

Computer-Assisted Drug Design

Computer-Assisted Drug Design

Edward C. Olson, EDITOR

The Upjohn Company

Ralph E. Christoffersen, EDITOR

University of Kansas

Based on a symposium
sponsored by the Divisions of
Computers in Chemistry and
Medicinal Chemistry at the
ACS/CSJ Chemical Congress,
Honolulu, Hawaii,
April 2-6, 1979.

A C S S Y M P O S I U M S E R I E S **112**

AMERICAN CHEMICAL SOCIETY

WASHINGTON, D. C. 1979

RS 401 .S95 1979 copy 1



Symposium on Computer
Assisted Drug Design (1979)

Computer-assisted drug
design

Library of Congress \square Data

Symposium on Computer-Assisted Drug Design, Honolulu, 1979.

Computer-assisted drug design.
(ACS symposium series; 112 ISSN 0097-6156)

"Based on a symposium sponsored by the Division of Computers in Chemistry [and the Medicinal Chemistry Division] at the ACS/CSJ Chemical Congress, Honolulu, Hawaii, April 2-6, 1979."

Includes bibliographies and index.

1. Chemistry, Pharmaceutical—Data processing—Congresses. 2. Structure-activity relationship (Pharmacology)—Data processing—Congresses. I. Olson, Edward C. II. Christoffersen, Ralph E., 1937— . III. American Chemical Society. Division of Computers in Chemistry. IV. American Chemical Society. Division of Medicinal Chemistry. V. ACS/CSJ Chemical Congress, Honolulu, 1979. VI. Title. VII. Series: American Chemical Society. ACS symposium series; 112.

RS401.S95 1979 615'.191'02854 79-21038
ISBN 0-8412-0521-3 ASCMC8 112 1-619 1979

Copyright © 1979

American Chemical Society

All Rights Reserved. The appearance of the code at the bottom of the first page of each article in this volume indicates the copyright owner's consent that reprographic copies of the article 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. 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 new collective works, for resale, or for information storage and retrieval systems.

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.

PRINTED IN THE UNITED STATES OF AMERICA

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

ACS Symposium Series

M. Joan Comstock, *Series Editor*

Advisory Board

Kenneth B. Bischoff

Donald G. Crosby

Robert E. Feeney

Jeremiah P. Freeman

E. Desmond Goddard

Jack Halpern

Robert A. Hofstader

James D. Idol, Jr.

James P. Lodge

John L. Margrave

Leon Petrakis

F. Sherwood Rowland

Alan C. Sartorelli

Raymond B. Seymour

Aaron Wold

Gunter Zweig

FOREWORD

The ACS SYMPOSIUM SERIES was founded in 1974 to provide a medium for publishing symposia quickly in book form. The format of the Series parallels that of the continuing ADVANCES IN CHEMISTRY SERIES except that in order to save time the papers are not typeset but are reproduced as they are submitted by the authors in camera-ready form. Papers are reviewed under the supervision of the Editors with the assistance of the Series Advisory Board and are selected to maintain the integrity of the symposia; however, verbatim reproductions of previously published papers are not accepted. Both reviews and reports of research are acceptable since symposia may embrace both types of presentation.

PREFACE

The existence of structure–activity relationships was recognized as early as 1868 by Crum–Brown and Frazer at Edinburgh University, and this postulate led, quite naturally, to attempts to modify (design) structures via chemical procedures to demonstrate increased biological activity. For many years this type of drug design proceeded empirically through the “chemical intuition” of organic and medicinal chemists. As the empirical data base was created, relationships were observed and methods for quantitating these correlations were developed.

The advent of the “age of computers” has made possible the computation of many indices not previously accessible and the use of sophisticated correlation methods to discover correlations among these indices in multivariable systems.

The diverseness of the methods that are currently available and their application is indicated by the far ranging content of this symposium on Computer-Assisted Drug Design. The goal of the symposium was to bring together experts in many of the specific disciplines that have been applied to drug design, to explore the interaction between these areas, and to examine the role of the computer in the overall process of drug design. The major areas treated by participants include receptor modeling, organic syntheses, quantitative structure–activity relations, quantum mechanics, pattern recognition, and molecular mechanics.

It is obviously not possible to cover in depth such a diverse array of topics in a single symposium. Therefore, it was the intent to establish the current state of the art in each of the previously mentioned fields and to show the necessity for the application of all methods in any integrated program of drug design.

There can be no doubt that the art or science (as the case may be) of drug design is still in its infancy, and that significant improvements in all technologies are essential to permit progress toward the goal of designing therapeutic agents for specific maladies. However, the potential rewards for success are enormous, both from a practical and fundamental point of view, which increases the urgency of such efforts. Indeed, efforts to sharpen the tools of both “lead-finding” and “lead-optimization” techniques and to apply them early-on in the development

of novel or improved therapeutic agents may well be essential for the continued development of new drugs as the costs of drug development escalate far more rapidly than the general inflationary increase.

The Upjohn Company
Kalamazoo, MI 49001

EDWARD C. OLSON

University of Kansas
Lawrence, KS 66045
June 8, 1979.

RALPH E. CHRISTOFFERSEN

Quantum Pharmacology: Recent Progress and Current Status

RALPH E. CHRISTOFFERSEN

Chemistry Department, University of Kansas, Lawrence, KS 66045

As even a casual perusal of the current literature will indicate, the development of theoretical techniques and application to problems in biology in general and drug design in particular has expanded substantially in recent years. This has been due in part to the appearance of techniques and computing equipment that allow examination of large molecular systems, but has also been motivated substantially by the potential practical use of these techniques in the design of new medicinal agents. When combined with similar advances in experimental biochemistry and pharmacology, the beginnings of a new field can be identified, in which problems of interest to pharmacology can be examined at the molecular level.

It is the status of these efforts that are of interest in the current symposium. More particularly, in this opening review we shall focus primarily upon the use of theoretical techniques to study problems in pharmacology. In order to produce a study of manageable size, this review will include only those studies involved with drug design *per se*, i.e., theoretical studies of isolated drug molecules, solvated drug molecules, or model drug-receptor interactions of potential use in the design of new medicinal agents. It will therefore omit a large number of important studies on other biological problems such as photosynthesis, the structure and function of nucleic acids, polypeptides and proteins, and many others. However, as will become apparent below, the activity in the area of "theoretical drug design" itself has been high, with significant progress occurring in several aspects.

Literature Review

Since the work in this area has been reviewed previously (1, 2) for the period up through the end of 1975, only studies published during 1976-1978 will be included here. In this period,

0-8412-0521-3/79/47-112-003\$05.00/0

© 1979 American Chemical Society

80 studies of interest to this review appeared and were found in 13 different journals and books. Combined with the 130 articles prior to 1974 (1) and the 24 studies reported in 1975 (2), it is seen that over 230 studies have now appeared in this area.

The theoretical techniques used for the studies in this review include ab initio techniques (3-6), semi-empirical techniques such as PCILO (7), CNDO and INDO (8, 9), IEHT (10-12), MINDO (13), and HMO (14), empirical potentials (15-17), and other techniques (18-21). In addition, several review articles and books in this area have appeared (22-24) during this period.

The 80 studies themselves (25-104) are summarized in Table 1, where they have been categorized by both the kind of agents that were studied and the methodology used. Before discussing these further, it is important to note several things. First, the assignment of a study to a particular type of agent frequently cannot be done unambiguously due to the fact that drugs frequently have more than one kind of activity. Thus, the entries in Table 1 may have applicability in more than simply the category in which they have been listed. Also, in some cases several different theoretical techniques have been employed, in which case a given study may have more than one entry in Table 1.

Next, it is important to remember that the action of a drug is determined in general by a variety of factors, including:

1. Absorption.
2. Excretion.
3. Catabolism.
4. Binding to Plasma Protein.
5. Penetration of the Blood-Brain Barrier.
6. Affinity for the Receptor.
7. Intrinsic Activity.

Each of these factors needs to be considered if a detailed understanding of drug action is to be obtained. However, the theoretical studies described in this review will deal typically with only one or two of the factors described above (e.g., receptor affinity and/or intrinsic activity), and will be of direct aid in drug design only if the other factors have a relatively constant contribution in the series of drugs studied.

Turning to the specific studies reported during 1976-78, there are several general comments of interest. Of greatest importance among these is the demonstrable change in the sophistication of the techniques used in theoretical studies. For example, only 8% of the studies reported prior to 1975 (1) used ab initio quantum mechanical techniques, while over 25% of the studies reported in the 1975-78 period used ab initio techniques. While the use of ab initio techniques does not per se indicate an improvement in the accuracy of the results (due to limited basis sets and other limitations), it does indicate that a "state-of-the-art" that was previously thought to be obtainable for "small" or "medium-sized" chemical systems is now available for use on the substantially larger systems of typical interest

Table 1. Theoretical Studies on Systems of Pharmacological Interest.

Agent	<u>Method Used</u>					
	IEHT	CNDO/2	INDO	PCILO	Ab Initio	Other
<u>A. Agents Affecting Nerve Function</u>						
1. Adrenergics			68 104		46	28,38 87
2. Analgesics	80	35,66	80	30,40, 62,76, 80,94, 96,98	34,36, 74,75, 99	61,63, 78,98, 100
3. Analeptics						
4. Anaesthetics		84				27,31
5. Anti-anginal Agents						
6. Anti-anxiety Agents						
7. Anti-arrhythmic Agents						
8. Anticonvulsants and Muscle Relaxants		48				53
9. Antidepressants and Stimulants		32				
10. Antiparkinsonian Agents						
11. Antipsychotics						
12. Antitussives						
13. Cholinergics and Anticholinergics		37	44	39		56
14. Emetics and Anti-emetics						
15. Hallucinogens		97,70			25,43, 64,71, 73,77, 85	79,97
16. Hypnotics and Depressants		69				

Table 1. (cont.)

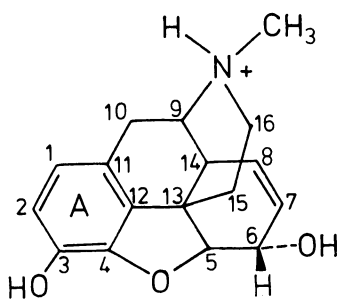
Agent	IEHT	CNDO/2	INDO	PCILO	Ab Initio	Other
<u>B. Agents Affecting Cardiovascular Function</u>						
1. Vasodilators and Vasoconstrictors						42,102
2. Hypotensive and Hypertensive Agents						45
3. Antihypercholesterol Agents						
4. Anticoagulants						
5. Cardiac Glycosides						
6. Vasopressor Agents						
<u>C. Agents Affecting Endocrine Function</u>						
1. Steroid Hormones						
2. Prostaglandins						
3. Polypeptide, Protein, and Other Hormones						88,95
4. Thyroid Agents						
5. Hypoglycaemic Agents			49			
<u>D. Anti-infective and Chemotherapeutic Agents</u>						
1. Anti-amoebic Agents						
2. Antibiotics		58		51,101	92	54,60,65,93

Table 1. (cont.)

Agent	IEHT	CNDO/2	INDO	PCILO	Ab Initio	Other
3. Anticancer Agents		52,57, 67,89, 90,91			55,81, 83,90	41,86
4. Antifungal Agents						
5. Antimalarial Agents				50		
6. Antimycobacterial Agents						
7. Antiparasitics						
8. Antiprotozoan Agents						
9. Antiseptics						
10. Antivirals	33 59 82					
<u>E. Miscellaneous</u>						
1. Antihistaminic and Antiallergenic Agents and Histamine					26,29, 72	
2. Carbohydrates					103	
3. Vitamins						
4. Insecticides and Insect Hormones						
5. Antiinflammatory Agents		47				
6. Antisecretory Agents						

- a. References are listed chronologically by year. Chronological ordering within a given year was not attempted. While it was intended that all relevant references in the period 1976-1978 should be included in this table, there are undoubtedly many inadvertent omissions for which the author apologizes in advance. Additional references from this period are welcomed.
- b. Entries in the table refer to the reference that describes the study that was carried out.

Figure 1. Morphine structure and atom numbering



in drug design.

The techniques have evolved in several other ways as well. For example, electrostatic potential maps have been found to be a quite useful and sensitive tool (25, 26, 41, 44, 46, 57, 71, 75, 99) for identifying reactivity characteristics of drugs, at least in electrophilic reactions. Since these maps can be calculated directly from the wavefunction, and are more sensitive measures of the electronic charge distribution for chemical reactions than, e.g., the net charges and bond orders from a population analysis, their introduction and use represent an important development in the kinds of theoretical tools available for use. In addition, the use of statistical mechanics in the interpretation of conformational energies has finally begun to be considered (28, 79). Also, the use of pseudopotential theory, at least for cases when second-row atoms are involved, has been tried in several cases (43, 64, 83, 100), and shows good potential.

Not only has the methodology developed significantly, but the models used to represent biological events have improved substantially in many cases. For example, not only have isolated molecules been studied, but models of the drug-receptor complex have been developed in several cases (37, 43, 52, 56, 71, 77, 81, 89, 90, 97). In other cases, models for the transition state itself have been investigated (60). In addition, solvent and other environmental effects are included more commonly than before (28, 38, 39, 40, 49, 55, 79), even though it is recognized that a fully satisfactory model of solvent effects that is computationally feasible is not yet available.

On the other hand, there are a variety of areas where substantial developmental efforts in methodology and/or biological models are needed. Perhaps most striking is the need for inclusion of substantially larger numbers of geometric degrees of freedom when studying conformational characteristics of molecules. While the study of the energetic dependence of one or two dihedral angle variations is perhaps useful as a starting point, it is clear that such studies are limited in the conclusions that can be extracted from them, and may in some cases lead to incorrect conclusions.

For example, even in the case of "rigid" opiates, it is clear that the flexibility present in the strained, multi-ring structure is significant and cannot be ignored. Even the C₁₂-C₁₃ bond in morphine (see Figure 1), which forms the nexus of each of the five rings, can be shown to allow a 30° variation within 3 kcal/mole or less, using either experimental or calculated results (74). This is an important flexibility that should not be ignored for several reasons. First, changes in this angle will change substantially the distance from the quaternary nitrogen to the aromatic moiety (Ring A in Figure 1), which is a feature that is thought by many to be important in opiate-receptor interactions. Next, if comparisons to other opiates

are made and the molecular flexibilities of morphine and the comparison molecule are ignored, it is possible to conclude that no structural similarity exists when in fact such similarities do exist. For example, when only two dihedral angle variations were considered(105, 106), it was concluded that meperidine and morphine could not be matched from a geometric point of view. On the other hand, if all bond distances, bond angles and dihedral angles are allowed to relax (74), a good fit of the two molecules can be achieved at only modest energetic cost (~3.6 kcal/mole for the axial form of meperidine, and 6.9 kcal/mole for the equatorial form of meperidine).

In a related issue, the development of reliable, fast, and generally applicable empirical potentials (or other procedures) that will allow examination of large numbers of degrees of freedom in a wide variety of classes of organic compounds remains an important unsolved problem. Just as in the case of neglect of geometric degrees of freedom, the use of naive approximations or potential functions that omit obvious terms of importance cannot be expected to provide reliable results (78). More generally, while there have been substantial contributions to the development of empirical potentials for specialized kinds of molecules, e.g., polypeptides (107), corresponding potentials for the broad range of molecules that are typically encountered in drug design are not currently available.

New Directions

In addition to the developments already referred to above, there are several other new directions that can be identified in the recent literature, that indicate the kinds of additional advances that can be expected in the next several years.

As a first example in this regard, it is of interest to consider a study reported by E. Clementi and collaborators (108) in 1978. In this paper, the manner of treating enzymic reactions using theoretical techniques was generalized substantially from earlier work. Since the description of enzyme-substrate interactions encounters the same kinds of difficulties as the description of drug-receptor interactions, this study may provide a good indication of new directions and capabilities that can be expected to develop further in the near future.

In this approach, enzymic reactions are considered to be divisible into two separate but related steps called "macrodeformations" and "microdeformations." Macrodeformations are intended to include the structural alterations that affect the entire enzyme plus substrate during the approach and reaction of the substrate and departure of the products. On the other hand, microdeformations recognize that, after the overall enzyme-substrate structure is approximately determined after approach of the substrate, a detailed description of the chemical reaction needs to be considered. These considerations assume that the

reaction is a relatively "localized" event, and that only a relatively small number of moieties in and around the active site plus the substrate need to be considered. Of course, the two processes cannot be rigorously separated as described, and a sequence of macrodeformations plus microdeformations is used to model the various steps along the reaction path.

For the description of the two kinds of processes, different kinds of theoretical techniques were used. For macrodeformations, the primary processes that are assumed to be occurring are a large number of non-bonded interactions that lead to the introduction of the substrate into the active site (or departure of the products). For such processes, empirical potentials are appropriate to use for several reasons. First, evaluation of the total energy can be done rapidly enough so that a large number of atoms and degrees of freedom can be examined. Next, as long as no bonds are made or broken, it is possible for empirical potentials to provide a satisfactory theoretical framework.

In the case of microdeformations, the actual chemical reaction at the active site is considered, and a detailed ab initio quantum mechanical formulation is used. For the initial studies described in 1978, a minimum basis set SCF calculation was carried out, but the concept is quite general.

As a first example, the cysteine proteinase, papain, was studied in its catalytic cleavage of peptide substrates. The enzyme contains over 200 residues, and the particular mechanism that has been proposed by Drenth, et al (109) was chosen for examination.

In the macrodeformation steps, no less than 73 residues plus substrate were considered, starting essentially from papain coordinates from X-ray studies. In the main, the results of macrodeformation studies of substrate entry are structures that are relaxed energetically from the X-ray crystal coordinates, including the formation of several new hydrogen bonds and the movement of several atoms in the active site in a manner that will facilitate the subsequent enzyme-substrate reaction. In the microdeformation studies, low energy structures for various intermediates were sought, including the form of the attacking species and the structure of the tetrahedral intermediate.

It is not expected that either the structure "refinements" of the enzyme or substrate that occurred during macrodeformation steps or the energetic or structural features found as a result of microdeformation steps are reliable enough at this stage to provide conclusive new insight into the enzymatic behavior of papain. However, these studies indicate clearly that the "state-of-the-art" has now changed substantially, and that meaningful models of enzyme-substrate or drug-receptor interactions can now be examined using theoretical techniques.

Another example that illustrates new directions of interest relates to the experimental characterization at the molecular

level of drug-receptor complexes. Specifically, the X-ray studies of Sobell and collaborators (110-112) on the binding of drugs to DNA models provide a good example of the level of detail in describing drug-receptor interactions that is now possible experimentally.

The particular system that was studied involved the binding of ethidium, a known DNA intercalator, to two different model DNA structures [5-iodouridylyl (3'-5')adenosine and 5-iodocytidylyl (3'-5') guanosine], followed by crystallization of the complexes and acquisition plus analysis of sufficient X-ray data to allow atomic resolution (i.e., standard deviations of $\pm 0.1\text{\AA}$ for bond lengths and $\pm 5^\circ$ for bond angles). The DNA-drug models used embody a large number of the features known to be important, i.e., double stranded DNA "mini-helices" are formed, with ethidium intercalated between two pairs of hydrogen bonded bases that are connected by the usual sugar-phosphate backbone.

These studies give rise to a variety of important structural features of such complexes, as well as to provide several mechanistic hypotheses that can be investigated further using theoretical techniques. For example, these studies suggest that drug intercalation results in a helical screw dislocation in the DNA double helix, whose magnitude determines the relative ring overlap between an intercalated drug molecule and adjacent base pairs. Also, conformational changes in the sugar-phosphate backbone accompany drug intercalation, which led to the suggestion that DNA "kinks" first, followed by intercalation of the drug. Such observations and suggestions, while very important contributions, also raise a series of questions which need further study, and for which theoretical techniques are now available.

Thus, both theoretical and experimental studies indicate that significant new directions can be expected in the next several years that will enhance substantially the ability to carry out drug design using techniques heretofore not available. Furthermore, current and likely developments in both hardware and software for minicomputers will extend even further the opportunities for use of sophisticated theoretical techniques on problems in drug design.

Acknowledgements

The author would like to express his deep appreciation to the Upjohn Company for their continuing support of this research, and to Drs. E.C. Olson, B.V. Cheney, D. Duchamp and many others at the Upjohn Company for their excellent comments and collaboration over many years. In addition, support for this symposium and to this author from the Harris Corporation is gratefully acknowledged.

Literature Cited

1. Christoffersen, R.E. in "Quantum Mechanics of Molecular Conformations," B. Pullman, ed., J. Wiley and Sons, New York, 1976, Chapter 3. "Molecules of Pharmacological Interest," 194-285.
2. Christoffersen, R.E. and Angeli, R.P. in "The New World of Quantum Chemistry," B. Pullman and R. Parr, eds., 189-210, D. Reidel Publishing Co., Dordrecht-Holland, 1976.
3. Whitman, D.R. and Hornback, C.J., J. Chem. Phys., (1969), 51, 398-402.
4. Hehre, W.J., Stewart, R.F., and Pople, J.A., J. Chem. Phys., (1969), 51, 2657-2664.
5. Clementi, E., Proc. Nat. Acad. Sci.(USA), (1972), 69, 2942-2944.
6. Christoffersen, R.E., Adv. Quantum Chem., (1972), 6, 333-393.
7. Diner, S., Malrieu, J.P., Jordan, F. and Gilbert, M., Theoret. Chim. Acta, (1969), 15, 100-111.
8. Pople, J.A., Santry, D.P., and Segal, G.A., J. Chem. Phys., (1965), 43S, S129-S135.
9. Pople, J.A., Beveridge, D.L., and Dobosh, P.A., J. Chem. Phys., (1967), 47, 2026-2033.
10. Rein, R., Fukuda, N., Win, H., Clarke, G.E., and Harris, F., J. Chem. Phys., (1966), 45, 4743-4744.
11. Carrol, D.G., Armstrong, A.T., and McGlynn, S.P., J. Chem. Phys., (1966), 44, 1865-1870.
12. Zerner, M. and Gouterman, M., Theoret. Chim. Acta, (1966), 4, 44-63.
13. Dewar, M.J.S. and Haselbach, E., J. Am. Chem. Soc., (1970), 92, 590-598.
14. See "Quantum Biochemistry" by B. Pullman and A. Pullman, Interscience, N.Y., 1963.
15. Caillet, J. and Claverie, P., Acta Cryst., (1975), A31, 448-460.

16. Weintraub, H.J.R. and Hopfinger, A.J., J. Theoret. Biol., (1973), 41, 53-75.
17. Momany, F.A., McGuire, R.F., Burgess, A.W., and Scheraga, H.A., J. Phys. Chem., (1975), 79, 2361-2381.
18. Popkie, H.E. and Kaufman, J.J., Int. J. Quantum Chem., (1975), QB2, 279-288; (1976), 10, 47-52.
19. Bonifacic, V. and Huzinaga, S., J. Chem. Phys., (1974), 60, 2779-2786; (1975), 62, 1507-1508; (1975), 62, 1509-1512.
20. Ewig, C.S. and van Wazer, J.R., J. Chem. Phys., (1975), 63, 4035-4041.
21. Davies, R.H., Bagnall, R.D., and Jones, W.G.M., Int. J. Quantum Chem., (1974), QB1, 201-212.
22. Pullman, B., Adv. Quantum Chem., (1977), 10, 251-328.
23. Kaufman, J.J., Int. J. Quantum Chem., (1976), QB3, 187-216; (1977), QB4, 375-412.
24. Richards, W.G., "Quantum Pharmacology," Butterworths, London, 1977.
25. Weinstein, H., Chou, D., Kang, S., Johnson, C.L., and Green, J.P., Int. J. Quantum Chem., Quantum Biol. Symp., (1976), 3, 135-150.
26. Weinstein, H., Chou, D., Johnson, C.L., Kang, S., and Green, J.P., Molec. Pharmacol., (1976), 12, 738-745.
27. Davies, R.H., Bagnall, R.D., Bell, W., and Jones, W.G.M., Int. J. Quantum Chem., Quantum Biol. Symp., (1976), 3, 171-186.
28. Kumbar, M., J. Med. Chem., (1976), 19, 1232-1238.
29. Richards, W.G. and Wallis, J., J. Med. Chem., (1976), 19, 1250-1252.
30. Loew, G.H., Berkowitz, D.S. and Newth, R.C., J. Med. Chem., (1976), 19, 863-869.
31. Hopfinger, A.J. and Battershell, R.D., J. Med. Chem., (1976), 19, 569-573.
32. Johnson, C.L., J. Med. Chem., (1976), 19, 600-605.

33. Miles, D.L., Miles, D.W., Redington, P. and Eyring, H., Proc. Nat. Acad. Sci. (USA), (1976), 73, 4257-4260.
34. Popkie, H.E., Koski, W.S. and Kaufman, J.J., J. Am. Chem. Soc., (1976), 98, 1340-1343.
35. Kaufman, J.J. and Kerman, E., Int. J. Quantum Chem., (1976), 10, 559-567.
36. Popkie, H.E. and Kaufman, J.J., Int. J. Quantum Chem., (1976), 10, 569-580.
37. Hall, L.H. and Kier, L.B., J. Theoret. Biol., (1976), 58, 177-195.
38. Caillet, J., Claverie, P. and Pullman, B., Acta Cryst., (1976), B32, 2740-2745.
39. Pullman, B., "Environmental Effects On Molecular Structure and Properties," B. Pullman, ed., 55-80, D. Reidel Publishing Co., Dordrecht-Holland, 1976.
40. Loew, G., Weinstein, H. and Berkowitz, D., "Environmental Effects on Molecular Structure and Properties," B. Pullman, ed., 239-258, D. Reidel Publishing Co., Dordrecht-Holland, 1976.
41. Giessner-Prette, C. and Pullman, B., Comp. Rend., (1976), D, 675-677.
42. Khaled, Md.A., Renugopalakrishnan, V. and Urry, D.W., J. Am. Chem. Soc., (1976), 98, 7547-7553.
43. Weinstein, H. and Osman, R., Int. J. Quantum Chem., Quantum Biol. Symp., (1977), 4, 253-268.
44. Weinstein, H., Srebrenik, S., Maayani, S. and Sokolovsky, M., J. Theoret. Biol., (1977), 64, 295-309.
45. Timmermans, P.B.M.W.M. and Van Zwieten, P.A., J. Med. Chem., (1977), 20, 1636-1644.
46. Petrongolo, C., Macchia, B., Macchia, F. and Martinelli, A., J. Med. Chem., (1977), 20, 1645-1653.
47. Gund, P. and Shen, T.Y., J. Med. Chem., (1977), 20, 1146-1152.
48. Blair, T. and Webb, G.A., J. Med. Chem., (1977), 20, 1206-1210.

