüthrich, K. *Science* **1989**, *243*, 45-50.
2. Liu, Y., Zhao, D., Altman, R. & Jardetsky, O. *J. Biomol. NMR.* **1992**, *2*, 373-388.
3. Sherman, S. and Johnson, M. *J. Magn. Reson.* **1992**, *96*, 457-472.
4. Sherman, S. A. and Johnson, M. E. *Prog. Biophys. Mol. Biol.*, **1993**, *59*, 283-339.
5. Kar, L., Sherman, S. A. and Johnson, M.E. *J. Biomolec. Str. Dyn.*, in press.
6. Kar, L., Lai, C.-S., Wolff, C.E., Nettesheim, D., Sherman, S.A. & Johnson, M.E. *J. Biol. Chem.* **1993**, *268*, 8580-8589.
7. Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F., Jr., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T., Tasumi, T. *J. Melec. Biol.* **1977**, *112*, 535-542.
8. Wand, A. J., Di Stefano, D. L., Feng, Y., Roder, H. and Englander, S. W. *Biochemistry* **1989**, *28*, 186-194.
9. Feng, Y., Roder, H., Englander, S.W., Wand, A.J. & Di Stefano, D.L. *Biochemistry* **1989**, *28*, 195-203.
10. Takano, T. and Dickerson, R. E. *J. Mol. Biol.* **1981**, *153*, 79-94; *ibid.* 95-115.
11. Richards, F. M. *Ann. Rev. Biophys. Bioeng.* **1977**, *6*, 151-176.
12. Bushnell, G. W., Louie, G. V. and Brayer, G. D. *J. Mol. Biol.* **1990**, *214*, 585-595.
13. Ziang, N., Mauk, A.G., Pielak, G.J., Johnson, J.A., Smith, M. & Hoffman, B.M. *Science* **1988**, *240*, 311-313.

14. Feng, Y., Roder, H. & Englander, S.W. *Biochemistry* **1990**, *29*, 3494-3504.
15. Koppenol, W.H. & Margoliash, E. *J. Biol. Chem.* **1982**, *257*, 4426-4437.
16. Koppenol, W.H., Rush, J.D., Mills, J.D. & Margoliash, E. *Mol. Biol. Evol.* **1991**, *8*, 545-558.
17. Kraulis, P. *J. Appl. Crystallogr.* **1991**, *24*, 946-950.
18. Gao, Y., Boyd, J., Williams, R.J.P. & Pielak, G.J. *Biochemistry* **1990**, *29*, 6994-7003.
19. Wishart, D.S., Sykes, B.D. & Richards, F.M. *Biochemistry* **1992**, *31*, 1647-1651.
20. States, D. J., Haberkorn, R. A. and Ruben, D. J. *J. Magn. Reson.* **1982**, *48*, 286-292.
21. Jeener, J., Meier, B. H., Bachman, P. and Ernst, R. R. *J. Chem. Phys.* **1979**, *71*, 4546-4553.
22. Wider, G., Macura, S., Anil-Kumar, Ernst, R. R. and Wüthrich, K. *J. Magn. Reson.* **1984**, *56*, 207-234.
23. Wagner, G. *Quart. Rev. Biophys.* **1983**, *16*, 1-57.
24. Nerdal, W., Hare, D. and Reid, B. R. *J. Mol. Biol.* **1988**, *201*, 717-739.
25. Braunschweiler, R., Blackledge, M. and Ernst, R. R. *J. Biomol. NMR.* **1991**, *1*, 3-11.
26. Brooks, B. R., Bruccoleri, R. E., Olafson, B. D., States, D. J., Swaminathan, S. and Karplus, M. *J. Comput. Chem.* **1983**, *4*, 187-217.
27. Billeter, M., Braun, W. and Wüthrich, K. *J. Molec. Biol.* **1982**, *155*, 321-346.
28. Bundi, A. and Wüthrich, K. *Biopolymers* **1979**, *18*, 285-298.
29. Skorstengaard, K., Jensen, M. S., Petersen, T. E. and Magnusson, S. *Eur. J. Biochem.* **1986**, *161*, 441-453.

RECEIVED July 14, 1994

Chapter 25

Structural Elements Involved in Allosteric Switch in Mammalian Pyruvate Kinase

Thomas G. Consler^{1,3}, Michael N. Liebman², and James C. Lee¹

¹Department of Human Biological Chemistry and Genetics,
University of Texas Medical Branch, Galveston, TX 77555

²Bioinformatics Program, Amoco Technology Company, Mail Code F-2,
150 West Warrenville Road, Naperville, IL 60563-8460

Pyruvate kinase is a key glycolytic enzyme which is regulated by an allosteric mechanism. In search of the major structural features that are involved in the allosteric switch a multifaceted approach was applied to the system. Based on results on computer modeling and steady state kinetics of isozymic forms of pyruvate kinase from an earlier study, it is concluded that there are two major structural elements involved in intersubunit contact. The difference in the primary sequence of muscle and kidney isozymes, which exhibit significantly different allosteric behavior, is located in one of these structural elements. Hence, it is proposed that the region between residues 385 and 425 must constitute an important component involved in intersubunit communication and in turn, the allosteric mechanism of pyruvate kinase.

Rabbit muscle pyruvate kinase (PK) is an allosteric enzyme under intensive investigations by a combination of approaches in order to elucidate the molecular mechanism of regulation (1-15). In earlier studies (1-2), it was shown that the hydrodynamic properties of the enzyme are altered by metabolites. Binding of substrate and metal ions required for activity causes the enzyme to assume a more symmetric structure, whereas the allosteric inhibitor would induce the enzyme to become more asymmetric. A more detailed probing of the structural features associated with the activation and inactivation of the enzyme was monitored by small angle neutron scattering (4). Results from this study indicate a "contraction" and "expansion" of the enzyme when it transforms between its active and inactive forms. The structural features associated with these global conformational changes may be related to the protein domains detected by x-ray crystallography (9-10).

³Current address: Glaxo Inc., Research Triangle Park, NC 27709

0097-6156/94/0576-0466\$08.00/0
© 1994 American Chemical Society

Domain Movement in the Activation of Pyruvate Kinase

Each subunit is composed of three major domains, one of which protrudes out into the solvent. This exposed domain, identified as domain B, forms a cleft with domain A adjacent to the active site. Domain B is attached to domain A by an apparently flexible hinge region. Domain C is located on the side of domain A opposite of domain B and it is this domain C which apparently is involved in significant amount of intersubunit contact. Chemical denaturation and proteolysis were employed as probes of the independence of these protein domains in PK (3). Results from these studies infer that a change in domain-domain interaction is involved in the transition between active and inactive enzymatic forms. To provide better understanding of the structural change involved in the activation/inactivation of PK, small angle neutron scattering (SANS) was employed to map the changes that occur upon binding of ligands to the enzyme. The SANS data were analyzed to yield $P(r)$, the length distribution function, which is defined as

$$P(r) = \frac{2r}{\pi} \int_0^{\infty} kI(k)\sin(kr)dk \quad (1)$$

where $k=4(\pi/\lambda) \sin \theta$, λ is the wavelength of 4.57 \AA , 2θ is the scattering angle, and

$$\ln I(k) = -k^2 R_G^2 / 3 + \ln I(0) \quad (2)$$

where $I(k)$ and $I(0)$ are the scattering intensities at angles 2θ and 0 , respectively, and R_G is the radius of gyration. The function of $P(r)$ can be approximated by the indirect transform method of Moore (16) and yields information which describes the size, shape, and frequency distribution of all the point-to-point pair distances between scattering centers of the particle.

Comparison between the active and inactive PK conformations expressed in the form of length distribution function is shown in Figure 1. These results suggest that the changes in conformation which accompany activation most notably affect distances in the range of approximately 72-120 Å, although much less significant perturbations are also observed around 35 Å. The bimodal nature of the distance distribution most likely results from the intra- and interdomain distances being readily separable in magnitude for the combined B domains of the tetramer.

The fact that the observed scattering data contain information pertaining to inter-atomic distances enables one to generate length distributions from both SANS and X-ray crystallographic data. The changes in length distributions, reflecting conformational changes induced by ligands, can then be compared between the two sets of experimental observations. A valid direct comparison between the SANS and X-ray crystallographic data demands the availability of data sets containing contributions from all scattering centers including that of the side chains. In the absence of complete data sets from both approaches, only separate comparisons can be conducted to yield valid conclusions for the PK system. Our approach involves

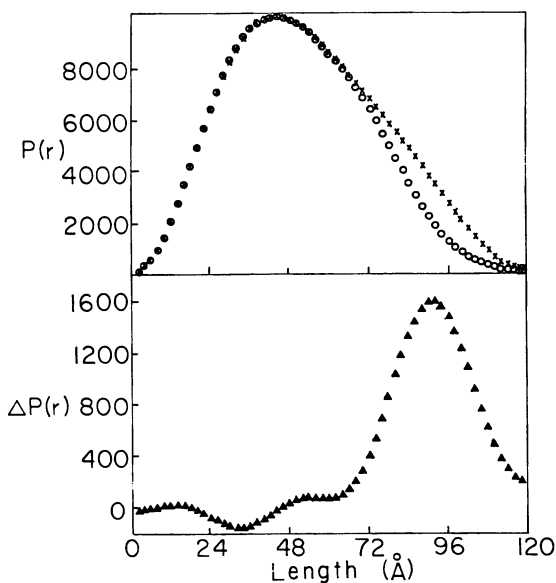


Figure 1. Comparison between active and inactive pyruvate kinase conformation. Upper, experimental length distribution functions; lower, difference length distribution function. X and O, enzyme in the presence of 15mM Phe (inhibitor) and 2mM phospho-enolpyruvate (substrate), respectively. (Reproduced with permission from ref. 4. Copyright 1988 American Society for Biochemistry and Molecular Biology, Inc.)

generating the difference distribution, $\Delta P(r)$, as shown in Figure 1. Having determined these changes in solution, they are used as a guideline for the modeling which involves the manipulation of the α -carbon coordinates defined by the X-ray crystallographic data. This is possible because for an object assumed to consist of discrete density points the Debye (17) relations can be used to calculate directly the scattering curve since

$$I(k) = \sum_{i=1}^n \sum_{j=1}^n \rho_i \rho_j \frac{\sin kr_{ij}}{kr_{ij}} \quad (3)$$

where ρ_i and ρ_j are the scattering length densities at position i and j , respectively, and r_{ij} is the distance between points i and j .

Simulation of conformational changes was conducted by manipulation of the X-ray crystal coordinates as indicated by the step of computer modeling, as shown in Figure 2. This simulation step employed computer graphics, which enabled one to conduct interactive rotation and translation, symmetry operations, and inter- α -carbon distance calculations. Having decided on a simulated conformational change, the information was then reintroduced into the path of data analysis in the form of altered α -carbon coordinates. All manipulations on the α -carbon coordinates are performed on isolated monomers. Subsequently, these newly modeled structures are used to reconstruct the tetramer by the same symmetry operations that yielded the original tetramer. The point at which the comparison between experimental and simulated data takes place is at the level of $P(r)$ distribution. The difference between solution conformations, $\Delta P(r)$ solution, and crystal conformations, $\Delta P(r)$ crystal, was compared; when $\Delta P(r)$ solution = $\Delta P(r)$ crystal, the conformational change was considered adequately modeled. It is especially useful for the comparison of data sets obtained under different solution conditions. Hence, $P(r)$ distributions were employed to illustrate changes in molecular dimensions that are the result of these solution variations.

Based on results from chemical studies (3) the hinge region between the A and B domains was chosen as a pivotal point for rotation of the B domain. The length distribution profile, as shown in Figure 3, indicates that a molecular contraction results from the rotation performed, as evidenced by the decreased maximum dimension and reduction of the shoulder seen at longer lengths (80-120 Å). These features of the length distribution profile are very similar to those obtained experimentally by SANS (Figure 1). The difference $P(r)$ plots shown in Figures 1 and 3 illustrate this. It is seen that the peak representing the difference between the two solution conformations of PK is nearly identical with that of the crystal and computer modeled form. Hence, it is concluded that a rotation of the B domain can qualitatively account for the change in conformation observed by SANS.

Structural Organization of Pyruvate Kinase

The results from SANS experiments and computer modeling indicate that the change in the hydrodynamic properties can be characterized by a rotation of the

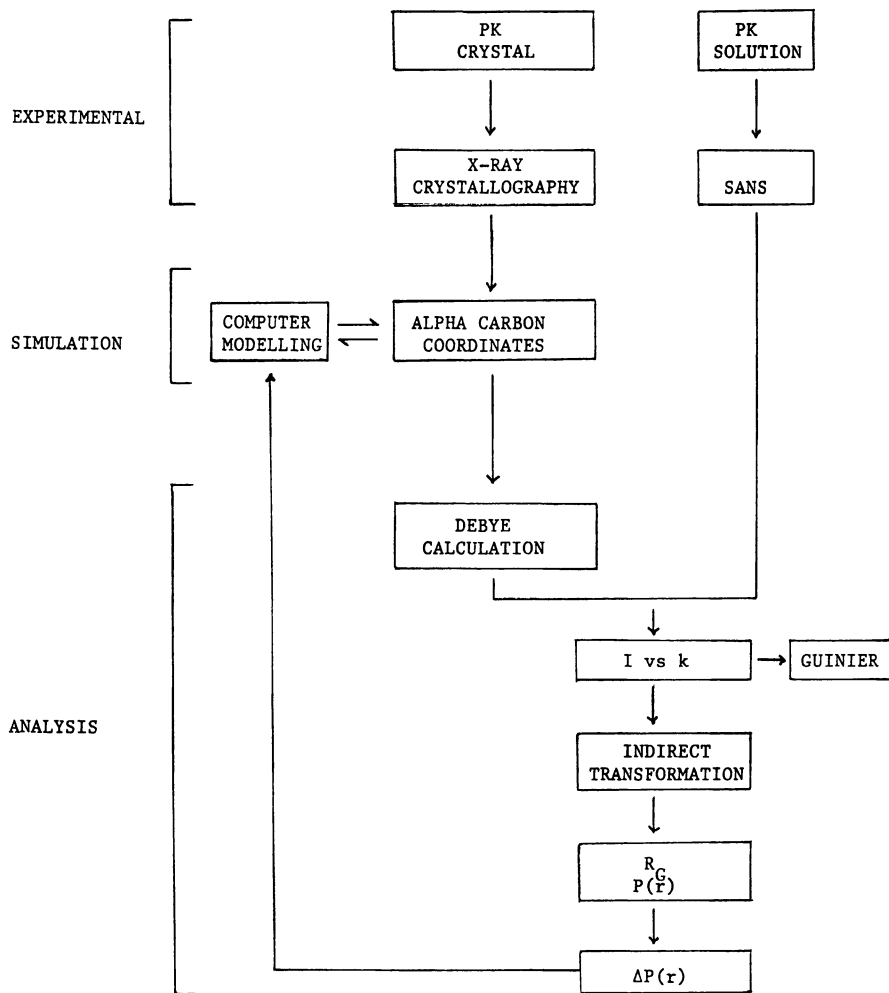


Figure 2. Algorithm for comparison of solution and crystalline structure. (Reproduced with permission from the ref. 4. Copyright 1988 American Society for Biochemistry and Molecular Biology, Inc.)

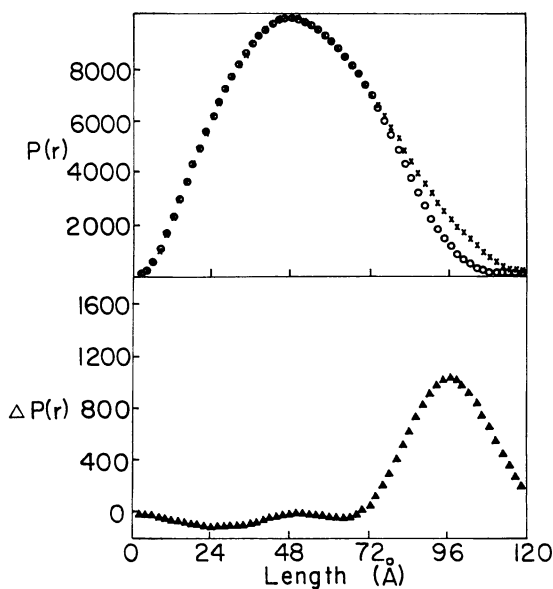


Figure 3. Comparison between active and inactive pyruvate kinase conformation. Upper, length distribution functions; lower, difference length distribution function. X, enzyme crystallographic data; o, enzyme with 35° rotation of the B domain along the X axis. (Reproduced with permission from ref. 4. Copyright 1988 American Society for Biochemistry and Molecular Biology, Inc.)

B domain relative to the A domain in each subunit. PEP and the cations, Mg^{++} and K^+ , cause the domain to move together, "priming" the enzyme for catalytic activity. Phe and high pH, on the other hand, cause this interdomain to open up, rendering PK inactive. However, these domain movements can only represent the local structural changes around the active site in each subunit. The domains involved in the cleft closure do not participate in intersubunit contact directly. It does not explain the cooperative behavior in the structural change of PK as indicated by the observation that binding one Phe molecule per PK tetramer can induce approximately 80% of the gross structural changes (1). Hence, additional conformational changes must accompany the cleft movement and these must involve intersubunit contact sites so that communication is transmitted among the four subunits. In order to identify the structural feature that is involved in intersubunit contacts, the crystalline structure of cat muscle PK (10-12) will be examined by distance matrix analysis.

The method of distance matrix analysis has been described in detail elsewhere (18-20), both in its general applications to the representation of a single protein, including the identification of structural domains and their interactions within the protein, and its application to the analysis of inter-macromolecular complexes. The first application is utilized to represent and examine the domain organization of PK. The second application is utilized to represent and examine the orientation and interaction between the individual subunits of the tetrameric form of PK whose structure is known from X-ray crystallography.

The basic construction of the distance matrix involves the generation of a square matrix, of order 'n', where 'n' is the total number of amino acids in the polypeptide. Each 'i-j' element of the matrix is filled with the inter-C(alpha) distance of the residues 'i' and 'j' in the amino acid sequence. This matrix is thus square and symmetric, and all elements which occur along the diagonal, when a polypeptide is compared with itself, are all 0.0. Similar matrices can be constructed involving two polypeptide chains that are observed in a specific orientation to study the organizational patterns which might be present. Typically these matrices are graphically represented by generating graphic elements (symbols) representing equi-distance contours within certain preset distance ranges, i.e., a different symbol or a different shading intensity for each range of r_{ij} values has been described (21). Typical boundaries selected for these values are 5.0 Å, 10.0 Å, and 15.0 Å, resulting in a contouring that highlights regions of contact within the structure that are close together in three-dimensional space, although they may be distant in amino acid sequence. Contours for the distance ranges given above will concentrate near the diagonal, with i and j of approximately equal in magnitude, and the patterns formed by the shading of these areas are indicative of secondary structural elements. Short distances between distant residues result from the tertiary folding of the protein. The contours that appear farther from the diagonal represent close distances between sequence-distant i and j residues and are, therefore, indicative of the tertiary structure and the folding of the protein. Regions contoured with intermediate density and located in contiguous positions along the diagonal were recognized as indications of folding domains. The same procedure can be extended to monitor intersubunit contacts as being described in the present case for PK.

Organization of the Structure of Pyruvate Kinase Monomer

The three-dimensional structure of cat muscle pyruvate kinase is represented in the distance matrix (Figure 4) based on the alpha carbon coordinates supplied by Professor Muirhead as an update of the file available within the Protein Data Bank (22). This figure has been computed to only present those inter-alpha carbon distances which are within 15.0 Å as a more detailed analysis of the structure awaits the atomic coordinates of the full structure. This analysis represents a low-resolution study which attempts to incorporate data from a variety of independent observations of the three-dimensional structure, amino acid sequence, and physical properties, both observed and computed. The monomer organization displays the characteristic patterns associated with internal folding domains as has been previously noted (12), with the domains readily identifiable by the clusters of contoured features which occur along the diagonal of the distance matrix (Figure 4). The boundaries previously proposed appear consistent with those suggested from the distance matrix using the guidelines outlined in the analysis of carboxypeptidase (23). Several important observations can be made concerning the organization of the monomer which are readily apparent in this form of representation:

I. The features which predominate in the distance matrix, in terms of frequency of observation, are the small, discontinuous features which occur at some distance from the diagonal and appear along lines that are horizontal or vertical in construct. These features are indicative of the orientational characteristics produced by the alpha-beta barrel structure which has been observed in PK, triose phosphate isomerase (TIM) and KDPG-aldolase (13, 24, 25). Use of the single-contour representation in Figures 4, and the large number of amino acids present in PK, the resolution of the contoured distance matrix precludes a detailed analysis of the secondary structural features which give rise to the interactions in the off-diagonal elements, however, even at such low resolution much information can be gleaned from such presentation. For example, the feature identified by the arrow in Figure 4, represents the interaction between amino acids whose sequence ranges are 440-465 and 90-105, which are a beta-alpha-beta and beta conformation, respectively. It can also be noted that this orientation of structure occurs in the interaction between the A1 and C domains, and is not the only potential interaction involving the 440-465 region of the sequence and residues in the A1 domain.

II. The features which appear adjacent to the diagonal and are more extensive (e.g., residues 40-110) represent secondary structural features which may, themselves, represent folding-domains or segments of such domains (e.g., domain A2 consists of segments 220-270, 271-340 and 341-389). Where the folding domain is large, as in A2, the same features as noted in (I) above, can be used to analyze the tertiary structure produced by the orientation of the separate structural elements, as observed for A2.