49. Hall, W.R., J. Med. Chem., (1977), 20, 275-279.
50. Loew, G.H. and Sahakian, R., J. Med. Chem., (1977), 20, 103-106.
51. Saran, A., Mitra, C.K. and Pullman, B., Int. J. Quantum Chem., Quantum Biol. Symp., (1977), 4, 43-54.
52. Pack, G.R. and Loew, G.H., Int. J. Quantum Chem., Quantum Biol. Symp., (1977), 4, 87-96.
53. Weintraub, H.J.R., Int. J. Quantum Chem., Quantum Biol. Symp., (1977), 4, 111-126.
54. Boyd, D.B., Int. J. Quantum Chem., Quantum Biol. Symp., (1977), 4, 161-168.
55. Thomson, C., Provan, D. and Clark, S., Int. J. Quantum Chem., Quantum Biol. Symp., (1977), 4, 205-216.
56. Holtje, H.-D., Int. J. Quantum Chem., Quantum Biol. Symp., (1977), 4, 245-252.
57. Politzer, P. and Daiker, K.C., Int. J. Quantum Chem., Quantum Biol. Symp., (1977), 4, 317-326.
58. Brown, R.E., Simas, A.M. and Bruns, R.E., Int. J. Quantum Chem., Quantum Biol. Symp., (1977), 4, 357-362.
59. Miles, D.L., Miles, D.W., Redington, P. and Eyring, H., J. Theoret. Biol., (1977), 67, 499-514.
60. Boyd, D.B., Proc. Nat. Acad. Sci. (USA), (1977), 74, 5239-5243.
61. Isogai, Y., Nemethy, G. and Scheraga, H., Proc. Nat. Acad. Sci. (USA), (1977), 74, 414-418.
62. Kaufman, J.J. and Kerman, E., Int. J. Quantum Chem., (1977), 11, 181-184.
63. Momany, F.A., Biochem. Biophys. Res. Commun., (1977), 75, 1098-1103.
64. Osman, R. and Weinstein, H., Chem. Phys. Lett., (1977), 49, 69-74.
65. Ughetto, G. and Waring, M.J., Mol. Pharmacol., (1977), 13, 579-584.

66. Riga, J., Verbist, J.J., Degelaen, J., Tollenaere, J.P. and Koch, M.H.J., Mol. Pharmacol., (1977), 13, 892-900.
67. Gund, P., Poe, M. and Hoogsteen, K.H., Mol. Pharmacol., (1977), 13, 1111-1115.
68. Grol, C.J. and Rollema, H., J. Pharm. Pharmacol., (1977), 29, 153-156.
69. Domelsmith, L.N., Munchausen, L.L. and Houk, K.N., J. Am. Chem. Soc., (1977), 99, 6506-6514.
70. Domelsmith, L.N., Munchausen, L.L. and Houk, K.N., J. Am. Chem. Soc., (1977), 99, 4311-4321.
71. Weinstein, H., Osman, R., Edwards, W.D. and Green, J.P., Int. J. Quantum Chem., Quantum Biol. Symp., (1978), 5, 449-461.
72. Cheney, B.V., Wright, J.B., Hall, C.M., Johnson, H.G. and Christoffersen, R.E., J. Med. Chem., (1978), 21, 936.
73. Anderson, G.M. III, Castagnoli, N., Jr. and Kollman P.A., "QuaSAR," Res. Monog. 22, G. Barnett, M. Trsic and R. Willette, eds., 199-217, Nat. Inst. on Drug Abuse, 1978.
74. Cheney, B.V., Duchamp, D.J. and Christoffersen, R.E., "QuaSAR" Res. Monog. 22, G. Barnett, M. Trsic and R. Willette, eds., 218-249, Nat. Inst. on Drug Abuse, 1978.
75. Kaufman, J.J., "QuaSAR" Res. Monog. 22, G. Barnett, M. Trsic, and R. Willette, eds., 250-277, Nat. Inst. on Drug Abuse, 1978.
76. Loew, G.H., Berkowitz, D.S. and Bart, S.K., "QuaSAR" Res. Monog. 22, G. Barnett, M. Trsic, and R. Willette, eds., 278-316, Nat. Inst. on Drug Abuse, 1978.
77. Weinstein, H., Green, J.P., Osman, R. and Edwards, W.D., "QuaSAR" Res. Monog. 22, G. Barnett, M. Trsic, and R. Willette, eds., 333-358, Nat. Inst. on Drug Abuse, 1978.
78. Froimowitz, M., "QuaSAR" Res. Monog. 22, G. Barnett, M. Trsic, and R. Willette, eds., 359-373, Nat. Inst. on Drug Abuse, 1978.
79. Kumbar, M. "QuaSAR" Res. Monog. 22, G. Barnett, M. Trsic, and R. Willette, eds., 374-407, Nat. Inst. on Drug Abuse, 1978.
80. Katz, R., Osborne, S., Ionescu, F., Andrusis, P. Jr., Bates, R., Beavees, W., Chou, P.C.C., Loew, G. and Berkowitz, D., "QuaSAR" Res. Monog. 22, G. Barnett, M. Trsic, and R. Willette, eds., 441-463, Nat. Inst. on Drug Abuse, 1978.

81. Lavery, R. Pullman, A., and Pullman, B., Int. J. Quantum Chem., Quantum Biol. Symp., (1978), 5, 15-20.
82. Miles, D.L. and Eyring, H., Int. J. Quantum Chem., Quantum Biol. Symp., (1978), 5, 173-190.
83. Kaufman, J.J., Popkie, H.E. and Preston, H.J.T., Int. J. Quantum Chem., Quantum Biol. Symp., (1978), 5, 201-218.
84. Davies, R.H., Mason, R.C., Smith, D.A., McNeillie, D.J., and James, R., Int. J. Quantum Chem., Quantum Biol. Symp., (1978), 5, 221-244.
85. Domelsmith, L.N. and Kouk, K.N., Int. J. Quantum Chem., Quantum Biol. Symp., (1978), 5, 257-268.
86. Smith, I.A. and Seybold, P.G., Int. J. Quantum Chem., Quantum Biol. Symp., (1978), 5, 311-320.
87. Weintraub, H.J.R. and Nichols, D.E., Int. J. Quantum Chem., Quantum Biol. Symp., (1978), 5, 321-344.
88. Momany, F.A., Drake, L.G. and AuBuchon, J.R., Int. J. Quantum Chem., Quantum Biol. Symp., (1978), 5, 381-392.
89. Del Conde, G., Estrada, M. and Cardenas, A., Int. J. Quantum Chem., Quantum Biol. Symp., (1978), 5, 393-402.
90. Pack, G.R., Loew, G.H., Yamabe, S., and Morokuma, K., Int. J. Quantum Chem., Quantum Biol. Symp., (1978), 5, 417-432.
91. Marsh, M.M. and Jerina, D.M., J. Med. Chem., (1978), 21, 1298-1301.
92. Testa, B. and Mihailova, D., J. Med. Chem., (1978), 21, 683-686.
93. Gerson, S.H., Worley, S.D., Bodor, N., and Kaminski, J.J., J. Med. Chem., (1978), 21, 686-688.
94. De Graw, J.I., Lawson, J.A., Crase, J.L., Johnson, H.L., Ellis, M., Uyeno, E.T., Loew, G.H., and Berkowitz, D.S., J. Med. Chem., (1978), 21, 415-422.
95. Momany, F.A., J. Med. Chem., (1978), 21, 63-68.
96. Loew, G.H. and Berkowitz, D.S., J. Med. Chem., (1978), 21, 101-106.

97. Dipaolo, T., Hall, L.H. and Kier, L.B., *J. Theoret. Biol.*, (1978), 71, 295-309.
98. Loew, G.H. and Burt, S.K., *Proc. Nat. Acad. Sci.(USA)*, (1978), 75, 7-11, and 6337.
99. Petrongolo, C., Preston, H.J.T. and Kaufman, J.J., *Int. J. Quantum Chem.*, (1978), 13, 457-468.
100. Kaufman, J.J., Popkie, H.E., Palalikit, S. and Hariharan, P.C., *Int. J. Quantum Chem.*, (1978), 14, 793-800.
101. Samuni, A. and Meyer, A.Y., *Mol. Pharmacol.*, (1978), 14, 704-709.
102. Urry, D.W., Khaled, M.A., Renugopalakrishnan, V. and Rapaka, R.S., *J. Am. Chem. Soc.*, (1978), 100, 696-705.
103. Jeffrey, G.A., Pople, J.A., Binkley, J.S. and Vishveshwara, S., *J. Am. Chem. Soc.*, (1978), 100, 373-379.
104. Shinagawa, Y. and Shinagawa, Y., *J. Am. Chem. Soc.*, (1978), 100, 67-72.
105. Loew, G. and Jester, J., *J. Med. Chem.*, (1975), 18, 1051-1056.
106. Loew, G., Jester, J., Berkowitz, D. and Newth, R., *Int. J. Quantum Chem., Quantum Biol. Symp.*, (1975), 2, 25-34.
107. Momany, F.A., McGuire, R.F., Burgess, A.W. and Scheraga, H.A., *J. Phys. Chem.*, (1975), 79, 2361-2381.
108. Bolis, G., Ragazzi, M., Salvaderi, D., Ferro, D.R. and Clementi, E., *Gazz. Chim. Ital.*, (1978), 108, 425-443.
109. Drenth, J., Kalk, K.H. and Swen, H.M., *Biochem.*, (1976), 15, 3731-3738.
110. Tsai, C., Jain, S.C. and Sobell, H.M., *J. Mol. Biol.*, (1977), 114, 301-315.
111. Jain, S.C., Tsai, C. and Sobell, H.M., *J. Mol. Biol.*, (1977), 114, 317-331.
112. Sobell, H.M., Tsai, C., Jain, S.C. and Gilbert, S.G., *J. Mol. Biol.*, (1977), 114, 333-365.

RECEIVED June 8, 1979.

Parameters and Methods in Quantitative Structure-Activity Relationships

ROMAN OSMAN, HAREL WEINSTEIN, and JACK PETER GREEN

Department of Pharmacology, Mount Sinai School of Medicine
of The City University of New York, New York, NY 10029

In studies of quantitative structure activity relationships (QSAR), the relative potencies of a series of drugs are subjected to analysis with the hope that biological potency will be described by a mathematical equation. QSAR is an actuarial or statistical method in which only objective data are used with no intrusion of models or mechanistic hypotheses. The equation that is obtained not only accounts for the relative potencies of the compounds, but from it are deduced predictions of the potencies of untested compounds; if the equation is valid, the predictions are ineluctable. The method thus has the capacity of yielding new (structurally related) drugs with desired potency, perhaps drugs with enhanced selectivity or fewer side effects.

Further, in the method that is most widely used, the equation contains a parameter(s) that reflects a property related to the structural characteristics of the molecules. Therefore, the equation has often been assumed to provide insight into mechanisms that determine the action of drugs. In fact, the parameters that account for the biological potency are usually not truly empirical, observed entities, but rather abstractions. These parameters are frequently reified, burdened with a concrete meaning they were not designed to bear, and in this process the true molecular mechanisms may become obfuscated. Further, to infer a mechanism from a correlation is to assume that the correlation implies a causal relationship between the dependent and independent variables, when in fact the two may be conjoint but not causally related. Causality can be inferred only if a testable mechanistic hypothesis is offered that relates the correlation parameters to the observed biological activity.

The attempt to correlate biological activity with chemical structure in quantitative terms assumes that a functional dependence exists between the observed biological response and certain physicochemical properties of molecules. Without a mechanistic model one obtains it by an empirical correlation. A rational way to define such empirical relationships is within the extrathermodynamic approach (1). Although extrathermodynamic relationships

0-8412-0521-3/79/47-112-021\$13.55/0

© 1979 American Chemical Society

do not directly follow from the axioms of thermodynamics, they are still able to express, in a simple way, relations among free energies and other thermodynamic properties. Furthermore, extrathermodynamic relationships also implicitly contain the relation between microscopic properties (e.g., electronic densities, atomic structure, etc.) and macroscopic properties (e.g., free energies). Hence, extrathermodynamic relationships can be used to relate observable properties to structure.

A. The extrathermodynamic derivation of QSAR parameters

We derive the extrathermodynamic relationships as a basis for the definition and understanding of the parameters for QSAR. This derivation is similar to that of Grunwald and Leffler (1).

The theoretical derivation of extrathermodynamical relationships is based on the arbitrary division of a molecule into regions, usually two: the basic structure R, and the substituent X (1, 2). An additivity postulate then enables one to express the partial free energy of the substance (or other related property) as a sum of independent contributions, each representing the arbitrarily defined parts, and an interaction term between the two parts (equation 1).

$$\bar{F}_{RX} = F_R + F_X + I_{R,X} \quad (1)$$

The additivity postulate is acceptable when the change in the property of the basic structure R upon substitution with X is relatively small (i.e., small perturbations).

The subject of interest in relating reactivity (or biological activity) to structure is the change in the property of the compound upon interaction with a system, chemical or biological. The change in the property is the difference between its final and initial value. Applying the additivity approximation results in equation 2.

$$\begin{aligned} \Delta \bar{F}_{RX} &= \bar{F}_{RX}^f - \bar{F}_{RX}^i \\ &= (F_R^f - F_R^i) + (F_X^f - F_X^i) + (I_{R,X}^f - I_{R,X}^i) \end{aligned} \quad (2)$$

Since the substituent is located at some distance from the reactive center we can assume that the change in its contribution, $F_X^f - F_X^i$, will be negligible. This leaves equation 3.

$$\Delta \bar{F}_{RX} = (F_R^f - F_R^i) + (I_{R,X}^f - I_{R,X}^i) \quad (3)$$

In the unsubstituted compound the change is

$$\Delta \bar{F}_{RH} = (F_R^f - F_R^i) + (I_{R,H}^f - I_{R,H}^i) \quad (4)$$

The effect of the substituent on this reaction or interaction is therefore

$$\begin{aligned}\delta_R \Delta \bar{F}_{RX} &= \Delta \bar{F}_{RX} - \Delta \bar{F}_{RH} \\ &= (I_{R,X}^f - I_{R,X}^i) - (I_{R,H}^f - I_{R,H}^i)\end{aligned}\quad (5)$$

which consists only of the interaction terms.

At this stage it is impossible to express the effect of the substituent independently of the reaction. To overcome this deficiency one applies a separability postulate (equation 6) which assumes that the interaction terms are factorable.

$$I_{R,X} = I_R \cdot I_X \quad (6)$$

It is this interaction between R and X (i.e., $I_{R,X}$) that mediates the effects of X on the reactivity of RX. Using the separability postulate and taking into consideration that the change in I_X during the reaction is negligibly small, we rewrite equation 5

$$\delta_R \Delta \bar{F}_{RX} = (I_R^f - I_R^i)(I_X - I_H) \quad (7)$$

Equation 7 clearly shows that the separation of the effect of the substituent on the reaction has been achieved in terms of a substituent constant, $I_X - I_H$, and a parameter related to the reaction, $I_R^f - I_R^i$.

1. Substituent constants for electronic and steric effects.

The well known Hammett parameters (3, 18) are obtained directly from equation 7 by arbitrarily setting I_H to zero, and defining $\rho = I_R^f - I_R^i$. Identifying $\delta_R \Delta \bar{F}_{RX}$ with $\log \frac{K_X}{K_H}$ (where K_X and K_H are the dissociation constants of X-substituted and unsubstituted benzoic acids, respectively) we obtain the Hammett equation (3, 4)

$$\log \frac{K_X}{K_H} = \rho \sigma \quad (8)$$

where σ is the substituent constant and ρ is the reaction constant. The effect of a given substituent in different reactions will be expressed by different ρ values.

The additivity and separability postulates have two interesting corollaries. The additivity postulate permits us to extend equation 1 to cases where many substituents exist on R. As an example, for the substituents X and Y,

$$\bar{F}_{RXY} = F_R + F_X + F_Y + I_{R,X} + I_{R,Y} + I_{X,Y} \quad (9)$$

If the interaction between the substituents $I_{X,Y}$ is small, equation 7 can be written as

$$\delta_R \Delta \bar{F}_{RXY} = (I_R^f - I_R^i) ((I_X - I_H) + (I_Y - I_H)) \quad (10)$$

Again setting I_H to zero we may write the Hammett equation as

$$\log \frac{K_{XY}}{K_H} = \rho(\sigma_X + \sigma_Y) \quad (11)$$

or in general form

$$\log \frac{K_{XY\dots}}{K_H} = \rho \Sigma \sigma \quad (12)$$

The additive nature of the substituent constants is expressed in equation 12, which states that the influence of several substituents is equal to the sum of the individual effects of each substituent alone. The additivity rule is usually reasonably well obeyed without any significant deviations (2, 5, 6, 7, 8).

The substituents may have different effects in different molecules or when in different positions on the same molecule and are hence called constitutive. The constitutive nature of σ is reflected for example in the different values for σ_{meta} (i.e., σ_m) and σ_{para} (i.e., σ_p) for the same substituent. The separability postulate implicitly assumes that there is a single interaction mechanism ($I_{R,X}$) between R and X, described by equation 6. This may not be true. For example, the Hammett equation is limited to only para and meta substituents, because an ortho substituent would add a steric interactions to the $I_{R,X}$ term. Many examples in which the simple Hammett equation does not hold because the assumption of a single interaction is not valid are known, and will be discussed below.

The second corollary is useful when equation 6 cannot describe $I_{R,X}$. Invoking additivity of interaction terms equation 6 becomes

$$I_{R,X} = I_R \cdot I_X + J_R \cdot J_X + \text{higher order interactions} \quad (13)$$

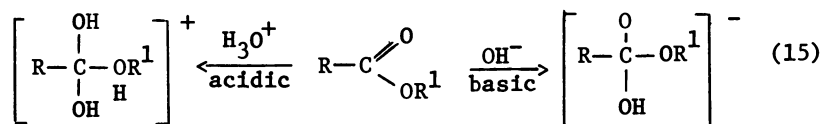
where J_R and J_X describe additional interaction variables between R and X. Following the derivation in equations 2-7 and neglecting higher order interactions, we can write

$$\begin{aligned} \delta_R \Delta \bar{F}_{RX} &= (I_R^f - I_R^i)(I_X - I_H) + (J_R^f - J_R^i)(J_X - J_H) \\ &= \rho_1 \sigma_1 + \rho_2 \sigma_2 \end{aligned} \quad (14)$$

This treatment can be generalized to many interactions. In equation 14 the different σ 's represent substituent constants that are believed to be related to independent properties of the system and the different ρ 's indicate that different reactivities (e.g.,

polar and steric) are associated with these properties (see below).

In principle, extrathermodynamic relationships that deviate from the simple Hammett equation (equation 8) can be treated by equation 14. The major problem is the determination of the different sets of σ 's, (e.g., σ_1 set and σ_2 set) in a way that will indeed reflect their relation to independent properties. An example of such a procedure is the separation of polar and steric effects (10). The need for such a separation arose when a nearly complete lack of correlation was observed between substituent effects represented by the Hammett σ constants and the rates for alkaline hydrolysis of aliphatic systems (12). Inspection of the structures indicated that the proximity of the substituents to the reaction site was a common feature. The steric interaction between R and X had to be considered separately from the electronic effects. Polar substituent constants were thus defined as the difference between the rate constants of base and acid catalyzed hydrolysis of esters. From the structural similarity of the transition states for these reactions (equation 15) it was assumed that the difference in their charge reflects only the polar effect of the substituent



The polar substituent constant, σ^* , is therefore defined (11, 12) as

$$\begin{aligned} \sigma^* &= \frac{1}{2.48} \left(\log \left(\frac{k}{k_0} \right)_B - \log \left(\frac{k}{k_0} \right)_A \right) \\ &= \frac{1}{2.48} \left(\log \frac{k_B}{k_A} - \log \frac{k_{0B}}{k_{0A}} \right) \end{aligned} \quad (16)$$

where B and A refer to base and acid catalyzed hydrolysis, k represents rate constants of hydrolysis of substituted acetates, and k_0 the rate constants of the unsubstituted acetate ester which serves as a reference. The factor 2.48 was arbitrarily introduced so that σ^* and σ have the same range of numerical values. The reactivity related to the polar substituent constants can be expressed as in equation 17.

$$\log \left(\frac{k}{k_0} \right) = \rho^* \sigma^* \quad (17)$$

The observation that polar effects in acid hydrolysis are much less important than the steric effects prompted introduction of the steric substituent constants, E_s (13, 14).

$$E_s = \log \left(\frac{k}{k_0} \right)_A \quad (18)$$

where A and k_0 have the same meaning as in equation 16. With the definition of σ^* and E_s , rate constants of reactions in which polar and steric effects were operative could be correlated by equation 19.

$$\log\left(\frac{k}{k_0}\right) = \rho^*\sigma^* + sE_s \quad (19)$$

This example illustrates the possibility of separating a multitude of interaction mechanisms into apparently independent terms each of which is meant to reflect distinct properties of the system. The importance of such a separation is not only in its ability to correlate observed properties with substituent constants but in any mechanistic implications it may possess. Further identification of the relation between measured properties (e.g., rate or equilibrium constant) and distinct molecular properties (e.g., electronic densities, molecular structure, etc.) provides a basis for formulating the mechanism of the reaction.

The validity of this separation is determined by the statistical validity of equation 19 and by the validity of the assumption that substituent constants (e.g., σ^* and E_s) represent single and independent interaction mechanisms between the substituent, X , and the molecule, R , (e.g., polar and steric effects). The validity of equation 19 in empirically correlating the observed rate constants with the polar and steric parameters is determined by the independence of the parameters. It can be shown that σ^* does not correlate with E_s to any significant degree. Regression of σ^* on E_s (10) gives the following equation

$$\sigma^* = (-0.2 \pm 0.2)E_s + (0.3 \pm 0.2) \quad (20)$$

$$n=25, r=0.04, s=0.801$$

It is clear that σ^* and E_s are indeed almost orthogonal.

The assumption that a substituent constant represents a single interaction mechanism is supported by the large number of structure-activity relationships described by equation 8 (4, 17). However, the same linear relationship can also hold for cases in which several interaction mechanisms are important if their relative contributions remain constant (1). Some of the risks in identifying a substituent constant with a single interaction mechanism become clear in examining the relation between σ_p and σ_m .

If only a single interaction mechanism is represented by σ it follows (1) that the difference, $\sigma_m - \sigma_p$, should be a constant. Table I (compiled from data in (1)) shows that it is not.

Table I. Dependence of $\sigma_m - \sigma_p$ on the substituent.

Substituent	σ_m	σ_p	$\sigma_m - \sigma_p$
NH ₂	-.16	-.66	.50
C(CH ₃) ₃	-.10	-.197	.097
CH ₃	-.069	-.17	.101
C ₆ H ₅	.06	-.01	.07
OCH ₃	.115	-.268	.383
OH	.121	-.37	.491
OC ₆ H ₅	.252	-.32	.572
F	.337	.062	.275
Cl	.373	.227	.146
COCH ₃	.376	.502	-.126
Br	.391	.232	.159
CN	.56	.66	-.10
NO ₂	.71	.778	-.068

Hence, the assumption of a single interaction mechanism is not valid. Hammett (18) recognized this discrepancy and attributed it to inductive and resonance effects (4). It was argued that σ_m represents the inductive effect and the difference ($\sigma_p - \sigma_m$) the resonance (i.e., mesomeric effect) (19, 20, 21). Further improvements in the separation into inductive and mesomeric effects are discussed below.

2. Parameters representing the effect of the medium

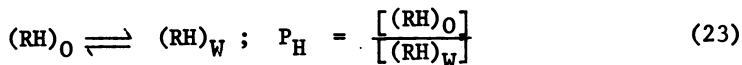
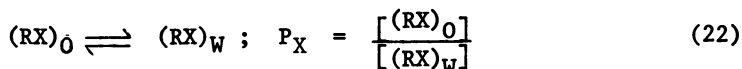
Early in this century, Meyer (109) and Overton (110) showed that the relative potencies of drugs that affect the nervous system correlated with their oil/water partition coefficients. Fifty years later it was shown that partition coefficients in different solvent systems were correlated (111), thus establishing the basis for an extrathermodynamical treatment of partition coefficients. As a result a new substituent constant, π , emerged (23, 24). For a substituent X, π_X is defined as

$$\pi_X = \log P_X - \log P_H \quad (21)$$

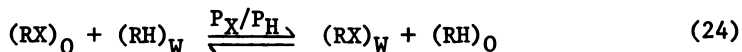
where P_X and P_H are the partition coefficients of the substituted and unsubstituted molecules, respectively, in the octanol/water system.

Since the properties (i.e., partition coefficients) from which π_X is obtained are related to the different solvation of the molecules in different solvents (i.e., octanol and water), we can analyze the substituent constant by an extrathermodynamic treatment of substituent and solvent effects.

We define the terms on the right side of equation 21 from the respective equations 22 and 23.



where O and W represent the octanol and aqueous phases respectively and P_X and P_H the equilibrium constants or the partition coefficients for the substituted and unsubstituted molecules. Combining the two equations, we can write



for which a new equilibrium constant exists, P_X/P_H . Consequently, the change in free energy for equation 24 is

$$\Delta F = -2.3RT(\log P_X - \log P_H) \quad (25)$$

The change in free energy for equation 24 is the sum of partial free energies of the constituting species.

$$\Delta F = (\bar{F}_{RX})_W + (\bar{F}_{RH})_O - (\bar{F}_{RX})_O - (\bar{F}_{RH})_W \quad (26)$$

Within the extrathermodynamic approximation, the partial free energy of a molecule in a solvent (e.g., in water, W) can be written as

$$(\bar{F}_{RX})_W = (F_R)_W + (F_X)_W + (I_{R,X})_W \quad (27)$$

Here we use the usual assumption of additivity of terms, representing the parts of the molecule, R and X, and the interaction between them, I_{RX} .

We now introduce the first assumption about the effect of the solvent: it can be represented as an interaction term for its effect on each of the terms in equation 27.

$$(F_R)_W = F_R + I_{R,W} \quad (28)$$

$$(F_X)_W = F_X + I_{X,W} \quad (29)$$

$$(I_{R,X})_W = I_{R,X} + J_{R,X,W} \quad (30)$$

The terms $I_{R,W}$ and $I_{X,W}$ represent respectively the solvation of the R part and the X part of the molecule by the solvent, i.e., water. The term $J_{R,X,W}$ represents the effect of solvation on the interaction between R and X. Since this is a second order term, we can assume that it is small compared to first order terms and neglect it for the time being. We define this as the first

approximation (1).

Applying equations 27 and 30 with the appropriate subscripts for the substituent and the solvent and taking into account the first approximation, we can write equation 26 as

$$\Delta F = \{(F_X)_W - (F_X)_O\} - \{(F_H)_W - (F_H)_O\} + \{(I_{R,X})_W - (I_{R,X})_O\} - \{(I_{R,H})_W - (I_{R,H})_O\} = (I_{X,W} - I_{X,O}) - (I_{H,W} - I_{H,O}) \quad (31)$$

Using the assumption of factorability of interaction terms (e.g., $I_{X,W} = I_X \cdot I_W$, see above) equation 31 becomes

$$\Delta F = I_X(I_W - I_O) - I_H(I_W - I_O) \quad (32)$$

Arbitrarily setting the term $I_H(I_W - I_O)$ to zero, we see that π_X will be proportional to $I_X(I_W - I_O)$. The conclusion from this derivation is that π_X is independent of the molecular system from which it is determined, and it depends on the solvents in which the partition was determined. If we take the octanol/water system as a standard we can arbitrarily set $(I_W - I_O) = -2.3RT$ and obtain equation 21.