III. The secondary structure involved in producing the A-domain of PK (i.e., A1 and A2) consists of four structural features which do not appear to be identical in their respective conformations, although all are organized about an alpha-beta motif. This distinguishes the organization of the alpha-beta barrel of PK

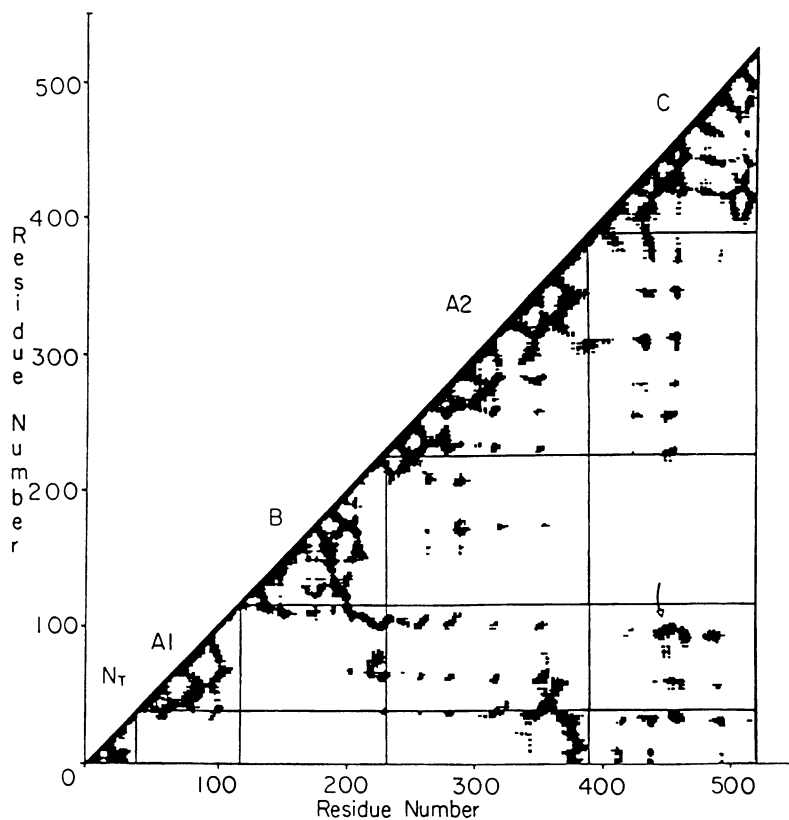


Figure 4. Intra-Subunit Distance Matrix Analysis for pyruvate kinase. This analysis is identical for all four subunits of pyruvate kinase. The contour level is 15 Å. The matrix is subdivided into domains, which are labeled along the diagonal; NT is the N-terminal domain, A1 is the first half of the A domain, B is the B domain, A2 is the second half of the A domain, and C is the C domain.

from that observed in TIM, where the secondary structure appears to repeat in an analogous manner among the segments involved, although altered in size if the respective segments (i.e., see TIM distance matrix in ref. 25).

IV. The feature of the distance matrix which depicts interactions between residues 330-389 and 1-60 represents a combination of two features - one common to PK, TIM and KDPG aldolase, and another unique to PK. The common feature is represented by the interaction between residues 40-60 and 340-370. It is indicative of the "closing" of the alpha-beta barrel such that the amino-terminal region of the barrel comes into contact with the carboxyl-terminal of the barrel. PK exhibits an additional interaction of this region with the amino-terminal tail of the intact polypeptide, residues 1-40, which is distinct from TIM and KDPG aldolase, and as discussed below, may bear significantly in establishing the cooperativity observed in some evolutionary forms of PK.

V. The region which extends between the second and third super-secondary structures of domain A2 appears to provide the contacts between the A2 and C domains of PK, through a helix-helix interaction. The actual feature of the distance matrix suggests that the helix which is at the amino-terminus of domain C (i.e., residues 389-405) actually is part of a helix which extends from residues 368-405 but is conformationally broken by a twist which occurs following residue 384. This will be shown to be an important distinction between the M_1 and M_2 isozymes from rat.

Organization of the Structure of Pyruvate Kinase Tetramer

The crystal structure of the tetramer of PK isolated from cat muscle has been shown to be a symmetrically exact tetramer (12), although presumably representing the inactive form of the enzyme. In this study the interfaces between individual subunits of the crystalline tetramer are examined to probe the source of the cooperativity which is observed in PK-Phe interactions (1) and in various evolutionary forms of PK (26). Thus the results and the analysis presented here, include the orientation and interaction between a single subunit and, successively, each of the other three subunits in the observed tetramer. The individual subunits of the tetramer are identified basing on the two-fold axis used in their generation from the original subunit structure. Thus subunit 2 has been generated using the two-fold axis that is parallel to the z-axis, subunit 3 uses that parallel to the x-axis and subunit 4 uses that parallel to the y-axis.

I. Interaction across the individual dimer interfaces within the tetramer are represented in the distance matrices computed as intermacromolecular distance matrices (27) shown in Figures 5-7. These figures utilize the same representation of the folding domains as used in Figure 4. Examination of the inter-subunit distance matrices reveals that the major dimer interface involves subunit 1 and 2, with the interaction between 1 and 4 also appearing potentially significant based on the orientation of the alpha carbon backbones.

II. The interface between subunits 1 and 2 primarily involves the direct contact between the A2-domains of the two subunits, as evidenced by the features which appear within the diagonal partition which corresponds to the A2-A2 domains (Figure 5). The interaction centers about the orientation of the structure

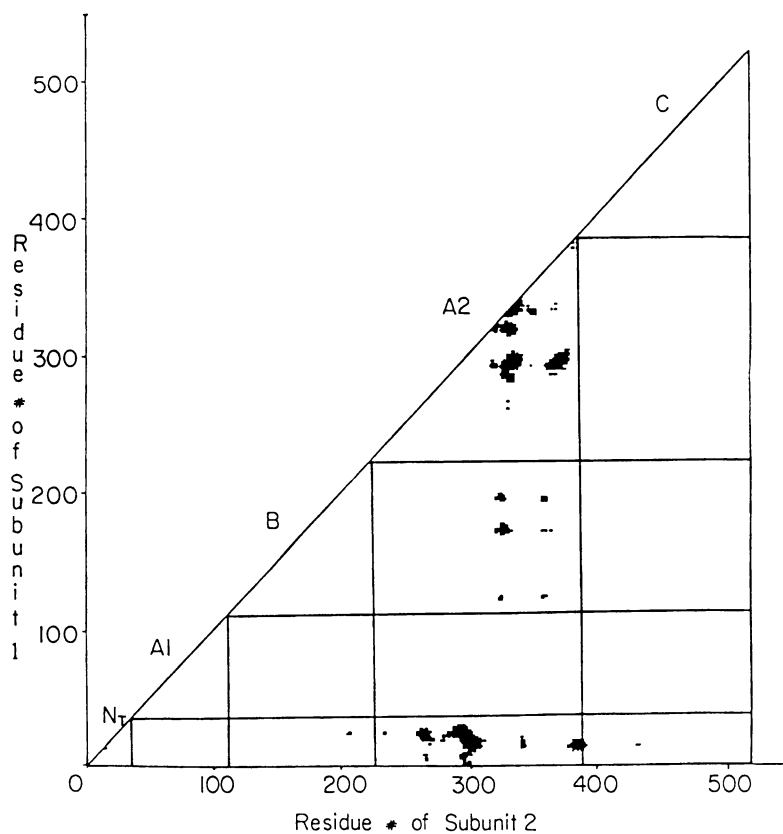


Figure 5. Inter-Subunit Distance Matrix Analysis for pyruvate kinase. This analysis is for subunits 1 and 2. Contour and subdivisions are as in Figure 4.

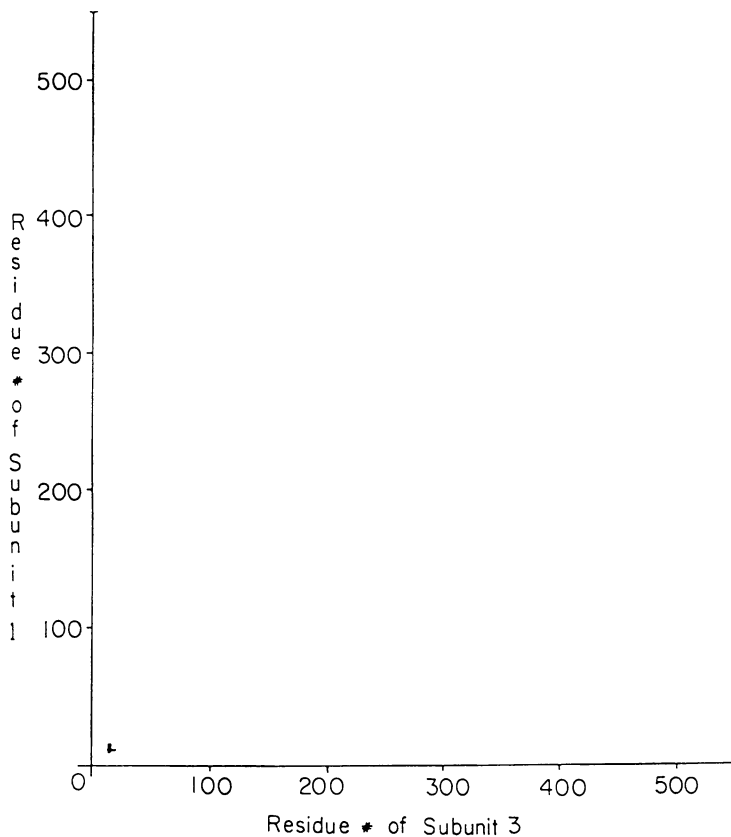


Figure 6. Inter-Subunit Distance Matrix Analysis for pyruvate kinase. This analysis is for subunits 1 and 3. Contour and subdivisions are as in Figure 4.

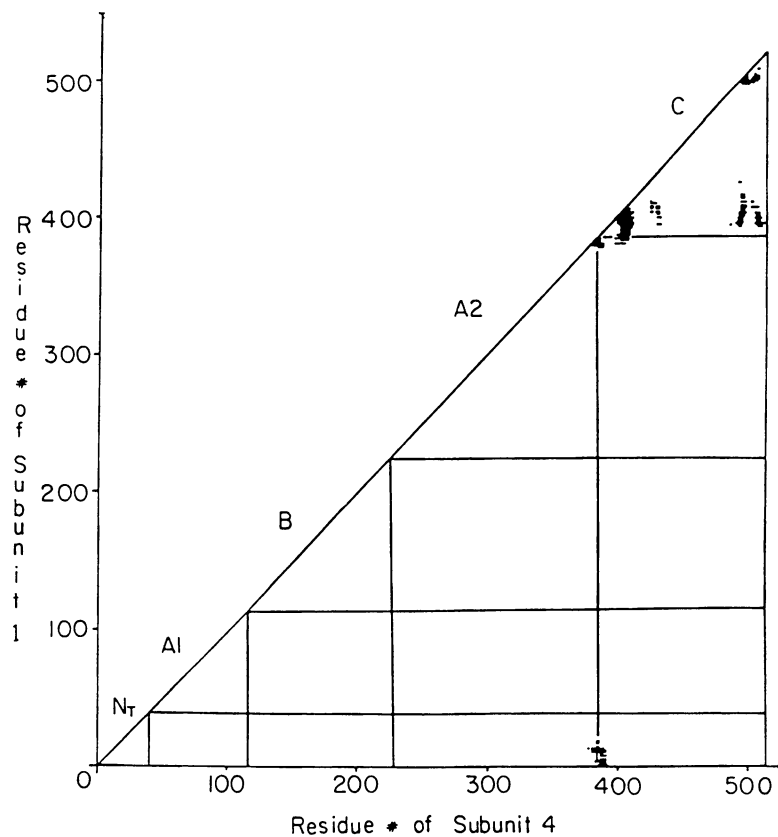


Figure 7. Inter-Subunit Distance Matrix Analysis for pyruvate kinase. This analysis is for subunits 1 and 4. Contour and subdivisions are as in Figure 4.

between residues 330-350 which includes a helical region beginning at residue 340. In addition, the amino-terminus of this helix appears oriented towards the B-domain of the opposite molecule. A second interaction involves a strand-strand interaction between residues 290-300 (A2-domain) and residues 10-30 of the N-domain of the opposing molecule. A potentially significant observation involves the helical region, residues 370-380, which form the carboxy-terminus of the A2-domain, and which potentially interact across the dimer 1/2 interface with the N-domain of the opposite molecule. This region of the subunit is also involved in the A2-C domain interface within the subunit itself.

III. The interface between subunits 1 and 3 (Figure 6) represents the potentially weaker inter-subunit interaction within the tetramer, involving only a small strand-strand orientation of the N-domains and few residues.

It should be mentioned that the first 9 residues were not well defined crystallographically and thus, were omitted from the coordinate file. These N terminal residues are part of the N terminal domain and could potentially enhance the interaction of this structural feature with others, especially between N terminal domains located on subunits 1 and 3, but also between this domain and the C α 1 and C α 2 of the opposing subunit in the dimer, as mentioned above in (II).

IV. The interface between subunits 1 and 4 (Figure 7) represents the second most extensive interface and almost solely involves the symmetrical orientation between the C-domains of the two subunits. Even within the C-domain interactions, the folding of the individual C-domains yields potential interaction only at the amino- and/or carboxy-terminal regions of these domains within the interface. It is notable, however, that a significant interaction may involve residues 370-380, the same region described above as interacting within the dimer 1-2 interface and between domains of the subunit. An extension of the interaction in this vicinity involves the distorted helix which extends from residues 380-400 and also provides for significant potential interaction in the dimer 1-4 interface. The interaction of this region is further suggested by its interaction with the N-terminal amino acids, again potentially linking inter-subunit and inter-domain regions of the tetramer. The strand-strand interaction in the interface, involving residues 500-520, is indicative of the C-domain structure as it is potentially interacting with the 380-400 helix already described, both within the subunit and in the dimer 1-4 interface.

Intersubunit Contacts and Potential Functional Roles.

Having identified the intersubunit contacts, it is interesting to assess the functional importance of these contacts. Hence, the structure function relations between pyruvate kinase isozymes are investigated. These isozymes are functionally identical and are different only due to small perturbations at key structural locations. As a consequence of these changes, the isozymes exhibit quantitatively different kinetic behaviors under the same experimental conditions. In this study a comparison between the rat muscle (M_1) and kidney (M_2) isozymes are made. Noguchi et al. (28) reported that M_1 and M_2 isozymes of rat PK are produced from the same gene by alternate RNA splicing. They sequenced the cDNA derived from both mRNA species and found that the coding regions were

identical, except for a 160 nucleotide stretch. This corresponds to 54 amino acid residues, 21 of which differ between the isozymes. These residues are located in the C domain and comprise the C α 1 and C α 2 helical segments. These are precisely the units of secondary structure that are demonstrated to be prominently involved in intersubunit contact between subunits 1 and 2 and subunits 1 and 4, as was discussed for the cat M₁ structure.

Using only the sequence data and the algorithm of Chou and Fasman (29), one can perform calculations in an effort to predict secondary structure. Results of such calculations on the sequence of the cat muscle enzyme consistently agree with the assignments of secondary structure made for this isozyme. Focusing on the region of sequence that differs in the M₁ and M₂ isozymes, computational results show that the structure spanning residues 371 to 396 can be represented by an extended alpha helical segment interrupted by beta strand precisely at the interdomain region (Figures 8 A and B). This helical region lies at the interface between the A- and C-domains, and comprises the units of secondary structure A α B and C α 1. In addition to breaking the helix at the domain interface, the M₂ sequence is predicted to have an extended C α 1 helix, as 2 additional residues (399-400) are predicted to be alpha helix (Figure 8B). The end of C α 1 is converted to random coil in the M₂ isozyme, primarily due to the change of residue 402 from serine to proline. This change may allow these helices of the M₂ isozyme to have additional orientational freedom in the subunit interface. C α 2 is strengthened in the M₂ isozyme, as 3 residues to this helix (412, 414-415) are predicted to be alpha helix, whereas they were predicted to be beta turns in the M₁ sequence. The turn following C α 2 is also strengthened in the M₂ isozyme, as 5 residues (422-426) that were helical in the M₁ isozyme are converted to beta turn in the M₂ sequence. This change is probably due to the conversion of a leucine to a cysteine residue (423). The overall prediction from these calculations is that in the M₂ isozyme, C α 1 is made more distinct from A α B at the A/C domain interface; C α 2 is enriched in helical propensity; and the residues following C α 2 have more propensity to form a beta turn. Obviously, the secondary structural analysis serves as a working hypothesis to provide a focal point to focus our future effort to determine experimentally the structures of both PK isozymes.

Having identified the structural differences in these isozymes, it is possible to provide a rationale for the differences in the kinetic behavior of these two isozymes. The fact that M₁ and M₂ have different allosteric properties as detected by kinetic methods has been known for some time (for a review, see 30). M₂ isozyme exhibits a sigmoidal dependency on substrate concentration, whereas it is hyperbolic for the M₁ isozyme. M₂ is subjected to allosteric activation by fructose 1,6 bisphosphate (FBP), but M₁ is insensitive to this activator in the absence of any other effectors.

Based upon earlier studies of M₁ isozymes from this laboratory, it was shown that PK can exist in two alternate conformations and indeed conforms to a two-state allosteric model (2). The usual hyperbolic kinetic behavior of M₁ can be shifted to a sigmoidal relation by the addition of the allosteric inhibitor, Phe. This kinetic behavior is the consequence of the equilibrium constant that governs the distribution of two states for M₁. This equilibrium constant is almost entirely in favor of the active state in the absence of allosteric effectors. Phe, pH, or

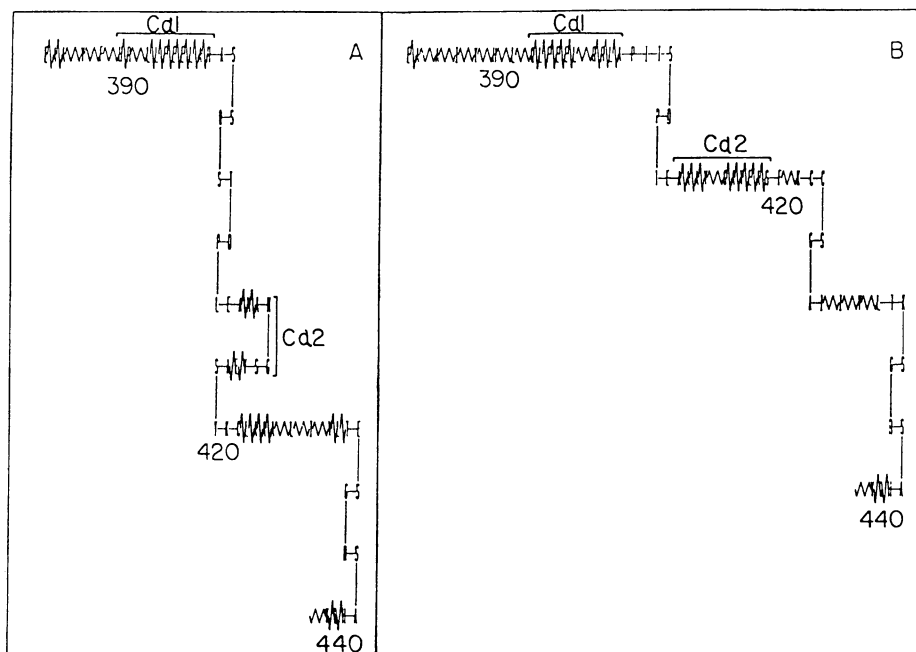


Figure 8. Predicted Secondary Structure for the Amino Acid sequence in the region of intersubunit contact for (A) rat M₁ pyruvate kinase and (B) rat M₂ pyruvate kinase. Alpha helical segments are represented by wide, curved lines. Beta strands are represented as elongated, zig-zag lines. Beta turns are represented as vertical lines. Random coil is represented as horizontal lines. Amino acid residue number is annotated below the structural representation. Important units of secondary structure are labelled. The structural representations were drawn with the aid of the computer program MSEQ.

temperature can be used to perturb this equilibrium, so as to favor the formation of the inactive form, thus allowing the allosteric properties to be exhibited. For M_2 , however, the equilibrium between the two states is conceivably poised closer to the inactive state, thus, even in the absence of inhibitors, the cooperative behavior can be observed. Hence, the difference between these two isozymes is most likely due to the difference in the equilibrium constant governing the distribution of the active and inactive forms of PK. As has been shown, the only difference between these two proteins, structurally, is one of the key structural units involved in intersubunit contact. This region must play an important role in the establishment of the equilibrium between alternate conformations of the enzyme. Subsequent study shows that indeed the various equilibrium constants governing the interactions between ligands and the enzyme are different between these two isozymes (5).

In developing these proposals, it is assumed that the structures of these isozymes are very similar and that units of structure, i.e., secondary, tertiary, and quaternary, will occupy relatively the same position in all PK isozymes. This assumption is most likely valid due to the high degree of sequence homology between the M_1 and M_2 PK isozymes. Even for the other isozymes, all evidence seems to indicate that the structure of the vertebrate isozymes of PK is species independent. Peptide and sequence analysis indicates strong homology between species as different as cat and trout (31), although these two isozymes have different kinetic properties. Hence, the primary assumption in this discussion is valid, thus allowing an initial attempt to identify structural differences between the isozymes and to correlate these distinctions with the characteristic kinetic properties of each isozyme.

If the N-terminal sequence and $C\alpha_1$, $C\alpha_2$ are intersubunit contacts intimately associated with isomerization between conformational states, then one can expect that any perturbation in those contacts will affect the allosteric properties of PK. Actually, there are two reports that support such a proposal. The liver isozyme activity is subjected to regulation by phosphorylation (for a review, see 32). As a result of this covalent modification, the properties of the enzyme are altered. The kinetic activity of phosphorylated PK is inhibited compared to its unphosphorylated predecessor, but the activating effect of FBP is maintained. The site of phosphorylation has been elucidated by isolating and sequencing a phosphorylated peptide (33). A serine residue 12 amino acid residues from the amino terminus of liver PK is the only site of phosphorylation. Comparison of amino acid sequences of PK isozymes reveals that M_1 and M_2 isozymes lack the first 13 amino acid residues found in the liver isozyme, and thus lack the phosphorylation site. Neither M_1 nor M_2 have been shown to be phosphorylated. In liver PK, the phosphorylated site is close in primary structure to the N-terminal domain, which has been shown to be an important intersubunit contact site. This phosphorylation then can be implicated in altering the intersubunit communication. The altered subunit communication is manifested as shifted into a sigmoidal dependency on substrate concentration, thus reflecting an inhibition of activity at lower substrate concentrations. The fact that FBP activates the phosphorylated form of liver PK to the same extent as dephosphorylated PK shows that the covalent modification does not irreversibly inhibit the enzyme. In

fact, it argues for the existence of at least two conformations of the enzyme, phosphorylation merely shifts the equilibrium between the two states toward the "inactive" form. Again, one of the key structural units in intersubunit communication, namely, the N-terminal domain, is shown to be of importance in the determination of the characteristic regulatory properties of an isozyme. Another report shows that the M_1 isozyme can be kinetically altered by treatment with an acid extract of rabbit liver (34). This modification results in a PK which exhibits a different dependency of enzyme activity on substrate concentration: it is shifted from hyperbolic to sigmoidal. The conversion in kinetic activity is concomitant with the proteolytic removal of a peptide (or peptides) totalling MW of 1,500, as revealed by change in migration rate of the subunit on SDS-PAGE. The site(s) of cleavage must be at the termini since the predominant peptide observed has an apparent MW of $\sim 55,000$, whereas the subunit MW of PK is 57,000. As has been pointed out, both the N and C-termini are in close proximity of each other. This change in kinetic activity is similar, qualitatively, to that seen with M_1 in the presence of L-phenylalanine, an allosteric inhibitor (1). The altered kinetics also resemble those exhibited normally by liver and M_2 isozymes. The evidence again is consistent with the proposal that these intersubunit contacts are involved in intersubunit communication and most likely play an important role in determining the equilibrium distribution of different conformational states of PK.