The independence of π on the molecular system endows it with the characteristic of a substituent constant. This property is a direct consequence of the first approximation, namely that π reflects only the different solvation of substituent X compared to a reference substituent H. The validity of this approximation can be tested by obtaining π_X from different molecular systems. It is well known (112) that π_X values derived from aromatic systems π_{Ar} differ from those derived from aliphatic systems π_{Al} . The following table presents a few selected examples (112).

Table II. π constants derived from aromatic (π_{Ar}) and from aliphatic (π_{Al}) systems.

Substituent	π_{Ar} $\log P_{C_6H_5X} - \log P_{C_6H_6}$	π_{Al} $\log P_{RX} - \log P_{RH}$	$\pi_{Ar} - \pi_{Al}$
I	1.12	1.00	0.12
COCH ₃	-0.55	-0.71	0.16
Br	0.86	0.60	0.26
CN	-0.57	-0.84	0.27
Cl	0.71	0.39	0.32
OCH ₃	-0.02	-0.47	0.45
N(CH ₃) ₂	0.18	-0.30	0.48
OH	-0.67	-1.16	0.49
NO ₂	-0.28	-0.85	0.57

It is clear from the differences (up to 0.57 for NO₂) between π_{Ar} and π_{Al} that the first approximation is inadequate.

A somewhat smaller variability is obtained for $\pi_{\text{CH}=\text{CH}}$ and $\pi_{\text{C}\equiv\text{C}}$ (112):

$$\begin{aligned} \log P_{\text{CH}_3\text{CO}(\text{CH}_2)_2\text{CH}=\text{CH}_2} - \log P_{\text{CH}_3\text{COCH}_2\text{CH}_3} &= 0.73 \\ \log P_{\text{C}_6\text{H}_5\text{OCH}_2\text{CH}=\text{CH}_2} - \log P_{\text{C}_6\text{H}_5\text{OCH}_3} &= 0.83 \end{aligned} \quad (33)$$

$$\begin{aligned} \log P_{\text{CH}_3\text{CH}_2\text{CH}_2\text{C}\equiv\text{CH}} - \log P_{\text{CH}_3\text{CH}_2\text{CH}_3} &= 0.48 \\ \log P_{\text{C}_6\text{H}_5\text{C}\equiv\text{CH}} - \log P_{\text{C}_6\text{H}_6} &= 0.40 \end{aligned} \quad (34)$$

The relation between these discrepancies and the nature of the substituents (Table II) can be attributed to electronic effects (112). Similar discrepancies have been observed in systems where steric effects are present, e.g., π_{CH_3} from 2-methylphenoxyacetic acid is 0.84 whereas from 3-methylphenoxyacetic acid it is 0.52. Another example is the π_{OCH_3} obtained from equation 35

$$\begin{aligned} \pi_{\text{OCH}_3} &= \log P_{1,2,3-(\text{CH}_3\text{O})_3\text{C}_6\text{H}_3} - \log P_{1,2-(\text{CH}_3\text{O})_2\text{C}_6\text{H}_4} \\ &= -0.56 \end{aligned} \quad (35)$$

This value should be compared to -0.02 obtained from the aromatic system and to -0.47 from the aliphatic system (see Table II). The steric interaction in 1,2,3-trimethoxybenzene probably causes a rotation of the methoxy group out of plane, bringing it close to an aliphatic-nonconjugated group.

Clearly the first approximation does not generate a universal π scale for substituents that is independent of the system from which they have been derived. Just as multiple σ scales have been derived to represent different electronic effects in different molecular systems (section A.1 above), so several π scales are needed to represent different solvations in different molecular systems. Alternatively, a separation of the effects (electronic and steric) that could contribute to the observed π values can be attempted, assuming that they are independent and additive.

An example of the separation of electronic effects that contribute to observed π values in substituted quinazolinones has been described (113). It was found that the observed hydrophobic constant, π_n , obtained from substituted quinazolinones can be correlated with the π constant obtained from the phenoxyacetic acid system (23, 24) and the Hammett σ :

$$\pi_n = \pi + k_n \sigma \quad (36)$$

where k_n is a constant that depends on the system from which π_n was obtained. This shows again that the first approximation, which assumed that the effect of solvation on the interaction term $I_{R,X}$ is negligibly small, is inadequate.

We define the second approximation (1) as the one in which the

term $J_{R,X,S}$ in equation 30 is not neglected (S stand for any solvent, e.g., water or octanol). Taking account of this term, we rederive equation 31 to obtain equation 37.

$$\Delta F = (I_{X,W} - I_{X,O}) - (I_{H,W} - I_{H,O}) + (J_{R,X,W} - J_{R,X,O}) - (J_{R,H,W} - J_{R,H,O}) \quad (37)$$

Applying the factorability principle to $J_{R,X,S}$

$$J_{R,X,S} = J_R \cdot J_X \cdot J_S \quad (38)$$

and substituting the proper subscripts for the solvent and the substituent we obtain

$$\Delta F = (I_X - I_H)(I_W - I_O) + J_R (J_X - J_H)(J_W - J_O) \quad (39)$$

Equation 39 clearly indicates the dependence of π_X on the system R through the interaction variable J_R and points to the origin of the variability in π is the result of the influence of the solvent on the interaction term between R and X . The second approximation which includes also the second order term $J_{R,X,S}$ casts a serious doubt on the applicability of a universal π scale which will adequately represent the effect of a substituent on the solvation process. However, individual π scales derived from specific molecular systems (e.g., aromatic, aliphatic, etc.) or from procedures to separate π from other effects (e.g., electronic, steric, etc.) could prove useful.

Further analysis of equation 39 reveals a reason for the validity of Collander's equation that correlates partition coefficients in different solvent systems (111). Thus, in equation 39 the terms $(I_W - I_O)$ and $(J_W - J_O)$ both reflect a variable characteristic of the solvents. For small changes in this variable they should be proportional to each other

$$J_W - J_O = f(I_W - I_O) \quad (40)$$

Substituting equation 40 in equation 39 we obtain

$$\Delta F = (I_W - I_O) \{ (I_X - I_H) + f \cdot J_R (J_X - J_H) \} \quad (41)$$

Thus, for the π constants for the same molecular system RX depends linearly on the solvent systems in which they were determined. Since π is an additive-constitutive property of the molecule we can obtain $\log P$ from the π values of the constituting parts (23, 24). The linear relation between $\log P$ values for the same compound in different solvents should also hold. This is reflected in the linear relation between partition coefficients in different solvent systems as expressed by Collander (111)

$$\log P_2 = a \log P_1 + b \quad (42)$$

Thus, Collander's equation 42, which was established from empirical studies (111), is theoretically derived from the extrathermodynamic treatment of solvent effects and is obtained as a result of the second approximation which takes into account the effects of solvation on the interaction term between R and X, $J_{R,X,S}$. If we now regard the interaction of drugs (in water) with the biophase as a partitioning phenomenon, equation 42 becomes the basis for the use of the partition coefficient in octanol/water, P_1 , as a model for the partition coefficient between biophase and water, P_2 .

B. The empirical framework for definition of QSAR parameters.

Numerical values for the structural parameters of QSAR are obtained from an evaluation of the effect of the substituent on the properties (e.g., the rate or equilibrium constants) of a model reaction. The classification of these parameters is therefore model dependent. The model reactions are chosen to represent the most pervasive types of physicochemical phenomena (e.g., dissociation reactions, hydrolysis, substitution reactions, partition between solvents).

1. The classification of σ constants

The derivation of σ values from the model reaction used by Hammett (18) assumes that the substituted benzoic acids dissolve in water and do not undergo side reactions. Furthermore, when a strong interaction occurs between the substituent and the benzene ring, the basic assumption of the extrathermodynamic relationships is violated. This may result in large σ values which do not obey the Hammett equation. One example is the presence of a strongly electron withdrawing or electron donating group in the para position. In order to overcome this problem a different model reaction is chosen. Taft (43, 44) suggested the use of benzoic acids in which the substituent was isolated from the ring by a CH_2 group to determine a new scale denoted σ° , which to a large extent overlaps with the σ scale. The major drawback of this method is that the ρ for this reaction is small, thus increasing the error in σ° . To overcome this problem the hydrolysis of substituted phenylacetic esters was proposed as a model reaction because in this series the reactivity is more sensitive to substitution (45, 46, 47). The advantage in this new scale of σ° is its uniformity and its unbiased treatment of substituents. Its shortcoming is that the scale is assumed to be free from a model error; the only error is in the experimental measurement of the dissociation constants. To minimize the error it was suggested (4) that σ should be defined as a parameter that best fits the whole body of data from different model reactions. This can be obtained by iterative methods. For lack of appropriate statistical and computational methods, this approach was not used until recently (48, 50). In

this work (48, 50) the model was defined in analogy to the Hammett equation as

$$\log K_{ik} = \phi_k + \rho_k \sigma_i + e_{ik} \quad (43)$$

The subscripts i refer to substituents and k to reactions (the models). The parameter ϕ_k coincides with $\log K_{0k}$, i.e., the rate or equilibrium constant for the unsubstituted compound, and e_{ik} contains both the experimental and model errors. In the statistical procedure (49) the parameters ϕ_k , ρ_k , and σ_i are estimated with the requirement that

$$\sum_i e_{ik} = \min \quad k=1,2,\dots,N \text{ (reactions)} \quad (44)$$

and

$$\sum_k e_{ik} = \min \quad i=1,2,\dots,M \text{ (substituents)} \quad (45)$$

This procedure usually gives a set of σ° for which the significance of the correlations is better. Obviously the σ° values will depend on the data bank and may change as the data bank expands. An important advantage is the rigorous definition of the reaction types that are represented by this scale of parameters.

The dissociation of phenols and anilines does not obey the simple Hammett equation. A new scale of σ_p^- was proposed (4, 19) for this type of reaction, based on the argument that a strong interaction exists between an electron withdrawing group and the reactive site. This interaction was considered to result from a direct conjugation of the substituent with the reaction center. No theoretical basis for this effect was offered. Also, σ_p^- is applicable only for electron withdrawing groups; σ_p had to be used for electron donating groups (19). This indicates that σ_p^- cannot represent a single interaction mechanism.

A reverse phenomenon occurs in electrophilic aromatic substitutions: they are more sensitive to electron donating substituents than to electron withdrawing substituents. A σ_p^+ scale was therefore defined from studies on the solvolysis of dimethylphenylcarbinyl chlorides in 90 per cent aqueous acetone (52, 53). Just as there is a discrepancy between σ_p^- and σ_p , so there is a discrepancy between σ_p^+ and σ_p for strong electron donating groups but not for electron withdrawing groups.

As σ is an electronic parameter, one can suspect that it can be separated into inductive, polar, and resonance (i.e., mesomeric) effects. Separation assumes that the effects are indeed independent and that appropriate model systems can be identified in which these effects can be isolated (232).

Several attempts have been made to achieve the separation. It was assumed that σ_p represents the inductive and resonance effects in an additive fashion (43, 61, 62) with equal weights

$$\sigma_p = \sigma_I + \sigma_R \quad (46)$$

where σ_I is the inductive and σ_R the resonance substituent constant. As resonance effects are different for meta substituents but the inductive effects remain practically the same, the relation between the two contributions for σ_m can be written as

$$\sigma_m = \sigma_I + \alpha \sigma_R \quad (47)$$

The same procedure was refined by using σ° instead of $\sigma_{m,p}$, with the resulting σ_R denoted as σ_R° (44). The resulting α has the value of 0.5 which reduces to approximately 0.1 when σ_p^- was used instead of σ_p (63, 64). This indicated that the separation reflected the larger contribution of the resonance effect in σ_p^- as compared to σ_p .

The construction of a good scale of inductive constants, σ_I , was successful; the scale for the resonance constants presents many problems. The inductive scale was constructed from several molecular reference systems: such as 4-substituted bicyclo [2.2.2]octane-1-carboxylic acids (65, 66, 67), α -substituted meta- and para-toluic acids (68, 69), and from comparison of base- and acid-catalysed hydrolysis of substituted acetates (43, 11) (i.e., the polar substituent constants σ^* which is related to σ_I). The σ_R scale can hence be obtained from equations 46 and 47. The precision of such determinations is inadequate because the factor α in equation 47 is too small. Values of σ_R were also obtained from NMR (70) and IR (71) measurements. The major problem with the σ_R scale is that the σ_R values are usually small with large standard deviations. Equation 48 is used to correlate rate or equilibrium constants with the double scale of σ_I and σ_R

$$\log\left(\frac{k}{k_0}\right) = \rho_I \sigma_I + \rho_R \sigma_R \quad (48)$$

As a multiparametric equation, it requires more precise measurements and larger samples to attain statistical validity. The variations in σ_R obtained from different molecular systems and their small values indicate the inadequacy of such separations to construct a reliable σ_R .

A more general way to express the separation into inductive and resonance contributions was proposed by Swain and Lupton (72). They claimed that any kind of electronic substituent constant can be represented as

$$\sigma = f \cdot F + r \cdot R + i \quad (49)$$

where F represents the field constants and R the resonance constants. The constant i was added in order not to overemphasize the relative weight of the unsubstituted compound in the least-squares fitting procedure. In fact, the values of the residual constant i are very close to zero.

An interesting approach was proposed by Dewar and Gridale (73) in which the parameters were obtained by semiempirical quantum mechanical methods. The separation was into electrostatic, F/r_{ij} and resonance combined with inductive effects (Mq_{ij} and $M'\pi_{ij}$)

$$\sigma_{ij} = F/r_{ij} + Mq_{ij} \quad (50)$$

$$\sigma_{ij} = F'/r_{ij} + M'\pi_{ij} \quad (51)$$

where r_{ij} is the distance between atoms i and j , q_{ij} is the charge overlap, and π_{ij} the π -system charge overlap. This model contains an oversimplified representation of the electrostatic term and may lack the desired physical meaning. This approach is different from other definitions of σ parameters in that it can be improved by a better definition of the terms in equations 50 and 51. Such definitions should follow the basic physical principles of the model and can be rigorously represented by quantum chemical formalisms.

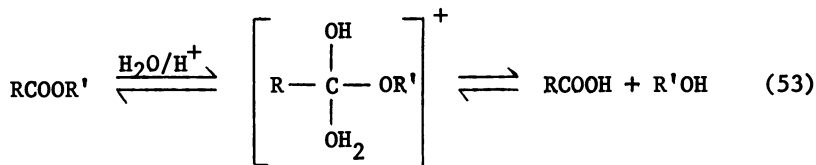
2. Parameters representing steric effects

In the early 1950s Taft (10) outlined a sound quantitative basis for the estimation of steric effects and for separating them from polar and resonance effects. This derivation follows the standard extrathermodynamic approach and is therefore empirical. The definition of steric substituent constants is closely related to polar substituent constants, for they are obtained from the same reference system (10). The polar substituent constants, however, have been shown to arise from electronic effects, and they strongly correlate with the inductive substituent constants (10, 43, 61).

$$\sigma_I = 0.45 \sigma^* \quad (52)$$

where σ_I is the inductive and σ^* the polar substituent constant.

The steric substituent constants, E_s , is defined as the ratio of rate constants of acid hydrolysis of substituted and unsubstituted acetates.



$$\log\left(\frac{k}{k_0}\right) = E_s \quad (54)$$

where k and k_0 are the rate constants for the acid catalyzed hydrolysis of the substituted and unsubstituted esters, respectively. The reference molecule for which the steric substituent constant is zero is therefore the ester of acetic acid; consequently $E_s^{\text{CH}_3} = 0$. In this procedure, it is assumed that (a) the relative free energy of activation (i.e., the difference between the free energy of activation of the substituted ester and the unsubstituted one) reflects only the steric effects of the substituent on the rate of acid hydrolysis of esters; and (b) the steric substituent constant is independent of R' and of solvent.

The validity of these assumptions was tested by applying the derived steric substituent constants to reactions in which the transition states resembles that of the acid hydrolysis of esters. It was found that very few reactions obey equation 54. These are exemplified by the acid catalyzed methanolysis of 2-naphthyl esters in methanol (86) and the reaction of 2-alkylpyridines with methyl iodide (87) and with trimethylboron (88) in nitrobenzene. The paucity of reactions that depend solely on steric effects is not surprising. To include electronic contributions, Pavelich and Taft (89) proposed equation 55

$$\log\left(\frac{k}{k_0}\right) = \rho^*\sigma^* + sE_s \quad (55)$$

Many reaction obey equation 55 (90) implying that the steric effect can be treated as an extrathermodynamic parameter.

Realizing the shortcomings of the steric substituent constants, several workers designed other scales, each of which represents a specific component of the overall steric effect. Charton (98) defined a steric parameter U as the difference in van der Waals radii of a substituent X and hydrogen

$$U_x = r_{vx} - r_{vh} = r_{vx} - 1.2 \quad (56)$$

The advantage of this definition of substituent constants is that they are free of the electronic effects that are present in E_s (see below, section C.2). Their use is limited to symmetrical substituents, such as CH_3 , CCl_3 (i.e., XY_3). In asymmetric groups the definition of the radius r_v is not possible. The similarity of these parameters to the E_s scale would suggest similar problems, but since they are not based on specific experimental measurements, improvement of the scale is more difficult.

The steric parameters E_s^e introduced by Unger and Hansch (99) are simply a shift of the original scale of E_s . In the former, $E_s^e = 0$ is for the hydrogen atom rather than for the methyl as in the latter scale.

The connectivity indices χ and χ_v (102, 103), quantitatively characterize the size of the molecule and the degree of branching. In that sense, and also because they are derived from a theoretical procedure (104), they lack the usual extrathermodynamic meaning. The strong correlation of χ with molar refractivity, M_R ,

suggests that these parameters measure the general steric bulk.

The minimal steric difference (MSD) is another example of a specifically designed steric constant (105) in which the parameters are defined as the non-overlapping volume of a certain molecule with a standard molecule possessing the biological activity under study. The reference molecule is usually the natural substrate, which often must be arbitrarily chosen. For conformationally flexible molecules the minimum energy conformations are chosen. This eliminates considerations of other conformers that may be important. Further, this method does not take into consideration that the steric interaction may cause a conformational change in the molecule or cause a change in its reactivity. The subjective way by which the molecules are superimposed on the reference molecule is also a shortcoming of the method. This method, however, offers a potential solution to the problem of different activities of stereoisomers.

A promising approach to the development of steric substituent parameters, STERIMOL, has been presented by Verloop (106). Concluding that a single steric substituent parameter is unable to account for all the steric effects that influence activity, he devised five parameters to describe the length and various widths in well defined directions of the substituent. The length, L , is defined as the longest axis of the substituent from its point of attachment. This parameter is in many respects equivalent to the length of a substituent, defined by Bowden and Young (107). The other four parameters, B_1 - B_4 , are related to the width of the substituent. B_1 represents the minimum width much in the same sense as the usual Taft's E_s (101) or Charton's U (98). The minimal width represents the tendency of conformationally flexible substituents to change their conformation in a way that minimizes the steric interaction. Thus, the smallest protruding group (i.e., the minimal width) will face the region where a maximal steric interaction occurs. The other width parameters are derived so that the four B parameters would measure the width in four rectangular directions. The major problem of this method remains the determination of the conformation of a flexible substituent for which the parameters will be determined. Usually the most reasonable conformation is chosen, i.e., the one of minimum energy, but in some cases, e.g., long hydrocarbon chains, more compact conformations are preferred because differences in energy are not too large and environmental effects may stabilize the compact geometry. From the five steric parameters the degree of deviation from spherical shape may be estimated by inspection of the ratios L/B_1 and B_4/B_1 (i.e., the ratio of length to minimal width and maximal to minimal width). The largest deviations observed for the 243 substituents that have been compiled (106) are $L/B_1 \approx 9$ and $B_4/B_1 \approx 7$. This method results in different sets of parameters for geometrical isomers around a non-flexible bond (e.g., E and Z isomers around a double bond). By defining the relative relationship of two of the parameters as orthogonal or opposite-colinear, a

distinction can be made between different stereoisomers.

3. Parameters representing hydrophobic interactions

The analysis of the parameters representing the effect of the medium in terms of extrathermodynamical relations has shown (see above, section A.2) that the hydrophobic substituent constant π depends on the molecules from which it is derived. For example, hydrophobic constants derived from the octanol/water partition coefficients of aromatic molecules differ from those obtained from aliphatic molecules (112). Collander's equation (111) provides the empirical basis for the evaluation of logP values for the same molecule in different solvents. However, solvents with markedly different solvation properties (e.g., hydrogen bonding ability) do not conform to Collander's equation (see below, section C.3).

It has been shown by Tute and collaborators (228) that the theoretical evaluation of logP on the basis of additivity principles can lead to misleading results for certain molecules (e.g., heterocyclics, sterically crowded or conformationally flexible molecules). This difficulty has been recently rationalized for rigid structures (229). Experimentally measured logP values are therefore more reliable than those estimated from component π values.

Other molecular properties have been also proposed to model the hydrophobic interactions. The parachor, which is related to the surface tension of a compound (139, 140) represents mainly the intermolecular interactions in a liquid. The Hildebrand-Scott solubility parameter, δ , (141) is related to intermolecular van der Waals forces and the closely related molar attraction constant, F , is obtained by multiplying δ by the molar volume (142). The partition coefficient between two solvents can be obtained from the solubility parameters and the molar volumes of the solute and the solvents (193). This relationship is based on regular solution theory (194) and the assumption that the partial molar volumes of the solute is not different from its molar volume. Recently this has been criticized and a new derivation was proposed (195) in which the partial molar volumes are taken into account. The molar refractivity, MR, is related to dispersion forces and can be obtained as a sum of the partial molar refractivities assigned to atoms and bonds (140, 143). These parameters have been compared (144) to establish their relative applicability to correlations with biological activity. The conclusion was that logP and molecular refractivity were the best parameters. Parameters obtained from high pressure liquid chromatography (144, 145, 146) and from thin layer chromatography (147) have also been used.

C. The physicochemical basis of QSAR parameters

The impressive success of the Hammett equation in correlating literally hundreds of observed properties (17) (e.g., rate and equilibrium constants, spectroscopic properties, etc.) may be attributable to the multitude of interaction mechanisms that is implicitly embedded in the values of σ . The validity of the separability and additivity axioms used in the derivation of extra-thermodynamic relationships is confirmed by the ability to separate experimentally multiple interaction mechanisms (e.g., inductive and resonance (19, 20, 21), polar and steric (10), enthalpic and entropic (22)). This separation fostered significant progress in the application of quantitative structure-activity relationships to the study of chemical mechanisms. For these relationships can now be expressed in terms of more basic properties of the molecules under study.

It is assumed that the molecular properties responsible for biological activity can be separated into hydrophobic, electronic and steric effects, all of which are independent. It is further assumed that the hydrophobic effects are fully described by the parameter π (23, 24, 25), the electronic effects by σ (1, 4, 26), and the steric effects by E_s (10, 27). The relationship between biological activity and the parameters is expressed by equation 57 (28).

$$BA = a\pi + \rho\sigma + sE_s + \text{const.} \quad (57)$$

Higher order terms in π must often be included in the equation (29, 20) which then becomes

$$BA = a\pi + b\pi^2 + \rho\sigma + sE_s + \text{const.} \quad (58)$$

$\log P$ (i.e., log of the octanol/water partition coefficient) is frequently interchanged with π . Equation 58 is the most widely used one in QSAR.

The model for equation 58 can be considered only as a zero order approximation. Proof, or at least substantial indication, is needed for the assumption that hydrophobic interactions are well represented by octanol/water partition coefficients. Moreover, the nature of the hydrophobic effect is complicated (see below), and π (or $\log P$) may be concealing a host of interactions (e.g., solute-solvent, substituent-molecule). Electronic effects, although relatively well understood in chemical systems (mainly due to the successful separations of effects), have a vague meaning in biological systems because the mechanism of interaction of the drug with the biological system is not known. Several sites on the drug may be important for the interaction, and more than one electronic parameter may be needed to account for the interaction. Further, overall chemical reactivity is not necessarily a simple sum of the reactivities at these several sites. Most

problematic are the steric parameters. They lack the property of additivity, and being related to the bulkiness of a substituent, they lack directional information, e.g., they cannot account for stereospecific interactions. These uncertainties inherent in the interpretation of the physical meaning of the parameters limit the scope of mechanistic conclusions from the correlations. Most importantly, a lack of correlation with a certain parameter cannot be interpreted to mean that the type of interaction linked to that parameter is not relevant to biological activity.

A fundamental requirement for any correlation is that a qualitatively constant mechanism exists. For example, in kinetic studies, the rate determining step must be the same for all molecules, and more importantly, the transition state for the reaction for all molecules must be of the same type. The same demand of biological data means that, for example, the drugs bind to the same receptor through the same types of interactions. An apparent lack of correlation may be due to the fact that more than one mechanism can bring about the same effect. The analysis of the deviations could identify differences in mechanism within a series of compounds or differences in the effects of substituents on the reactivity of specific compounds in this series.

Pervasive is also the problem of parameter intercorrelation. To what extent, for example, do hydrophobic parameters depend on electronic changes in the molecule? Or how much does the steric parameter, being related to the size of a substituent, contribute to the lipophilicity of the molecule? Without understanding of the physicochemical basis of the parameters used in QSAR it is impossible to assess their significance and the validity of the models from which they are derived. In subsequent sections we discuss the physicochemical basis of these parameters.

1. The electronic substituent constant

The numerous reactions and physical properties that can be correlated with electronic substituent constants suggest that they reflect a basic mechanism for the effect of the substituent on the reaction or the property. Many reactions were correlated by the Hammett equation, such as electrophilic (4, 31, 32, 33), nucleophilic (4, 34) and radical (35, 36, 37). Spectroscopic properties, such as infrared and ultraviolet absorption frequencies and intensities (38), nuclear magnetic resonance shifts of various nuclei, e.g., ^1H , ^{19}F , ^{13}C (39), and mass spectroscopic ionization potentials, fragmentations and distributions (40) were also correlated. Enzymological (41, 42) and pharmacological observations (9) were also correlated with electronic substituent constants. A discussion of the acceptable range of the Hammett equation, the treatment of deviations, and the meaning of the separation of interaction mechanism is of prime importance in assessing the contribution of electronic effects to biological activity.