Besides the intersubunit contacts described, there is another interesting and potentially critical region of intersubunit contact within a dimer. It involves only a few residues, but these residues are part of the N-terminal domain of subunit 1 and precisely the hinge region of domain B of subunit 2. The hinge region is proposed to be involved in the isomerization between active and inactive forms (3, 4). Thus, the N-terminal domain, having been described as potentially an important structural unit involved in intersubunit contact, is directly adjacent to the site of a major conformational change of its neighboring subunit in the dimer. The two subunits in a dimer are thus linked by structural features involved in the conformational change at the active site and the contact between subunits.

Conclusion

The evidence presented herein, taken together, supports the hypothesis that the structural elements constituting the intersubunit contact regions of PK play critical roles in the regulation of PK activity. These elements of structure being the N-terminal domain (residues 10-40) and the two alpha-helices of the C domain (residues 384-425). Each subunit communicates to its neighboring subunits via these structural units, in response to the binding of an effector or substrate molecule, which triggers a conformational change in that subunit. This intersubunit communication must be propagated towards the active site in order for the regulatory behavior to be detected. A conformational change has been described involving the active site region of rabbit muscle PK (3, 4). How is the intersubunit communication related to this conformational change at the active site? At present, there is no specific information to elucidate the pathway of communication, although potentially a direct transmission between the N-terminal domain and B domain is possible. It is evident much investigation is needed to elucidate the molecular mechanism of allosteric regulation of PK.

Acknowledgments

This work was supported by U.S. Public Service Grants DK-21489 and GM-45579 and by The Robert A. Welch Foundation grants H-0013 and H-1238.

Literature Cited

1. Oberfelder, R.W.; Lee, L.L.-Y. and Lee, J.C. *Biochemistry*, **1984**, *23*, 3813-3821.
2. Oberfelder, R.W.; Barisas, G. and Lee, J.C. *Biochemistry*, **1984**, *23*, 3822-3826.
3. Consler, T.G. and Lee, J.C. *J. Biol. Chem.*, **1988**, *263*, 2787-2793.
4. Consler, T.G.; Uberbacher, E.C.; Bunick, G.J.; Liebman, M.N. and Lee, J.C. *J. Biol. Chem.*, **1988**, *263*, 2794-2801.
5. Consler, T.G.; Woodard, S.H. and Lee, J.C. *Biochemistry*, **1989**, *28*, 8756-8764.
6. Consler, T.G.; Jennewein, M.J.; Cai, G.-Z. and Lee, J.C. *Biochemistry*, **1990**, *29*, 10765-10771.
7. Consler, T.G.; Jennewein, M.J.; Cai, G.-Z. and Lee, J.C. *Biochemistry*, **1992**, *31*, 7870-7878.
8. Heyduk, E.; Heyduk, T. and Lee, J.C. *J. Biol. Chem.*, **1992**, *267*, 3200-3204.
9. Muirhead, H. *Trends Biochem. Sci.*, **1983**, *8*, 326-330.
10. Stammers, D.K. and Muirhead, H. *J. Mol. Biol.*, **1975**, *95*, 213-225.
11. Stammers, D.K. and Muirhead, H. *J. Mol. Biol.*, **1977**, *112*, 309-316.
12. Stuart, D.I.; Levine, M.; Muirhead, H. and Stammers, D.K. *J. Mol. Biol.*, **1979**, *134*, 109-142.
13. Muirhead, H.; Clayden, D.A.; Barford, D.; Lorimer, C.G.; Fothergill-Gilmore, L.A.; Schiltz, E. and Schmitt, W. *EMBO J.*, **1986**, *5*, 475-481.
14. Phillips, F.C. and Ainsworth, S. *Int. J. Biochem.*, **1977**, *8*, 729-735.
15. Lonberg, N. and Gilbert, W. *Cell*, **1985**, *40*, 81-90.
16. Moore, P.J. *J. Appl. Cryst.*, **1980**, *13*, 168-175.
17. Debye, P. *Ann. Phys. (Leipzig)*, **1915**, *46*, 809-823.
18. Liebman, M.N. *Ph.D. Thesis*, Department of Chemistry, Michigan State University, **1977**.
19. Liebman, M.N. and Weinstein, H. In *Structure and Motion: Membranes, Nucleic Acids and Proteins*; Clementi, E., Corongin, G., Sarma, M.K. and Sarma, R.H., ed.; Adenine, Guilderland, New York; 339-359.
20. Rossmann, M.G. and Argos, P. *J. Biol. Chem.*, **1975**, *250*, 7525-7532.
21. Ooi, T. and Nishikawa, K. In *Conformation of Biological Molecules and Polymers*; Bergman, E.D. and Pullman, B., ed.; Academic Press, New York; 173-187.
22. Brenstein, F.C.; Koetzle, T.F.; Williams, G.J.B.; Meyer, E.F. Jr.; Brice,

- M.D.; Rodgers, J.R.; Kennard, O.; Shimanouchi, T. and Tasumi, M. *J. Mol. Biol.*, **1977**, *112*, 535-542.
23. Liebman, M.N.; Venanzi, C.A. and Weinstein, H. *Biopolymers*, **1985**, *24*, 1721-1758.
24. Fothergill-Gilmore, L.A. *Trends Biochem. Sci.*, **1986**, *11*, 47-51.
25. Lebioda, L.; Hatada, M.H.; Tulinsky, A. and Mavridis, I.M. *J. Mol. Biol.*, **1982**, *162*, 445-458.
26. Hall, E.R. and Cottam, G.L. *Int. J. Biochem.*, **1978**, *9*, 785-793.
27. Liebman, M.N. *Enzyme*, **1986**, *36*, 115-140.
28. Noguchi, T.; Inoue, H. and Tanaka, T. *J. Biol. Chem.*, **1986**, *261*, 13807-13812.
29. Chou, P.Y. and Fasman, G.D. *Ann. Rev. Biochem.*, **1978**, *47*, 251-276.
30. Imamura, K. and Tanaka, T. *Methods Enzymol.*, **1982**, *90*, 150-165.
31. Harkins, R.N.; Nocton, J.C.; Russell, M.P.; Fothergill, L.A. and Muirhead, H. *Eur. J. Biochem.*, **1983**, *136*, 341-346.
32. Engstrom, L. *Curr Top. Cell. Reg.*, **1978**, *13*, 29-51.
33. Hjelmquist, G.; Andersson, J.; Edlund, B. and Engstrom, L. *Biochem. Biophys. Res. Commun.*, **1974**, *61*, 559-563.
34. Fujii, Y.; Kobashi, K. and Nakai, N. *Arch. Biochem. Biophys.*, **1984**, *233*, 310-313.

RECEIVED January 13, 1994

Chapter 26

Representation of Biochemistry for Modeling Organisms

Toni Kazic

**Institute for Biomedical Computing, Washington University,
St. Louis, MO 63110**

Before one makes a database, one must needs understand the fundamental structure of that portion of the “real world” the database is intended to model. This understanding inevitably guides decisions on representation and implementation, so its fidelity to reality is critical: an accurate model is easier to change as knowledge evolves, and appropriate representational choices simplify the process of database revision. In this paper I describe the basic representational principles we have reached in our attempt to model portions of cellular biochemistry, and sketch some of their consequences for representation and our implementations.

From physiology to behavior, biology arises from functions determined by molecular interactions. Structural studies, whether of molecules, organisms, or ecosystems, are motivated by “what does it do?” Mentally constructing an integrated view of structure and function for small, isolated systems is relatively simple because there are few components and interactions. But mental modeling techniques, even with the assistance of paper, are not easily transferred to systems of tens or hundreds of species and reactions. Though the test of understanding is accurate prediction of mechanism and experimental outcomes, the number of components and the emergence of new phenomena militates against the success of attempts at even modest systems. Yet experience in modeling increasingly complex systems is likely both to improve performance in prediction and to help discover new biological principles. It is also critical to speeding the design of new biological entities for specific purposes, such as the amelioration of disease, the production of useful molecules, and the improvement of food crops.

Biological phenomena and our conceptions about them are extremely complex, so building appropriate databases is not a trivial goal (1). Databases intended to support modeling of complex phenomena need not only to describe structures, but also to represent functions and to automate reasoning. It is clearly advantageous if the database can carry out all three tasks as seamlessly and easily as trained biologists, the user unaware of any computational distinctions among them. Pre-

0097-6156/94/0576-0486\$08.00/0
© 1994 American Chemical Society

dicting the metabolic fate of a compound or the metabolic changes produced by an altered enzyme — examples of high-level inferences — can be thought of as queries to databases which contain both functional and structural information on the metabolic machine. Since the database contains the rules for computing the queries, it becomes coextensive with the model (and I will use the nouns database and model interchangeably). Direct experience building more restricted databases is a good way to learn how more complicated systems can be realized.

The first step is to design the database's representations. By *representation* I mean how information or ideas are expressed in the database. For example, the word "hand" is a token which stands for the appendage below the wrist. A database might represent the appendage by the token ("hand"), a structural description ("four fingers and a thumb"), or a functional one ("grips"). An anatomical database might have images of the hand and its underlying structures, while a radiology database might include X-rays and MRI scans of individuals' hands. All of these are valid representations for the appendage, and are clearly linked by fundamental relationships. Databases with more than one representation of the appendage might also describe the relationships among them, so that users and programs could reliably retrieve that representation most suitable for a particular computation (1).

We have tackled the representational problem by considering what data are required to answer a class of queries, beginning with tracing the atoms of a compound through its products in a metabolic pathway. This question is a simplification of *in vivo* metabolic tracer experiments, and requires representations of the compounds' structures, their reactions, and the enzymes' specificities. We have focused on biochemistry first because understanding molecular interactions is fundamental to our modern synthesis of biology and because it offers a wide range of data for testing. Here I briefly sketch some of the most fundamental aspects of the representations; more complete descriptions are in preparation. Since the biochemical questions are ultimately about the fates of molecules, we have named the project *Moirai*, after the trio of Greek Fates described by Hesiod (2,3). *Klotho*, who cards and spins the wool, includes the structural descriptions of the compounds as raw material; *Atropos*, who weaves the fabric of life, includes the reactions; and *Lachesis*, who bounds life by cutting the thread, includes the constraints and dynamics of the system of reactions.

The Representational Rationale

The first question one faces when trying to integrate many different types of information is how they are related, both to each other and to the phenomena reflected in the information. Answering this question constructs a coherent and consistent abstraction of the universe of discourse, the subject under discussion. Ideally, the abstraction should be homologous to the inherent organization of the universe of discourse without prejudice. We have interpreted this desiderium to mean that the model should store basic data — the "facts" — separately from the classification schemes, generalizations, or opinions of the community; and that the latter category of ideas should be explicitly represented in their own right, rather than folded into the basic data (1,4). The abstraction *per se* can be implemented in any computer language.

The Structure of the Informational Dimensions. We have attempted to organize the model to correspond with those human ideas and experimental results which describe cellular physiology; loosely, the cell's "viewpoint". We divide the

cell into three fundamental dimensions of information: *structures* — any molecule or a portion thereof; *transformations* — any reaction or process, or groups of reactions or processes; and *constraints* — terms which summarize the results of physiological, structure-function, or kinetic studies, and which are used to restrict computations on the other two dimensions. For each aspect, there may be several interconvertible representations of the same information, each optimized for a particular class of computation (1). In essence, the molecules and processes form a multidimensional network. If we imagine that structures are the nodes, transformations the arcs¹, and constraints the information attached to nodes and arcs, the network is multidimensional in two senses. First, the number of arcs for each node (its *degree*) can be quite high; second, the network evolves over time and physiological state. This evolution will include both changes in parameters associated with nodes and arcs and in the topological structure of the network.

A consequence of this view is that all structures and transformations are equally important. Reactions are described independently of their position in the traditional groupings of biochemistry textbooks and classification schemes. Thus, there are no pointers from one reaction to the “next”, as occur in some databases (5–7). The groupings arise from historical accident, such as the naming of pathways, or from the application of classification schemes, such as the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology’s (8) or Walsh’s (9). But what matters to the cell is not organized human convenience, but the logical, functional, spatial, and temporal associations among molecules and reactions. We store information on the composition of the traditional groupings as desirable. Flattening the hierarchical, tree-like groupings acknowledges that in cells all the reactions of a given physiological state coexist with molecules equilibrating among the various pools, reflecting more accurately the dynamics of subcellular processes. The flattening also reduces redundancy in the network, since reactions which are repeated in many groupings now need only be represented once. The database becomes a more accurate model of the cell’s biochemical organization, and allows one to test hypotheses about the latter’s inherent structure and the consequences of a particular arrangement. Finally, flattening reduces the use of implicit, subjective information in the model, providing an improved data substrate to test any classification scheme (10).

Expressing Relationships among the Elements. Yet if flattening allows us to discover more, it also forces us to think clearly how molecules and reactions are characterized. After all, humans classify and group for good reason — to discover common motifs, to simplify masses of data, and to infer from simpler to more complex systems. There are two basic relationships we need to express: the part-whole (enzyme X is composed of two subunits each of polypeptides A and B); and the collection (redox reactions). For both, we describe the underlying structural or functional properties, and then use *rules* — logical specifications of the answer’s requirements — to display molecules or reactions with these relationships.

Recursion in the Architecture of the Model. We use recursive data structures to express part-whole relationships. Intuitively we know that larger entities are built of smaller ones, and experimentally we seek to reduce a phenomenon to the simplest elements possible. If one diagrams part-whole relationships, they resemble trees, with the whole being the root, and the leaves being the parts. Usually the parts are arranged in subassemblies, so that the tree is hierarchical.

¹Making transformations the nodes and structures the arcs is equally valid and does not effect the discussion. The one used corresponds with the way pathways are usually drawn.

The essence of recursion is to describe an object in terms of itself: for example, trees in terms of branches and leaves. The object can be either an algorithm for a computation or a data structure. Recursion can be applied to any object if one can define a condition for stopping and a condition for continuing on to the next round. To fulfill these conditions, the object must consist of subobjects which are smaller in some sense than the object, and the structure of at least one subobject is identical to that of the object. The "smaller" criterion contributes to the stopping condition by guaranteeing the object will terminate at some finite smallest object rather than tunneling indefinitely through ever-smaller objects, while the "identity" criterion allows the computation to continue until the object is fully expanded by providing the transition to the next instance of the object. For example, if a molecule is built of smaller molecules, substituent groups and atoms, or an enzymatic reaction of ligand binding, catalytic and product-release steps, then these criteria will be fulfilled. While necessary, the criteria are not sufficient to specify the stopping condition. The stopping condition takes advantage of the fact that any object can be said to be built of two subobjects — itself and an empty object. When used in other operations, the empty object is null in the sense that the other object will be produced. Thus 0 and the empty set are empty objects for addition/subtraction and set union/intersection, respectively. Hence one can define a stopping condition as the empty molecule or reaction. Two consequences of using empty objects are that any object can be described recursively, and that the parts of an object need not be identical to each other for the recursion to be successful.

The part-whole relationships lead naturally to the idea of hierarchies: of substituent groups, compounds, and macromolecular complexes; and reactions, pathways and processes. Recursion is the fundamental tool used to express and navigate these hierarchies. The elements of the hierarchies — the leaves on the trees — are still the nodes and arcs of the network, so that the hierarchies are an additional layer which is used only as needed. For example, enzymatic reactions are frequently broken down into a series of biochemical steps; the overall reaction, or root of the tree, stands "higher" than the details of substrate binding or proton rearrangements. There are some interesting and useful consequences of these optional hierarchies, which will be treated more extensively elsewhere (Dunford-Shore and Kazic, and Kazic, in preparation); for the moment it suffices to say that a wide variety of computations can exploit these natural hierarchies without prejudicing other computations which would find them a hindrance. For example, it provides one way of navigating through the network without interfering with rules designed to recognize functional patterns (*metabolic motifs*). In general, recursion is the fundamental technique for expressing the relationship of a part to the whole in a hierarchy. Since so much of biology is hierarchical, it makes sense to choose a representation which can express this in the simplest and most natural way, and which can exploit recursion in manipulating information. Using recursion to express the hierarchical relationships among cellular components and processes is a good example of making the model homologous to the universe of discourse.

Forming Collections. The second relationship is that of the collection defined by user criteria. These can incorporate a classification scheme (all phosphorylation reactions); a navigational description (all noncyclic routes involving five or fewer enzymatic steps from 1,3-bisphosphoglycerate to succinate excluding glucose); or a query (how many compounds share precursors with lysine). These collections can be formed on the basic nodes and arcs of the network, on its hierarchies, or on any combination of the two. The operations which specify the

collection need not be restricted to Boolean, or even second-order logical operations: a wide variety of computations, from numerical algorithms to neural nets, can be used. The specification rule serves as a high-level, uniform interface to a collection of specialized computational devices, which if desired can be invisible to the user. Further, the form of the user interface is not restricted to a command-line interface, but can include graphical or hypertext presentations. The specifications provide navigational and discovery tools limited only by the imagination of the users.

Extension and Intension. If we model something well, then applying the model of its functions to the model of its structure should produce an accurate model of the result. The consequence of this inference property is that one need not explicitly encode every possible combination of transformation and structure, but can use the data and rules to generate these combinations, just as one could write a program which inferred the number and shape of the faces of a Platonic solid from a fact giving the number of vertices. One can distinguish between extensional representations of information, in which a fact is explicitly encoded, and intensional representations, in which a fact is represented by a rule and a list of basic facts used to compute the fact. Although we represent many facts extensionally, many others use a rule. For example pathways and processes are indicated implicitly, and their component hierarchies — much of these also implicit — are navigated by a rule. Another rule computes the overall reaction using the collected components. Similarly, our sample query of tracing the atoms uses brief descriptions of the substrates, the enzyme's specificity, and the type of chemistry, and rules involved to intensionally specify the mechanism of an enzymatic reaction and to compute the product structures.

Methods

In principle, representations using these operations can be achieved with many combinations of database management systems and computer languages. However hybrid systems are extremely clumsy, particularly if many different data types, algorithms, or queries must be integrated. This clumsiness is usually called impedance mismatch, and is very expensive both in expressive power and computational ease (11). We have chosen instead to integrate data and queries in a single high-level language, Prolog. Prolog is a declarative language which permits us to simply write complex queries by specifying the criteria the result must fulfill, rather than giving a procedure for finding the answer. The queries operate on the data encoded in the same language; data representations are both extensional and intensional, and the distinction is invisible to the user. Using a single language helps us to maintain symmetry in the representations, so that the same code can use facts or rules about molecules, reactions and constraints in any type of computation. Constructing a complete cellular model is a considerable task, but building intermediate models on the same foundation considerably shortens the development time for more complex models and queries.

Results

We have been experimenting with alternative representations for compounds, reactions and constraints in Prolog, testing them by comparing the results of our computations to the known biochemistry. To begin with we have focused on glycolysis, but the representations developed are quite general. We have developed

a database of stereochemically correct compound structures in several representations, each appropriate to specific reasoning tasks, and a layered graph grammar to transform them (*Klotho*). We are revising and expanding a previous prototype database of carbohydrate metabolism (12) to reflect an improved understanding of the relationships among various representations of reactions, and are implementing representations which express the biochemical and chemical mechanisms for product structure prediction (*Atropos*).

Representation of Compounds. Since we need to recognize and manipulate the relevant substituent groups in order to express biochemical function, we have focused first on compounds. For a preliminary account see (4); a fuller treatment is in preparation (Dunford-Shore and Kazic). The closer the correspondence between the natural biochemical language and the representations, the more easily the latter will communicate with mortals; hence the representation is based on existing biochemical language and concepts. For compounds, the guiding principle is that they are composed of hierarchical substituent groups: clusters of atoms which have distinct chemical properties and which are manipulated by the enzymes according to the "rules" of organic chemistry. Expressing compound structures in terms of substituent groups thus leads naturally to the description of the chemistry. Indeed, the narrative language contains many terms "Xylate" which translate "attach X group to a compound" (4). Since more than one copy of a particular group can exist in a compound, the enzyme catalyzing the reaction imposes specificity constraints. For example, in the phosphorylation of glucose ATP has three phosphates and glucose five hydroxyls; but only the γ -phosphate and the C₆ hydroxyl are involved in the reaction. The representation has been designed to allow us to point easily to particular groups.

The compound is briefly described in terms of its overall configuration (chain, ring or ring system) and substituent groups in a configuration fact. This representation allows simple, high-level specification of many compounds, and can be used to represent features which are common to classes of compounds while including specification of the variable groups or structures; thus coding of Markush structures is elementary. Since humans are trained to view compounds in this way, the representation effortlessly mirrors most of the common views. The configuration fact is expanded by the transformational graph grammar to produce several alternative representations, including a complete specification of the compound as a series of terminals detailing the atom-by-atom, bond-by-bond connections. For compound classes, the result is the generation of all compounds conforming to the configuration fact, with the exception that variable groups and bonds are expressed. This gives one route to the searching and recognition of Markush structures. The various representations of compounds and the grammar form the *Klotho* component of the system.