A contribution to the verification of the separability of inductive and resonance effects comes from the thermodynamic approach (74). Since

$$\Delta G^\circ = -2.3RT \log K \quad (59)$$

we can write the Hammett equation as

$$\Delta G^\circ = -2.3RT \rho \sigma \quad (60)$$

Assuming that the free energy ΔG° is separable into inductive and resonance free energies, we can write

$$\Delta G^\circ = \Delta G_I^\circ + \Delta G_R^\circ = a_I t_I + a_R t_R \quad (61)$$

where a is the substituent constants and t the transmission coefficient of the specific effect. Combining equations 60 and 61 we can write

$$a_I t_I + a_R t_R = -2.3RT \rho \sigma \quad (62)$$

Since the transmission coefficients t_I and t_R are assumed to be independent of the substituent in a reaction where only the substituents are changing (the reaction conditions remain the same), the ratio between the transmission coefficients is a constant

$$\frac{t_R}{t_I} = b \quad (63)$$

Substituting equation 63 into equation 62 and rearranging terms we obtain equation 64

$$\left(\frac{t_I}{T}\right) \left(-\frac{a_I + a_R \cdot b}{2.3R}\right) = \rho \sigma \quad (64)$$

It immediately becomes evident that although we assumed the existence of two independent effects (i.e., inductive and resonance), under certain conditions this relationship can be expressed by a single ρ identified as (t_I/T) and a single σ identified as

$\left(-\frac{a_I + a_R \cdot b}{2.3R}\right)$. Before proceeding with the analysis of the conditions under which equation 64 is valid, we note the remarkable similarity of this equation to the equations proposed by Yukawa and Tsuno (75) and by Yoshioka, Hamamoto and Kubota (76), equations 65 and 66. These equations were offered to express the correlation in electrophilic and nucleophilic processes, respectively, in which direct conjugation was of importance.

$$\log \frac{k}{k_0} = \rho (\sigma_{m,p} + r(\sigma_p^+ - \sigma_p^-)) \quad (65)$$

$$\log \frac{k}{k_0} = \rho (\sigma_{m,p} + r(\sigma_p^- - \sigma_p^+)) \quad (66)$$

If we identify r in equations 65 and 66 as $\frac{b}{2.3R}$ in equation 64, the argument (75) that ρ and r in equations 65 and 66 are independent parameters seems to be invalid. In fact a claim has been made that ρ and r (77, 78) are linearly dependent which is confirmed by equation 63. It should be noted that whereas equations 65 and 66 were proposed on a purely empirical basis in order to improve the correlation in directly conjugated systems, equation 64 resulted from the assumption of the separability of the inductive and resonance effects expressed in equation 62.

The analysis of equation 64 offers important insight into the nature of ρ and σ in the Hammett equation. Although the correct general dependence of ρ on temperature, as predicted from the isokinetic relationship between the enthalpy and the entropy of a reaction is (1, 79, 80)

$$\rho = \beta_\infty (1 - \frac{\beta}{T}) \quad (67)$$

where β is the isokinetic temperature, the proportionality between ρ and $1/T$ is maintained in the term (t_1/T) . In fact, this corresponds to a case (80) when the change in the entropy is zero. The analysis of the expression of σ in equation 64 shows that σ does indeed depend on the reaction conditions reflected in b . Thus, in principle σ is not a pure substituent constant independent of the reaction. However, in cases where b (equation 64) remains constant at different reaction conditions or is close to zero (i.e., the resonance effect is much smaller than the inductive effect) the Hammett equation is expected to be valid. Under these conditions ($b \rightarrow 0$) σ represents almost entirely the inductive effect of the substituent. This conclusion is in clear agreement with the small values of σ_R obtained from various molecular reference systems, because in attempting to separate σ_I and σ_R the molecular systems chosen usually obeyed the Hammett equation quite well, i.e., b was small as was shown above.

An illuminating thermodynamic analysis of the separation of the free energy relationships into enthalpic and entropic contributions was presented by Krygowski and Fawcett (22). They assumed that the linear dependence of the free energy on substitutions can be represented by contributions from enthalpic and entropic linear relationships as follows.

$$\delta\Delta H = \alpha_H(\Delta H^\circ) + \beta_H \quad (68)$$

$$\delta\Delta S = \alpha_S(\Delta S^\circ) + \beta_S \quad (69)$$

The free energy is then expressed as

$$\delta\Delta F = \alpha_H(\Delta H^\circ) - \alpha_S(T\Delta S^\circ) + \beta_H - \beta_S T \quad (70)$$

and the Hammett equation can be written as

$$\log \frac{K}{K_0} = \rho_H \sigma_H + \rho_S \sigma_S \quad (71)$$

where σ_H and σ_S are the enthalpic and entropic substituent constants, respectively, and ρ_H and ρ_S the respective reaction constants. This separation was successfully applied to various thermodynamic data (22). The usual meaning attributed to the Hammett σ as representing electron withdrawal (positive σ) or electron donation (negative σ) is not reflected in the individual σ_H and σ_S . In general σ_H is negative and small whereas σ_S is positive. An interesting correlation was found between σ_H and σ_S (equation 72) which indicates the predominant role of entropy in the substituent effect.

$$\sigma_S = -3.6\sigma_H + 0.15 \quad (r = 0.953) \quad (72)$$

This correlation is also in agreement with the isoequilibrium (isokinetic) relationship (1).

$$\delta\Delta H^\circ = \beta_1 \delta\Delta S^\circ \quad (73)$$

The relationship in equation 72 indicates that the enthalpic and the entropic contributions to the substituent effect are not independent. Therefore, the assumption that " σ_H and σ_S are independent substituent constants accounting for the enthalpic and entropic contributions" (22) is not valid. However, for the separation of these contributions, this assumption is not needed, since the enthalpic and entropic contributions to the free energy of dissociation can be obtained experimentally in an independent fashion. In fact, the relationship between the enthalpic and the entropic contributions on one hand with the inductive and resonance effects (which are assumed to be independent) on the other hand can be obtained only by using the relationship between enthalpic and entropic contributions.

According to Hepler (81) one can attribute the changes in ΔH° with the substituent to intramolecular (int) and intermolecular (ext) changes.

$$\delta\Delta H^\circ = \delta\Delta H_{int}^\circ + \delta\Delta H_{ext}^\circ \quad (74)$$

The intermolecular changes are meant to reflect changes in solute-solvent interactions due to the change in the electronic charge distribution with changes in substitution. The intramolecular

changes, however, are assumed to reflect the changes in electronic charge distribution and therefore they will be related to the two electronic effects of a substituent, inductive and resonance.

$$\delta\Delta H_{\text{int}}^{\circ} = \delta\Delta H_{\text{I}}^{\circ} + \delta\Delta H_{\text{R}}^{\circ} \quad (75)$$

Combining equations 73, 74 and 75 we can write

$$\delta\Delta S^{\circ} = \alpha_{\text{I}}\delta\Delta H_{\text{I}}^{\circ} + \alpha_{\text{R}}\delta\Delta H_{\text{R}}^{\circ} + \gamma \quad (76)$$

where γ represents all the other contributions that are not reflected in the intramolecular changes, e.g., $\delta\Delta H_{\text{ext}}^{\circ}$. Assuming that $\delta\Delta H_{\text{I}}^{\circ}$ and $\delta\Delta H_{\text{R}}^{\circ}$ are related to σ_{I} and σ_{R} and $\delta\Delta S^{\circ}$ to σ_{S} we can write

$$\sigma_{\text{S}} = \rho_{\text{I}}\sigma_{\text{I}} + \rho_{\text{R}}\sigma_{\text{R}} + \gamma \quad (77)$$

Krygowski (22) used a multiple regression analysis to obtain good correlations with $\sigma_{\text{S}}^{\text{para}}$ and $\sigma_{\text{S}}^{\text{meta}}$

$$\sigma_{\text{S}}^{\text{para}} = (1.2 \pm 0.2)\sigma_{\text{I}} + (0.8 \pm 0.2)\sigma_{\text{R}} + (0.0 \pm 0.1) \quad (78)$$

$$\sigma_{\text{S}}^{\text{meta}} = (1.2 \pm 0.1)\sigma_{\text{I}} + (0.4 \pm 0.1)\sigma_{\text{R}} + (0.0 \pm 0.1) \quad (79)$$

It is apparent from equations 78 and 79 that ρ_{I} does not depend on the position of the substituent, whereas ρ_{R} is significantly smaller for meta substitution than for para substitution. Thus, the inductive effect remains a constant contribution while the resonance contribution changes with position, as expected.

Several conclusions emerge from this analysis. The entropic contribution constitutes the major part in the change in free energy and amounts, in "normal" Hammett relationships, to approximately 80 per cent, the rest coming from the enthalpic effect. In cases where a strong resonance is operative the relative contribution of the enthalpic effect increases up to approximately 40 per cent. A similar result is obtained from the analysis of inductive and resonance contributions to σ_{S} , suggesting a parallelism between entropic and inductive effects on the one hand and enthalpic and resonance effects on the other.

The large contribution of the inductive effect to the substituent constant σ , and its relation to the entropic contribution (22) offer mechanistic insight into Hammett relationships. These facts show that in most cases the substituent mainly affects the molecular changes in solute-solvent interactions between the final and initial stages (thermodynamic data) of a reaction or between the transition and the ground states (kinetic data). These changes are of course due to intramolecular charge redistribution (isoequilibrium and isokinetic relationships) but these intrinsic changes, as reflected in the enthalpic contributions, are much smaller than those reflected in the entropic contribution.

Thus, although the basic source of the electronic effect of a substituent on chemical and biological activity comes from changes in fundamental properties such as electronic charge distributions, it is observed mainly through the solvent-attenuated effects.

These conclusions are of utmost importance in studying chemical structure-biological activity relationships. In most cases the specific molecular interactions between a biologically active compound and a receptor are not identified. Therefore, a judicious choice of a specific type of electronic substituent constants (e.g., σ° , σ_I , σ_R , σ_P^+ , etc.) is usually impossible. It appears that the best choice for such correlations would be the σ° or the σ_I scale. It should be kept in mind, however, that these σ sets represent the solvent-attenuated effects of substituents on molecularly distinct chemical interactions. When a successful correlation between biological activity and σ is obtained, lacking the identification of a specific molecular event, the only conclusion possible is that this correlation reflects mainly the changes in solvation due to the interaction. Conversely, if a correlation with σ is not obtained, one is not allowed to conclude that electronic effects do not play any role in that specific biological interaction. Rather, the acceptable conclusion is that the specific choice of σ does not reflect the true electronic effect of the substituent on the biological activity. For there is no doubt that electronic structure and its change upon substitution or interaction are always the basis for any effect.

2. The steric substituent constant

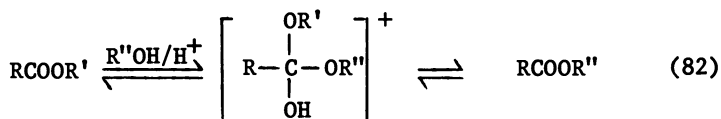
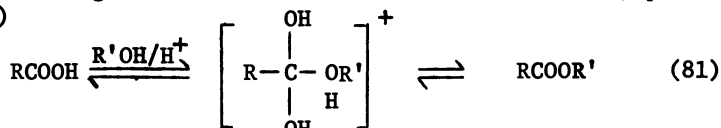
Several examples exist of correlations between biological activity and steric substituent constants: the hydrolysis of p-nitrophenylalkanoates by esterases (82) and the inhibition of cholinesterase by ethyl-p-nitrophenylalkylphosphonates (83) were found to depend only on steric substituent constants. The anti-adrenergic effects of 2-bromophenylethylamines on cat blood pressure were found to depend on steric as well as polar effects (84).

The search for an influence of steric effects on biological activity results from their well-established effects in chemical systems. As far back as 1894 (85), "steric hindrance" was invoked to account for the difficult esterification of 2,6-disubstituted benzoic acids. However, inspection of equation 53 suggests that the substituent constant, E_s , should include contributions from electronic effects as well. Indeed, acidic hydrolysis of esters of benzoic acids (4) have ρ values in the range -0.2 to +0.5. These ρ values are small and lie on both sides of zero, leading to the conclusion that electronic contributions to E_s are negligible (90). Proof for the existence of electronic effects in E_s comes from Hancock's work (15) in which the electronic effects were attributed to a hyperconjugative effect of α -hydrogen atoms.

$$E_s = E_s^c + h(n-3) \quad (80)$$

Here n is the number of α -hydrogen atoms, and h was obtained from quantum mechanical calculations (16) and represents the hyperconjugation constant. The improved correlations for some reactions (15) in which E_s was substituted with equation 80 supports the argument that electronic effects contribute to the steric parameter.

The assumption, that E_s is independent of the alcohol part of the ester and of solvent can also be questioned. Different scales of E_s were obtained for acid catalyzed esterification of RCOOH with R'OH (equation 81), for acid catalyzed transesterification of RCOOR' with R''OH (equation 82), and for acid hydrolysis (equation 53), although all have similar transition states (equations 81 and 82)



Although experimental errors can account for different E_s scales, it is likely that the transition states in equations 81 and 82 are indeed different enough to account for them. Both reasons were used to justify averaging the E_s values in order to "reduce small specific effects and experimental errors" (91).

Because of the different E_s values obtained from different molecular systems with similar transition states, it seems appropriate to analyze the model reaction from which these constants were derived. Since rate constants of the acid hydrolysis are used in the definition of E_s , these should reflect the changes in free energy of the transition and ground states with substitution. The transition state, however, has a tetrahedral geometry whereas the ground state is planar trigonal. Hence, the steric parameter reflects the additional steric interaction that is encountered by a reaction site upon changing the hybridization from sp^2 to sp^3 . Therefore, reactions exhibiting such a change (or the reverse one) should include a significant contribution from steric effects. In this category are included reactions such as addition to double bonds (92), hydration of aldehydes and ketones (93) and polymerisation of vinylic monomers (94). Reactions not exhibiting such a change between ground and transition (or initial and final) states should lack this contribution. Nucleophilic reactions, for example, do not obey equation 55 and a special scale for the steric effect in such reactions had to be obtained (95). Okamoto and his colleagues (96) also showed that E_s values are not

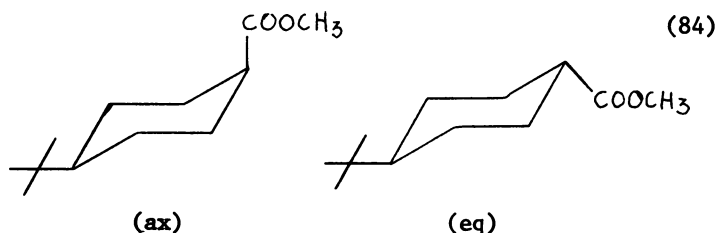
applicable to S_N2 reactions, attributing this to the fact that the ground state in S_N2 reaction is tetrahedral and the transition state assumes a trigonal bipyramidal geometry.

The limited range of validity of the steric substituent constant can be further illustrated by its relatively small contribution to the "ortho effect". Ortho substituted derivatives of benzoic acid do not obey the usual Hammett equation and the deviations were usually attributed to steric effects. Charton, after analyzing the ortho effect (97), described the value of an observed property, Q_X , as a sum of inductive (σ_I), resonance (σ_R) and steric (r_v) effects.

$$Q_X = \alpha\sigma_I + \beta\sigma_R = \psi r_v + h \quad (83)$$

As usual, h is introduced in case the observed property of the unsubstituted molecule, (Q_H), is not equal to zero. The steric parameter r_v was estimated from van der Waals radii in various ways. For ortho substituents it was found that the contributions from inductive and resonance terms outweigh by far the steric contribution, indicating the relative unimportance of the latter in the ortho effect. By contrast in aliphatic systems the contribution from r_v has been found to dominate all the others. The validity of this substituent constant has to be evaluated, but the correspondence between the r_v from aliphatic compounds and E_s was shown by the good correlation between them (98, 99). Since the van der Waals radii do not contain directional information, the correlation between r_v and E_s indicates that the E_s constants reflect the size and bulk of the substituent and lack the important property of directionality.

The directionality property of the parameters is most relevant to conformationally flexible molecules and to the description of differences in reactivity of optical isomers. The steric substituent constant reflects the weighted average of the steric effects in variously populated conformations that contribute to the observed property. For example, an indication of different steric effects in conformationally locked compounds can be obtained from the hydrolysis of methyl trans- and cis-4-*t*-butyl cyclohexanecarboxylates (100).



American Chemical
Society Library
1155 16th St. N. W.

In acid hydrolysis the ratio of the rate constants for the equatorial and axial conformers is found to be $k_{eq}/k_{ax} \approx 4.8$, indicating that the steric effect of the equatorial conformation is smaller than that of the axial. Similar measurements of the conformationally flexible unsubstituted cyclohexane carboxylate result in an averaged value that depends on the relative populations of the two conformations in the ground and transition states. The differential reactivity of optical isomers should exhibit a similar difficulty (but see the discussion on minimal steric difference, MSD).

A bothersome property of the steric substituent constants is their lack of additivity, as illustrated in the following table.

Table III. Nonuniform dependence of E_s on systematic increase in size of substituent

Substituent	E_s	Δ	Substituent	E_s	Δ	Substituent	E_s	Δ
CH ₃	0.00		CH ₃	0.00		cyclobutyl	-0.06	
		-0.07			-0.24			-0.45
CH ₃ CH ₂	-0.07		ClCH ₂	-0.24		cyclopentyl	-0.51	
		-0.40			-1.30			-0.28
(CH ₃) ₂ CH	-0.47		Cl ₂ CH	-1.54		cyclohexyl	-0.79	
		-1.07			-0.52			-0.31
(CH ₃) ₃ C	-1.54		Cl ₃ C	-2.06		cycloheptyl	-1.10	

The non-uniform change in the steric substituent constant with systematic changes in the structure of the substituent clearly suggests that the substituent affects the reactivity of the molecule by a complex mechanism. Few attempts have been made to analyze this effect.

One approach, presented by Taft (10) and later confirmed by Becker (101), is based on the separation of the free energy of activation into contributions coming from polar and steric effects (10). The contribution from the steric effect was then postulated to contain both "steric strain" or repulsion and "steric hindrance of motions" obtained from the statistical thermodynamic partition function Q^\ddagger . This separation therefore represents the changes in potential and kinetic energies related to the steric effect. Since the total steric effect $\Delta\Delta E_s^\ddagger$ is related to E_s

$$\Delta\Delta E_s^\ddagger = -2.3RT(E_s) \quad (85)$$

the separation into $\Delta\Delta E_R^\ddagger$ (repulsion) and $-RT\ln(\pi Q^\ddagger)$ (hindrance of motions), presented in equation 86, should indicate the relative contributions from these two mechanisms.

$$\Delta\Delta E_s^\ddagger = \Delta\Delta E_R^\ddagger - RT\ln(\pi Q^\ddagger) \quad (86)$$

It was also shown that the potential energy repulsion term is related to the change in enthalpy (equation 87).

$$\Delta\Delta H^\ddagger = \Delta\Delta E_\sigma^\ddagger - \Delta\Delta E_R^\ddagger \quad (87)$$

where $\Delta\Delta E_\sigma^\ddagger$, the contribution from polar-electronic effects, can be obtained from equation 88.

$$\Delta\Delta E_\sigma^\ddagger = -2.3RT(\rho^*\sigma^*) \quad (88)$$

Consequently, the entropic contribution is directly related to the steric hindrance of motions.

$$-T\Delta\Delta S^\ddagger = -RT\ln(\pi Q^\ddagger) \quad (89)$$

which can be obtained from statistical mechanical partition function terms.

This analysis showed that steric repulsion of small substituents (e.g., H, CH₃, CH₃CH₂, (CH₃)₂CH, and (CH₃CH₂)(CH₃)CH) makes negligible contributions, the steric effect being fully represented by the kinetic energy term, i.e., $-RT\ln(\pi Q^\ddagger)$. Steric repulsion became significant only for larger groups (e.g., (CH₃)₃C, neopentyl, (CH₃CH₂)₂CH) and was characterized by a similar non-uniform behavior as was observed for E_S. The steric hindrance of motions, on the other hand, showed an asymptotic behavior which did not seem to increase further with the size of the substituent, as shown in Table IV.

Table IV. Separation of the steric effect $\Delta\Delta E_S^\ddagger$ into steric repulsion, $\Delta\Delta E_R^\ddagger$, and hindrance of motions, $-RT\ln(\pi Q^\ddagger)$, in hydrolysis of esters (90).

Substituent	$\Delta\Delta E_S^\ddagger$	$\Delta\Delta E_R^\ddagger$	$-RT\ln(\pi Q^\ddagger)$
	(kcal/mole)		
H	0.0	0.0	0.0
CH ₃	+1.7	0.0	+1.6
CH ₃ CH ₂	+1.8	0.0	+1.8
(CH ₃) ₂ CH	+2.4	0.0	+2.5
(CH ₃) ₂ CHCH ₂	+3.0	0.0	+2.8
(CH ₃)(CH ₃ CH ₂)CH	+3.1	+0.5	+2.6
(CH ₃) ₃ C	+3.8	+1.4	+2.2
(CH ₃) ₃ CCH ₂	+4.1	+1.8	+2.3
(CH ₃ CH ₂) ₂ CH	+4.4	+2.0	+2.4

Interpretation of these findings is complicated: although the steric repulsion could be attributed to intrinsic molecular phenomena, the lack of dependence of the steric hindrance of motions on solvent effects was not in general agreement with the environmental effects as represented by the entropy. However because the larger substituents have many degrees of conformational

freedom, the constant value for $-RT\ln(\pi Q^\ddagger)$ reflects the conformational averaging of substituent effects that obscure specific solvent-solute interactions.

The profoundly complex influence of the steric effect on reactivity shows the need for caution in using steric substituent constants to correlate chemical activity with structure because the reaction model from which these constants were obtained may not apply to the interactions that occur in biological systems.

The STERIMOL parameters (see above, section B.2) have yielded good correlation coefficients and small standard deviations in QSAR (106). Comparison (108) between STERIMOL and other steric parameters frequently used in QSAR (e.g., E_s , U, MSD, etc.) reveals correlations between members of the two sets. Correlations of biological activity with STERIMOL parameters (108) reveal their applicability. For most of the correlations two parameters usually suffice, thus increasing the likelihood of statistical significance of the correlations. Moreover, selection of specific substituent constants by multiple regression analysis can help to identify specific steric interactions that influence the biological activity. Still, obtaining improved steric parameters from theoretical and experimental studies remains a major challenge.

It is clear that a single steric substituent constant is of limited applicability as was shown by Verloop (106, 108). The complex nature of the steric parameters reflects the fact that they represent both intra- and intermolecular steric interactions. Even though attempts to separate these contributions have been made (10), as described above, they have not been as fruitful as analogous efforts in separating electronic effects. Not only is the relation between intra- and intermolecular steric interactions complicated, but the experimental models from which the steric parameters are derived represent only a limited aspect of the steric effect on reactivity (e.g., the difference between steric interaction in planar-trigonal and tetrahedral geometries represented by E_s (107)). In this respect Verloop's approach (106) using many steric parameters, each representing a different aspect of molecular structure, offers promise. When linked to the statistical selection of the most important and significant contributions of steric parameters, it has predictive value and is potentially applicable to the elucidation of mechanisms of drug action.

3. The partition coefficient and the hydrophobic substituent constant

In the case of parameters related to the effects of the medium Collander's relation (equation 42) has been shown to hold for many different compounds in many solvent systems (112-118). For example, partition coefficients between water and oleyl alcohol, isobutyl alcohol, and xylene correlate well with partition

coefficients between water and octanol. Important to the understanding of the physicochemical basis of these parameters is the analysis of deviations from the linear relationship expressed in equation 42 (112, 119). These usually occur when the solvation mechanisms of the solvents are markedly different. For example, partition in solvents capable of hydrogen bonding do not correlate well with partitions in solvents that lack this property. Thus the correlation between $\log P_{\text{octanol/water}}$ and $\log P_{\text{cyclohexane/water}}$ is characterized by a correlation coefficient of 0.79 and a standard deviation of 0.39 for nine phenols (120). When a small amount of tributyl phosphate, a hydrogen bond acceptor, was added to the cyclohexane/water system, a new relationship was expressed

$$\log P_{\text{octanol}} = 1.0 \log P_{\text{cyclohexane}} + 1.2 \log K_{\text{HB}} + 2.35 \quad (90)$$

where $\log P_{\text{octanol}}$ and $\log P_{\text{cyclohexane}}$ represent the partition coefficients between the respective solvents and water and K_{HB} is the calculated hydrogen bonding equilibrium constant. The correlation of equation 90 increased to 0.98 and the standard deviation decreased to 0.14.

Deviations from equation 42 are not attributed solely to solvent effects. Applying Collander's equation to many partition coefficients for various compounds in various solvent systems, Hansch (112, 121) found that deviations from the correlation can be classified on the basis of known chemical properties as hydrogen bond donors for those that were below the regression line and as hydrogen bond acceptors for those above the line.

In analogy to the analysis of the deviations of electronic effects from the simple Hammett equation (18) (see above, section C.1), one can argue that $\log P$ is not free of electronic and steric effects that influence the solvation properties of the compounds. These, in turn, are reflected in the observed partition coefficients. For example, the different contributions of electronic effects to the hydrophobic substituent constant, π , becomes immediately evident from the comparison of values derived from aromatic and aliphatic systems (112) (see Table II). The different partition coefficients for ortho substituted compounds compared to meta and para substituted derivatives (112) is in agreement with the hypothesis that steric effects are responsible for this behavior. The dependence of the partition coefficient on conformational effects and on folding has also been offered to explain the deviations in $\log P$ values for compounds of the type $\text{C}_6\text{H}_5\text{CH}_2\text{CH}_2\text{CH}_2\text{X}$. For substituents (X) such as hydroxy, halogens, carboxylate, amine, nitrile and methoxy the values of π_{X} obtained from $\log P_{\text{C}_6\text{H}_5\text{CH}_2\text{CH}_2\text{CH}_2\text{X}} - \log P_{\text{C}_6\text{H}_5\text{CH}_2\text{CH}_2\text{CH}_3}$ were consistently lower by approximately 0.6 (an averaged value for 12 substituents (112)) than those obtained from $\log P_{\text{R}_\text{X}} - \log P_{\text{R}_\text{H}}$, where R is an alkyl chain. To explain this difference an intramolecular interaction was postulated between the polar group X and the benzene

ring as follows



This was supported by evidence from nuclear magnetic resonance studies (148).

In addition to intramolecular effects, intermolecular interactions have been shown to modify the observed partition coefficient. These include association of polar molecules in apolar systems, or dissociation of acids (e.g., carboxylic acids or protonated forms of amines, which are the conjugated acids of the amines). The analysis of these effects has been presented in an explicit form (136, 149, 150). A distinction has to be made however, between the types of intermolecular interactions. Association of neutral polar molecules in apolar solvents may influence the observed partition coefficient, but there is little doubt that the monomer is the compound that influences the biological activity through its interaction with the biological system. For the dissociation of acids or conjugated acids the problem becomes more complicated. It is often difficult to establish whether the protonated or the deprotonated forms or both are responsible for the biological response. Another consideration of this dissociation is related to the different pH values in different regions of the body, which changes the ratio between the protonated and unprotonated forms. Obviously, this may have an important effect on the properties of the compound, as well as on the possible ratio between active and inactive forms. A more detailed analysis of these effect is presented below (section C.3b).