Grammars for natural languages such as English specify the structural relationships among sentence elements such as nouns and verbs, and are used both to generate and recognize well-formed sentences. Similarly, there are rules which govern how molecules are put together: valency, the notion of polymerization, the relationships among the various enantiomers, etc. Obviously chemical compounds can be considered as mathematical, connect-the-dots type graphs, and many authors use these ideas in their representations (13–15 are randomly chosen examples). However in most instances the graph is stored as one or more connection tables and manipulated by matrix operations. What is different here is that the graph is severely contracted to its most basic elements, and then manipulated by a grammar which incorporates the pertinent chemical and nomenclatural rules.

This grammar is a graph grammar since it encodes formal operations on a graph; we were unaware of the graph grammar literature when we started (see 16 for a formal introduction to this area), though we did consciously begin with work of Searls (17). Using a grammar to expand and manipulate compound representations is exactly using a set of rules to navigate a hierarchy, and has all the advantages described above: here the hierarchy is composed of the structural elements of the compound. For example the grammar explicitly traces the relationships among atoms through all transformations, so atoms need not be "labeled" as they are physically in tracer experiments.

Representation of Reactions. We have been using our previous prototypic model of *E. coli* genetics, biochemistry and physiology (12) as a testbed for representation and algorithm development. It contains reactions in the anaerobic catabolism of glucose and its derivatives, and related reactions of inhibition, activation, transport, and regulation. Present historical pathways represented include the glycolytic, Embden-Meyerhof, hexose monophosphate shunt, pentose phosphate, methylglyoxal and phosphotransferase (PTS) pathways. Some enzymological parameters measured *in vitro* are also included.

A brief description of the representational methods follows; an abbreviated description has been published (1), and a more extensive treatment is in preparation (Kazic). Reactions are represented either extensionally, by listing their substrates, products and any catalysts, or intensionally, by referring to other reactions which when linearly summed will produce the desired equation. The intensional representation of reactions often relies on describing them in terms of their component reactions. The overall reaction is computed from the most basic reactions (the leaves on the hierarchical tree), a brief fact specifying the tree's branching, and a rule for linearly summing the branches in the correct sequence. The basic parts are equivalent to the nodes and arcs in the network described so far, and the overall reaction is an example of the optional hierarchy. Traversal of this hierarchical structure is accomplished by a recursive rule. As the computation proceeds, appropriate calculations are carried out for K_{eq} and $\Delta G'_0$. Zooming in on the network reveals two types of nodes, one each for reactions and compounds, and three types of arcs, for substrate, product and catalyst relationships. Cofactors, ions, etc. are included as substrates and products in the reactions' representation, producing an accurate accounting of their arrangement (topology). Those reactions not having catalysts omit that arc and its corresponding node; otherwise, this organization describes the basic reactions from which trees can be built. The prokaryotic cellular compartments are also distinguished. The representation is general enough to encode any type of reaction, including metabolism, transport, isomerization, regulation and polymerization, and can be recursively combined to produce more complex reactions, pathways and processes. Predicates which check reactions for charge and mass balance and flag unbalanced reactions have been implemented and are used routinely.

Much of our effort has been to accurately describe the variety of biochemical reactions which occur, and the ways biochemists study and describe them. Clearly listing substrates, products, and catalysts is only a first step in describing biochemical reactions. For *Atropos*, we are really interested in describing the biochemical and chemical mechanisms of the reactions so we can use these rules to predict product structures. Humans frequently mix these descriptions, often inconsistently (1,10). Such mixing can be disastrous for computer programs which are trying to use the information in executing inferences, because the program has no way of distinguishing the nuances that signal which description predominates.

Preventing this mixing requires clear definition of each descriptive type and their interrelationships. We have recently clarified these for chemical, kinetic and mechanistic descriptions, and are now revising the prototype carbohydrate database to test these definitions. A fuller treatment of these issues is in preparation (Kazic).

Prospects

Designing new biochemical processes or rational modification of organisms requires an accurate, global perspective of cellular physiology. The representational principles and their implementation sketched here have as their primary aims flexibility and expressivity. Both are critical to permit the model to evolve as the field grows, modifying old ideas and introducing new ones. As we continue to test these tools by using them to describe ever more complicated ideas, it is reasonable that additional principles will emerge. However the evidence suggests our current ideas will serve well to provide this perspective for computational models for the foreseeable future.

Acknowledgments. This work was begun while I was a visiting scientist in the Mathematics and Computer Science Division of Argonne National Laboratory. Without their hospitality and the generosity of Ross Overbeek it might never have started, and without the help of Dan Hartl and David States it certainly wouldn't have continued. It has benefited from discussions with many people, including Lloyd Barr, Christopher Beecher, Mary Berlyn, Charles Cantor, Daniel Davison, Brian Dunford-Shore, Richard Feldmann, Christopher Fields, Robert Futrelle, Daniel Hartl, Stephen Henikoff, Lawrence Hunter, Peter Karp, Michael Liebman, Ronald Loui, Michael Mavrovouniotis, George Michaels, Harold Morowitz, Frederick Neidhardt, G. Christian Overton, Ross Overbeek, Monica Riley, Paul Schlesinger, David Searls, David States, Gary Stormo, Shalom Tsur, William Wise, and Maria Zemankova. I am grateful for the support of the National Science Foundation via IRI-9117005. Much of the long gestation of this paper occurred while I was a student at the Aspen Center for Physics Workshops on Recognizing Genes, 1992 – 1994.

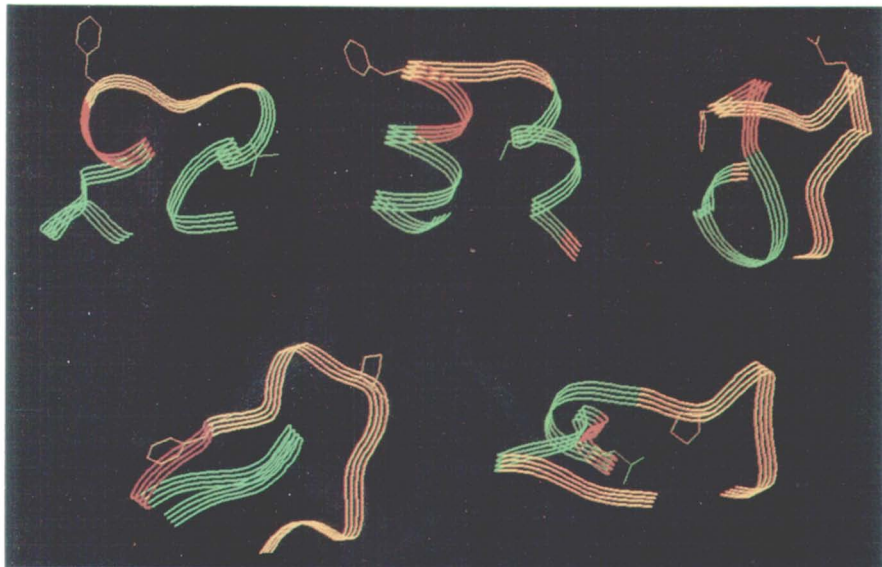
Literature Cited

1. Kazic, T. Representation, reasoning and the intermediary metabolism of *Escherichia coli*. In *Proceedings of the Twenty-Sixth Annual Hawaii International Conference on System Sciences*; Mudge, T. N.; Milutinovic, V.; Hunter, L., Eds., volume 1, pp 853–862, Los Alamitos CA, 1993. IEEE Computer Society Press.
2. Bell, R. E. *Women of Classical Mythology*. Oxford University Press, New York, 1991.
3. Hesiod. *Theogony*. In *Hesiodus Carmina*; Rzach, A., Ed., Leipzig (1908). B. G. Teubner.
4. Kazic, T. Reasoning about biochemical compounds and processes. In *Second International Conference on Bioinformatics, Supercomputing and the Human Genome Project. Proceedings*; Lim, H. W.; Fickett, J. W.; Cantor, C. R.; Robbins, R. J., Eds., pp 35–49, Singapore, 1993. World Scientific.
5. Barcza, S.; Kelly, L. A.; Lenz, C. D. *J. Chem. Info. Comp. Sci.* **1990**, *30*, 243–251.

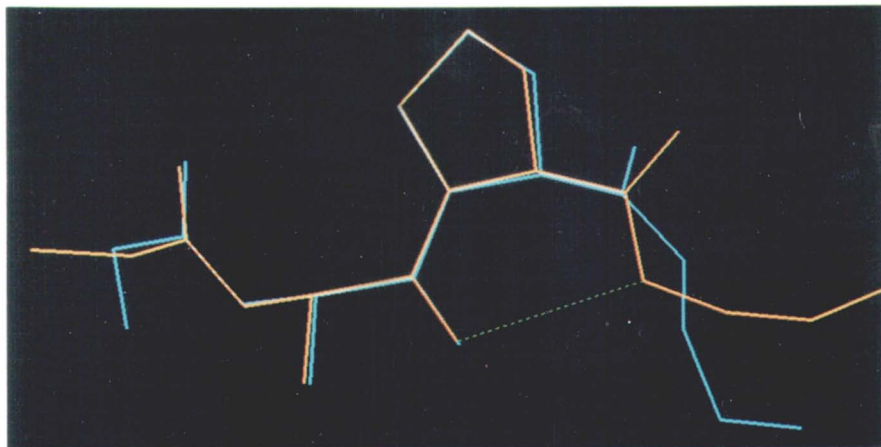
6. Karp, P.; Riley, M. Representations of metabolic knowledge. In *Proceedings of the First International Conference on Intelligent Systems for Molecular Biology*; Hunter, L.; Searls, D.; Shavlik, J., Eds., pp 207–215, Bethesda MD, 1993. Morgan Kauffman.
7. Ochs, R. S.; Conrow, K. *J. Chem. Info. Comp. Sci.* **1991**, *31*, 132–137.
8. International Union of Biochemistry and Molecular Biology. *Enzyme Nomenclature. Recommendations (1992) of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology*. Academic Press, Inc., London, 1992.
9. Walsh, C. *Enzymatic Reaction Mechanisms*. W. H. Freeman and Co., San Francisco, 1979.
10. Kazic, T. Biochemical databases: challenges and opportunities. In *New Data Challenges in Our Information Age. Proceedings of the Thirteenth International CODATA Conference*; Glaeser, P. S.; Millward, M. T. L., Eds., pp C133–C140, Paris, 1994. CODATA Secretariat.
11. Naqvi, S.; Tsur, S. *LDL: A Logical Language for Data and Knowledge Bases*. Computer Science Press, Rockville MD, 1989.
12. Kazic, T.; Liebman, M. N.; Overbeek, R. A. Steps towards a computational model of *Escherichia coli* physiology. In *Bioinformatics, Integration of Organismic and Molecular Databases, and Use of Expert Systems in Biology*; Morowitz, H., Ed., Fairfax, VA, 1990. George Mason University.
13. Balaban, A., Ed. *Chemical Applications of Graph Theory*. Academic Press, London, 1976.
14. Hartsfield, N.; Ringel, G. *Pearls in Graph Theory. A Comprehensive Introduction*. Academic Press, Inc., New York, 1990.
15. Trinajstić, N.; Nikolić, S.; Knop, J.; Müller, W.; Szymanski, K. *Computational Chemical Graph Theory: Characterization, Enumeration, and Generation of Chemical Structures by Computer Methods*. Ellis Horwood, New York, 1991.
16. Maggiolo-Schettini, A.; Winkowski, J. A programming language for deriving hypergraphs. In *Graph-theoretic concepts in computer science. Lecture notes in computer science no. 484*; Möhring, R., Ed., pp 221–231, Berlin, 1991. Springer-Verlag.
17. Searls, D. A Prospectus for a Molecular Logic. UniSys internal report, Paoli PA (undated).

RECEIVED August 2, 1994

These color plates are for Chapter 4.

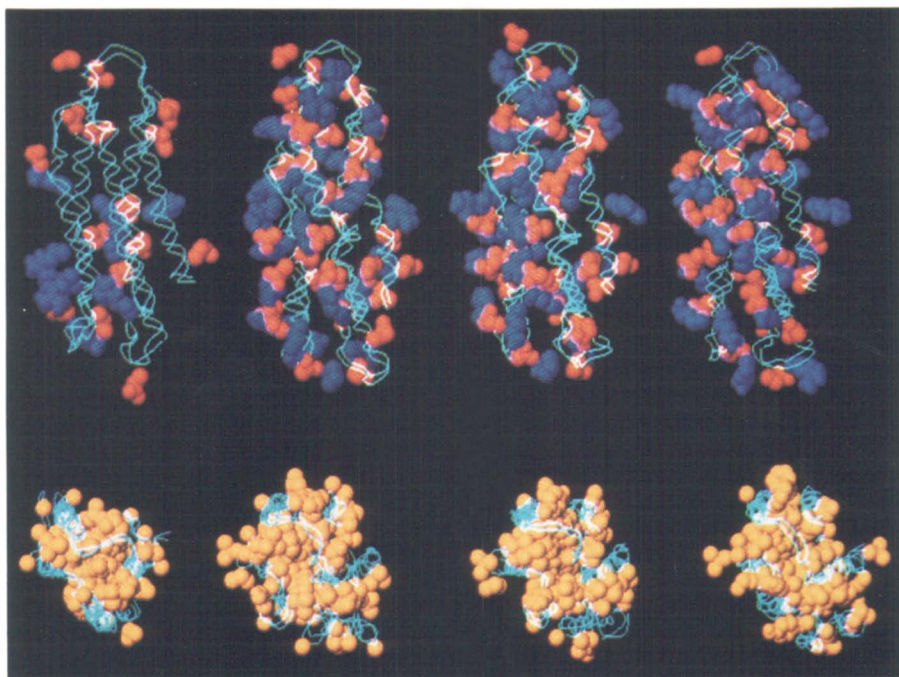


Color plate 1. Molecular models of FL (top left), FS (top middle), FQ (top right), FP (bottom left) and LP (bottom right), shown as ribbons with the side-chains of the substituted residues displayed. Compact subunits, green; cell attachment site, red.



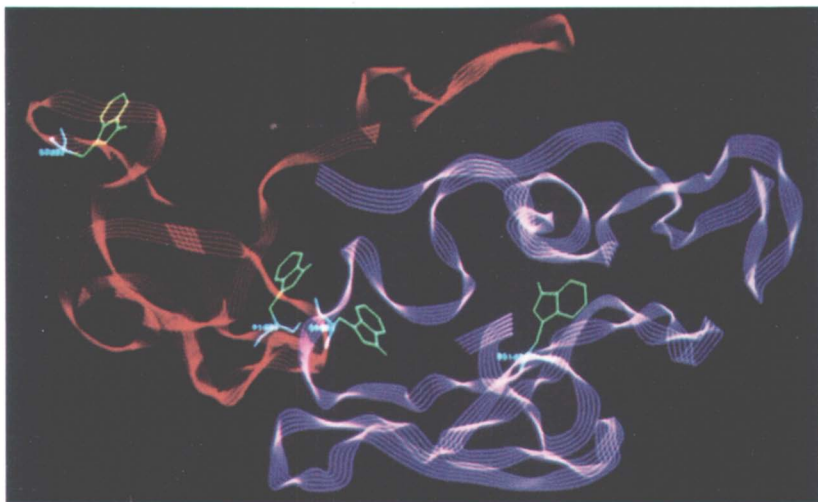
Color plate 2. Superposition of all the heavy atoms from the alpha carbon of Leu-151 to the amide nitrogen of Arg-154, inclusive, of the FP (blue) and LP (yellow) molecular models. Hydrogen bond, green dashed line.

This color plate is for Chapter 7.

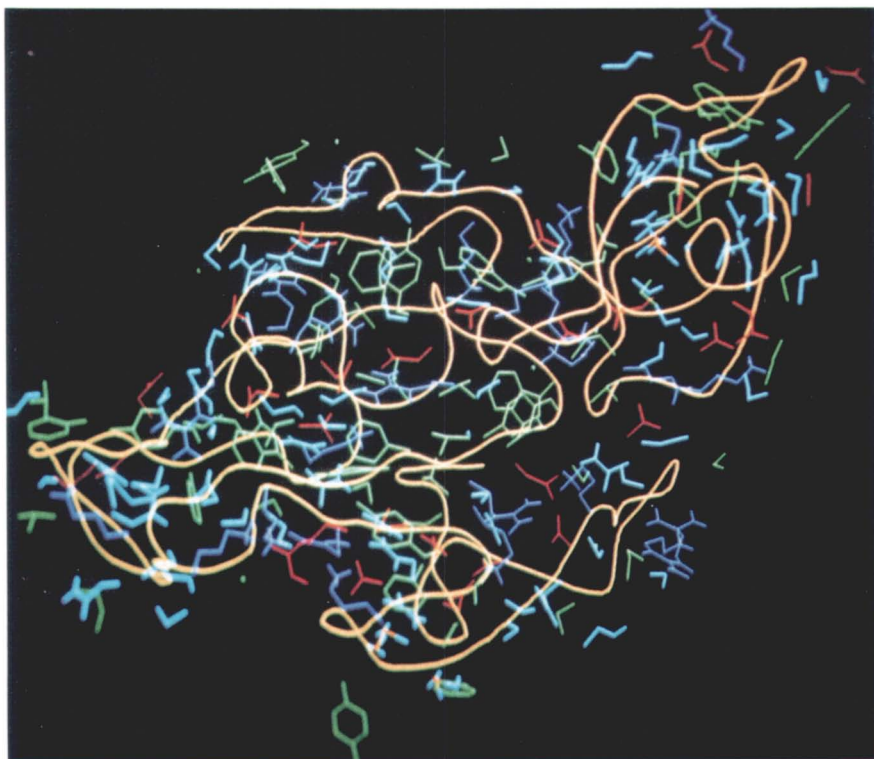


Color Plate 3. Energy refined models of from left to right apo lipophorin-III (residues 7 - 156), canine apolipoprotein A-I (residues 72 - 236), human apolipoprotein A-I (residues 73 - 237) and chicken apolipoprotein A-I (residues 72 - 236). Peptide backbones are represented by a double stranded ribbon. In the lateral view, only ionizable sidechains are displayed with acidic groups (Glu, Asp) in red and basic groups (Lys, Arg) in blue. In the end-on view only hydrophobic sidechains (Ala, Ile, Leu, Met, Phe, Tyr, Trp, Val) are displayed in orange.

These color plates are for Chapter 8.



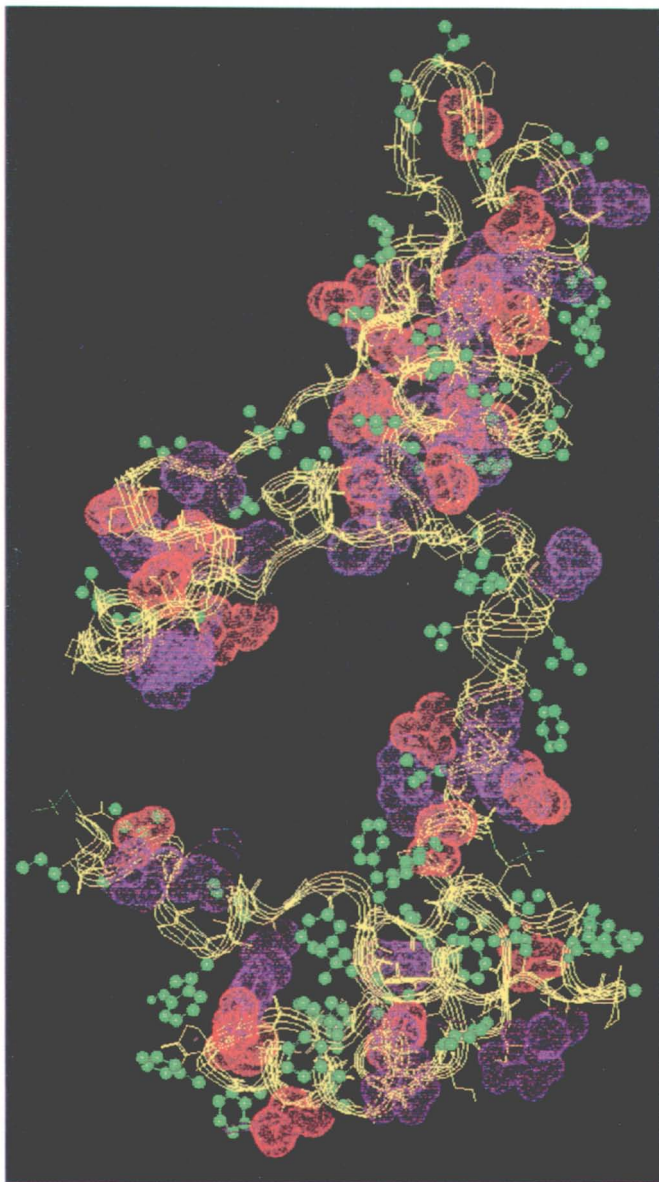
Color Plate 4. Ribbon tracing of the polypeptide backbone of the theoretical working model for sTF. The four tryptophan residues (14, 25, 45, and 158) are shown as well as the subtilisin cleavage between residues 86 and 87.



Color Plate 5. Line tracing of the backbone of the working model for sTF showing side chains.