The deviations are therefore attributed to either large changes in the properties of the solvents or to large interactions between the solute and the solvent. Both effects challenge the assumption in the extrathermodynamic treatment that substituent and solvent effects are small perturbations and are therefore additive properties of the constituting parts and the interaction between them (equation 28-30).

a. The partitioning between octanol and water as a model for the interaction of small molecules with biological macromolecules.

The study of solution theory (112, 123) has produced a significant understanding of the determinants for solubility and

partition between water and non-polar solvents. However, the information about the interaction of drugs with biological systems comes from indirect sources (12). For example, the stability of protein-protein complexes comes from the replacement of protein-water by protein-protein interactions (125). The stabilization therefore results from the changes in the entropy of the system related mainly to the reorganization of the aqueous phase; the enthalpic contribution is relatively small. Nevertheless, the specificity of such interactions is entirely dependent on the molecular complementarity of the interacting protein surfaces.

The "hydrophobic interaction" model. Taking protein-protein interactions as a model for the interaction of small molecules with biological macromolecules, one should note that the desolvation of the interacting species need not be complete. The desolvation can occur only from the interacting molecular surfaces (126, 127, 128), leaving the rest of the surrounding water to act as a dielectric continuum which has an attenuating effect on the intermolecular interaction (128). This model is also supported by crystallographic studies of enzymes and enzyme-substrate complexes (129, 130). For example, the crystal structure of carboxypeptidase shows a water molecule bound to the zinc ion in the active site. The comparison of crystallographic electron densities in the native enzyme and the enzyme-substrate complex (the substrate is glycyltyrosine) indicates that the water molecule in the active site was displaced by the substrate, while several other water molecules located in various regions of the enzyme remained intact (129, 130).

Similar behavior characterizes the process of partition between water and non-aqueous solvents. The thermodynamic changes following the partitioning of alkanes from water to nonpolar solvents (e.g., benzene, ether, or CCl_4) are characterized by a large positive change in entropy and small negative change in enthalpy, resulting in a negative change in the free energy (123). This was attributed to the reorganization of the structure of the aqueous phase. According to Frank and Evans (122), the presence of apolar solutes in the aqueous phase imparts a highly structured envelope of water molecules around the solute molecules. These structures, referred to as "flickering clusters" to indicate their lack of stability, are destroyed when the apolar solute is eliminated. The resulting large entropy change is due to the randomization of the water molecules. The negative change in the free energy and the clear preference of the apolar solute molecules to reside in an apolar environment was termed the "hydrophobic effects" (122, 123, 131). This term was recently criticized by Hildebrand (132, 133) who stated that it is not the hydrophobia of apolar solutes that is responsible for the observed effects but rather the hydrophilicity of the water molecules which cannot be replaced by equivalently potent interactions between water and apolar molecules. In analogy to the "hydrophobic effect" in partitioning and solubility phenomena, the thermodynamic aspects of

drug-receptor interactions (e.g., enthalpies, entropies and free energies) were termed "hydrophobic interactions". The use of this term in describing drug-receptor interactions is justified as far as the desolvation of the drug and the receptor surface is concerned. But, these interactions are characterized by their outstanding specificity, which is related to the molecular complementarity of the drug and the receptor.

Specificity in drug-receptor interactions. Following the analogy with protein-protein interactions where molecular complementarity was considered necessary to maximize the hydrophobic binding (125), drug-receptor specificity is defined as a requirement for complementarity between the drug and receptor molecules. Because short-range forces must be important in the drug-receptor complex, specificity must be reflected in the electronic and steric properties of the drug and the receptor. Thus, drug-receptor interactions can be characterized by discrete physico-chemical processes related to the solvation and to the electro-steric properties of the drug and the receptor. Because of their known structure and means to learn their properties, drugs are the variable that is studied and from which, inferences are made of their interactions with receptors and the effects they produce. The receptor, or more generally, the biological system, does not present itself as a distinct and well defined molecular structure and is therefore treated as a medium, albeit with high selectivity. The need for an appropriate model therefore becomes urgent (see discussion in this symposium (230)).

The choice of a solvent model for the hydrophobic interaction. Obviously the selectivity of the biological system cannot be modeled by a solvent. However, solvents can be chosen to model the partial desolvation observed in protein-protein interactions (see above) because there is a qualitative analogy between water and non-polar solvents and the interaction of small molecules with biological systems. This can be expressed as a quantitative relation between partition coefficients and the binding to biological systems (receptors, proteins, membranes, etc.): It has been pointed out (134) that the best way to express the partition between the aqueous phase and biological system is by the ratio of bound and free ligand concentration (B/F). According to Collander's equation (111), the relationship between B/F and P is given by equation 92.

$$\log(B/F) = a \log P + b \quad (92)$$

Not all solvent systems behave similarly, and usually low quality correlations result with solvents such as hydrocarbons, carbon tetrachloride and chloroform. This was attributed to the extensive desolvation of water upon partitioning into these solvents. Long chain aliphatic alcohols have the tendency to dissolve

limited amounts of water. For example, the concentration of water at saturation in selected solvents is compared in the following table (compiled from (112)).

Table V. Solubility of water in apolar solvents

Solvent	Water concentration at saturation (M)
cyclohexane	0.0025
heptane	0.0033
CCl ₄	0.010
CHCl ₃	0.0684
oleyl alcohol	0.712
octanol	2.3
n-butanol	9.44

It is clear that the large difference in water content between the two groups of solvents shown in Table V will influence the degree of desolvation upon partitioning from water to these solvents. This effect will become even more important for compounds that have polar groups and a high affinity for hydrogen bonds. The choice of octanol as a solvent to model the interaction of small molecules with biological systems is based on its apolar nature, thus representing the hydrophobic interaction (see above). Its ability to dissolve limited quantities (2.3M) of water models the partial desolvation of the interacting species.

Analysis of the solvent model for hydrophobic interactions.

Using the octanol/water partition coefficients, Hansch (114) analysed the binding of various compounds to proteins and found a good agreement with equation 92. One example is the binding of barbiturates to bovine serum albumin and to proteins from various tissues. The correlations are usually good (e.g., $r = 0.95$) and they show a similar behavior characterized by a slope close to 0.5 and a negative intercept. The averaged equation is:

$$\log(B/F) = (0.55 \pm 0.22) \log P - (1.38 \pm 0.44) \quad (93)$$

We analyzed the correlation of the membrane/buffer partition coefficient ($P_{m/b}$) obtained by Roth and Seeman (135) for 14 selected anesthetics with octanol/water partition coefficients (112)

$$\log P_{m/b} = (0.72 \pm 0.08) \log P_{o/w} - (0.35 \pm 0.18) \quad (94)$$

Interestingly, Roth and Seeman (135) found that the membrane/buffer partition coefficient is invariably one fifth of the octanol/water partition coefficient, in disagreement with equation 94.

Within the model represented by equation 92, the slope, a , and the intercept, b , may be interpreted as characterizing the biological system (e.g., proteins and membranes) with reference to the properties of octanol. Slopes smaller than unity would indicate that octanol is more hydrophobic than the biological system. By "more hydrophobic" we mean that its discriminative power between compounds with different $\log P$ values is larger than that of the biological system. The opposite is true for slopes greater than unity. The intercept is obtained for a hypothetical molecule with $\log P = 0$ (i.e., $P = 1$), which indicates that its solubility in water and in octanol is the same. A negative intercept, therefore, indicates that the biological system prefers more hydrophobic molecules than octanol. The slope and the intercept, therefore, relate respectively to the discriminative power and the intrinsic hydrophobicity of the membrane compared to the same properties of octanol. Both equations 93 and 94 show essentially the same behavior. Both the proteins and the membranes are less discriminative than octanol but prefer more hydrophobic molecules.

In like manner Rekker (136) developed the concepts of iso-, hyper-, and hypodiscriminative solvents. He found that only the lower alcohols such as sec-butanol, n-butanol, and n-pentanol are hypodiscriminative with respect to octanol. Other apolar solvents such as oleyl alcohol, ether, chloroform, benzene, and cyclohexane are hyperdiscriminative with respect to octanol. In his analysis of the applicability of this concept to the classification of biological membranes, he concluded that the buccal membrane, for example, is hypodiscriminative, the gastrointestinal tract might be regarded as isodiscriminative, and the blood-brain barrier is hyperdiscriminative. This is a general classification and may heavily depend on the structures of the studied molecules as well as on the observed processes (e.g., permeation through membranes rather than binding; see also below).

It thus seems appropriate to use partition coefficients between aqueous and organic phases to represent molecular properties related to the hydrophobic interactions between small molecules and the biophase. This choice is well established at the level of a second order approximation in the extrathermodynamic derivation (see section A.3). Numerous examples (137, 138) illustrate the utility of the partition coefficient in correlations of biological activity with chemical structure.

b. The partitioning between octanol and water as a model for drug distribution.

Many correlations of biological activity with $\log P$ are non-linear, and show maximal activity at an optimal value for the partition coefficient, $\log P_0$. This prompted Hansch (114) to suggest that the partition coefficient not only describes the hydrophobic interaction between molecules and the biophase (e.g., membranes, receptors) but also their ability to permeate through membranes.

The biological activity of a molecule is therefore postulated to depend on two processes: its passage through membranes to reach the site of action, and its interaction with the site. The course of passing through membranes can be considered as a partitioning between alternating hydrophilic and hydrophobic phases. Hydrophilic molecules, possessing very small partition coefficients, tend to concentrate in hydrophilic phases, whereas lipophilic molecules, having very large partition coefficients will be localized in hydrophobic phases. In both circumstances, ability to be transported to the site of action is impaired, resulting in reduced biological activity. Only compounds with an optimal partition coefficient will be efficiently transported to the site of action and thus show high biological activity. Hansch argued that a parabolic term in $\log P$ will properly account for this phenomenon. The resulting equation, describing the dependence of the biological activity on the partition coefficient, is therefore

$$\log(1/C) = -a(\log P)^2 + b\log P + c \quad (95)$$

Here C represents the concentrations of a series of compounds that elicit the same biological response. The negative coefficient of the $(\log P)^2$ term indicates that the biological activity, $\log(1/C)$, passes through a maximum for a certain optimal partition coefficient, $\log P_0$. This intuitive suggestion was largely based on the postulate (151) that the biological activity of a compound is related to its probability of reaching the receptor site in a specific interval of time. If $\log P$ is the only variable, then the logarithm of this probability is proportional to $(\log P - \log P_0)^2$. The assumption that there is only one rate determining process associated with biological response results in equation 95.

A more elaborate kinetic model was developed (152) in which the transport of the drug was described as a time dependent partitioning process in alternating aqueous and lipophilic compartments. The solution of the set of kinetical equations becomes extremely complicated for increasing number of compartments, but it was shown (114) that the numerical solution for a 20-compartment model can be fitted accurately by equation 95. Several other models have been proposed to account for the nonlinear dependence of the biological activity on the partition coefficient. They can be characterized either as kinetic models (153, 154), or as equilibrium models (155, 160). Although the basic assumptions in the various models differ slightly most of them lead to the conclusion that the parabolic dependence of the biological activity on $\log P$ is only an approximation.

As opposed to the models mentioned above, which are based on both transport and hydrophobic binding, Franke (161) proposed an empirical model to account for the nonlinear behavior based on hydrophobic binding only. This model distinguishes between binding to a hydrophobic surface and binding to a hydrophobic pocket. Both binding processes are characterized by a linear dependence

of the biological activity on $\log P$ in agreement with Collander's equation (111). In binding to a hydrophobic surface, the slope of the linear relation is 0.5, indicating only partial desolvation as described above. Binding to a hydrophobic pocket is characterized by a slope of 1.0 and is attributed to a transfer into the interior of the membrane. Intermediate cases are described by parabolic relationships. The overall dependence of the biological activity on $\log P$ can be expressed as a polynomial of the fourth degree (i.e., $\log(1/C) = a(\log P)^4 + b(\log P)^3 + c(\log P)^2 + d(\log P) + e$). Depending on the span of $\log P$ values for each series, one may obtain linear, parabolic or even more complicated expressions. This model was successfully applied to the anticholinergic potency of polymethylene-bis-trimethylammonium compounds which was expressed as an equation of the fourth order in π (161).

The most inclusive model to account for the nonlinear dependence of biological activity on $\log P$ is that derived by Martin and Hackbarth (162). It is an equilibrium model based on partition between multiple aqueous and non-aqueous compartments, similar to the one presented by Hyde (158). The inclusion of ionization processes of acids and conjugated acids (e.g., protonated amines) in the model and the analysis of the effect of these processes on the partition coefficient and the interactions with receptors makes this model generally applicable.

Even at the level of the simplest model, composed of a receptor and an aqueous phase in which an equilibrium exists between the protonated and deprotonated forms, the dependence of biological activity on $\log P$ is nonlinear. Assuming that only the neutral form of the drug interacts with the receptor leads to the following equation.

$$\log(1/C) = \log(1-\alpha) + b(\log P) - \log(a \cdot P^b(1-\alpha) + 1) + c \quad (96)$$

Here α is the ionization fraction and a and b are constants characteristic of the system. Investigation of equation 96 shows that for small $\log P$ values the term $\log(a \cdot P^b(1-\alpha) + 1)$ becomes negligibly small and the equation becomes linear in $\log P$.

$$\log(1/C) = \log(1-\alpha) + b(\log P) + c \quad (97)$$

For large $\log P$ values (i.e., $a \cdot P^b(1-\alpha) \gg 1$) the biological activity becomes independent of $\log P$ and is characterized by a constant related only to the properties of the system.

$$\log(1/C) = -\log a + c \quad (98)$$

It is interesting to note that the relation of biological activity to the ionization fraction, α , is independent of its relation to $\log P$, therefore the two equilibria of partitioning and ionization have to be considered separately. The independence of

these two effects also suggests that a highly ionized compound can be active providing it has a high enough partition coefficient.

The extension of the previous model to include a nonaqueous phase into which another partitioning can take place results in a dependence of $\log(1/C)$ on $\log P$ which resembles the parabolic dependence. The most interesting consequence of this model, which assumes that the neutral form interacts with the receptor, is that the optimal value of $\log P$ ($\log P_0$), for which the activity is maximal, depends on the ionization fraction, α . The higher the ionization fraction, the larger is $\log P_0$, indicating that for highly ionized drugs, a larger partition constant is needed to obtain a maximal biological activity. If in the same model the ionized form of the drug is assumed to interact with the receptor rather than the neutral form, the dependence of $\log P_0$ on α is different. Now $\log P_0$ no longer depends on α but rather the maximal biological activity at $\log P_0$ is small for slightly ionized compounds and is highest for completely ionized species.

The advantage of this model is that it can describe analytically the process of partition and dissociation and their interrelation as expressed in biological activity. A further advantage of this model is that it can be easily extended to account for electronic and steric effects. In the simple case, the electronic and steric effects are assumed to influence directly the biological activity and are added as independent contributions to the equations. Thus, for example equation 96 becomes

$$\log(1/C) = \log(1-\alpha) + b(\log P) - \log(a \cdot P^b(1-\alpha) + \rho\sigma + sE_s + c) \quad (99)$$

where the notation is similar to equation 96 and ρ, σ, s , and E_s have their usual meaning. More complicated examples in which electronic and steric effects influence the partition coefficient or the dissociation process were also presented (162).

In summary, the partition between aqueous and nonaqueous solvents seems to be a good model for both the hydrophobic binding to biological systems and the transport of molecules through membranes. The basis for this model in relation to binding is based on the Collander equation 42 which can be derived from an extra-thermodynamic analysis of solvent effects at the level of a second order approximation (section A.3). The hydrophobic substituent constant, π , is however the result of a first order approximation. It may be useful in correlations with biological activity, but its limitations should be recognized. Thus, it can be used for compounds with relatively small variations in structure (e.g., separate π values for aromatic and aliphatic systems) or when the π constants are free from other effects such as electronic and steric. The use of the more fundamental property, the partition coefficient, rather than the hydrophobic substituent constant is justified on a theoretical level. Partition coefficients are also good models for drug distribution phenomena. The use of testable

model systems to describe drug distribution as a phenomenon related to partition has made considerable contribution to the understanding of the factors that play a role in distribution and binding of both neutral and charged molecules.

4. Parameters derived from quantum chemical calculations

The reactivity of molecules represents their ability to undergo certain interactions. In the extrathermodynamic approach, the reactivity characteristics of molecules are described as changes in the reactivity of a reference molecule upon substitution. It is known from chemistry, however, that molecular interactions are determined by properties of the entire molecule. Quantum chemistry offers the means of obtaining molecular properties from first principles of physics and chemistry; the quantum chemical computation methods are now able to predict good relative values of physicochemical properties that can be determined by experiment either with great difficulty or only by inference. These include multipole moments, molecular polarizabilities, ionization potentials, electron affinities, charge distributions, scattering potentials, spectroscopic transitions, geometries and energies of transition states, and the relative populations of various conformations of molecules. Some of these properties are directly related to molecular reactivity (e.g., charge distribution, molecular polarizabilities, scattering potentials), and they can be implemented in QSAR studies. Quantum mechanical methods can therefore be used to obtain reactivity characteristics in order to relate molecular structure to the observed biological activity (183, 230).

In principle, the molecular wavefunctions obtained from quantum chemical calculations contain all the information about the reactivity characteristics of molecules. The choice of a specific reactivity characteristic that can account for the observed biological activity should be guided by a mechanistic hypothesis regarding the relevant type of reactions. The rational way of using quantum mechanical methods in drug design is therefore to understand the molecular determinants for a specific biological activity from the pertinent reactivity characteristics and to use this understanding to design new active structure (230). This potential for the application of quantum chemistry to drug design has not been fully realized or recognized. Instead, quantum chemical methods have been primarily used as a means to obtain new types of parameters to be used in correlations with biological activity.

The quantum mechanical parameters used in QSAR can be divided into two categories: those related to charge and those related to energy. In the first category the following parameters have been used: atomic charge densities, ϵ (82, 163, 164, 165); atomic net charges q and Q^T (166-171); atomic σ and π charges, q^σ and q^π (167, 171, 172); frontier electron density,

i.e., the density in the highest occupied molecular orbital, HOMO, (166, 173, 174, 175, 196); and free valence (176). The energy related parameters that have been used in QSAR are energies of highest occupied (E_{HOMO}) and lowest empty (E_{LEMO}) molecular orbitals (164, 171, 172, 174, 177, 196); differences in eigenvalues of molecular orbitals (164); total interaction energies (178); and Coulombic interaction energies (171). Parameters that are related to both charge densities and energies have also been used. These include the nucleophilic and electrophilic superdelocalizabilities, S^{N} and S^{E} (29, 167, 168, 172, 176, 179, 180, 181), frontier polarizabilities (171) and molecular electrostatic field potentials obtained from point charges (171, 182, 183) or from the electronic charge distribution described by the molecular wavefunction (183-191, 230).

The quantum mechanical indices described above are obtained from calculated molecular wavefunctions. The quality of the wavefunction and, consequently, of the indices depends entirely on the formalism and the level of approximation one uses. Because the molecular wavefunction is often described as a linear combination of atomic orbitals (LCAO), one can easily obtain some of the indices mentioned above. These include the atomic net charges, sigma and pi charges, frontier electron densities, E_{HOMO} and E_{LEMO} and the superdelocalizability parameters. Within the LCAO approximation one can assign the reactivity indices to specific atoms or bonds in the molecule. These indices reflect the stationary reactivity properties of the atoms and bonds in the molecule as described by charges or orbital energies and can therefore serve only as indicators of the reactivity.

The use of quantum mechanical indices of reactivity in QSAR has become increasingly popular, mainly due to the development of reliable computational methods and the availability of computer programs from which these indices can be easily obtained. In many of these studies quantum chemical indices are used instead of the extrathermodynamic parameters. A few attempts have been made to relate the quantum chemical indices to the extrathermodynamic parameters (76, 179, 180, 192). In a revealing example (192), transition state energies and charge redistribution in S_{N}^2 reactions of substituted benzyl chlorides were examined as a function of the substituent. It was found that changes in the computed energy of activation were highly correlated with σ values of the substituents. However, there were "no differences between the charge perturbations produced by the substituents in benzyl chloride and in the model (unsubstituted benzyl chloride) transition states" (192). It was concluded that the substituent effect results from the interaction of a relatively constant charge distribution on the benzyl portion with a varying charge distribution on the substituent. Thus, the common assumption that substituent effects arise from the changes in electron distribution in the vicinity of the reactive center is not supported.

Many good correlations between biological activity and quantum chemical indices have been found, and these can be used to predict the activity of other untested molecules. These indices must be used with caution in inferring mechanisms of drug action, because they have none of the advantages of the extrathermodynamic parameters but are burdened by the nature of the approximations used in their calculation. Inappropriate interpretations of the mechanistic meaning of these correlations and of correlations with parameters describing electronic effects (231) have led to criticism of the methods of quantum mechanics rather than of the abuse of these methods (197). The use of quantum chemistry in gaining insight into drug mechanisms and drug design is discussed in this symposium (230).

D. Methods in QSAR

Most structure-activity data have been analyzed by two methods: the empirical method of Hansch (114) and the mathematical method of Free and Wilson (198). Other methods have also been developed and applied to the study of QSAR (102), including pattern recognition (199-206), cluster analysis (207, 208), and discriminant analysis (204, 209, 210, 211). Pattern recognition refers to a collection of methods used to detect relationships (i.e., patterns) within a large number of observations which simultaneously depend on many variables. These methods usually do not assume any explicit form of relationship between the observables and the independent variables, and thus are not subject to a statistical test of their quality. They are very helpful in classification and reduction of large amounts of information, e.g., biological activity and chemical structure, to patterns that can be then treated by standard regression methods. In cluster analysis, the relationships between the observations that are associated with several properties can be qualified. For example, many substituents can be analyzed to determine which of the subgroups of substituents represents similar properties (207). Cluster analysis can be helpful in minimizing the number of compounds that need to be synthesized in order to obtain the maximum variation in the effects of substituents on biological activity. Discriminant analysis distinguishes between groups that confer the predictor variables different properties. It could be used, for example, to identify among a group of compounds, those structural parameters that contribute to activity or lack of activity or to agonism and antagonism. These methods are helpful in classification of drugs and in the construction of a set of compounds to which one of the quantitative methods (i.e., Hansch or Free-Wilson) can be applied.

The Hansch method (114) relates the observed biological activity to extrathermodynamic parameters (see above, section A) that are assumed to represent the electronic, steric and hydrophobic properties of the compounds responsible for the biological

effect. The extrathermodynamic basis and the physicochemical meaning of these parameters have been discussed in section C. The relationship between the biological activity and the extrathermodynamic parameters can be expressed in equation 100

$$\log(1/C) = -a(\log P)^2 + b(\log P) + \rho\sigma + sE_s + \text{const.} \quad (100)$$

where C is the concentration of a drug that elicits a specific biological response and $\log P$, σ , and E_s are the octanol/water partition coefficient, Hammett's substituent constant, and the steric substituent constant as described in sections B and C. The inclusion of the $(\log P)^2$ variable is discussed in section C. The purpose of the Hansch relationship is to find a quantitative functional dependence of the biological activity on independently obtained parameters. From this one can predict biological activities of compounds that have not been tested.

An entirely different approach was taken by Free and Wilson (198). They assumed that the biological activity of a molecule can be represented as the sum of the activity contributions of a definite substructure and the corresponding substituents. This implies that the contributions of the parent fragment (a hypothetical structure that has no substituents) and of the substituents are constant and independent of substituents at other positions on the fragment. This can be expressed mathematically as

$$BA = \sum_{ij} G_{ij} X_{ij} + \mu \quad (101)$$

Where G_{ij} is the activity contribution of a substituent X_i in position j , and μ represents the overall average of biological activity values. For example, $X_{ij} = 1$ if the substituent X_i is in position j , otherwise it is equal to zero. The most convenient way to construct equation 101 is by generating a matrix in which each row represents a specific compound. The columns are divided into as many groups as there are substituted positions on the molecule and each column within a group represents a certain substituent. For a compound in which two sites, a and b, have substituents $X_{1a}, X_{2a}, \dots, X_{na}$, and $Y_{1b}, Y_{2b}, \dots, Y_{mb}$, the following matrix can be constructed:

Compound number	Site a				Site b				Biological activity
	X_1	X_2	$X_3 \dots X_n$	Y_1	Y_2	$Y_3 \dots Y_m$			
1	1	0	0 ... 0	0	1	0 ... 0	a_1		
2	0	1	0 ... 0	1	0	0 ... 0	a_2		
3	1	0	0 ... 0	0	0	0 ... 1	a_3		
.		
.		
.		
k	0	0	0 ... 1	1	0	0 ... 0	a_k		

The matrix cannot be solved in this form because for each row the sum of X_a or Y_b values is equal to one. These linear dependences make the matrix singular and therefore an infinite number of solutions is possible. To eliminate singularities, the calculated activity contributions G_{ij} for a specific substituent X_i are arbitrarily defined as 0, independently of position (e.g., X_i can be H). Consequently the calculated biological activity of the unsubstituted compound is equal to μ in equation 101. This correction to the Free-Wilson method (198), proposed by the Fujita-Ban model, assigns the constant term μ to the calculated value of the unsubstituted molecule whereas the Free-Wilson models assigns it to the hypothetical structure which bears neither substituents nor hydrogens. The Fujita-Ban method is therefore a simple linear transformation of the Free-Wilson method.

A different modification was proposed by Cammarata and Yau (213). They arbitrarily assigned the observed biological activity of the unsubstituted compound to be the constant term. By this they assumed that there is no experimental error in determining the biological activity of the unsubstituted compound. However, since all the measured values contain experimental error, the weights given to the different activity contributions, G_{ij} , obtained from a least-squares procedure will not be equal. This is inconsistent with the definition of the least-squares method (214).

The Free-Wilson method (198) and its modifications (212, 213) are all based on the linear additivity assumption. This was criticized by Bocek et al. (215, 216), and the possibility of interactions between substituents was introduced. For two substituents the Bocek-Kopecky interaction model can be expressed by equation 102

$$BA = b_X + b_Y + e_X e_Y + k \quad (102)$$

where b_X and b_Y are the additive contributions of substituents X and Y and the $e_X e_Y$ term accounts for the interaction between them. Although this model properly accounts for the possibility of interactions, its practical use is limited because the number of parameters greatly increases as the number of substituents increases.

Several attempts have been made (217, 218, 219, 220) to establish the relationship between the extrathermodynamic approach of Hansch (114) and the mathematical approaches (198, 212, 213, 215, 216). The numerical equivalence and the theoretical interrelation between the two approaches derives from the assumption that the contributions of the individual substituents to the biological activity is represented as a weighted sum of several physicochemical properties (e.g., the extrathermodynamic parameters in the Hansch equation) as expressed in equation 103

$$G_{ij} = \sum_k b^k \phi_{ij}^k \quad (103)$$

where ϕ_{ij}^k is a specific physicochemical property (parameter) k of substituent i in position j . Substitution of equation 103 in equation 101 results in

$$BA = \sum_{ij} \sum_k b^k \phi_{ij}^k + \mu \quad (104)$$

Using the principle of additivity of extrathermodynamic parameters (see section A), we can sum the same parameters for various substituents at different position and write

$$BA = b^{\sigma} \sum_{ij} \phi_{ij}^{\sigma} + b^{\pi} \sum_{ij} \phi_{ij}^{\pi} + b^{Es} \sum_{ij} \phi_{ij}^{Es} + \mu \quad (105)$$

This equation shows the relationship between the Hansch and the Free-Wilson equations. It has been suggested (217, 220) that the Bocek-Kopecky interaction model holds for the parabolic dependence of biological activity on the partition coefficient.