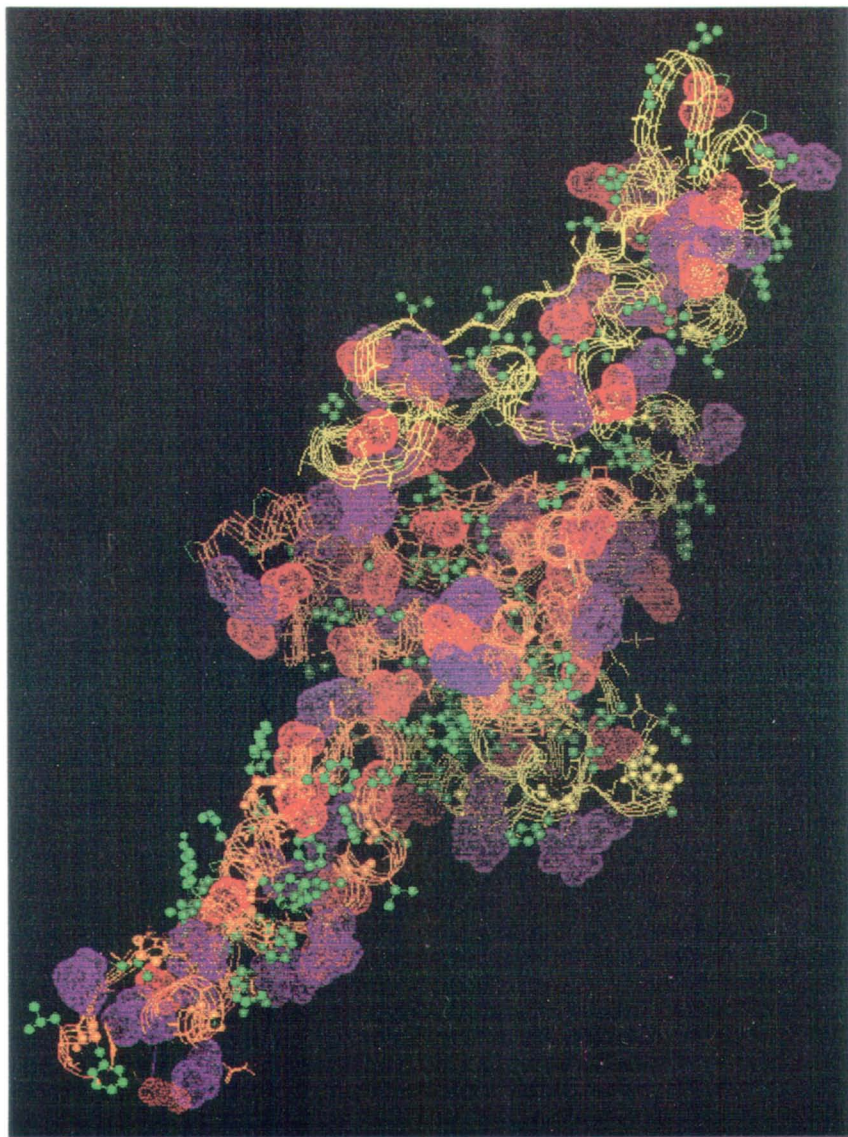
In Molecular Modeling; Kumosinski, T., et al.;
ACS Symposium Series; American Chemical Society: Washington, DC, 1994.

This color plate is for Chapter 9.



Color plate 6. Model of αB_2 backbone with added sidechains. Color code: yellow, backbone; red and purple, negative and positive charged sidechains, respectively; green, hydrophobic sidechains.

This color plate is for Chapter 9.



Color plate 7. "Working" model of the αA_2 and αB_2 complex. The backbone of αA_2 (left) and αB_2 (right) is orange and yellow, respectively.

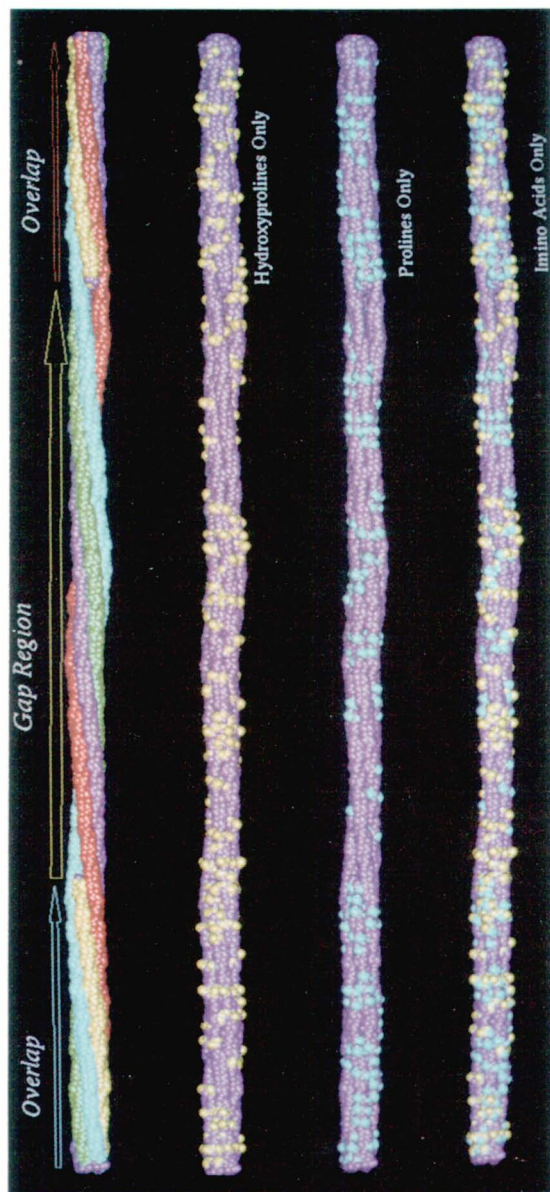
This color plate is for Chapter 9.



Color plate 8. An attempt to superimpose the Greek key folding pattern and β -pleated sheet secondary structure of bovine γ -crystallin (left) on the α A primary structure resulted in a nonsense molecule (right). Color code: white, backbone of respective molecules; blue, proline residues; red and purple, negative and positive charged sidechains, respectively; green, hydrophobic sidechains.

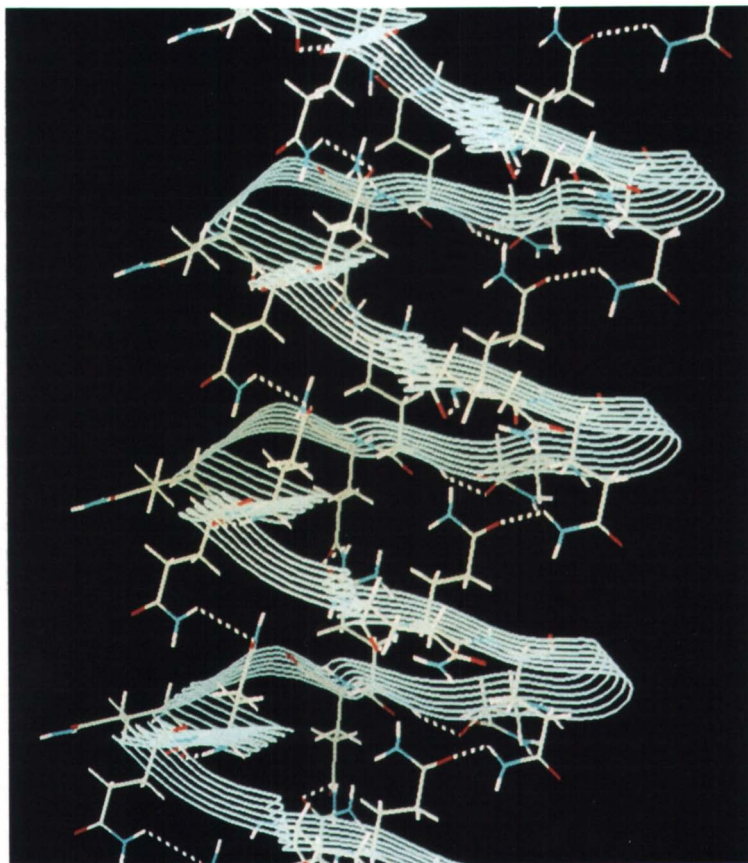
This color plate is for Chapter 10.

Imino Acid Distribution in Type II Microfibril



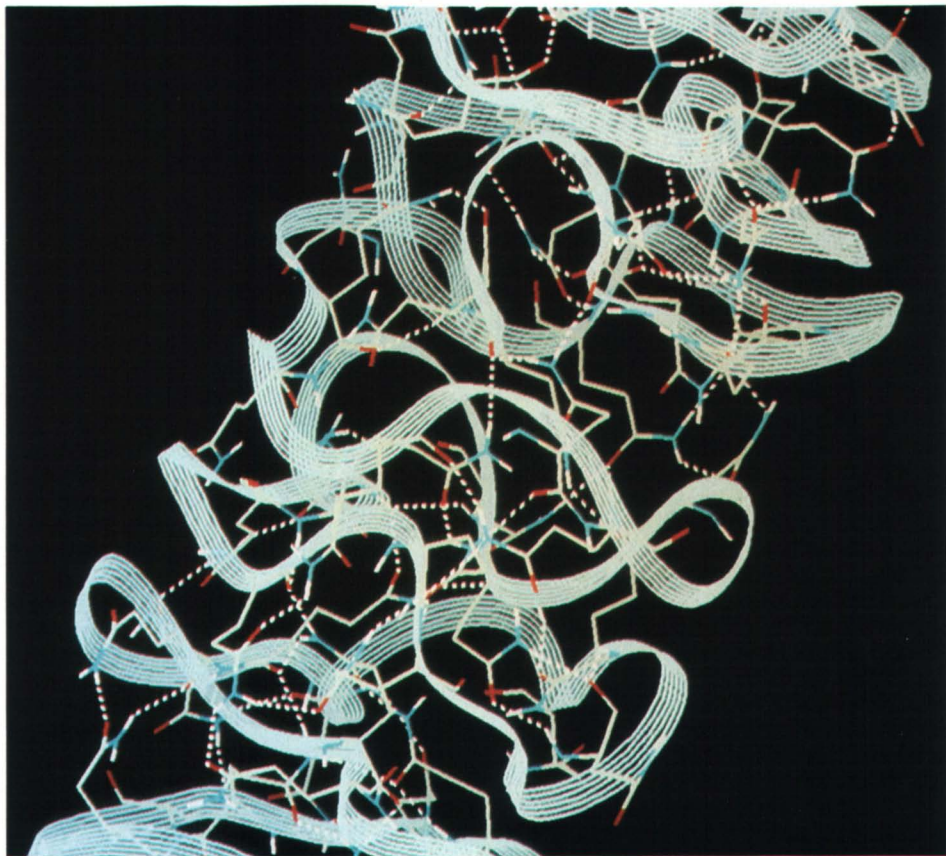
Color Plate 10. Type II collagen microfibril space-filling models. Top. Backbone features of model indicating the boundaries of two overlap and one gap region, with each triple helix labelled with a different color; upper middle. Hydroxyproline distribution (colored yellow) within the microfibril; lower middle. Proline distribution (colored blue) within the microfibril; bottom. Combined proline and hydroxyproline distributions.

This color plate is for Chapter 13.



Color Plate No. 11. 120-residue γ spiral resulting from the 6-residue repeating sequence prior to energy minimization and dynamics calculations. Side view of spiral with backbone represented by a ribbon structure and hydrogen bonds shown as dashed lines. Only glutamine side chains shown.

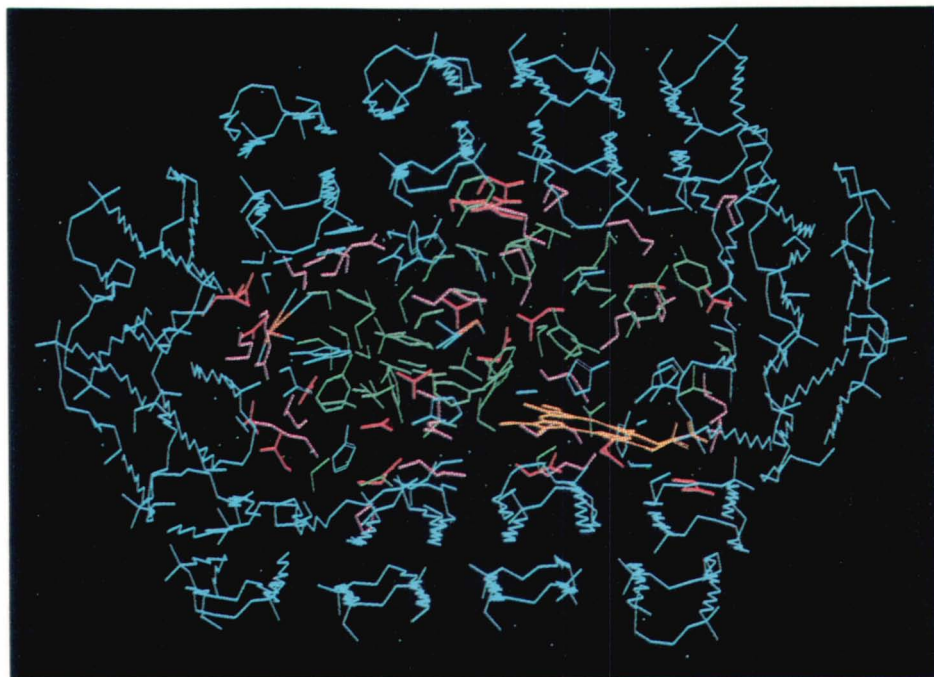
This color plate is for Chapter 13.



Color Plate No. 12. Structure as in Color Plate No. 11, but after 30 ps of dynamics calculations.

This color plate is for Chapter 16.

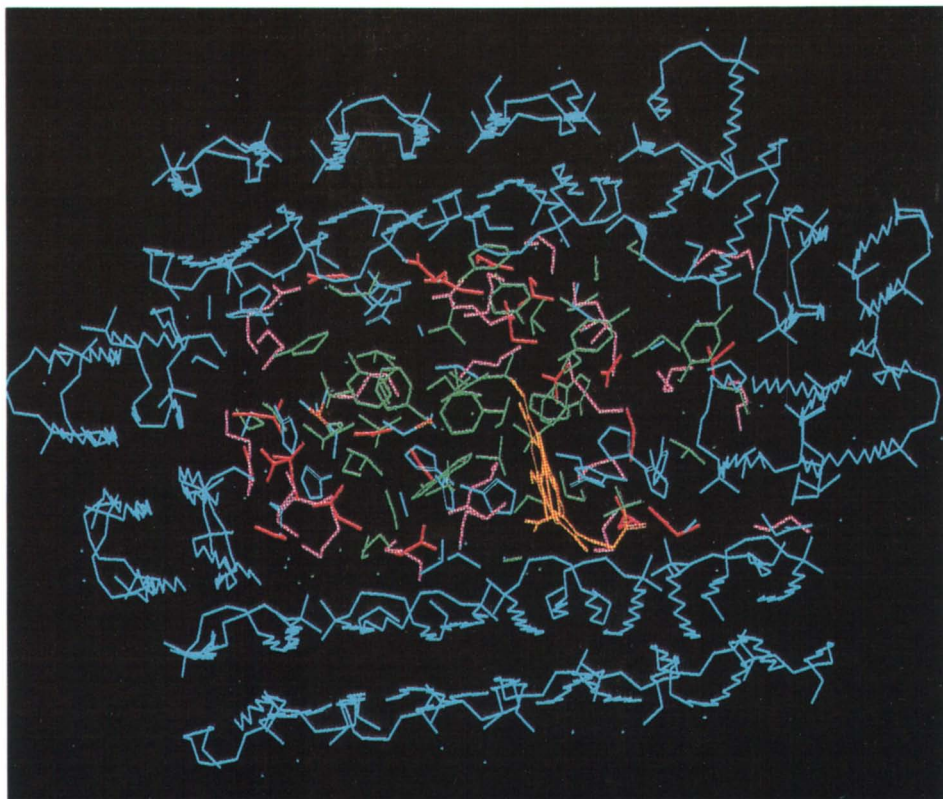
MYGLOBIN DDAB 48



Color Plate 13. Top view of model of Mb-DDAB after 40 ps dynamics. Mb histidine residues unprotonated. Only amino acid residues shown for Mb in the center of the 48-DDAB bilayer (blue). For Mb, purple = positive; red = negative; green = hydrophobic; yellow = heme.

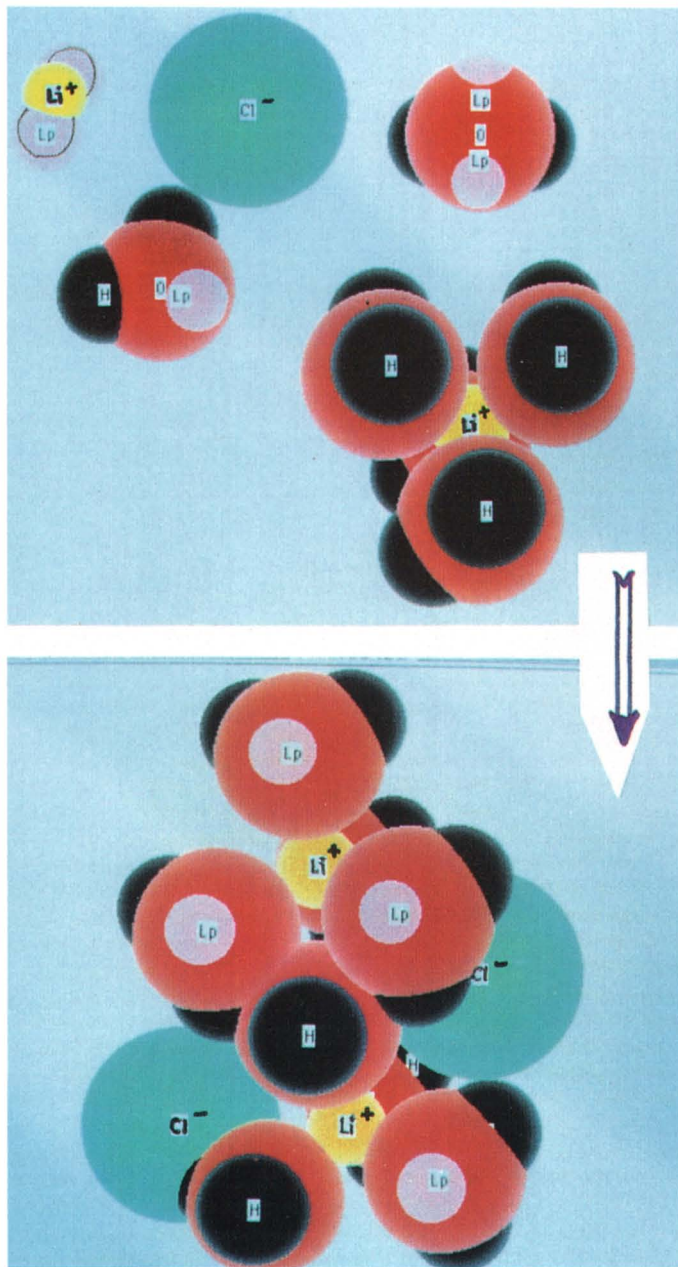
This color plate is for Chapter 16.

MYGLOBIN HIP DDAB 48



Color Plate 14. Top view of model of Mb-DDAB after 40 ps dynamics. Mb histidine residues protonated. Only amino acid residues shown for Mb in the center of a 48-DDAB bilayer (blue). For Mb, purple = positive; red = negative; green = hydrophobic; yellow = heme.

This color plate is for Chapter 17.



COLOR PLATE 15. Idealized 3D Model of $\text{Li}^+(\text{nH}_2\text{O})\text{Cl}^-$ tetramer clusters in concentrated aqueous solutions of lithium chloride at 298 K.

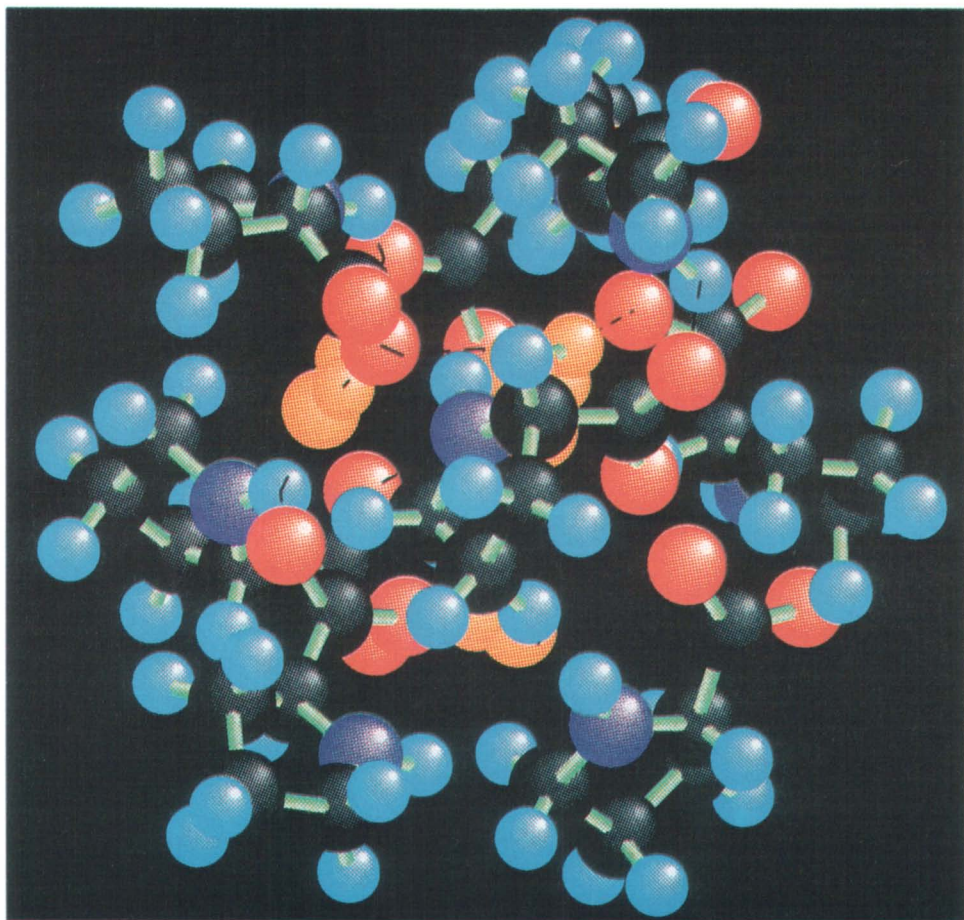
This color plate is for Chapter 17.

0.1 nm



COLOR PLATE 16. Molecular model of $\text{Li}^+(\text{nH}_2\text{O})\text{Cl}^-$, water-bridged ion clusters in glasses at 100 K, derived from pulsed ^1H NMR data.

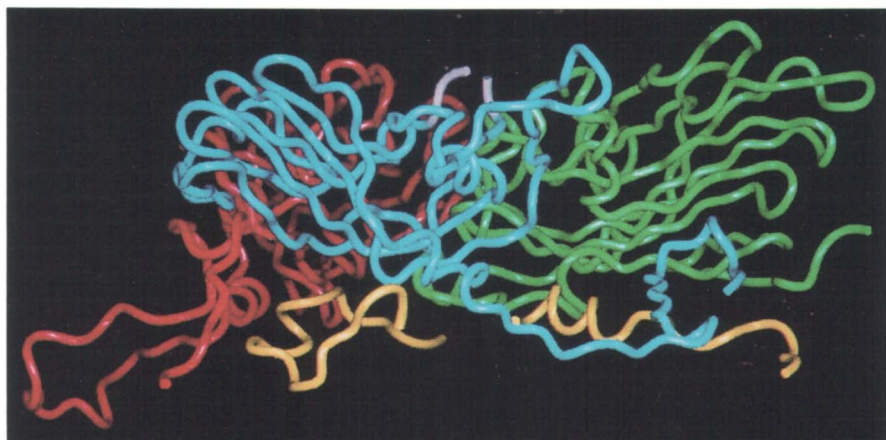
This color plate is for Chapter 18.



Proline 8 Water 4; $t = 50$ ps

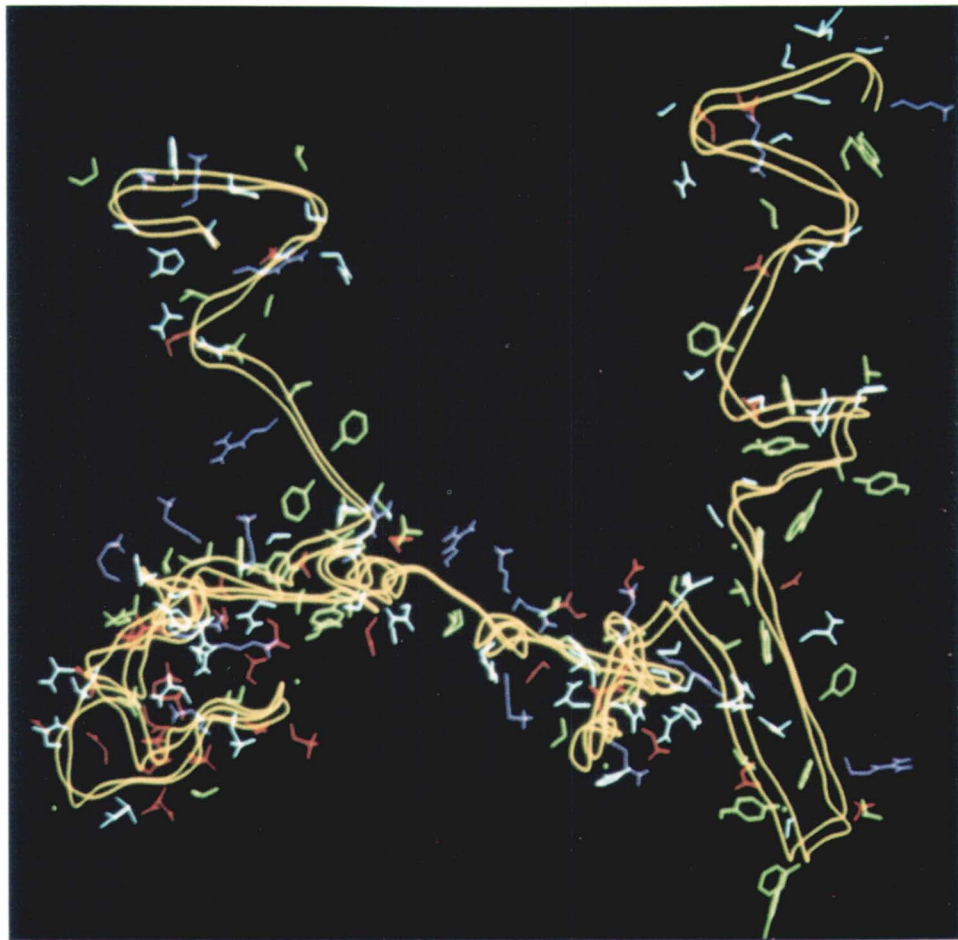
COLOR PLATE 17. The configuration of water and proline molecules are shown after 50 ps from the start of simulation.

This color plate is for Chapter 20.



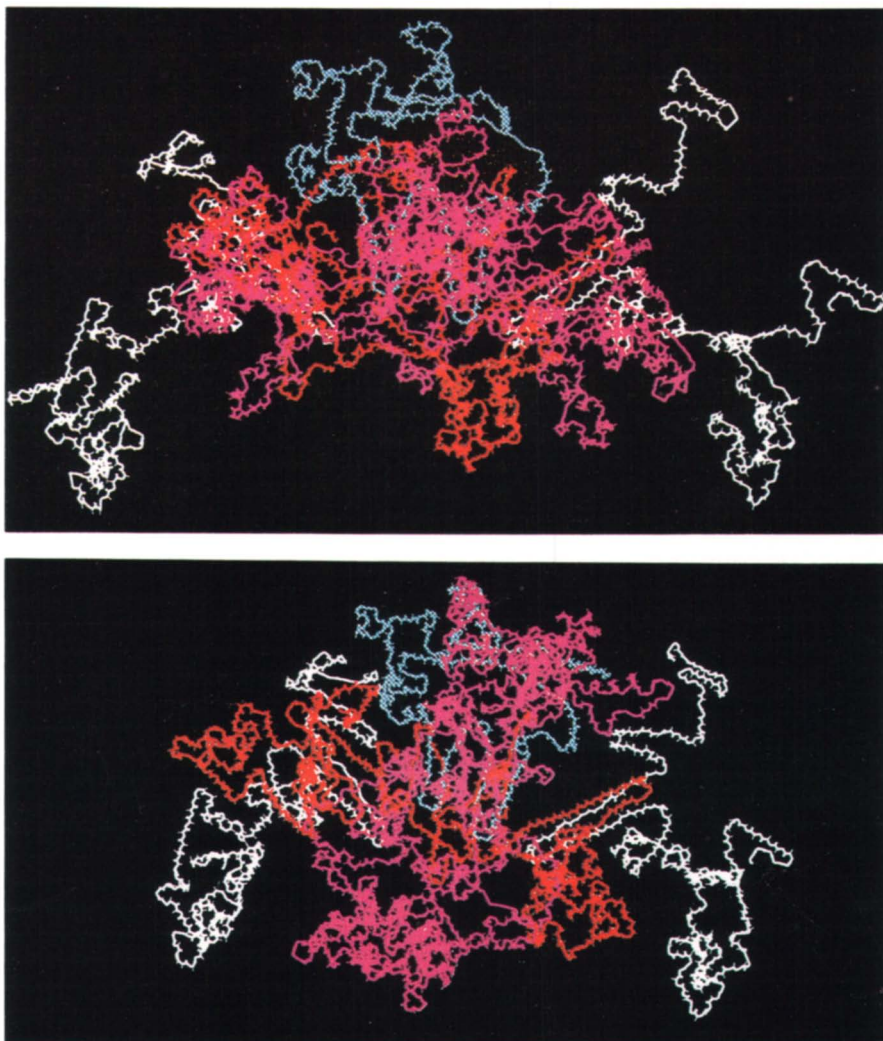
Color Plate 18. A side view of one protomer, shown as a solid ribbon, with the external side of the capsid at the top and the internal side at the bottom. Residues 132-134 and 157-159 of VP1 at the base of the immunodominant loop are shown in violet. VP1, blue; VP2, green; VP3, red; VP4, yellow.

This color plate is for Chapter 21.



Color Plate 19. Energy minimized three dimensional molecular model of α_{s1} -casein. The peptide backbone has been replaced by a double yellow ribbon. Neutral side chains are colored cyan, hydrophobic side chains green, acidic side chains red, and basic side chains purple.

This color plate is for Chapter 22.



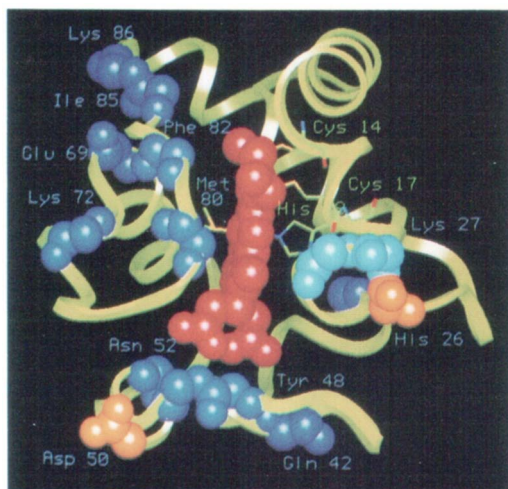
Color Plate 20. (Top) Energy minimized casein asymmetric submicelle structure, i.e., one κ -casein B, four α_{s1} -casein B and two β -casein A² asymmetric dimers. Ribboned backbones without side chains; κ -casein B in cyan, α_{s1} -casein B in red and white; β -casein A² backbone colored in magenta. (Bottom) Energy minimized casein symmetric submicelle structure, i.e., one κ -casein, four α_{s1} -casein B and two β -casein symmetric dimers. Ribboned backbones without side chains; κ -casein B in cyan, α_{s1} -casein B in red and white β -casein A² in magenta.

This color plate is for Chapter 23.



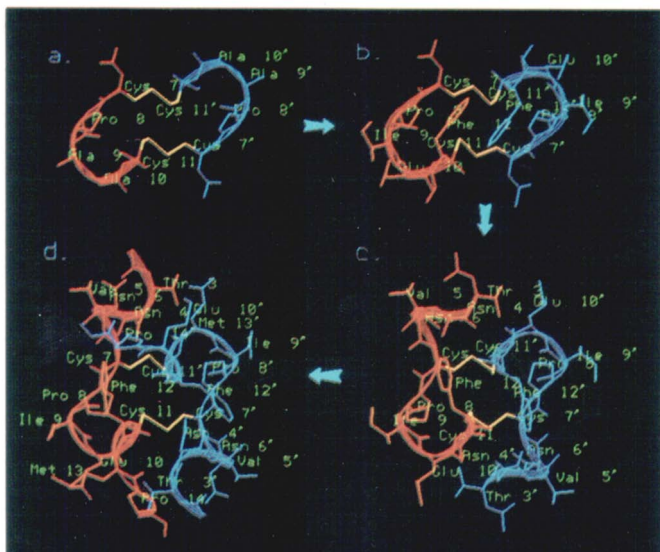
Color Plate 21. Spacefill model of 180 residues of gly-ala-hpro template energy minimized built structure of the triple helix of tropocollagen. Each chain is colored individually (magenta, green, and red-orange), to easily visualize the super-helix surface motif.

This color plate is for Chapter 24.

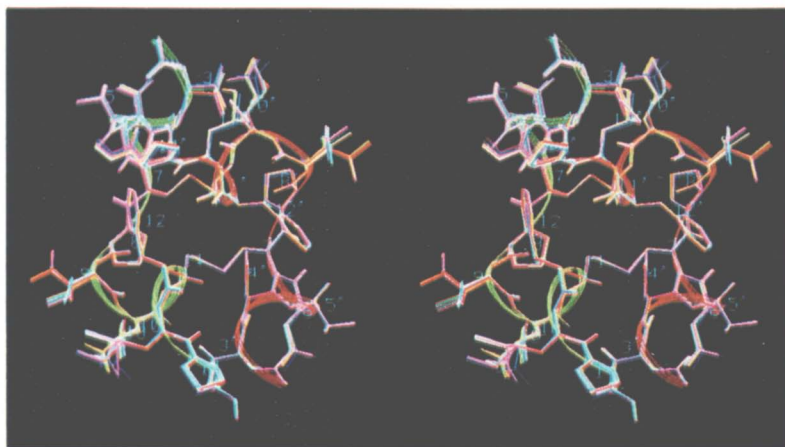


Color Plate 22. The regions of oxidation state dependent conformational change observed in solution and crystal forms of cytochrome *c* are visualized using the coordinates of tuna ferrocyanochrome *c*. The exposed heme edge faces the viewer. Side chain atoms are shown for the heme bound Cys14, Cys17, His18 and Met80, and also for Asp50. Residues with significantly different backbone conformations between the two oxidation states are shown in space-filling representation and color coded as follows: *dark blue*: those observed in the solution structures for the horse cytochromes (32, 42, 47-49, 52, 69, 72, 80, 82, 86); *orange*: those observed in the crystal structures of the tuna cytochromes (26, 50); *light blue*: those observed in both solution and crystal structures (27, 28). The heme is also shown in space-filling representation (in red), while the unaltered residues are shown using a solid ribbon tracing.

These color plates are for Chapter 24.



Color Plate 23. The main build-up steps used in the BUILD strategy are shown for the antiparallel dimer structures: (a) Cys⁷-Pro-Ala-Ala-Cys¹¹; (b) Cys⁷-Pro-Ile-Glu-Cys-Phe¹²; (c) Thr³-Asn-Val-Asn-Cys-Pro-Ile-Glu-Cys-Phe¹²; (d) Thr³-Asn-Val-Asn-Cys-Pro-Ile-Glu-Cys-Phe-Met-Pro¹⁴. The selected orientation shows the inter-chain disulfide bonds at the center of each figure. Hydrogens are not displayed. Residue numbers with primes belong to the second (symmetrical) chain. Ribbon traces have been added to illustrate the symmetry in the backbone conformation of the two chains.



Color Plate 24. Stereo diagram of the final 16 structures for the Thr³-Pro¹⁴ dimer segment in the same orientation as shown in Color Plate 23.

Author Index

- Armitage, I. M., 29
Baianu, Ion C., 269,325
Barford, Robert A., 172
Baumruk, Vladimir, 61
Bott, Richard, 18
Brown, Eleanor M., 100,368
Brown, F., 45,362
Chen, James M., 139
Consler, Thomas G., 466
Doner, L., 342
Dukor, Rina K., 61,235
Farnsworth, Patricia N., 123
Farrell, Harold M., Jr., 100,368,392
France, L. L., 45
Groth-Vasselli, Barbara, 123
Guha, A., 113
Gupta, Vijai P., 61
Hasselbacher, C. A., 113
Huo, Dongfang, 61
Irwin, P., 342
Jiwani, Nilofer G., 185
Johnson, Michael E., 446
Kakalis, L. T., 29
Kar, Leela, 446
Kasarda, Donald D., 209
Kazic, Toni, 488
Keiderling, Timothy A., 61
King, Gregory, 113,123,172,209,342,392
Klein, J., 342
Konigsberg, W. H., 113
Kumosinski, Thomas F., 71,100,113,123,
209,250,269,325,342,368,392,420
Laue, T. M., 113
Lee, James C., 466
Liebman, Michael N., 1,185,221,235,466
Mavrovouniotis, Michael L., 221
Nassar, Alaa-Eldin F., 250
Nemerson, Y., 113
Newman, J. F. E., 45,362
Ozu, E. M., 269,325
Pancoska, Petr, 61
Piatti, P. G., 45,362
Pieffer, P., 342
Reddy, Venkatramana N., 221
Ross, J. B. A., 113
Rusinova, E., 113
Rusling, James F., 250
Sheldon, Adrian, 139
Sherman, Simon A., 446
Toth, I., 45,362
Unruh, Joseph J., 71,420
Urbanova, Marie, 61
Waxman, E., 113
Wei, T. C., 269,325

Affiliation Index

- Amoco Technology Company,
1,185,221,235,466
Charles University, 61
Genencor International, 18
Loyola University of Chicago, 185
Mount Sinai School of Medicine, 113
Northwestern University, 221
OsteoArthritis Sciences, Inc., 139
School of Pharmacy, The, 45,362
University of Connecticut, 250
University of Illinois at Chicago, 61,446
University of Illinois at Urbana, 269,325
University of Maryland at College Park,
221
University of Medicine and Dentistry–
New Jersey Medical School, 123
University of New Hampshire, 113
University of Texas Medical Branch, 466
U.S. Department of Agriculture, 45,71,100,
113,123,172,209,250,269,325,342,362,
368,392,420
Washington University, 488
Yale University School of Medicine, 29
Yale University, 113

Subject Index

A

A-P peptide bond, *trans* and *cis* forms, 36*f*

Abstraction, Petri nets, 229

Accessible surface area analysis

procedure, 193

structure-derived, 198

Aggregate structures

assembly, 371

construction, 394–396

Agricultural applications

analogy with pharmaceutical needs, 2

molecular modeling, 1–16

product development pathway, 3*f*

Alanine residues, insertion into a helical segment, 110

Algorithms

correlation with FT-IR results, 89

crystal structure determination, 22

Alpha–beta barrel structure, pyruvate

kinase, 473,475

Amide groups, spiral stabilization, 218

Amide I and II envelopes, errors in calculations, 87

Amide I bands, analysis in Mb films, 255*t*

Amide vibrational modes, regularities of VCD spectra, 62

Amino acid(s)

categorized by hydrophobicity values, 187

changes in loop region of FMDV, 365*t*

hydration and activity in aqueous solutions, 325–419

interaction within local environment in protein, 203

interchange of nonequivalent acids, 204

partitioning tendency between polar and nonpolar phases, 204

range of dipole moment values, 203

structure, fiber interactions, and stability, 151,153

wheat flour doughs, 209–220

Amino acid residues

accessible surface area frequencies, 201*f*

Amino acid residues—*Continued*

dipole moment frequencies, 200*f*

van der Waals volume frequencies, 200*f*

Amino acid sequences

collagen polypeptides, 140

determination of protein conformation, 195

models for apo A-I, 110–111

Amino acid substitutions in proteins,

structure–function analysis, 185–208

Amphipathic potentials, apolipoprotein

models, 103

Antigenic variation, foot-and-mouth

disease virus, 362–367

Antiparallel sheet dimer, α -carbon chain

trace of backbone, 386*f*

Apolipoprotein A-I

alignment of sequences, 104*f*

energy-refined models, 106*f*,CP3

molecular modeling, 100–112

Aqueous electrolyte solutions

LiCl and other alkali halides, 307–322

multinuclear spin relaxation

measurements, 305,307

Area percent, proteins, 87–89

Arginine, behavior in aqueous solutions, 329

Asparagine (ASN), large bands due to side chains, 87–89

Association free energies

breakdown of electrostatic and

hydrophobic differences, 182*t*

calculations, 172–184

Asymmetric protein

soluble tissue factor, 118–119

submicelle structure, 405

Atom–atom pair correlation functions,

molten LiOH, 281*f*

Atomic temperature factor, relative

vibrational motion of atoms, 22

Avian pancreatic polypeptide

ribboned backbone dynamic structure, 425*f*

NOTE: CP refers to the color plates in the center of this volume.

- Avian pancreatic polypeptide—*Continued*
 ribboned backbone models, 374*f*
 three-dimensional structure, 424
- B**
- β turns, stable spiral structures, 213
- Backbone conformations
 α -crystallin structure, 126–128
 determination, 448
 estimation of precision, 461–462
 Ramachandran plot of dihedrals, 463*f*
 solution vs. crystal structures, 455
- Band shape, variation, 63–64
- Barrel structure, β -trypsin and
 α -chymotrypsin, 95
- Basis sets, X-ray crystal structure of
 proteins, 72
- Bias
 compensation in data selection, 13–14
 data or methodology, 11
 historical origin, 12
 identifying and overcoming, 11–14
 parallel data models, 15*f*
 problem-solving methodologies, 12–14
 unknown or unattainable information, 15*f*
- Binding conformations, prediction, 181*t*
- Binding free energy, salting-out and
 salting-in, 442,444
- Binding mechanism, calcium or
 magnesium chloride, 436
- Binding site region
 exposed hydrophobic regions, 134
 N-terminus telopeptides, 162
- Biochemical reaction systems, modeling
 discrete event systems, 221–234
- Biochemistry, structural modeling,
 488–496
- Biological pathways, representation by
 Petri nets, 227
- Biologically relevant peptides, modeling,
 45–60
- Biomembranes, diffusion, 265
- Biomolecular structural studies, VCD, 61
- Bound water molecules, myosin in
 solution, 335–336
- Boundedness, Petri nets, 228
- Bovine caseins
 energy-minimized, 368–390
 submicelle structure, Lattman
 methodology, 412–417
 three-dimensional molecular modeling,
 392–419
- Bovine pancreatic trypsin inhibitor
 comparison of energy-minimized
 structures with SAXS profiles, 408–411
 SAXS profile, 410*f*
 solution and crystal structures, 455–456
 temperature factors, 409*t*
- Brookhaven Protein Data Bank
 crystallographic structures, 18
 definitions of conformation, 89
 three-dimensional structures of
 proteins, 187
- BUILD strategy
 main build-up steps, CP23
 software and procedure, 457
 spatial structure, 459–460
- Build-up strategy, local conformations
 estimated by probabilistic approach, 447
- C**
- C-terminal dimer fragment
 arrangement of monomer chains, 460
 conformational analysis, 460–461
 primary sequence, 457–462
 structure consistency with NMR data, 461
- C-terminal extensions, α -crystallin
 aggregates, 134
- C-terminal tryptophan, α_{s1} -casein, 388
- CaCl₂, molecular dynamics in water,
 434–442
- Calcium binding sites, time constant, 329*t*
- Carbohydrate complex, entrapment of
 water, 342–361
- Casein
 asymmetric and symmetric submicelles,
 407,CP20
 backbone structure with labeled prolines,
 398*f*–399*f*
 molecular dynamics of salt interactions,
 420–445

α_{s1} -Casein

- backbone and stereo view of refined model, 378*f*,CP19
 - backbone ribbon structure, 437*f*,438*f*
 - α -carbon chain trace, 387*f*
 - comparison of hydrophilic domain with dephosphorylated compound, 441*f*
 - double ribbon structure, 402*f*
 - energy calculations for N-terminal segment, 376*t*
 - energy calculations for refined structure, 377*t*
 - energy-minimized working model, 368–390
 - hydrophilic domain, 434–442
 - initial and final secondary structures, 382*t*
 - initial backbone and stereo chain trace of initial structure, 379*f*,CP19
 - molecular dynamics of hydrophilic half, 439*t*,443*t*
 - phosphorylation sites, 383
 - physicochemical studies, 384–388
 - refined structure of synthetic submicelle framework, 406*f*
 - refined three-dimensional structure, 377,380
 - secondary structure analysis, 380–382
 - sequence, 372*f*–373*f*
 - structure–function relationships, 420–445
 - tetrameric structure, 385–388
- β -Casein
- backbone structure, 403*f*
 - energy-minimized casein submicelle structure, CP20
- Casein complex, structures, 396–405
- Casein dimers, calculated energy, 401*t*
- Casein micelle, structural theories, 393
- Casein submicelle
- energy for refined structures, 404*t*
 - hydration value, 407–408
 - structural rigidity, 416
 - structure, 404–408
 - three-dimensional structure, 392–319
- Cat muscle pyruvate kinase
- crystal structure, 475