In regard to the representation of biological activity in the various Free-Wilson models there has been discussion (221, 222) whether $\log(1/C)$ or C should be used. From the correspondence between the Hansch model and that of Free and Wilson, it is clear that $\log(1/C)$ should be used. Furthermore, when the linear scale of concentrations, C , is used to express biological activities, the statistical nature of the procedure can produce predictions with negative concentration values (!), a result that is impossible with a logarithmic scale.

A useful outcome of the correspondence between the two methods is the possibility of their combined application. The inclusion of indicator variables in the extrathermodynamic treatment can be regarded as an example of such an application. This is an acceptable procedure in regression analysis, and is applied in the following manner: If a specific biological activity has been measured in two congeneric series that differ only by a substituent in a position other than the position of the varying substituent, an indicator variable can be used to combine the two series into one single correlation. This is done by indicating the presence of a specific feature in one series and its absence in the other. Various biological responses measured for the same series can also be combined into one correlation by using indicator variables to distinguish between the biological responses. As with the Free-Wilson method (see above) in order to avoid singularities, the number of indicator variables has to be one less than the number of features to be distinguished. The similarity of the coefficients of the indicator variables for the different biological activities can indicate the extent of a similar dependence of these various biological activities on a certain physicochemical property. For example, a large number of erythromycin esters of type A and B with substituent variations at three positions (other than the position where a systematic substitution was introduced) was correlated with the partition coefficient and

three indicator variables. These indicator variables distinguished type A from type B and two of the three positions (9). In like manner the data reported by Hansch and Anderson (25) on the effect of barbiturates on inhibition of Arabacia cell division, inhibition of rat brain oxygen consumption, inhibition of NADH oxidation, and hypnotic activity were combined into one correlation with logP and three indicator variables to distinguish between the various biological activities (9).

It has been suggested (9) that both the Hansch and the Free-Wilson methods be used in QSAR analysis, but in general the methods are used exclusively. The Hansch method is best applied to a congeneric series in which a systematic change in a single substituent is made. The need for reliable physicochemical parameters for the substituents on the studied molecules is usually restrictive. The Free-Wilson method, on the other hand, is not limited by the physicochemical parameters (since only indicator variables are used) but needs large variations in substitution patterns to avoid singularities. Clearly the methods compliment each other.

Statistical methods. Certainly one of the most important considerations in QSAR is the statistical analysis of the correlation of the observed biological activity with structural parameters - either the extrathermodynamic (Hansch) or the indicator variables (Free-Wilson). The coefficients of the structural parameters that establish the correlation with the biological activity can be obtained by a regression analysis. Since the models are constructed in terms of multiple additive contributions the method of solution is also called multiple linear regression analysis. This method is based on three requirements (223): i) the independent variables (structural parameters) are fixed variates and the dependent variable (biological activity) is randomly produced, ii) the dependent variable is normally and independently distributed for any set of independent variables, and iii) the variance of the dependent variable must be the same for any set of independent variables.

The biological activity satisfies the second and third requirements, since each measurement is done independently (see, however, below). Upon repeated measurements the values will be normally distributed, and will have an equal variance. The independent variables in the Hansch approach are obtained from experimental measurements and contain therefore an experimental error. They cannot be considered as fixed variates. The experimental error in these parameters, however, is usually much smaller than the error in the dependent variable; they can be treated therefore as fixed variates. The complications arising from correlations between observations that include experimental errors have been thoroughly analyzed (224).

In the usual situation m observed values are fitted to an equation with n variables and p parameters ($m > p$) by a

least-squares procedure. This procedure defines the best fit by requiring that the sum of the squares of the differences between the observed and the calculated values be minimal. Goodness of fit can then be defined by three quantities: the multiple correlation coefficient, r , the standard deviation of regression, s , and the overall F value for the test of coefficient significance. The correlation coefficient gives an indication of the correspondence of the observed and the calculated values. Its square, r^2 , can be interpreted as the fraction of variance in the observed data which is explained by the correlation. The standard deviation of the regression is a measure of the scatter of the experimental values from the mean. The overall F value indicates how much better is the correlation (within a certain degree of significance level) from the correlation in which all the coefficients are taken as zero.

In order to obtain reliable QSAR the design of the series must adhere to the following requirements. The variance of all the variables should be high in order to establish the sensitivity of the biological activity to the structural parameters. The variation of the parameters should be systematic in order to minimize the effort and can be guided by computations (9) or by operational schemes (225). The parameters should be characterized by the absence of intercorrelations. Intercorrelation between parameters may lead to non-unique solutions and to misinterpretation of the correlations which may result in incorrect predictions. The analysis of the physicochemical basis for the extra-thermodynamic parameters (section C) showed that in certain instances these correlations are unavoidable due to the inherent relationships between the physical properties they represent.

Although not within the scope of this review, it must be emphasized that biological activity is the factor that is commonly treated with least rigor in analyses relating chemical structure to biological activity. When the principles of drug action and the axioms of receptor theory and enzyme kinetics are disregarded, the resultant observations are misleading, and they can yield no inferences about mechanisms. Measured *in vivo*, the biological potency of a drug is a composite of many interactions, including its rates of absorption, metabolism, excretion, and of access to the target organ; its affinities for plasma albumin (the bound drug is inactive), for the receptor associated with the effect, for other receptors that could enhance or nullify this effect. All these interactions must be regarded as independent, and it would be expected that the QSAR equation for each interaction would be different. Hence a correlation with biological potency *in vivo* either embodies the net result of all the interactions or reflects the preponderant rate limiting reaction(s). However satisfactory a correlation may be, mechanistic inferences are very limited indeed. When potency is measured *in vitro*, there are also complexities but these are tractable. In assessing potency on an isolated tissue, chemical or enzymatic

activation of the drug must be prevented. The drugs being compared must be shown to act on the same receptor and in the same manner. Full agonists cannot be compared with partial agonists, nor can competitive antagonists with noncompetitive antagonists. Inferences about mechanisms from analysis of noncompetitive antagonists are risky. Agonists and antagonists cannot be treated as one group in QSAR. These and other constraints have been well reviewed (226, 227).

E. Concluding Remarks

We have not attempted to evaluate practical applications of QSAR to the design of drugs or even the accuracy of the predictions that have been made; this subject has been extensively reviewed (9, 29, 30, 99, 106, 108, 114, 119, 126, 136, 180). Rather, we have presented the methods, their interrelationships, and the parameters, including those derived from quantum chemistry. We discussed the extrathermodynamic relationships that provide a common basis for most of the parameters used in QSAR. We have presented a critical analysis of the empirical framework for the derivation of these parameters and a rationale for choosing them. The analysis of the physicochemical basis of the parameters and of the methods shows the caution needed in interpreting molecular mechanisms from QSAR correlations.

Acknowledgements

This work was supported by the National Institute on Drug Abuse (DA 01875). The computations were supported in part by a grant of computer time from the University Computer Center of the City University of New York. Analysis of results in this work was carried out in part on the PROPHEET System, a national computer resource sponsored by the NIH through the Chemical/Biological Information - Handling Program, Division of Research. H.W. is recipient of a Career Scientist Award from the Irma T. Hirsch Trust.

Literature Cited

1. Leffler, J.E. and Grunwald, E., "Rates and Equilibria of Organic Reactions", J. Wiley & Sons, Inc., New York, 1963.
2. Exner, O., in "Advances in Linear Free Energy Relationships", Ch. 1, N.B. Chapman and J. Shorter, Eds., Plenum Press, London, 1972.
3. Hammett, L.P., J. Amer. Chem. Soc. (1937) 59, 96.
4. Jaffe, H.H., Chem. Rev. (1953) 53, 191.
5. Shorter, J. and Stubbs, F.J., J. Chem. Soc. (1949) 1180.
6. Jones, B. and Watkinson, J.B. J. Chem. Soc. (1958) 4064.
7. Mather, J.G. and Shorter, J., J. Chem. Soc. (1961) 4744.
8. Idoux, J.P. and Hancock, C.K., J. Org. Chem. (1967) 32, 1935.

9. Martin, Y.C., "Quantitative Drug Design, A Critical Introduction", Medicinal Research Series vol. 8, G.L. Grunewald, Ed., Marcel Dekker, Inc., New York, 1978.
10. Taft, R.W., in "Steric Effects in Organic Chemistry", Ch. 13, M.S. Newman, Ed., Wiley, New York, 1956.
11. Taft, R.W., J. Amer. Chem. Soc. (1952) 74, 3120.
12. Taft, R.W., J. Amer. Chem. Soc. (1953) 75, 4231.
13. Taft, R.W., J. Amer. Chem. Soc. (1952) 74, 3126.
14. Taft, R.W., J. Amer. Chem. Soc. (1953) 75, 4538.
15. Hancock, C.K., Meyers, E.A. and Yager, B.J., J. Amer. Chem. Soc. (1961) 83, 4211.
16. Kreevoy, M.M. and Eyring, H., J. Amer. Chem. Soc. (1957) 79, 5121.
17. Chapman, N.B. and Shorter, J., Eds. "Advances in Linear Free Energy Relationships", Plenum Press, New York, 1972.
18. Hammett, L.P., "Physical Organic Chemistry", McGraw Hill, New York, 1940.
19. Bordwell, F.G. and Cooper, G.D., J. Amer. Chem. Soc. (1952) 74, 1058.
20. Roberts, J.D., Webb, R.L. and McElhill, E.A., J. Amer. Chem. Soc. (1950) 72, 408.
21. Benkeser, R.A. and Krysiak, H.R., J. Amer. Chem. Soc. (1953) 75, 2421.
22. Krygowski, T.M. and Fawcett, W.R., Can. J. Chem. (1975) 53, 3622.
23. Fujita, T., Iwasa, J. and Hansch, C., J. Amer. Chem. Soc. (1964) 86, 5175.
24. Iwasa, J., Fujita, T. and Hansch, C., J. Med. Chem. (1965) 8, 150.
25. Hansch, C. and Anderson, S.M., J. Org. Chem. (1967) 32, 2583.
26. Ritchie, C.D. and Sager, W.F., Progr. Phys. Org. Chem. (1964) 2, 323.
27. Charton, M., J. Org. Chem. (1964) 29, 1222.
28. Hansch, C., Muir, R.M., Fujita, T., Maloney, P.P., Geiger, F. and Streich, M., J. Amer. Chem. Soc. (1963) 85, 2817.
29. Hansch, C., Accts. Chem. Res. (1969) 2, 232.
30. Hansch, C., in "Physico-Chemical Aspects of Drug Action", E.J. Ariens, Ed., Pergamon, New York, 1968.
31. Okamoto, Y. and Brown, H.C., J. Org. Chem. (1957) 22, 485.
32. Dickinson, J.D. and Eaborn, C., J. Chem. Soc. (1959) 3036.
33. Pearson, D.E., Baxter, J.F. and Martin, J.C., J. Org. Chem. (1952) 17, 1511.
34. Miller, J. and Parker, A.J., Austral. J. Chem. (1958) 11, 302.
35. Swain, C.G., Stockmayer, W.H. and Clarke, J.T., J. Amer. Chem. Soc. (1950) 72, 5426.
36. Russell, G.A., J. Org. Chem. (1958) 23, 1407.
37. Walter, R.I., J. Amer. Chem. Soc. (1966) 88, 1923.
38. Katritzky, A.R. and Topsom, R.D., in "Advances in Linear Free Energy Relationships", N.B. Chapman and J. Shorter, Eds., Ch. 3, Plenum Press, London, 1972.

39. Ewing, D.F., in "Correlation Analysis in Chemistry. Recent Advances", Ch. 8, N.B. Chapman and J. Shorter, Eds., Plenum Press, New York, 1978.
40. Bursery, M.M., in "Advances in Linear Free Energy Relationships", N.B. Chapman and J. Shorter, Eds., Ch. 10, Plenum Press, London, 1972.
41. Kirsch, J.F., in "Advances in Linear Free Energy Relationships", N.B. Chapman and J. Shorter, Eds., Ch. 8, Plenum Press, London, 1972.
42. Hansch, C., Grieco, C., Silipo, C. and Vittoria, A., J. Med. Chem. (1977) 20, 1420.
43. Taft, R.W., J. Phys. Chem. (1960) 64, 1805.
44. Taft, R.W., Ehrenson, S., Lewis, I.C. and Glick, R.E., J. Amer. Chem. Soc. (1959) 81, 5352.
45. Yukawa, Y., Tsuno, Y. and Sawada, M., Bull. Chem. Soc. Japan (1966) 39, 2274.
46. Norman, R.O.C., Radda, G.K., Brimacombe, D.A., Ralph, P.D. and Smith, E.M., J. Chem. Soc. (1961) 3247.
47. Norman, R.O.C. and Ralph, P.D., J. Chem. Soc. (1963) 5431.
48. Sjoström, M. and Wold, S., Chemica Scripta (1974) 6, 114.
49. Wold, S. and Sjoström, M., in "Correlation Analysis in Chemistry. Recent Advances", Ch. 1, N.B. Chapman and J. Shorter, Eds., Plenum, New York, 1978.
50. Sjoström, M. and Wold, S., Chemica Scripta (1976) 9, 200.
51. Schwarzenbach, G. and Rudin, E., Helv. Chim. Acta (1939) 22, 360.
52. Stock, L.M. and Brown, H.C., Adv. Phys. Org. Chem. (1963) 1, 35.
53. McGary, C.W., Okamoto, Y. and Brown, H.C., J. Amer. Chem. Soc. (1955) 77, 3037.
54. Lancelot, C.J., Harper, J.J. and Schleyer, P. von R., J. Amer. Chem. Soc. (1969) 91, 4294.
55. Lancelot, C.J. and Schleyer, P. von R., J. Amer. Chem. Soc. (1969) 91, 4291.
56. Ruasse, M.-F. and Dubois, J.-E., Tetrahedron Lett. (1970) 1163.
57. Bunnett, J.F. and Zahler, R.E., Chem. Rev. (1951) 49, 294.
58. Ogata, Y., Kawasaki, A. and Okumura, N., J. Org. Chem. (1964) 29, 1985.
59. Jencks, W.P., Progr. Phys. Org. Chem. (1964) 2, 63.
60. Beltrame, P., Veglio, C. and Simonetta, M., Chem. Commun. (1966) 433.
61. Taft, R.W. and Lewis, I.C., J. Amer. Chem. Soc. (1958) 80, 2436.
62. Taft, R.W. and Lewis, I.C., J. Amer. Chem. Soc. (1959) 81, 5343.
63. Taft, R.W. and Lewis, I.C., Tetrahedron (1959) 5, 210.
64. Roberts, J.L. and Jaffe, H.H., J. Amer. Chem. Soc. (1959) 81, 1635.

65. Roberts, J.D. and Moreland, W.T., J. Amer. Chem. Soc. (1953) 75, 2167.
66. Holtz, H.D. and Stock, L.M., J. Amer. Chem. Soc. (1964) 86, 5188.
67. Ritchie, C.D. and Lewis, E.S., J. Amer. Chem. Soc. (1962) 84, 591.
68. Exner, O., Coll. Czech. Chem. Comm. (1966) 31, 65.
69. Exner, O. and Jonas, J., Coll. Czech. Chem. Comm. (1962) 27, 2296.
70. Taft, R.W., Price, E., Fox, I.R., Lewis, I.C., Andersen, K.K. and Davis, G.T., J. Amer. Chem. Soc. (1963) 85, 709, 3146.
71. Brownlee, R.T.C., Katritzky, A.R. and Topsom, R.D., J. Amer. Chem. Soc. (1965) 87, 3260.
72. Swain, C.G. and Lupton, E.C., J. Amer. Chem. Soc. (1968) 90, 4328.
73. Dewar, M.J.S. and Gridale, P.J., J. Amer. Chem. Soc. (1962) 84, 3539, 3548.
74. Hansen, L.D. and Hepler, L.G., Can. J. Chem. (1972) 50, 1030.
75. Yukawa, Y. and Tsuno, Y., Bull. Chem. Soc. Japan (1959) 32, 965, 971.
76. Yoshioka, M., Hamamoto, K. and Kubota, T., Bull. Chem. Soc. Japan (1962) 35, 1723.
77. Norman, R.O.C. and Radda, G.K., Tetrahedron Lett. (1962) 125.
78. Knowles, J.R., Norman, R.O.C. and Radda, G.K., J. Chem. Soc. (1960) 4885.
79. Exner, O., Coll. Czech. Chem. Comm. (1964) 29, 1094.
80. Hepler, L.G., Can. J. Chem. (1971) 49, 2803.
81. Hepler, L.G., J. Amer. Chem. Soc. (1963) 85, 3089.
82. Hansch, C., Deutsch, E.W. and Smith, R.N., J. Amer. Chem. Soc. (1965) 87, 2738.
83. Hansch, C., Farmaco, Ed. Sci. (1968) 23, 293.
84. Hansch, C. and Lien, E., Biochemical Pharmacol. (1968) 17, 709.
85. Meyer, V. and Sudborough, J.J., Chem. Ber. (1894) 27, 1580.
86. Hafernist, M. and Baltzly, R., J. Amer. Chem. Soc. (1947) 69, 362.
87. Brown, H.C. and Domash, L., J. Amer. Chem. Soc. (1956) 78, 5384.
88. Brown, H.C., Gintis, D. and Domash, L., J. Amer. Chem. Soc. (1956) 78, 5387.
89. Pavelich, W.A. and Taft, R.W., J. Amer. Chem. Soc. (1957) 79, 4935.
90. Shorter, J., in "Advances in Linear Free Energy Relationships" Ch. 2, N.B. Chapman and J. Shorter, Eds., Plenum Press, London, 1972.
91. Taft, R.W., J. Amer. Chem. Soc. (1952) 74, 2729.
92. Mouvier, G. and Dubois, J.-E., Bull. Soc. Chim. France (1968) 1441.
93. Bell, R.P., Adv. Phys. Org. Chem. (1966) 4, 1.

94. Otsu, T., Ito, T., Fukumizu, T. and Imoto, M., Bull. Chem. Soc. Japan (1966) 39, 2257.
95. Bogatkov, S.V., Popov, A.F. and Litvinenko, L.M., Reakts. Sposobnost Org. Soedin (1969) 6, 1011; EE, 436.
96. Okamoto, K., Nitta, I., Imoto, T. and Shingu, H., Bull. Chem. Soc. Japan (1967) 40, 1905.
97. Charton, M., Progr. Phys. Org. Chem. (1971) 8, 235.
98. Charton, M., Progr. Phys. Org. Chem. (1973) 10, 81.
99. Unger, S.H. and Hansch, C., Progr. Phys. Org. Chem. (1976) 12, 91.
100. Chapman, N.B., Shorter, J. and Toyne, K.J., J. Chem. Soc. (1963) 1291.
101. Becker, F., Z. Naturforsch., (1959) 14a, 547; (1960) 15b, 251; (1961) 16b, 236.
102. Kier, L.B. and Hall, L.H., "Molecular Connectivity in Chemistry and Drug Research", Academic Press, New York, 1976.
103. Kier, L.B. and Hall, L.H., J. Pharmac. Sci. (1976) 65, 1806.
104. Raudic, M., J. Amer. Chem. Soc. (1975) 97, 6609.
105. Simon, Z., "Quantum Biochemistry and Specific Interactions", Abacus Press, Kent, England, 1976.
106. Verloop, A., Hoogenstraaten, W. and Tipker, J. in "Drug Design", E.J. Ariens, Ed., vol. 7, Academic Press, New York, 1976.
107. Bowden, K. and Young, R.C., J. Med. Chem. (1970) 13, 225.
108. Verloop, A. and Tipker, J. in "Biological Activity and Chemical Structure", J.A.K. Buisman, Ed., Elsevier, Amsterdam, 1977.
109. Meyer, H., Arch. Exptl. Pathol. Pharmacol. (1899) 42, 110.
110. Overton, E., "Studien uber die Narkose", Fischer, Jena, Germany, 1901.
111. Collander, R., Acta Chem. Scand. (1951) 5, 774.
112. Leo, A., Hansch, C. and Elkins, D., Chem. Rev. (1971) 71, 525.
113. Wulfert, E., Bolla, P. and Mathieu, J., Chim. Ther. (1969) 4, 257.
114. Hansch, C. in "Drug Design", E.J. Ariens, Ed., Vol. 1, Academic Press, New York, 1971.
115. Smith, H.W., J. Phys. Chem. (1921) 25, 204.
116. Smith, H.W., J. Phys. Chem. (1921) 25, 605.
117. Meyer, K.H. and Hemmi, H., Biochem. Z. (1935) 277, 39.
118. Koizumi, T., Arita, T. and Kakemi, K., Chem. Pharm. Bull. (1964) 12, 413.
119. Verloop, A., in "Drug Design", E.J. Ariens, Ed., Vol. 3, Academic Press, New York, 1972.
120. Higuchi, T., Richards, J., Davis, S., Kamada, A., Hou, J., Nakano, M., Nakono, N. and Pitman, I., J. Pharm. Sci. (1969) 58, 661.
121. Leo, A. and Hansch, C., J. Org. Chem. (1971) 36, 1539.
122. Frank, H.S. and Evans, M.W., J. Chem. Phys. (1945) 13, 507.

123. Kauzmann, W., Adv. Protein Chem. (1959) 14, 1.
124. Tanford, C., Science (1978) 200, 1012.
125. Chotia, C. and Janin, J., Nature (1975) 256, 705.
126. Rekker, R.F. in "Biological Activity and Chemical Structure", J.A.K. Buisman, ed., p. 107 Elsevier, Amsterdam, 1977.
127. Hopfinger, A.J., "Intermolecular Interactions and Biomolecular Organization", Wiley Interscience, New York, 1977.
128. Warshel, A. and Levitt, M., J. Mol. Biol. (1976) 103, 227.
129. Lipscomb, W.N., Hartsuck, J.A., Reeke, G.N., Quioco, F.A., Bethge, P.H., Ludwig, M.L., Steitz, T.A., Muirhead, H. and Coppola, J.C., Brookhaven Symp. Biol. (1968) 21, 24.
130. Lipscomb, W.N., Hartsuck, J.A., Reeke, G.N. and Quioco, F., Proc. Nat. Acad. Sci. USA, (1969) 64, 1389.
131. Tanford, C., "The Hydrophobic Effect: Formation of Micelles and Biological Membranes", Wiley, New York, 1973.
132. Hildebrand, J.H., J. Phys. Chem. (1968) 72, 1841.
133. Hildebrand, J.H., Proc. Natl. Acad. Sci. USA (1979) 76, 194.
134. Bird, A.E. and Marshall, A.C., Biochem. Pharmacol., (1967) 16, 2275.
135. Roth, S. and Seeman, P., Biochem. Biophys. Acta (1972) 255, 207.
136. Rekker, R.F., "The Hydrophobic Fragmental Constant; Its Derivation and Application, A Means of Characterizing Membrane Systems," Vol. I of Pharmacochimistry Library, W. Th. Nauta and R.F. Rekker, Eds., Elsevier, Amsterdam, 1977.
137. Hansch, C. and Dunn, III, W.J., J. Pharm. Sci. (1972) 61, 1.
138. Hansch, C., in "Correlation Analysis in Chemistry. Recent Advances", Ch. 9, N.B. Chapman and J. Shorter, Eds., Plenum, New York, 1978.
139. Quayle, O.R., Chem. Rev. (1953) 53, 484.
140. Vogel, A.I., Cresswell, W.T., Jeffery, G.H. and Leicester, J., J. Chem. Soc. (1952) 514.
141. Hildebrand, J.H. and Scott, R.L., "The Solubility of Nonelectrolytes", Reinhold, New York, 1950.
142. Small, P.A., J. Appl. Chem. (London) (1953) 3, 71.
143. Hansch, C., Leo, A., Unger, S.H., Kim, K.H., Nikaitani, D. and Lien, E., J. Med. Chem. (1973) 16, 1207.
144. Carlson, R.M., Carlson, R.E. and Kopperman, H.L., J. Chromatography (1975) 107, 219.
145. Twitchett, P.J. and Moffat, A.C., J. Chromatography (1975) 111, 149.
146. McCall, J., J. Med. Chem. (1975) 18, 549.
147. Tomlinson, E., J. Chromatography (1975) 113, 1.
148. Baker, B., Kawazu, M., Santi, D. and Schwan, T., J. Med. Chem. (1967) 10, 304.
149. Moelwyn-Hughes, E.A., "Physical Chemistry", Pergamon Press, New York, 1961.

150. Dippy, J.F., Hughes, S.R. and Laxton, J.W., J. Chem. Soc. (1956) 2995.
151. Hansch, C. and Fujita, T., J. Amer. Chem. Soc. (1964) 86, 1616.
152. Penniston, J.T., Beckett, D.-L. and Hansch, C., Mol. Pharmacol. (1969) 5, 333.
153. Flynn, G.L. and Yalkowsky, S.H., J. Pharm. Sci. (1972) 61, 838.
154. Yalkowsky, S.H. and Flynn, G.L., J. Pharm. Sci. (1973) 62, 210.
155. Higuchi, T. and Davis, S.S., J. Pharm. Sci. (1970) 59, 1376.
156. McFarland, J.W., J. Med. Chem. (1970) 13, 1192.
157. Flynn, G.L., Yalkowsky, S.H. and Weiner, N.D., J. Pharm. Sci. (1974) 63, 300.
158. Hyde, R.M., J. Med. Chem. (1975) 18, 231.
159. Kubinyi, H., Arzneim.-Forsch. (1976) 26, 1991.
160. Kubinyi, H., J. Med. Chem. (1977) 20, 625.
161. Franke, R. in "Biological Activity and Chemical Structure", J.A.K. Buisman, Ed., vol. 2 of Pharmacology Library, Elsevier, Amsterdam, 1977.
162. Martin, Y.C. and Hackbarth, J.J., J. Med. Chem. (1976) 19, 1033.
163. Hansch, C., Proc. 3rd Int. Pharmacol. Meet. (1968) 7, 141.
164. Hermann, R.B., Culp, H.W., McMahon, R.E. and Marsh, M.M., J. Med. Chem. (1969) 12, 749.
165. Lien, E.J., J. Med. Chem. (1970) 13, 1189.
166. Martin, Y.C., J. Med. Chem. (1970) 13, 145.
167. Cammarata, A. and Stein, R.L., J. Med. Chem. (1968) 11, 829.
168. Rogers, K.E. and Cammarata, A., J. Med. Chem. (1969) 12, 692.
169. Cammarata, A., J. Med. Chem. (1968) 11, 1111.
170. Rogers, K.E. and Cammarata, A., Biochim. Biophys. Acta (1969) 193, 22.
171. Wohl, A.J., Mol. Pharmacol. (1970) 6, 195.
172. Allen, R.C., Carlson, G.L. and Cavalito, C.J., J. Med. Chem. (1970) 13, 909.
173. Green, J.P. and Kang, S. in "Molecular Orbital Studies in Chemical Pharmacology", L.B. Kier, Ed., Springer Verlag, New York, 1970.
174. Johnson, C.L. and Green, J.P., Int. J. Quantum Chem. (1974) QBS1, 159.
175. Weinstein, H., Chou, D., Kang, S., Johnson, C.L. and Green, J.P., Int. J. Quantum Chem. (1976) QBS3, 134.
176. Sasaki, Y. and Suzuki, M., Chem. Pharm. Bull. (1970) 18, 1759, 1774.
177. Green, J.P., Johnson, C.L. and Weinstein, H., in "Psychopharmacology: A Generation of Progress", M.A. Lipton, A. DiMascio and K.F. Killam, Eds., Raven Press, New York, 1978.

178. Fuller, R.W., Marsh, M.M. and Mills, J., J. Med. Chem. (1968) 11, 397.
179. Cammarata, A., J. Med. Chem. (1969) 12, 314.
180. Cammarata, A., in "Molecular Orbital Studies in Chemical Pharmacology", L.B. Kier, Ed., Springer Verlag, New York, 1970.
181. Green, J.P. and Malrieu, J.P., Proc. Nat. Acad. Sci. USA (1965) 54, 659.
182. Hayes, D.M. and Kollman, P.A., J. Amer. Chem. Soc. (1976) 98, 3335, 7811.
183. Weinstein, H., Int. J. Quantum Chem. (1975) QBS2, 59.
184. Weinstein, H., Srebrenik, S., Maayani, S. and Sokolovsky, M., J. Theor. Biol. (1977) 64, 295.
185. Weinstein, H., Srebrenik, S., Pauncz, R., Maayani, S., Cohen, S. and Sokolovsky, M., in "Chemical and Biochemical Reactivity", E.D. Bergmann and B. Pullman, Eds. p. 493, D. Reidel Publ. Co., Dordrecht, Holland, 1974.
186. Scrocco, E. and Tomasi, J., in "Topic in Current Chemistry, New Concepts II", No. 42 p. 95, Springer Verlag, Berlin, 1973.
187. Politzer, P., Daiker, K.C. and Donnelly, R.A., Cancer Letters (1976) 2, 17.
188. Petrongolo, C., Machia, B., Machia, F. and Martinelli, A., J. Med. Chem. (1977) 20, 1645.
189. Petrongolo, C., Gazz. Chim. Ital. (1978) 108, 445.
190. Weinstein, H. and Osman, R., Int. J. Quantum Chem. (1977) QBS4, 253.
191. Weinstein, H., Osman, R., Edwards, W.D. and Green, J.P., Int. J. Quantum Chem. (1978) QBS5, 449.
192. Davidson, R.B. and Williams, C.R., J. Amer. Chem. Soc. (1978) 100, 73.
193. Davis, S.S., Experientia (1970) 26, 671.
194. Hildebrand, J.H., Prausnitz, J.M. and Scott, R.L. "Regular and Related Solutions", van Nostrand-Reinhold, New York, 1970.
195. Srebrenik, S. and Cohen, S., J. Phys. Chem. (1976) 80, 996.
196. Green, J.P., Johnson, C.L., Weinstein, H., Kang, S. and Chou, D., in "The Psychopharmacology of Hallucinogens", R.C. Stillman and R.E. Willette, Eds., p. 28, Pergamon, New York, 1978.
197. Hansch, C. in "Biological Activity and Chemical Structure", J.A.K. Buisman, Ed., p. 279, Pharmaco Chemistry Library, vol. 2, Elsevier, Amsterdam, 1977. (From "Summary of the Round Table Discussion").
198. Free, S.M., Jr. and Wilson, J.W., J. Med. Chem. (1964) 7, 395.
199. Ting, K.H., Lee, R.C.T., Milne, G.W.A., Shapiro, H. and Gaurino, A.M., Science (1973) 180, 417.

200. Kowalski, B.R. and Bender, C.F., J. Amer. Chem. Soc. (1974) 96, 916.
201. Stuper, A.J. and Jurs, P.C., J. Amer. Chem. Soc. (1975) 97, 182.
202. Chu, K.C., Feldman, R.J., Shapiro, M.B., Hazard, G.F. and Geran, R.I., J. Med. Chem. (1975) 18, 539.
203. Soltzberg, L.J. and Wilkens, C.L., J. Amer. Chem. Soc. (1977) 99, 439.
204. Martin, Y.C., Holland, J.B., Jarboe, C.H. and Plotnikov, N., J. Med. Chem. (1974) 17, 409.
205. Dunn, III, W.J., Wold, S. and Martin, Y.C., J. Med. Chem. (1978) 21, 922.
206. Dunn, III, W.J. and Wold, S., J. Med. Chem. (1978) 21, 1001.
207. Hansch, C., Unger, S.H. and Forsythe, A.B., J. Med. Chem. (1973) 16, 1217.
208. Dunn, III, W.J., Greenberg, M.J. and Callejas, S.S., J. Med. Chem. (1976) 19, 1299.
209. Martin, Y.C., Cancer Chemother. Rep. Part 2, (1974) 4, 35.
210. Prakash, G. and Hodnett, E.M., J. Med. Chem. (1978) 21, 369.
211. Dove, S., Franke, R., Mndshojan, O.L., Schuljev, W.A. and Chashakjan, L.W., J. Med. Chem. (1979) 22, 90.
212. Fujita, T. and Ban, T., J. Med. Chem. (1971) 14, 148.
213. Cammarata, A. and Yau, S.J., J. Med. Chem. (1970) 13, 93.
214. Kubinyi, H. and Kehrhn, O.-H., J. Med. Chem. (1976) 19, 578, 1040.
215. Bocek, K., Kopecky, J., Krivucova, M. and Vlachova, D., Experientia (1964) 20, 667.
216. Kopecky, J., Bocek, K. and Vlachova, D., Nature (London) (1965) 207, 981.
217. Singer, J. and Purcell, W.P., J. Med. Chem. (1967) 10, 1000.
218. Cammarata, A., J. Med. Chem. (1972) 15, 573.
219. Craig, P.N. in "Biological Correlation - The Hansch Approach", Advances in Chemistry Series (1972) 114, 115.
220. Kubinyi, H., J. Med. Chem. (1976) 19, 587.
221. Purcell, W.P. and Clayton, J.M., J. Med. Chem. (1968) 11, 199.
222. Ban, T. and Fujita, T., J. Med. Chem. (1969) 12, 353.
223. Anderson, R.L. and Bancroft, T.A., "Statistical Theory in Research", McGraw Hill Inc., New York 1952, p. 168.
224. Riggs, D.S., Guarnieri, J.A. and Addelman, S., Life Sciences (1978) 22, 1305.
225. Topliss, J.G. and Martin, Y.C. in "Drug Design", E.J. Ariens, Ed. vol. 5, Academic Press, New York, 1975, p. 1.
226. Ariens, E.J., ed. "Molecular Pharmacology" Vol. 1, Part 1, pp. 3-503, Academic Press, New York, 1964.
227. van Rossum, J.M. in "Molecular Pharmacology" E.J. Ariens, Ed. Vol. 2, Part 2, pp. 201-255, Academic Press, New York, 1964.

228. Canas-Rodriguez, A. and Tute, M.S. in "Biological Correlations - The Hansch Approach", Adv. Chem. Ser. (1972) 114, 41.
229. Osinga, M., J. Amer. Chem. Soc. (1979) 101, 1621.
230. Weinstein, H., Osman, R. and Green, J.P., this symposium.
231. Nichols, D.E., Shulgin, A.T. and Dyer, D.C., Life Sci. (1977) 21, 569.
232. Shorter, J., in "Correlation Analysis in Chemistry. Recent Advances", Ch. 4, N.B. Chapman and J. Shorter, Eds., Plenum, New York, 1978.

RECEIVED June 8, 1979.

Molecular Mechanics and Crystal Structure Analysis in Drug Design

DAVID J. DUCHAMP

Physical & Analytical Chemistry Research, The Upjohn Company,
Kalamazoo, MI 49001

Although most medicinal chemists use structure-activity relationships of various sorts in drug design, very few use quantitative three-dimensional structural information in designing potentially active molecules. Three-dimensional information that is being used, comes for the most part from qualitative work with molecular models. This situation exists, not because of any lack of interest or lack of awareness among medicinal chemists, but because: 1) availability of three-dimensional information most often occurs after the work of the medicinal chemist is done, and 2) when such information exists, for example, published results from crystal structure studies, it is not in a form which is very useful to the medicinal chemist.

Modern digital computer technology is making it increasingly possible to obtain three-dimensional information in a timely fashion, and to provide tools for useful interpretation of these results. This paper describes two methods, one experimental and one calculational, for obtaining and interpreting accurate three-dimensional information. Later papers in this symposium describe in detail the use of computer graphics, an essential tool in the timely determination and useful interpretation of such data.

Crystal Structure Analysis

The experimental technique which gives the most accurate structural information about molecules of the size of drug molecules is crystal structure analysis. The precision obtainable in a typical x-ray diffraction experiment is illustrated in Figure 1 (1). Bond distances are routinely determined to standard deviations of 0.002 to 0.005 Å; bond angles to 0.1 to 0.4°. The literature contains the results of thousands of crystal structure analyses (2). These results are contained in the unit cell parameters and the table of x,y,z coordinates of the atoms. Most research-oriented computer centers have computer

0-8412-0521-3/79/47-112-079\$06.00/0

© 1979 American Chemical Society

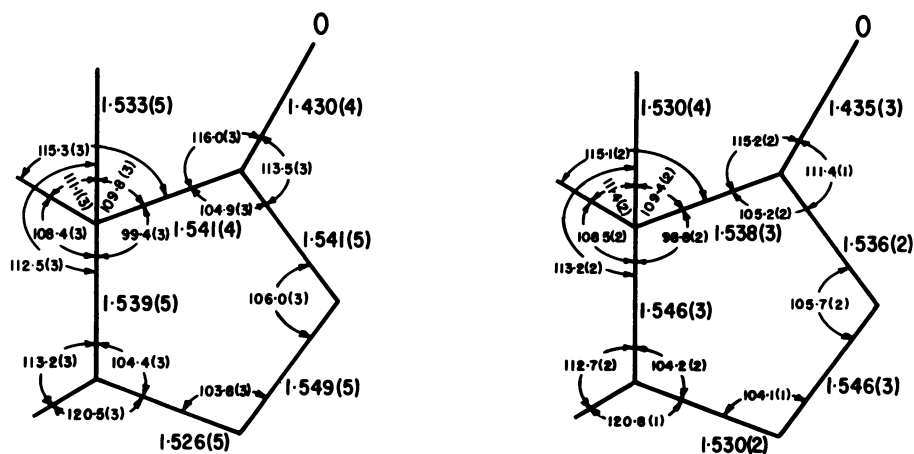


Figure 1. Precision of a typical x-ray diffraction experiment. The D ring of two steroids, 19-norandrostenediol (left) and 7- α -methyl 19-norandrostenediol (right) are shown. Bond distances (Å) and bond angles (deg) are followed by standard deviations in the last digit (in parentheses).

programs which will calculate almost any molecular parameter from this information -- bond distances, bond angles, torsion angles, deviations from planes, etc. An increasing number of computer centers have the facilities to produce drawings from this data on interactive graphics terminals and/or on hard copy graphics devices. Unfortunately, some journal editors have not recognized the importance of the coordinate results table and have insisted it be dropped from a publication or be relegated to a microfilm supplement, thus vastly decreasing the value of the publication. The coordinate table is far more important than tables of distances and angles, or even pretty pictures such as shown in Figure 1.

The easiest way to access the crystal structure literature is through "Molecular Structures and Dimensions", which provides an up-to-date well indexed bibliography of all organic and organometallic crystal structures that are available (2). This excellent series is the work of the Cambridge Crystallographic Data Centre, established by Professor Olga Kennard in the U.K. In addition to the bibliographies, the Cambridge Data Centre has affiliated data centers in various countries to which they supply bibliographic and structural information, including coordinate and unit cell information, in computer readable form, along with computer programs to access the file and display the three-dimensional structures. Copies of the data file and programs can be obtained at a low cost from the national data centers. For steroid structure, "The Atlas of Steroid Structure" (3) is a particularly good reference. This work contains a substantial amount of useful interpretative information as well as coordinate and unit cell information.

Although crystal structure analysis was once a very time consuming and very expensive process, this has not been the case for a number of years. Structural results are usually available within two weeks of when a crystal is placed on the diffractometer instrument. This change has been due to a number of factors. Computer-controlled diffractometers have made data collection more accurate and much easier. The process of obtaining a trial structure has been much facilitated by the use of computerized direct methods and by computer graphics. More powerful computers have made the entire crystallographic computing process much simpler. Several good crystallographic computing systems are being used, which handle such things as crystallographic symmetry and non-orthogonal coordinate systems fairly automatically. Whereas the price of computers and computing has decreased several fold in the last few years, the amount of computing needed for a crystal structure analysis has increased only slightly, leading to a dramatic reduction in the computing cost for a structure determination.

In fact, digital computers and automatic diffractometers have simplified the process of crystal structure analysis to the point where there is no reason, other than reluctance to

take time to learn a new technique, why most crystal structures cannot be determined by any chemist who is reasonably adept at running scientific instruments and working with molecular models. All that is needed is access to an automatic diffractometer for data collection, access to computer programs, and computer time for calculations. It is probably wise to arrange for consultation with someone who has previously done a structure determination or with a professionally trained crystallographer to smooth over the learning process. Even if a chemist never does another crystal structure, the experience of having performed one provides a very valuable background for using and interpreting crystal structure results in the literature.

A number of complications can arise in crystal structure analysis, in particular, when crystal structure results are used to aid in drug design. Because of the experimental nature of the technique, a sample, in the form of a single crystal, is needed; this sample is sometimes difficult or impossible to obtain. Crystal structures are still being found which do not yield readily to the computerized solutions mentioned above; the number of such structures is decreasing year by year, but they still do occur.

Scientists using crystal structure results need to be aware of several factors related to such data (4-7). Since x-rays interact with electrons in the crystal, x-ray diffraction does not give good experimental positions for hydrogen atoms which possess only one electron each. Also of major concern, is that crystal structure analysis measures the molecule in the conformation it assumes in the crystal, whereas for drug design, the chemist usually is interested in the the conformation at a receptor.

Diffraction data are spacially averaged over millions of unit cells, symmetry related by translation, and are time averaged over a period of several days. If all unit cells are not exactly the same, the structure is said to contain "disorder". Structural results in the disordered areas do not represent individual molecules, but some kind of average over different states. If a crystal is deteriorating during the data collection process, the resulting structural results will be distorted by an averaging due to this process. Thermal effects lead to structural results which represent average positions of atoms. In certain molecules, sizeable vibrations occur in the crystal at the temperature of the diffraction experiment. This leads to a shortening in some bond distances and a distortion of certain bond angles. Figure 2 shows an example of shortening of a measured bond length by thermal vibration in crystals of fentanyl citrate studied in our laboratory (8). The observed length of the C-C bond in the ethyl moiety was significantly shorter in a room temperature x-ray analysis, than it was when thermal vibrations were significantly reduced by taking the data at low temperature. Most thermal shortening of bond

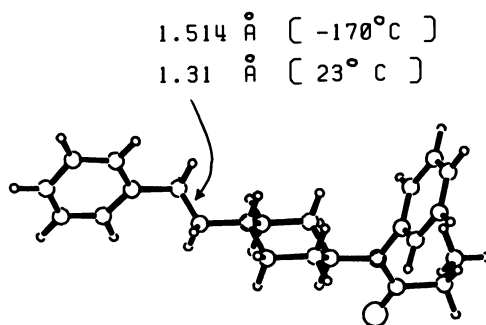


Figure 2. Bond shortening in room temperature x-ray diffraction results on fentanyl citrate

lengths is not as obvious as illustrated here. Low temperature data collection is very effective in eliminating thermal effects, and also deterioration effects. The disorder phenomenon persists at low temperature however.

Molecular Mechanics

Molecular mechanics may be regarded as quantitative model building. The calculational technique is simple in principle (9,10). The strain energy of a particular conformation of a molecule is expressed as the sum of terms of several types, each related to certain structural parameters; for example, bond length, non-bonded contact, torsion angle. Each term is a simple expression involving one or more empirically derived potential parameter, and one or more structural parameter. Usually, the structural parameters of the molecule are varied to minimize the strain energy, the potential parameters being held fixed. The number and variety of the terms used in the strain energy expression varies greatly with investigator, as do the thoroughness of the methods used to find minimum energy conformations, and the accuracy of molecular mechanics results.

The form of the potential used in my laboratory is given by

$$E_{\text{strain}} = E_b + E_\theta + E_{\text{nb}} + E_{\text{hb}} + E_{\text{op}} + E_{\text{tor}} + E_{\text{dd}}.$$

E_b is a bond distance deformation energy, defined as

$$E_b = f_d(r - r_0)^2,$$

where f_d and r_0 are the empirical potential parameters, and r is the bond distance. One such term is entered for each bond distance.

E_θ is a bond angle deformation energy, defined as

$$E_\theta = f_\theta(\Delta\theta^2 - 0.000002924\Delta\theta^3), \quad \Delta\theta = |\theta - \theta_0|,$$

where f_θ and θ_0 are the empirical potential parameters, and θ is the bond angle, one such term for each bond angle.

E_{nb} is a non-bonded interaction energy, defined as

$$E_{\text{nb}} = fe^{-gr} - (e/r^6),$$

where f , g , and e are the empirical potential parameters and r is the non-bonded distance. One such term is entered for each r , although for most calculations, those terms for r greater than a specified minimum (e.g., 10Å) are skipped.

E_{hb} is a hydrogen bond energy term, modified from the definition of Schroeder and Lippincott (11), to be a function with one potential parameter.

E_{op} is an out-of-plane deformation energy (12) (for bending at sp^2 planar hybridized atoms), defined as

$$E_{op} = f \cdot D^2,$$

where f is an empirical potential parameter, and D is the perpendicular distance from the sp^2 atom to the plane defined by the three atoms bonded to it. One such term is entered for each sp^2 atom bonded to three other atoms.

E_{tor} is a torsional strain energy term, defined as

$$E_{tor} = \sum_n v_n (1 \pm \cos n\phi),$$

where v_1, \dots, v_n are empirical potential parameters, and ϕ is the torsion angle. By choosing appropriate n , the proper form of the potential is selected. An expression like this is entered for each torsion angle.

E_{dd} is an electrostatic interaction energy term, defined as

$$E_{dd} = Kq_i q_j / Dr_{ij},$$

where K is a scaling factor, D is a damping factor analogous to the bulk dielectric constant, and q_i and q_j are the charges on atoms i and j separated by distance r_{ij} . A term is entered for each pair of charged atoms.

Since it is a completely computational technique, molecular mechanics requires no sample and allows the user to introduce calculational tools not possible for real molecules. For example, the introduction of "extra potentials" (13) is particularly useful when molecular conformations other than the minimum energy one must be studied. Extra potentials may be added which make it prohibitively expensive (in energy terms) for a molecule not to assume a structural feature. The minimum energy conformation of the molecule subject to the constraints imposed by the extra potentials may be obtained by minimizing the total energy -- strain energy plus extra potential energy.

$$E_{total} = E_{strain} + E_{extra}$$

By comparing the strain energy of the molecule so minimized with the strain energy of a molecule minimized free of any extra potentials, the energy cost of assuming a constrained conformation may be studied.

Molecular mechanics is not without its complications, some quite serious if one is interested in valid and accurate results. It is a technique where quality results are obtained only if one uses a well designed set of potential parameters and does a quality calculating job. Most molecular mechanics advocates have failed to inform other scientists of the critical need to verify potential parameters for the types of molecules being calculated. The balancing of these empirical parameters is also extremely important -- changing one parameter quite often results in a need to change several others in order to rebalance the set, since the potentials are not orthogonal to each other in parameter space. Molecular mechanics potential sets have been developed to produce accurate results in certain areas, including molecular structure, heats of formation, and vibrational spectra. When calculating molecular structure, it is important to use potential parameters which have been developed to produce accurate molecular structures.

The mathematical nature of the energy minimization calculation leads to calculational problems. There is no known mathematical method for obtaining all possible energy minima in the parameter space of a molecule. There are good mathematical techniques for moving downhill in energy to a minimum, but there is no mathematical way of determining if the minimum reached is global, that is, if it corresponds to the lowest energy conformation. If all low energy conformations of a molecule are to be explored, a trial and error searching process must be used. Such techniques are very costly in computer time because of the combinatorial aspects of the problem. In fairly flexible molecules, the number of combinations which must be explored is so vast, that approximations are frequently made which are so drastic that one must wonder if important conformations have not been missed.

Finally, the nature of the final minimization process must be noted in evaluating the quality of a result from molecular mechanics. Searching is frequently done by holding bond distances and angles fixed and only varying torsional angles, there being no other calculationally feasible way to handle the problem in many cases. It is well known experimentally and theoretically that small distortions in bond distances and especially in bond angles frequently can lower (greatly) the strain energy of a given conformation. Experience is that when atoms are allowed to vary freely (bond distances and angles are allowed to change) conformational energy changes of over 5 kcal/mole are not uncommon, and a different energy ordering of low energy conformations frequently results. Unless a free minimization calculation has been performed, reported molecular mechanics energy values must be regarded with suspicion.

Other molecular mechanics complications result from the nature of the strain energy model. The energy is usually evaluated for the molecule in a vacuum, that is, free of contacts

with other molecules. If the crystal structure is known, interactions in the crystal may be simulated. A reasonably good way of simulating a molecule in solution by molecular mechanics is yet to be developed, however. Also, thermal effects are accounted for only very indirectly in the empirical calculation of strain energy. "Strain energy" as calculated by molecular mechanics is not the free energy of the molecule in a given conformation. Strain energies correspond more closely to enthalpy differences (10). Whereas the enthalpy difference between two conformations of a molecule is significant and important, it is the free energy difference which determines the relative percentage of the two conformers occurring at a given temperature. This is especially important to remember when correlating percentages of conformers determined experimentally with the percentages predicted from molecular mechanics energy differences.

Combining Molecular Mechanics and Crystal Structure Analysis

There are a number of advantages to using a combined molecular mechanics/crystal structure analysis approach to studying three-dimensional molecular structure for drug design purposes. Crystal structure results can be used to verify and develop molecular mechanics potential parameters. Molecular mechanics can be used to test for crystal structure "artifacts", that is to test if a feature observed in a crystal structure is an inherent property of the molecule or if it results from the way in which the molecule is influenced by its neighbors in the crystal. Molecular mechanics results can be used to extrapolate crystal structure results to non-crystalline situations. And perhaps most important, molecular mechanics results are more apt to be valid if backed up by experimental data from crystal structure analysis. If a set of potential parameters can correctly reproduce the experimental results for a molecule in the crystal, this set of potential parameters should also calculate meaningful results for this molecule in other situations. One knows also that the conformation obtained by free minimization from the crystal conformation is a low energy conformation -- it is not always the lowest energy conformation, but is usually within 2-3 kcal/mole of the lowest. This can be extremely valuable information. The energy accessibility of a proposed active conformation may be evaluated by comparing its energy with the energy obtained by free minimization from the crystal conformation, thus alleviating the need to search conformational space for a global minimum energy.

Potential Parameter Verification

As mentioned above, when a set of potential function parameters is to be used in molecular mechanics calculations on

a certain class of molecules, it is important to verify that the parameters work for that class of molecules. The only way to do this is to compare calculated molecular mechanics results with experimental results. For molecules the size of most drugs, the only structural experimental results which are accurate enough for this are crystal structure results. When making such comparisons to verify a set of potential functions, molecules of similar size and chemical type to those being studied must be used. Experimentally it is well known, for example, that the amide group in formamide is a poor model for the amide group in peptides (14). Experience has shown us that molecular mechanics results should be compared with low temperature crystal structure results if at all possible. If this is not done, it is frequently difficult to discern whether differences in calculated and experimental molecular parameters are due to potential parameter deficiencies or thermal or deterioration effects in the crystal. Experience has also shown that for meaningful comparisons, experimental molecular parameters from the crystal must be compared with calculated parameters obtained by minimization in the crystal. In addition, important potential parameters should be checked against more than one crystal structure, preferably several of somewhat different chemical type.

The potential parameters which we employ were developed over a period of years use in calculating large molecule systems, such as drug molecules. For some parameters, values from the literature (10, 15-21, for example) were refined and balanced, in other cases, new values were chosen from study of experimental structural results. The values of all parameters were refined and tested by comparing calculated and observed parameters for large molecules in the crystal. The procedure which we follow for minimizing the molecule in the crystal is to surround an asymmetric unit of molecules in the crystal with symmetry related neighboring atoms out to a certain maximum distance (usually around 10 Angstroms). Then the minimization is carried out by holding the unit cell parameters fixed and varying the atomic positions, while preserving the symmetry of the space group. This simple method produces good structural results if calculation of structural parameters is desired. To save computer time, the first few iterations in the minimization are run with only hydrogen atom positions varying, since hydrogen positions from x-ray analysis are poor and sometimes need sizable adjustment. This simple method will not work if either unit cell parameters are to be varied, or if accurate lattice energies are to be calculated. For these purposes, lattice sums must be evaluated for accurate results (22).

The accuracy to be expected in such verifications is illustrated in Figure 3. Accurate low temperature crystal structure results on morphine monohydrate (8) are compared with results obtained by minimizing morphine monohydrate in the

crystal as described above. Some of the molecular parameters which were particularly hard to fit are tabulated. The agreement which can normally be obtained in such comparisons is summarized in the figure at the lower right.

Using Molecular Mechanics to Test Crystal Structure Features

In the crystal structure of N-acetyl phenylalanyl tyrosine (23) the amide group connecting the phenylalanine and the tyrosine was found to be considerably non-planar. In particular, as illustrated in Figure 4, the C-C-N-C torsion angle across the amide bond was 17.7° out of planar. Minimization in the crystal by molecular mechanics gave a value in good agreement with the observed. When the molecule was minimized free, however, the nonplanarity as indicated by the C-C-N-C torsion angle was reduced to 4.7° , still significant but far less than found in the crystal. Apparently the three hydrogen bonds which occur in the crystal strongly influence the conformation of the molecule in the crystal. In a case such as this, it is important that the molecular mechanics potentials, which are used to test if a feature persists outside of the crystal, be able to correctly reproduce the observed feature by calculation of the molecule in the crystal.

Another example of the use of molecular mechanics to test and interpret features found in crystal structure studies is given by a comparison of 19-nor androstenediol and 7- α -methyl-19-nor androstenediol (1). The introduction of the 7- α -methyl moiety into 19-nor androstenediol was found to increase its potency about 400-fold in laboratory studies.