Cat muscle pyruvate kinase—*Continued*

- secondary structure, 480
 - structural organization, 472
 - three-dimensional structure, 473,474*f*
- Catalysis, proposed mechanisms, 33–34
- Catalytic films, reductive dehalogenation of organohalide pollutants at electrodes, 251
- Cataractogenic processes, α -crystallin, 124
- Cell attachment site, propensity to form β turns, 56
- Cellular biochemistry, structural modeling, 488–496
- Central cavity, hydrophobic N-terminal domains, 134
- Chaperone protein, lens transparency, 123
- Charge hopping, electroactive films, 264
- Charged residues, redox state dependent conformational change, 452
- Chou–Fasman technique, secondary structure predictions, 53*f*
- Chymosin cleavage sites, α_{s1} -casein, 383
- α -Chymotrypsin, FT-IR results for percent extension, 94
- Circular dichroism (CD)
- protein secondary structure, 130
 - spectra of peptides in aqueous solution, 47–52
 - synchrotron radiation, 45–60
- Classification, Petri nets, 229–230
- Collagen
- arrangement relative to microfibril, 161*f*
 - atomic motions of triple-helical segment, 165,166*f*
 - functions and structures, 139–141
 - packing, crystalline properties, 140–141
 - triple-helical molecules, model of Smith microfibril, 139–170
 - two-dimensional molecular alignment, 151
- Collagen microfibril
- applications of 3-D model, 153
 - energy-minimized model, 141,151–153
 - space-filling models, CP10
 - telopeptide binding site region, CP9
- Collagenase cleavage site
- interactions, 162–165

- Collagenase cleavage site—*Continued*
model compounds designed to mimic, 163
type II microfibril structure, 163–165
- Collections, defined by user criteria,
491–492
- Compensation for bias, 13–14
- Component bands, nonlinear regression
analysis, 83
- Computational methods
analysis of crystallographic structure
data, 239
standard approach for molecular
dynamics, 269
tests, 179
- Computer-generated working models,
 α -crystallin subunits, 123–138
- Conformational assignments, sequence-
based prediction methods for turns,
212–213
- Conformational differences, solution and
crystal structures, 455–456
- Conformational space, high temperature
molecular dynamics, 57
- Consensus sequence
modeling of peptides, 217
structure after application of γ turns, 214f
three-lobed structure, 213
- Correct pathway, selecting in real-world
situation, 9f
- Correlation times, computed fluctuations
in solution, 270
- Coulombic interactions, Mb–DDAB
models, 263
- Crystal structure
apolipoproteins, 101
tetramer of PK, 475
trypsinogen, 240–244
- α -Crystallin
absorbance vs. concentration, 131f
backbone structure, 126–128
backbone with added side chains, CP6
cataractogenic processes, 124
maturation of lens fiber cells, 124
micellar quaternary structure, 134
modeling, 125–126
superimposed pattern and secondary
structure, CP8
- α -Crystallin—*Continued*
survey of literature, 133–135
working model of complex, CP7
- α -Crystallin subunits
computer-generated working models,
123–138
predicted secondary structure, 132
tertiary structure, 132–133
- Crystalline environment, three-dimensional
protein structure, 4–6
- Crystalline structure, algorithm for
comparison with solution structure, 470f
- Crystallization conditions, trypsinogen
and complexes, 238t
- Crystallographic analysis
accuracy and reliability of models, 23–27
error sources and methods of estima-
tion, 28
- Crystallographic structure determination
growing crystals, 19
X-ray, 18–28
- β -Cyclodextrin (β CD), possible use as
extraction agent, 174
- Cyclophilin
conformation of substrate, 35–40
peptidyl proline isomerase, 33–40
transferred NOE measurements, 29–44
- CyP-bound substrate conformation,
determination in solution, 34–40
- Cytochrome *c*
angular root mean square deviations, 454t
evolutionarily invariant residues,
functional implications, 452
local conformation, 451t
oxidation state dependent conformational
change, CP22
secondary structure of redox states,
453–455
solution and crystal forms, 455–456
- D
- D interval, repeating unit in fiber-forming
collagens, 151
- Database
protein structures, 4
structural accuracy, 488–496

- Deconvolution
 quantitative criteria, 75
 rationale, 75–82
- Denatured proteins, central cavity of
 α -crystallin aggregates, 134
- Deuterium (^2H) NMR, theory, 275,278
- Deuterium quadrupole coupling constants,
 solids containing ions and free HDO
 molecules, 309*t*
- Deuterium quadrupole-echo lineshapes,
 simulated, 279*f*
- Deuteron spin energy levels, static
 magnetic field, 276*f*
- Didodecyldimethylammonium bromide,
 spectroscopy and molecular modeling of
 electrochemically active films, 250–268
- Difference FT-IR derivative spectra,
 trypsinogen, 243–244
- Difference linear distance analysis,
 procedure, 191
- Difference method, local structure
 determination of electrolyte solutions,
 287–291
- Difference spectra, secondary structural
 changes due to environment, 78,82
- Diffraction data, crystal structure
 determination, 19–21
- Diffraction pattern, calculated vs.
 observed, 21–22
- Dimer formation, dog-leg structures, 400
- Dimer–tetramer model, LiCl solutions
 and glasses, 316
- Dipole moment
 amino acids in proteins, 206
 computation procedure, 191,193
 conformational changes in charged
 residues, 452
 structure-derived analysis, 198
 variability in hydrophobic amino
 acids, 203
- Dipoles in water, molecular dynamics and
 NMR studies, 269–324
- Discrete event systems approach
 modeling of biochemical reaction systems,
 221–234
- Petri net as a model, 225–227
 qualitative modeling of a pathway, 222
- Distance geometry approach, structure
 determination, 456–457
- Distance matrix analysis, structural
 organization of pyruvate kinase, 472
- Dog-leg structures, casein, 397–401
- Domain movement, activation of pyruvate
 kinase, 467–469
- Doughs, elasticity, 209
- Drug design, specificity and potency, 168
- Dynamics
 application followed by reminimization
 of energy, 213
 Mb–DDAB films, 258–259
- E
- Elastase
 effects of choosing too few component
 bands, 83
 FT-IR results for percent extension, 94
 hydrophobicity profiles and structure
 alignments, 197*f*
 linear distance plots and structure
 alignments, 196*f*
- Elasticity, doughs, 209
- Elastin, β spiral conformation, 209–220
- Elastin polypentapeptide
 β spiral, 217
 Urry structure, 218–219
- Electrochemical kinetics, Mb–DDAB
 films, 263–265
- Electrochemically active films, Mb–DDAB,
 250–268
- Electrolytes in water, molecular dynamics
 and NMR studies, 269–324
- Electron density map, visualization of
 scattering, 21
- Electron spin resonance (ESR) spectra,
 Mb–DDAB films, 255–257
- Electron transfer rates, Mb–DDAB
 films, 265
- Electronic spectra, myoglobin films,
 257–258
- Electrostatic contributions to association
 free energies, 172–184

- Electrostatic interactions
differences in function between
molecules, 111
Mb–DDAB model, 259
- Energetic evaluation, refined models, 105*t*
- Energy barriers, calculation of association
free energies, 173
- Energy minimization
complex aggregate structures, 394
defining structure–function relationships,
442
polypeptide backbone and side chains, 142
polypeptide chain, 212
- Energy-minimized models
 α_{s1} -casein
predicted, 368–390
Ramachandran plot, 381*f*
fiber-forming collagens, 167
Smith microfibril model, 150
synthetic submicelle, 405, CP20
three-dimensional, generation, 371–377
- Energy of trajectory, dynamics simulation, 213
- Energy-refined helices, amphipathic
analysis, 107*t*
- Entrapment of water, NMR and molecular
modeling evidence, 342–361
- Entropy, increase in free energy, 175
- Enzyme
biomembranes in living organisms, 250
prediction of structural features, 5
pyruvate kinase, 466–485
relationship of structure and function, 4
- Error
boundaries and confidence levels, 25
crystallographic analysis, 28
- Evolution
functional similarity, 186
proteins and amino acid, 185
- Evolutionarily invariant residues,
functional implications, 452
- Exchange behavior, *N*-phenyl uronamide
system, 348–356
- Extendability, Petri nets, 229
- Extended conformation, determined
from FT-IR amide I results, 93–94
- Extension, modeling, 492
- F
- Factor analysis, CD results, 72
- Fairness, Petri nets, 228
- Fe(III) heme, Mb–DDAB films, 251
- Ferri- and ferrocycytochrome *c*, estimated
local conformations, 447
- Fiber cell maturation, structure of
 α -crystallin, 124, 135
- Fiber-forming collagens, repeating motif,
145–146
- Fibronectin
build-up strategy to determine
structure, 447
primary structure and NOESY
connectivities for 6-kD C-terminal
fragment, 458*f*
spatial arrangements of binding domains, 462
- Fibrous proteins, molecular dynamics of
salt interactions, 420–445
- Film structure, Mb–DDAB films, 259–265
- Finite difference Poisson–Boltzmann
method, calculating electrostatic free
energies, 175
- First-order difference, NiCl₂ solutions, 293*f*
- Flattening, dynamics of subcellular
processes, 490
- Folded polypeptides, relative energies,
54–56
- Folding domain
pyruvate kinase, 473
pyruvate kinase tetramer, 475
- Folding pattern
 α -, β -, and γ -crystallin, 133
 α -lactalbumin and lysozyme, 67
- Foot-and-mouth disease virus (FMDV)
icosahedral capsid, 45
structure–serologic relationships of
immunodominant site, 362–367
- Force constraints, placement during
structural refinement using molecular
dynamics, 146*f*
- Force field calculation
molecular dynamics and mechanics,
422–423
nonlinear regression analysis, 75
unresolved spectra, 73

Fourier transform infrared spectroscopy (FT-IR)
 methods currently used, 236
 resolution-enhanced deconvolution, 71–98
 soluble tissue factor, 116–117
 Free energy calculations
 changes in van der Waals interactions, 176,178
 error bars, 179
 least-squares straight line fits, 179
 SCAAS model, 175–176
 Functional conservation, evolution, 185

G

γ turns, stable spiral structures, 213
 Gap region, collagen molecules, 151
 Geometry minimization, molecular dynamics, 212
 Glasses
 multinuclear spin relaxation measurements in relation to local structure and dynamics, 305,307
 Raman spectra, 301
 Global secondary structure analysis, proteins in solution, 71–98
 Globular proteins
 lens transparency, 123
 percent extended content, 84*t*
 percent helix content, 92*t*
 percent nonperiodic content, 93*t*
 roughness index, 55
 Glucuronic acid derivative, entrapment of water, 342–361
 Glutamine (GLN)
 cohesive nature of gluten, 209–210
 large bands due to side chains, 87–89
 Glutenin subunits
 repeating amino acid sequences, 211*f*
 spiral structures, 209–220
 15(Gly-Pro-Hyp)₃₀₀ microfibril, molecular modeling, 145–146
 Glycine
 conformations determined by probabilistic approach, 448

Glycine—*Continued*
 evolutionarily invariant residues, 452
 hydration and activity, 322–323
 α -Glycoamylase, helix content, 69
 Glycosylation sites
 atoms of side chains, 119*f*
 soluble tissue factor, 119
 Grammar, structural relationships, 493–494
 Growing crystals for structure determination, 19
 Growth factor proteins, α -helix content, 69
 Guest-CD inclusion complexation, hydrophobicity, 183

H

²H NMR, theory, 275,278
¹H NMR relaxation measurements, lysozyme in aqueous solutions, 333–334
 Heat shock proteins (hsp)
 α -crystallin, 124
 homology with subunits, 132
 structure of α -crystallin, 135
 α -Helical band shape, variation, 62–63
 Helical content, calculated and experimental results, 94–96
 Helical proteins, Ramachandran analysis, 90
 Helical segments in apolipoprotein models, amphipathic potentials, 103*t*
 Helical structure, stability, 109–111
 Helical wheel projection
 amphiphilic α -helix, 48
 residues, 49*f*
 α -Helix
 amphiphilic, 48
 percentage at each peptide concentration, 48,51
 serine proteases, 95
 Helix 1–5, substitution of apo A-I sequences, 107–109
 α -Helix content, peptide concentrations, 50*f*
 α -Helix formation, peptides, 52
 α -Helix-forming properties, serological data, 56–57

- Helix–helix interaction, isozymes from rat pyruvate kinase, 475
- α -Helix structure, α -crystallin, 134–135
- α -Helix terminator, helix-forming properties, 57
- Hen's egg white lysozyme, FT-IR analysis, 73–75
- Hierarchical tree, representation of reactions, 494
- Hierarchies, part–whole relationships, 491
- High density lipoprotein (HDL), apolipoprotein A-I, 100–112
- Hinge region, intersubunit contacts, 483
- Holy Grail of molecular biology, transformation of information, 5f
- Homologous proteins, evolutionarily related, 186
- Homology, α -crystallin subunits, 123
- Horse and rider model, casein structure, 397
- Horse ferro- and ferricytochrome *c* conformational differences, 450
- secondary structure, 454f
- Human α -lactalbumin, VCD spectra, 68f
- Human Genome Project
- protein sequence information, 186
- public domain sequence database, 4
- Human soluble tissue factor
- predicted structure, 118f
- testing an FT-IR-consistent model, 113–122
- Human type II collagen Smith microfibril, three-dimensional energy-minimized model, 139–170
- Hydrated asymmetric and symmetric models, energy-minimized, 413
- Hydrated ion clusters, structure, 301,305
- Hydrated proteins, molecular dynamics computations, 329
- Hydrated structures, construction, 395
- Hydration
- ion binding to soy protein, 339–340
- myoglobin microcrystals, 334
- myosin, 334–336
- protein model, 332–333
- soybean protein, 338–340
- wheat gluten proteins, 338
- Hydration numbers, Monte Carlo simulations, 270
- Hydration shell
- halides, 297
- ions, residence time of water, 299
- Hydration structure, Li^+ and Cl^- in concentrated solutions, 290f
- Hydrogen bonding, spiral stability, 218
- Hydrophobic potentials, apolipoprotein models, 103
- Hydrophilic domain
- α_{s1} -casein, 434–442
- casein interaction sites for colloid formation, 407
- Hydrophobic areas, proline turns, 375
- Hydrophobic contributions to association free energies, 172–184
- Hydrophobic effect, definition, 176
- Hydrophobic interactions
- α_{s1} -casein, 383–384
- dimerization residues, 385
- spurious structural changes, 71
- Hydrophobicity
- analysis procedure, 191
- hydrophobic nature of structurally similar regions, 202
- limited predictive capability, 187
- prediction of structure similarity, 206
- relative to particular scale, 51
- scales of free energy of transfer, 187
- structure analysis, 186–187
- vs. dipole moment, 205f
- vs. secondary and tertiary structure, 194
- Hydroxyproline, functional role within collagen, 165
- Hydroxyproline clusters, folding of collagen tripeptide complex, 167
- I
- Imino acid
- hydration and activity in aqueous solutions, 325–419
- microfibril model, 167

- Immunodominant site
 foot-and-mouth disease virus, 45,362–367
 substituted amino acid sequence, 46*t*
- Immunogenic peptide, biologically active conformation, 47–48
- Immunological specificity, steric constraints, 58
- Indole rings, tryptophan residue, 120–121
- Information, amount needed, 10–11
- Information processing
 structural biology, 4–6
 technology transfer, 2–4
- Informational dimensions, structure, 489–490
- Infrared absorption spectroscopy, history, 235–236
- Infrared measurement, proteins, 73
- Inhibitor-induced structural changes in serine proteases, 235–247
- Intension, modeling, 492
- Intensity curves, NiCl₂ solutions in water, 281*f*
- Interaction energy
 H₂O:*N*-phenyl uronamide complex, 357*t*
 ion–water clusters, 297
- Interhelical collagen interactions in fibers, molecular packing, 146–147
- Intersubunit contacts
 allosteric properties of PK, 482–483
 hinge region, 483
 potential functional roles, 479–483
 pyruvate kinase tetramer, 475,479
- Ion(s)
 independent-hydration model, 294
 velocity correlation function in aqueous solutions, 299–300
- Ion binding
 myosin solubility and self-association, 335–336
 soy proteins, 339–340
 wheat gluten proteins, 338
- Ion-pair cluster model, ⁷Li NMR transverse relaxation, 315*f*
- Ion-pair dimer, backbone structure, 386*f*
- Ion–solvent interactions, mean distance and mean square deviations, 295*t*
- Ionic hydration, distribution function, 288
- Ionic polymers, films loaded with electroactive counter ions, 264
- Irregular content, Ramachandran analysis, 96
- Isoelectric binding model, salt cations on soy protein, 339
- Isotopic substitution, partial pair-atom correlation functions, 280
- Isozymes
 pyruvate kinase, homology between species, 482
 structural differences and kinetic behavior, 480,482
- K
- Kinetic activity, rabbit liver PK, 483
- L
- α-Lactalbumin, crystal and aqueous solvent environment structures, 67
- Lateral packing arrangement, collagens, 140
- Lattman program, casein structure, 408–411
- Lens fiber cells, maturation, 124
- Lens transparency, globular proteins, 123
- Ligands, determining solution conformation, 29–44
- Linear dichroism
 myoglobin films, 257–258
 myoglobin orientation, 262–263
 order parameters from soret bands, 258*t*
- Linear distance plot (LDP), procedure, 189,191
- Lipid, biomembranes in living organisms, 250
- Lipid transport, plasma lipoprotein particles, 100
- Lipoproteins, protein components, 100–101
- Lithium chloride
 clusters in glasses, local structure, 272*f*–273*f*
 idealized 3D model of aqueous solutions, CP15
 water-bridged ion clusters in glasses, CP16
- Lithium ion, hydration, 293*t*

- Lithium–water configurations, LiCl solutions, 295*t*
- Liver pyruvate kinase, phosphorylation, 482–483
- Local conformation, proteins in solution, probabilistic approach, 446–465
- Local environment, analysis of dipole moment, 204
- Longitudinal magnetic relaxation rates
lithium and sodium perchlorates and tetrafluoroborates, 320*f*–321*f*
variation with concentration, 322
- Loop(s), helical segments, 109
- Loop region, antigenic variation, 362–367
- Low hydration waters, space-filled energy-minimized model, 417*f*
- Lysine, behavior in aqueous solutions, 329
- Lysozyme
best fit by nonlinear regression analysis, 85*f*–86*f*
best fit for Fourier deconvoluted FT-IR spectrum, 80*f*–81*f*
effects of choosing too few component bands, 83
Fourier deconvolution of FT-IR spectrum, 76*f*
FT-IR analysis, 73–75
Ramachandran plot from X-ray crystallographic structure, 91*f*
second derivative FT-IR spectrum of amide I and II bands, 77*f*
solutions and hydrated powders, 333–334
theoretical FT-IR spectrum, 78,79*f*
- M
- Mammalian pyruvate kinase
structural elements involved in allosteric switch, 466–485
See also Pyruvate kinase (PK)
- Markush structures, coding, 493
- Matrix metalloproteinases, definition, 162
- Matrix methods, Petri nets, 230
- Maturation, lens fiber cells, 124
- Mb–DDAB film
conceptual models of several bilayers, 266*f*
electrochemical kinetics, 263–265
ESR spectra, 256*f*
film structure and stability, 259–262
model, 260*f*,261*f*
molecular models, 258–259
RAIR spectra, 256*f*
top view of model, CP13,CP14
UV-VIS spectra, 260*f*
See also Didodecyldimethylammonium bromide, Myoglobin
- Metal complex multianions, DDAB films, 262
- MgCl₂, molecular dynamics in water, 434–442
- Micelle formation, protein–protein interactions, 393
- Microcrystals, hydration of myoglobin, 334
- Microfibril alignment, α 1 chains of type II collagen, 152*f*
- Microfibril extension, repeating motif, 145–146
- Microfibril model
arrangement of collagen triple helices, 148*f*
construction, 143
imino acids, 167
- Microfibril template, type II collagen model, 147
- Milk protein, molecular basis for structure–function relationships, 392–419
- Model systems, artificial boundaries, 176
- Modeling
biological pathways, discrete event systems approach, 221–234
biologically relevant peptides, 45–60
bovine caseins, 392–419
evaluation of results, 218–219
integration into problem-solving process, 14–15
structural accuracy, 488–496

- Molecular dynamics
aqueous solutions of electrolytes, 294,297
concentrated protein solutions, 325–419
defining structure–function relationships, 442
model modifications, 143
motion and molecular configuration as function of time, 423
water and selected ions, 269–324
- Molecular dynamics calculation
example of computer run, 318f
LiCl·8H₂O and NaCl·8H₂O, 317f
- Molecular force field energy minimization, casein, 394
- Molecular modeling
apolipoproteins
energy-minimized models, 105
software, 102–103
bovine caseins, 392–419
commercial packages, 8
from virtual tools to real problems, 1–16
high temperature molecular dynamics, 45–60
integration into problem-solving process, 14–15
low energy conformations of peptides, 54–56
Mb–DDAB films, 250–268
structure–function relationships, 420–445
theory, 421–422
type II microfibril model, 146–150
use of a template, 100–112
- Molecular orientation, role in electron transfer kinetics, 265
- Monomer chains, arrangement in dimer, 460
- Monte Carlo simulations
aqueous solutions of electrolytes, 270
ion–water clusters, 294,297
- Multibilayer films, double chain phosphate surfactants containing Mb, 266
- Multinuclear spin relaxation, concentrated protein solutions, 325–419
- Multiple conformations, variable properties, 203
- Muscle proteins, myofibrillar, 336–337
- Mutant sequence, conformation, stability, and properties of template, 110
- Myofibrillar proteins, solutions with electrolytes, 336–337
- Myoglobin
charge-paired amino acid partners, 259
component bands, effects of choosing too few, 83
electrochemical parameters, 264t
hydration of microcrystals, 334
increase of electron transfer rate, 251
orientation and secondary structures, 262–263
- Ramachandran plot from X-ray crystallographic structure, 91f
residence site in DDAB films, 266
spectroscopy and molecular modeling of electrochemically active films, 250–268
- Myosin, in electrolyte solutions, 334–336
- N
- N*-Phenyl uronamide
and water, molecular dynamics simulations, 356–357
change in relative concentration over time, 345f
concentration of water dependence, 354f
cross peak areas, 351f
cross peaks from experiment, 352f
entrapment of water, 342–361
formation scheme, 344f
¹H NMR spectra, 349f
inversion recovery experiments, 354f
proton NMR mass spectral assignments hydrated and dehydrated forms, 355t
spin-lattice relaxation times, 353t
reciprocal weighted average distance of water molecules, 358f
structure and conformation, 344f
variation of –NH↔H₂O distances, 360f
variation of –OH↔H₂O distances, 359f

- Native human type II sequence, incorporation, 144
- Native microfibril model, collagen type II, 141
- Natural mutants, interchange of nonequivalent acids, 204
- Near-neighbor interactions, magnetic field perturbation calculations, 63
- Neutron scattering intensities, isotope substitution effect, 286*f*
- Neutron scattering studies, local structure in aqueous solutions of electrolytes, 278–294
- Noncrystallizable proteins, prediction of secondary structures, 71
- Nonequivalent amino acids, overlap of dipole moment magnitude, 206
- Nonlinear regression analysis
concentration dependence of relaxation measurements and glass transition temperature, 319*f*
Fourier deconvolution, 75
rationale for parameters, 82–87
- Nuclear magnetic resonance (NMR) spectroscopy, theory, 270–271
- Nuclear Overhauser effect (NOE) assignments, intra-chain and inter-chain, 459
magnitude, 30
- Nuclear spin energy levels, spin-1/2 nucleus, 273*f*
- Nuclear spin magnetization, precession, 274*f*
- O**
- ¹⁷O NMR transverse relaxation rates
concentration dependence, 314*f*
nonlinear variation, 323*f*
variation with proline concentration, 326*f*
- Optical rotatory dispersion (ORD) curves, Gaussian bands, 71–72
- Organohalide pollutants, dehalogenation, 251
- Orientation, myoglobin in films, 262–263
- Orientational correlation functions, ionic solutions in water, 297–298
- Oxytocin
Fourier transform infrared spectroscopy, 426,427*f*
molecular dynamics, 430*f*–432*f*
molecular dynamics simulations, 426–429
structure–function relationships, 420–445
- P**
- Packing arrangement, collagens, 140
- Part–whole relationships, recursive data structures, 490–491
- Partial structure factor, variation with Q, 285*f*
- Peptide(s)
CD spectra, 49*f*,50*f*
immunogenic synthetic, structural properties, 45–60
loop region of capsid protein, 366*t*
modeling, 45–60
molecular dynamics of salt interactions, 420–445
molecular models, CP1
secondary structure content
aqueous solution, 47*t*
predicted from CD spectra, 51*t*
superposition of heavy atoms, CP2
values for parameters, 52*t*
- Peptide models
VCD spectra, 65*f*
VCD studies, 62–63
- Peptidyl proline isomerase, cyclophilin, 33–40
- Peptidyl prolyl isomerase, transferred NOE measurements, 29–44
- Petri nets
behavioral properties, 227–228
biological compounds and their activities, 226*f*
definitions, 222–223
execution, 223
features of models, 229–230
graph of places, transitions, and arcs, 222*f*

- Petri nets—*Continued*
marking, 223,224f
mathematical definition, 225
methods of analysis, 230–233
transition as representation of a subnet, 231f
use of qualitative methods, 233
- Pharmaceutical development path, 3f
- Pharmaceutical needs, analogy with agricultural needs, 2
- Phosphorylation, liver pyruvate kinase, 482–483
- Phosphorylation sites, α_{s1} -casein, 383–388
- Physical diffusion of electroactive species, electroactive films, 264
- PK, *See* Pyruvate kinase (PK)
- Plasma lipoprotein particles, lipid transport, 100
- Polarization term, ion–water clusters, 297
- Polycrystalline *S*-methyl- $^2\text{H}_3$ methionine, 335f
- Polymorphism, collagens, 140
- Polypeptide backbone and side chains, energy minimization, 142
- Polypeptide backbone fragments, algorithm to determine equivalence, 239
- Polypeptide chain
collagenlike conformation, 163
modeling software, 212
proton–proton distances, 459
- Population ratio, TRNOE, 32
- Potassium chloride, computer-generated configuration of ions, 272f
- PPIase, catalysis mechanisms, 33–34
- Predicted secondary structure, comparative study, 132f
- Preferred conformations, predicted and actual, 179,182
- Primary structure analysis, hydrophilic N-terminal region of peptides, 52–56
- Probabilistic approach
NMR-based protein structure determination, 446–465
overall spatial structure, 447
- Problem solving, evolving new methods, 14–15
- Product development, path and needs, 2
- Proline
local secondary structure in a molecule, 109
molecular configuration, CP17
molecular dynamics in water, 327f–328f
 ^{17}O NMR transverse relaxation of water, 325
structural motifs for casein, 371,375
- Proline–hydroxyproline distribution in microfibril, structure–function analysis, 165–167
- Protease digestion site, soluble tissue factor, 119
- Protein(s)
activity model, 332–333
amino acid substitutions, 185–208
average deviation in strand structure, 94
biomembranes in living organisms, 250
computer modeling of hydrated conformations, 329
fractional secondary structures, 66
linear regression analysis for area percent, 88f
local conformation, 462,464
molecular dynamics and multinuclear magnetic resonance studies, 325–419
molecular dynamics of salt interactions, 420–445
myofibrillar, 336–337
NMR-based determination of accurate local conformation and 3D structure in solution, 446–465
prediction of tertiary structures by using hydrophobicity, 202
qualitative VCD patterns, 63–66
salt-induced resolubilization, 420–421
soybean hydration, 338–340
structural biology, 4–6
water relaxation, 332
wheat, 338
- Protein Data Bank
structure information base, 4
three-dimensional structural protein database, 185–186

- Protein folding
 buried tryptophan, 120–121
 role of water, 121
- Protein in solution, global secondary structure analysis, 71–98
- Protein in water, correcting absorption spectra, 236
- Protein micelles, α -crystallin subunits, 134
- Protein oxygens, radial distribution around calcium ions, 330*f*
- Protein phosphorylation, fixed charges, 124
- Protein secondary structure
 CD spectroscopic studies, 130
 determination using VCD, 61–70
 qualitative aspects, 64
 standard deviations of fits and predictions, 67*t*
- Protein structure
 determination from function, 4
 geometric parameters, 239
 information bias, 12
 prediction, 10
 three-dimensional structure encoded in one-dimensional amino acids, 186
- Public domain sequence database, Human Genome Sequencing Project, 4
- Pure phase absorption, aliphatic region, 36*f*
- Pyruvate kinase (PK)
 comparison between active and inactive conformation, 468*f*, 471*f*
 crystalline structure of cat muscle PK, 472
 domain movement in activation, 467–469
 intrasubunit distance matrix analysis, 474*f*, 476*f*–478*f*
 mammalian, 466–485
 predicted secondary structure for amino acid sequence, 481*f*
 rabbit muscle, 466
 structural elements in allosteric switch, 466–485
 structural organization, 469, 472
- Pyruvate kinase monomer, organization of structure, 473–475
- Pyruvate kinase tetramer, organization of structure, 475–479
- Q
- Qualitative analysis, preliminary conclusions about biological pathway, 222
- Quantitative analysis, techniques with uncertain parameters, 222
- R
- Rabbit liver pyruvate kinase, dependency of enzyme activity on substrate concentration, 483
- Rabbit muscle pyruvate kinase
 conformational change, 483
 molecular mechanism of regulation, 466
- Radial distribution functions
 aqueous solutions of LiCl, 296*f*
 CaCl₂, 291*f*
 Monte Carlo simulations, 270
 NiCl₂ solutions in water, 282*f*
- Ramachandran analysis
 global secondary structure of proteins, 89–90
 molecular modeling software, 90
 serine proteases, 95
 turn and irregular content, 96
- Raman spectra
 LiCl, 304*f*, 306*f*
 LiCl·NH₂O solutions and glasses, 300–305
- Random coil peptides
 CD spectra, 47, 49*f*
 immunogenic peptides, 56–57
 polypeptides and proteins, 63
- Reachability, Petri nets, 228, 230, 231*f*
- Real problems
 defining, 8–14
 molecular modeling, 1–16
- Receptor-bound ligands, determining solution conformation, 29–44
- Recursion, architecture of the model, 490–491
- Redox state dependent conformational change
 charged residues, 452
 correlation with location of residues, 453

- Redox states, conformational differences, 450
- Reflectance–absorbance infrared (RAIR) spectroscopy
 myoglobin films, 254*f*,256*f*
 orientation of DDAB and secondary structure of myoglobin, 253–255
- Relaxation mechanisms, liquid, 271,275
- Reorientation time
 calculated, 298*t*
 water molecules, 318
- Repeating sequence
 after 30 ps of dynamics calculations, 215*f*,CP12
 fiber-forming collagens, 145–146
 γ spiral, 120-residue, CP11
 peptide chain, 214*f*
 Ramachandran plot for structure, 216*f*
 stick model of structure, 215*f*
- Representation of compounds, biochemical functions, 493–494
- Representation of reactions, methods, 494–495
- Representational rationale, construction of a database, 489–492
- Residence time
 monovalent ions and water, 300*t*
 water in hydration shell of ions, 299
- Resolution-enhanced deconvolution, FT-IR, 71–98
- Resolution enhancement factor (REF), nonlinear regression analysis, 82
- Resolution limits, relative rigidity of a molecule, 21
- Reversibility, Petri nets, 228
- Ribbon structure, ribonuclease, 95
- Rigid lattice, concentration dependence of ^1H NMR second moments, 310*f*–312*f*
- Rigidity, molecules in a crystal, 21
- Root mean square (RMS) value
 influence of number of Gaussian peaks, 83*t*
 nonlinear regression fit, 82–83
- Rotational motions, influence on powder spectrum of partially ordered solid, 277*f*
- Roughness index, globular proteins, 55
- S
- ^{35}S -methionine labeled viruses, radioimmunoprecipitation with antipeptide antisera, 366*t*
- Salt, molecular dynamics of interactions with peptides, fibrous proteins, and casein, 420–445
- Salt binding, effect on dynamic structure of hydrophilic protein, 440–442
- Salting-in
 molecular basis in proteins, 420
 soy protein, 339
 thermodynamics, 434–442
- Secondary structure
 analytical methods, 97
 bond angles in aqueous crystals, 375
 bovine α -crystallin, FT-IR studies, 128–130
 proteins, prediction, 66–67
- Sequence-based algorithm, α_{s1} -casein, 370–371
- Sequence comparisons
 apolipoproteins, 101–102
 functional identities, 103
- Sequence substitution, Smith microfibril model, 147–150
- Sequence-to-structure pathway, proteins and enzymes, 5
- Serine
 conservation of magnitude, 204
 dipole moment frequency plot, 199*f*
- Serine phosphates, hydrophilic domains of α_{s1} -casein, 401
- Serine phosphorylation sites, α -crystallin, 134
- Serine proteases
 information bias, 12
 inhibitor-induced structural changes, 235–247
 trypsinlike subfamily, 240–247
- Serological data
 correlation with major antigen, 56
 helix-forming parameters, 57

- β -Sheet structure
 α -crystallin, 134–135
 ribonuclease, 95
- Side chains, tryptophan residues, 120f
- Single-value property, reliability, 195,202–204
- Slow-growth free energy calculations, reproducibility, 178
- Small-angle X-ray scattering (SAXS) profiles, calculation, 395–396
- Smith microfibril
 energy-minimized model, 151–153
 geometric constraints of models, 167
 incorporation of native type II collagen sequence, 147–150
 initial buildup of full microfibril model, 145
 modeling strategy, 142
 negative staining banding patterns, 140
 packing arrangement, 140
 structural and energetic refinement, 144–145,149–150
 type II collagen, 139–170
- Software
 algorithms, 189
 backbone and side chains, 142
 backbone conformations, 448
 BUILD approach, 457
 distance geometry approach to structure determination, 456–457
 energy-minimization calculations, 394–395
 modeling of α -crystallin subunits and complex, 125
 molecular dynamics simulations, 143
 molecular modeling, 90
 molecular modeling of apolipoproteins, 102–103
 potential energy function, 142
 sequence-based prediction of conformation, 212
 solvation modeling studies, 347
 spatial structure in solution, 449
 sulfonamide and β CD molecules, 174
 three-dimensional structure of α_{s1} -casein, 369–370
 three-dimensional structure of sTF, 115
- Solubility data, soy proteins, 339
- Soluble domain, human tissue factor, 113–122
- Soluble tissue factor (sTF)
 amide I band and structure assignments, 117t
 FT-IR spectrum, 116–117
 model for three-dimensional structure, 113–122
 theoretical model
 development, 116–117
 testing, 118–121
- Solution structure
 comparison by using local conformations, 448–449
 determination methods, 447
 receptor-bound ligands, 29–44
- Soy protein
 molecular dynamics of hydration, 338–340
 salt association, 339
- Spacer residue, alanine, 110
- Spatial structure, methods of determination, 456–457
- Spectral analyses of secondary structure, main goal, 66–67
- Spectral features, correlation with secondary structure, 62
- Spectroscopic modeling, structural picture of Mb-DDAB films, 265–266
- Spectroscopy
 Mb-DDAB films, 250–268
 molecular dynamics of aqueous solutions of electrolytes, 270
- Spin-lattice measurements
 LiCl solutions, 308f
 LiI solutions, 306f
- Spiral stabilization, hydrogen bonding of glutamine side-chain groups to backbone groups, 218
- Spiral structure
 β and γ turns, 213
 comparison in wheat and elastin by molecular modeling, 209–220
 compatibility with energy and dynamics calculations, 217–218
 glutenin subunits, 210

- Stability, Mb-DDAB films, 259–262
- Stabilization energy, docking β -casein dimers, 405
- Statistical mechanics, combined with Monte Carlo simulations, 270
- Storage proteins, wheat, 338
- Strand content, determined from FT-IR amide I results, 93–94
- Strip of helix
amphipathic potentials of sequence fragments, 103
template to determine amphipathic potential, 102
- Structural biology, protein, 4–6
- Structural changes in proteins, analysis methods, 244,246
- Structural comparison, strategies, 449
- Structural conservation, evolution, 185
- Structural elements, expressing relationships, 490–492
- Structural mapping, Petri net models, 232–233
- Structural perturbation, similarity in hydrophobicity, 187
- Structural properties, Petri nets, 228–229
- Structural proteins, lens transparency, 123
- Structural reduction, Petri nets, 230,232*f*
- Structural relationships, combination of elements, 493–494
- Structural relaxation processes, observed onset, 313
- Structure alignment, similarity between two proteins, 189
- Structure determination, NMR-based methods, 456
- Structure–function relationships
amino acid residues in collagen sequences, 165–167
amino acid substitutions in proteins, 185–208
collagen fibers, 151,153
collagen microfibril models, 167–168
native collagen fiber systems, 168
protein, 420
- Structure modeling, procedures, 115–116
- Structure sequence, α_{s1} -casein, 370–371
- Structure–serologic relationships,
foot-and-mouth disease virus, 362–367
- Submicellar casein, SAXS profiles, 414*f*–415*f*
- Submicelle(s), bovine casein, 392–419
- Submicelle structure
bovine casein, 412–417
casein, 404–408
secondary structural analysis, 405,407
spectroscopic data, 406*t*
symmetric and asymmetric models, 405
temperature factors from SAXS, 414*t*
- Substrate, selection, 34–35
- Subtilisin
active site of native and variant strains, 26*f*
distances between equivalent atoms, 26*f*
electron density map, 24*f*
reliability of structure determination, 18–28
trace of native and variant strains, 24*f*
X-ray diffraction pattern, 20*f*
- suc-AAPF-pNA in buffered D₂O solutions, NOE difference spectra, 37*f*–39*f*
- Sulfathiazone– β CD inclusion complex,
ball-and-stick representation, 183*f*
- Sulfonamide(s), schematic structures, 180*f*
- Sulfonamide– β CD inclusion complexes
association free energies, 181*t*
relative association free energies, 172–184
- Supercomputer, Monte Carlo simulations, 270
- Superposition of alpha carbons, RMSD values, 57*t*
- Superposition procedure, determination of structural equivalences, 189
- Supramolecular models, theoretical viability, 258
- Surfactant bilayers, dynamic nature, 265
- Symmetric model, submicelle structure, 405
- Synthetic submicelle, framework, 396–404
- T
- Target-based drug design, specificity and potency, 168
- Technology transfer, agricultural and pharmaceutical applications, 2–4

- Telopeptides, structure and analysis, 153,162
- Template
alternating arrangement of residues, 217
microfibril model for type II collagen, 141
turn conformation, 90
use in molecular modeling, 100–112
- Tertiary structural homology, computer assisted molecular modeling, 133
- Tetrafluoroborate hydration, molecular model, 322*f*
- Theoretical FT-IR spectra
calculations for amide I, 84
deconvolved, 85
- Thermal neutrons, intensity vs. scattering angle, 284*f*
- Thermodynamic cycle
calculation of association free energies, 177*f*
free energy change, 173
- Thermodynamic integration method, electrostatic free energy differences, 175
- Thermodynamic linkage model
¹⁹F NMR relaxation data, 318
Li⁺, Cl⁻, and water, 313,316
- Thr3-Pro14 dimer segment, stereo diagram of final 16 structures, CP24
- Three-dimensional structures, proteins in solution, probabilistic approach, 446–465
- Time correlation functions, vectors, 331*f*
- Tissue factor (TF), testing an FT-IR-consistent model of the soluble domain, 113–122
- Transferred nuclear Overhauser effect (TRNOE)
experimental aspects of measurements, 32–33
macromolecular–ligand complexes, 31*t*
principle, 30–33
receptor-bound ligands, 29–44
- Transitions, VCD and FT-IR, 64
- Trifluoroethanol (TFE), intramolecular electrostatic interactions, 48
- Tripeptide scheme, three-dimensional collagen structure, 163
- Triple-helical collagen segment, molecular dynamics simulation, 165,166*f*
- Tropocollagen
molecular dynamics simulations, 429–434
spacefill model of 180 residues, CP21
structure–function relationships, 420–445
three-dimensional structure, 433*f*–435*f*
- Trypsin
component bands, effects of choosing too few, 83
crystallographic data, flow chart, 241*f*
FT-IR results for percent extension, 94
hydrophobicity profiles and structure alignments, 197*f*
linear distance plots and structure alignments, 196*f*
standard conformations, 192*f*
subfamily, conformational perturbations, 240–247
- Trypsinogen
crystal structure, 240–244
difference FT-IR derivative spectra, 244*f*,245*f*
difference linear distance plots, 242*f*
- Tryptophan residue, local environment, 120–121
- Turn content, Ramachandran analysis, 96
- Type II collagen
amino acid sequences and collagen triple helices, 154*t*–160*t*
Smith microfibril, 139–170
- Type II microfibril model, molecular modeling, 146–150
- Tyrosine, position in spiral, 217
- U
- UV–VIS spectra, Mb–DDAB film, 260*f*
- V
- Vaccination, control of FMDV, 362
- Vaccines, FMDV preparation, 362–363
- Valine residues, position in spiral, 217

- van der Waals volume analysis
 computational procedure, 193
 structure-derived, 198
- Velocity correlation function (VCF), ions in
 aqueous solutions, 299–300,302*f*–303*f*
- Vibrational circular dichroism (VCD)
 applications, 67–69
 comparison to IR absorbance spectra, 64
 comparison to other spectroscopic
 methods, 62
 dependence on local chirality of
 chromophores, 64
 protein secondary structure, 61–70
- Vibrational spectroscopies, established
 role, 61
- Vibrational transitions, measurement
 methods, 78
- Virtual tools, molecular modeling, 1–16
- Virus
 foot-and-mouth disease (FMDV), 362–367
 neutralization with antipeptide antisera,
 366*t*
- W**
- Water
 and *N*-phenyl uronamide dimer, molecular
 dynamics simulations, 356–357
 bound to proteins, relaxation, 332
 hydrogen bonding in spiral, 218
 residence time in hydration shell, 299
- Water bridges
 hydroxyproline-based hydrogen
 bonds, 165
 Li⁺ and Cl⁻ ions, 290,294
- Water entrapment, NMR and molecular
 modeling evidence, 342–361
- Water molecules
 coordination of Li⁺ and Cl⁻ ions, 296*f*
 hydration shell, 297
- Weighted distribution function
 CaCl₂ solution, 292*f*
 NiCl₂ solution, 289*f*
- Wheat, high-molecular-weight glutenin
 subunits, 209–220
- Wheat protein, hydration and ion-binding
 properties, 338
- Working model(s)
 agreement with literature, 133–135
 α -crystallin subunit
 backbone, 127*f*,129*f*
 evidence for 24% α -helix, 131*f*
 predicted structure, 129*f*
 secondary structure validation, 128–132
 three-dimensional structure, 126–128
- X**
- X-Pro peptide bonds
cis-trans isomerization, 33–34
 transferred NOE measurements, 29–44
- X-ray crystallographic structures,
 appropriate algorithm, 89
 reliability, 18–28
- X-ray diffraction patterns, crystal
 structure determination, 19
- X-ray intensity functions, Metglas 2826,
 283*f*
- X-ray interference functions, noncrystalline
 Co_{0.9}P_{0.1}, 283*f*
- X-ray studies, local structure in aqueous
 solutions of electrolytes, 278–294
- Z**
- Zipperlike mechanism, hydroxyproline
 cluster, 167
- Zwitterions, molecular dynamics and
 multinuclear magnetic resonance
 studies, 325–419