The crystal structures of the two were determined by x-ray diffraction, and the conformations of the A-rings were found to be significantly different, as shown in Figure 5. Molecular mechanics was used to test whether the difference in observed conformation of these molecules was due to a difference in preferred conformation or due to crystal packing. Molecular mechanics was used to vary the conformation using the C(1)-C(10)-C(5)-C(6) torsion angle as a parameter and calculating the strain energy at each conformation, as illustrated in Figure 6 for one of the molecules. A free minimization was performed at each torsion angle. When the energy was plotted against torsion angle (see Figure 7), the minima for the two molecules were found to occur at the same point. In fact, as Figure 6 shows, the A-ring may vary quite a bit in its orientation with respect to the remainder of the molecule. From this, one may conclude that 1) the effect of the 7- α -methyl on the activity of the molecules is not through any change in inherently preferred conformation, and 2) the 19-nor androstenediol molecule is fairly flexible. As Figure 5 illustrates, the 7- α -methyl group does protrude out from the plane of the molecule, and could definitely influence this flexible molecule to assume

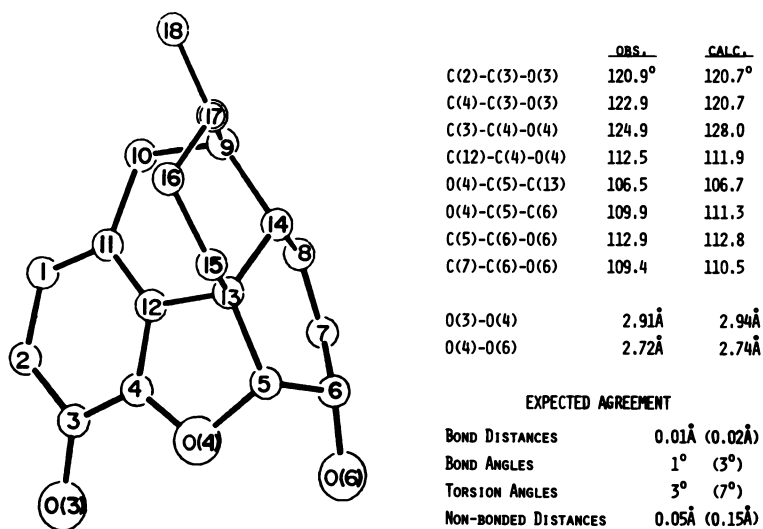


Figure 3. Comparison of calculated and observed values of selected structural parameters for morphine monohydrate in the crystal. Values listed under expected agreement are for minimization in the crystal in general; values in parentheses are maximum observed deviations.

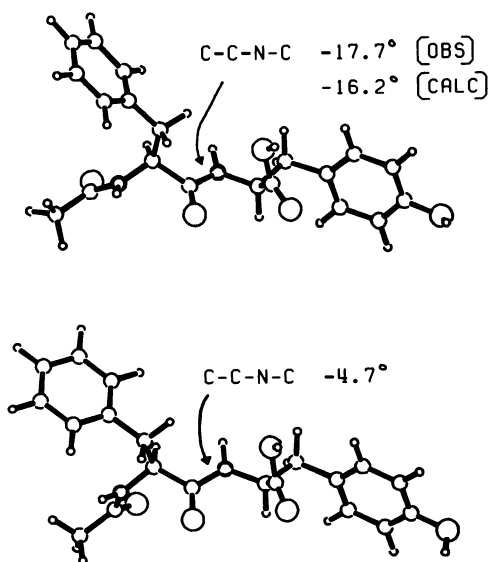


Figure 4. N-acetyl phenylalanyl tyrosine in the crystal (upper) and minimized free (lower)

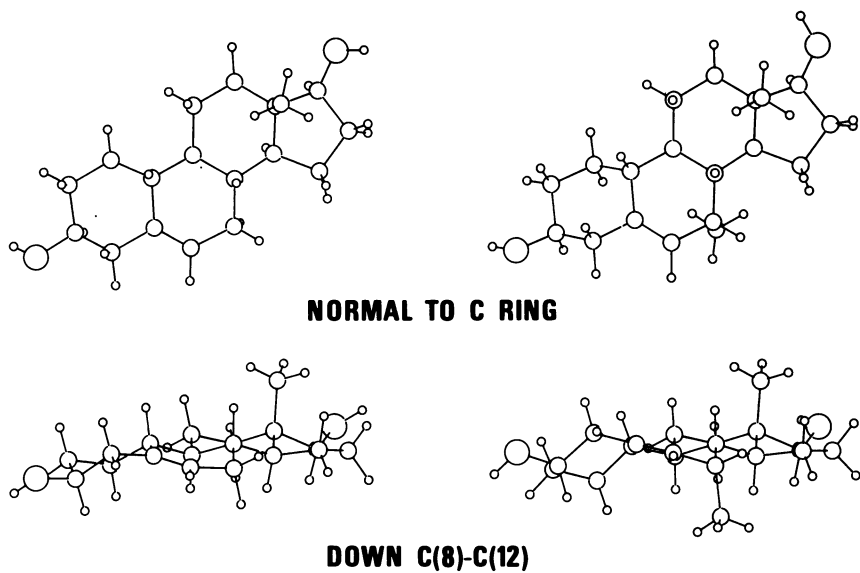


Figure 5. X-ray diffraction crystal structure results on 19-norandrostenediol (left) and 7- α -methyl-19-norandrostenediol (right)

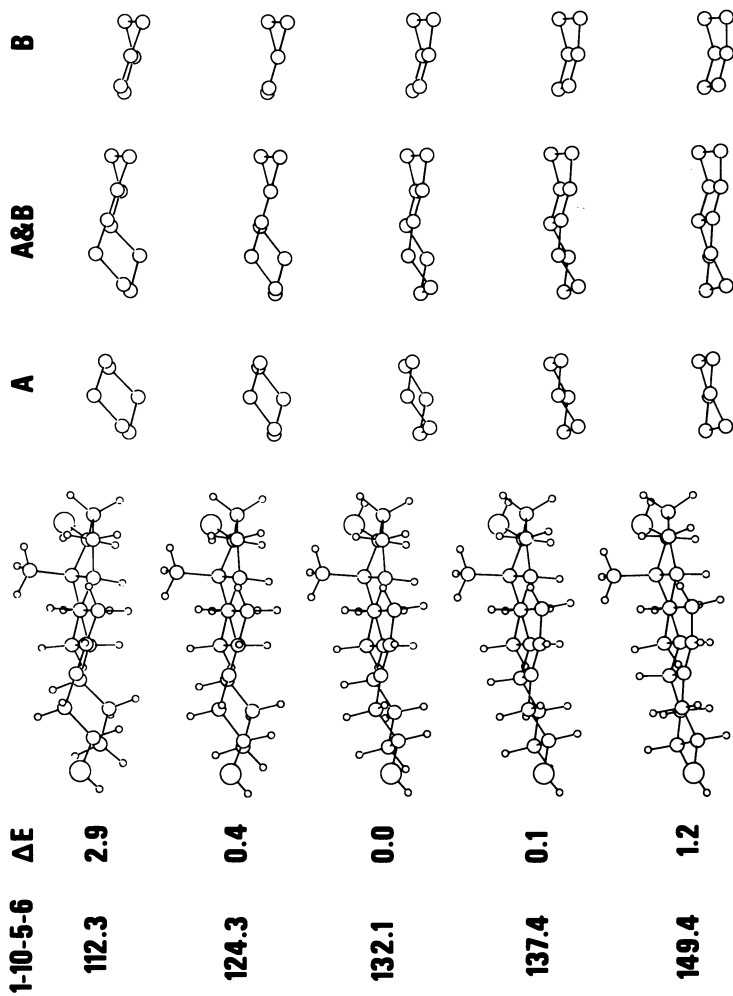


Figure 6. Conformations of the rings of 19-norandrostenediol as the $C(1)-C(10)-C(5)-C(6)$ torsion angle is varied

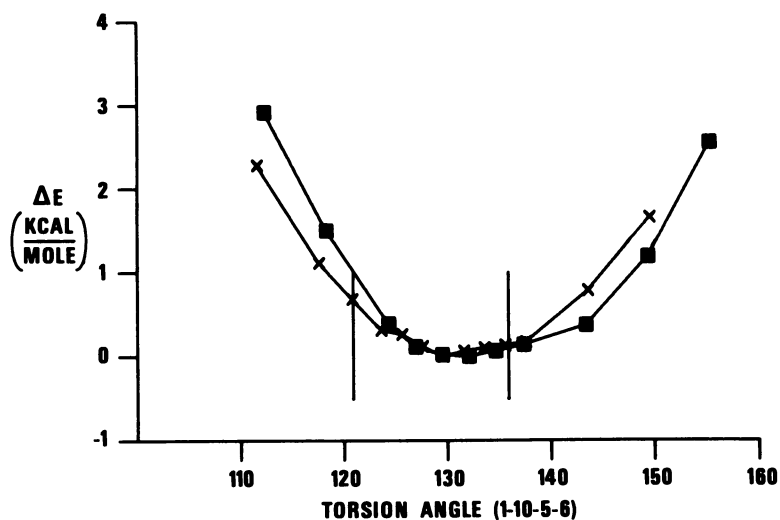


Figure 7. Calculated strain energy vs. C(1)—C(10)—C(5)—C(6) torsion angle in 19-norandrostenediol (■, R=H) and 7- α -methyl-19-norandrostenediol (×, R=CH₃)

a different conformation when it sits on a receptor, just as it ceded it to assume a different conformation in the crystal. This investigation has obviously not ruled out metabolism or transport effects which may be very important in this case.

Examples From the Study of Analgesics

Some examples from our recent study of opiate analgesic structures (8,24) serve to illustrate further the types of information which a combined approach can provide. The approach used in this study was 1) accumulate accurate crystal structure results on representative compounds by literature search and performing crystal structure determinations, then 2) develop and verify molecular mechanics potential parameters for use with analgesics, and 3) perform strain energy calculations to find active conformations by comparing different chemical structural types which act at a common receptor. The object of the study was to 1) better define the three-dimensional requirements of opiate analgesics, and 2) better understand the opiate receptor itself by indirectly investigating it.

Although the morphine molecule is usually considered very rigid because of its interlocking rings, molecular model studies had hinted that there might be some flexibility, and that this flexibility might be characterized by rotation about the C(12)-C(13) bond (see Figure 3). To study this flexibility by molecular mechanics, we used an "extra potential" to cause the C(4)-C(12)-C(13)-C(14) torsion angle to be constrained to various values and did a complete minimization at each constrained value. The resulting potential well is plotted in Figure 8. It is evident from this plot that there is a considerable amount of flexibility in the morphine molecule. The C(4)-C(12)-C(13)-C(14) torsion angle may be varied over a 40° range without introducing excessive strain in the molecule. As the drawings above the plot in Figure 8 illustrate, this is an important flexibility since it strongly influences the three-dimensional relationship between the basic nitrogen and the aromatic ring, greatly changing the distance of the basic nitrogen from the plane of the aromatic ring. The conclusions of interest here are that morphine has some flexibility, and when compared to other molecules, its atomic positions must be varied and should not be considered rigid.

The fentanyl molecule has considerable flexibility. In order to study the conformation of fentanyl at the receptor, we matched its conformation to that of morphine which is considerably less flexible. To perform the calculation, morphine and the test molecule are overlaid (in the computer) and connected at four points by extra potentials as shown in Figure 9. Three potentials align the aromatic ring, and one matches the position of the basic nitrogen. The gross alignment of the two molecules is done on our computer graphics system, including any large

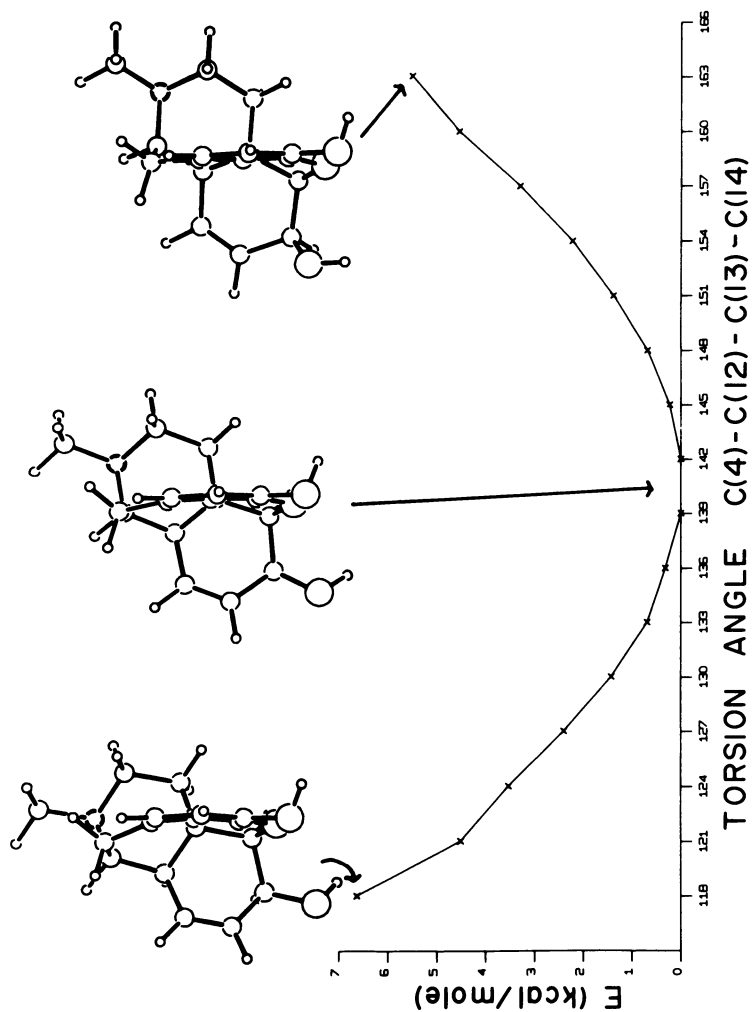


Figure 8. Flexibility of morphine to rotation about the C(12)—C(13) bond

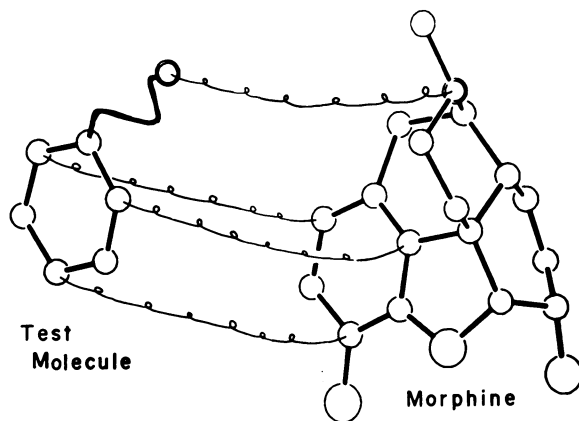


Figure 9. Method used to link morphine to other opiate molecules for conformational matching. The springs represent "extra potentials" of the form shown at the lower left ($E = cd^2$)

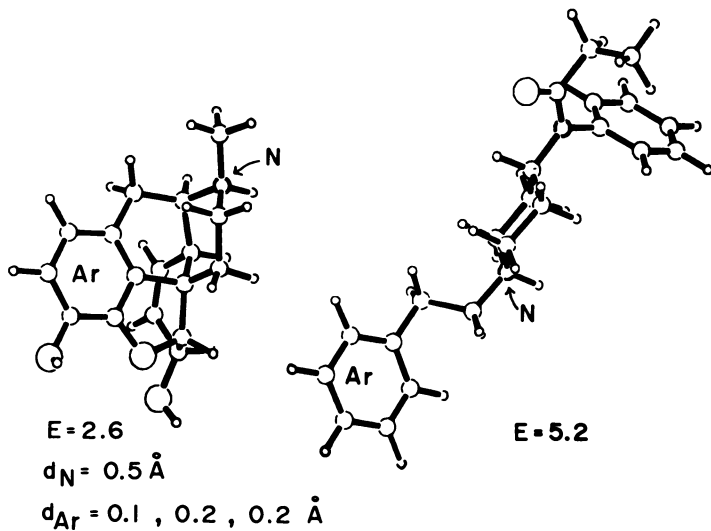


Figure 10. Best fit from conformational matching of morphine and fentanyl

changes in molecular conformation. Various ways of matching the two together were tried, including matching both aromatic rings of fentanyl to the morphine aromatic ring. For each match, the two molecules are minimized simultaneously with all x,y,z coordinates of all atoms being varied. The energy minimized is the sum of the strain energy of both molecules and the extra potential energy. The results of this experiment are illustrated in Figure 10. The energies shown are the differences between the calculated strain energies for the molecule minimized together as described above, and the strain energy obtained when the molecule is minimized free of all intermolecular interactions and extra potentials. The distances are the final lengths of the "springs" shown in Figure 10. Both the energies and distances indicate a good fit. The values obtained for the energies and distances must be interpreted somewhat qualitatively, since by changing the constants in the extra potentials, one can go to higher energies and shorter distances or lower energies and longer distances. The conclusion is that with fentanyl in this conformation, a reasonably good fit can be obtained at a reasonable energy, an energy well within the range that might be expected to be present in a drug-receptor interaction. The various conformations of fentanyl which have been discussed are shown in Figure 11. Note that the phenethyl group is rotated differently in the matched conformation from both the conformation obtained by free minimization and the conformation observed in the crystal.

When we performed a crystal structure analysis on meperidine hydrochloride, two symmetry independent molecules were found in the symmetry independent unit. The two molecules differ conformationally in the orientation of the ethyl group. In the structure of meperidine hydrobromide reported by Van Konigsveld (25), the ethyl group is found in yet a third conformation. Molecular mechanics calculations verified that there is no significant energy difference between the three orientations of the ethyl group. In all three observed molecules, the phenyl is equatorial to the 6-membered ring. It had earlier been suggested, however, that an axial phenyl orientation might be the active form (26). When minimized as a free molecule, the equatorial phenyl form is found to be favored by 0.8 kcal/mole over the axial phenyl form, a smaller energy difference than one might expect.

Both the axial phenyl and the equatorial phenyl conformers were matched to morphine using the 4-point matching technique described above. The results are illustrated in Figure 12 with the equatorial match on the left and the axial on the right. Both can be made to fit. The fit for the axial is somewhat better. The meperidine strain energy for the axial match is 3.6 kcal/mole, whereas 6.9 kcal/mole was required for a similar match in the equatorial case. Both energies are differences from the lowest energy conformer which is equatorial. Although

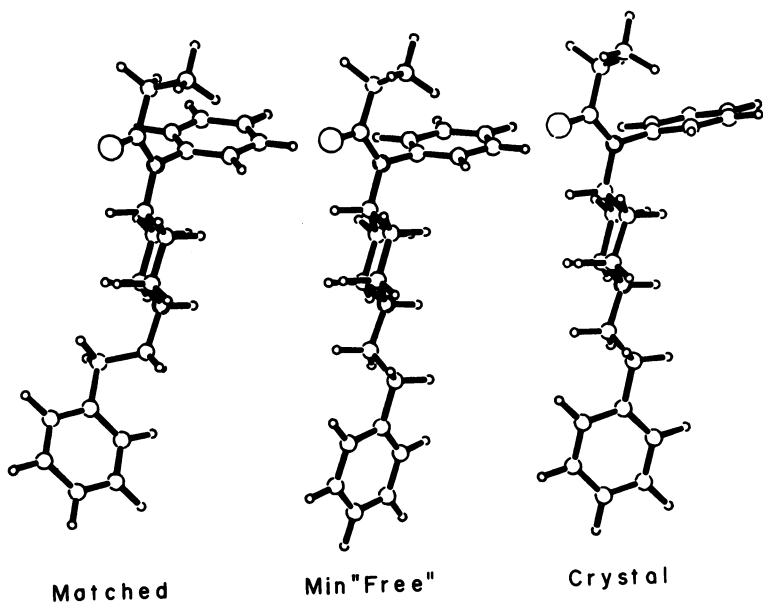


Figure 11. Three conformations of fentanyl

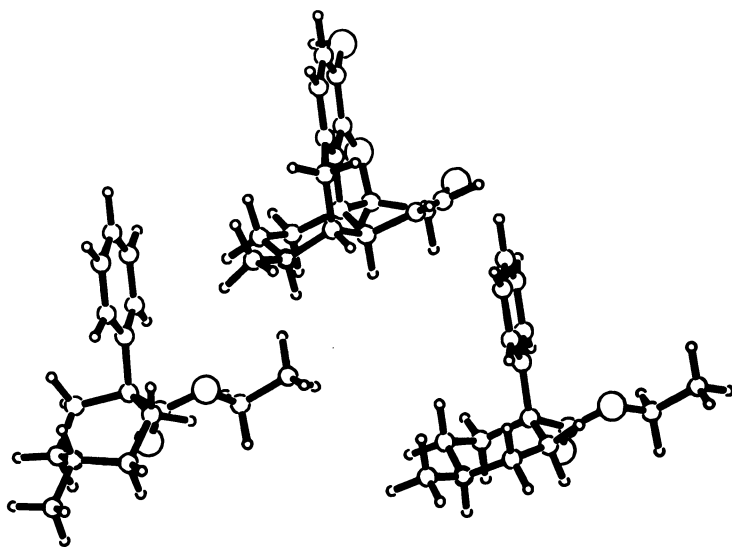


Figure 12. Best fit conformations from matching morphine (top) with meperidine equatorial (left) and axial (right) phenyl conformers

these data point to the axial conformer as being a better match, the conclusive point appears to be the orientation of the proton on the basic nitrogen. After matching, the proton on this nitrogen is oriented in the same way as that of morphine in the axial case, but in an unacceptable direction in the equatorial phenyl case.

Building and Testing a Model for Receptor Requirements

If structural information is available on a number of molecules which are known to act at a given receptor, a model for the requirements of the receptor may be developed. Different methods have been used for this; the one we use involves exact quantitative conformation matching using molecular mechanics. Extra potentials can be used to match conformations of more than two molecules at a time by simultaneous minimization. In our analgesic work, we frequently have matched morphine to two or three molecules simultaneously, and have introduced more extra potentials to extend the fitting requirements beyond the four points mentioned above. The principal disadvantage to matching many molecules simultaneously is that the computer time required goes up faster than the number of molecules. The technique, and our computer programs, allow a different approach to simplify the calculations, however. Once the simultaneous minimization of a number of representative molecules has achieved a good match at a number of points, the three-dimensional coordinates of these points constitute a model, which other molecules acting at the same receptor may be tested against. To assess the fit of a subsequent molecule to this model, it is then sufficient to minimize the molecule subject to the constraints imposed by extra potentials linking sites in the molecule to the points defining the model; the points in the model being held fixed during the minimizing process. This calculation is roughly equivalent in computer time to minimizing a single molecule.

Computer graphics is essential in this type of work. Initial matches of molecular conformations are found by computer graphics, reducing the amount of conformational searching needed. By varying the conformation of molecules via interactive graphics, the chemist can use his intuition to greatly reduce the amount of computer calculations. Other papers in this symposium expound further on the utility of this valuable tool in drug design.

Conclusions

The combination of crystal structure analysis and molecular mechanics is a valuable tool in drug design, since it can be used to provide three-dimensional information which can greatly help the medicinal chemist design new molecules with desirable

activities. Crystal structure results are not very hard or very expensive to obtain. The cost of determining a crystal structure has declined greatly in the past few years; and the literature abounds with useful crystal structure results. With the proper computer facilities, molecular mechanics calculations are easy to set up and run. Crystal structure results can be used to provide the experimental verification necessary for insuring that molecular mechanics results have validity in the molecules being calculated. Special computational techniques can be used to explore non-minimum energy conformations and to conformationally match molecules. The approach can be extended to allow the building and testing of a model for receptor requirements provided structural information is available on a number of molecules which interact at the receptor.

Advances in computer technology are reducing the cost of computers and are making this process less expensive. New, less expensive and functionally improved interactive graphics terminals are being introduced, including terminals with fairly high resolution color graphics. The trend toward less expensive computing and graphics will reduce the cost of this quantitative model building technique to the point that all medicinal chemists can have easy access to such devices. Indeed, pharmaceutical companies and other large research facilities will not be able to afford not to provide their chemists with such facilities.

Just as with other drug design tools, the study of three-dimensional structure as described in this paper does not give a complete picture of a drug molecule and its actions. Electronic effects are studied only in a peripheral way here; quantum mechanics calculations, as described by Christoffersen in the first paper of this symposium, can be used to obtain further valuable information. Neither of the above, however, give much direct information about the transport of drugs to the site of action. Drug design is a very complicated process, necessitating the use of a number of tools, each having strengths in certain areas. Crystal structure analysis and molecular mechanics occupy a prominent place among the tools of drug design.

LITERATURE CITED

1. Duchamp, D.J. and Campbell, J.A., *Acta Cryst.*, (1972), A28, S50.
2. Kennard, O. and Watson, D.G., "Molecular Structures and Dimensions", Volumes 1-9, Crystallographic Data Centre, Cambridge, England (1970-1979).
3. Duax, W.L., and Norton, D.A., "Atlas of Steroid Structure", (1975), New York, IFI/Plenum, (Volume 2, in press).
4. Shoemaker, D.P., in: "Critical Evaluation of Chemical and Physical Structural Information", Lide, D.R. and Paul, M.A. ed., Washington, National Academy of Sciences, 1974, p. 157-185.
5. Ibers, J.A., *ibid.*, p. 186-198.
6. Donahue, J., *ibid.*, p. 199-218.
7. Stout, G.H. and Jensen, L.H., "X-ray Structure Determination", Macmillan, New York (1968).
8. Duchamp, D.J., Chidester, C.G., Olson, E.C., *Amer. Cryst. Assoc. Abstr. Papers* (summer meeting), (1977), 5, 83.
9. Kitaigorodsky, A.I., "Molecular Crystals and Molecules", Chapter VII, Academic Press, New York (1973).
10. Engler, E.M., Andose, J.D., and von Schleyer, R.P., *JACS*, (1973), 95, 8005.
11. Schroeder, R. and Lippincott, E.R., *J. Phys. Chem.* (1957), 61, 921.
12. Duchamp, D.J., Cheney, B.V., unpublished work.
13. Duchamp, D.J., in: "Algorithms for Chemical Computations", Christoffersen, R., ed. Washington: ACS, 1977, p. 98-121.
14. For peptide amide parameters, see, for example, Marsh, R.E. and Donohue, J., *Adv. Protein Chem.* (1967), 22, 235.
15. Moelwyn-Hughes, E.A., "Physical Chemistry", 2nd Revised Edition, Pergamon Press, New York, (1961), p. 518-519.
16. Williams, D.E., *Acta Cryst.*, (1974), A30, 71.

17. Ermer, O. and Lifson, S., *JACS*, (1973), 95, 4121.
18. Govers, H.A.J., *Acta Cryst.*, (1975), A31, 380.
19. Bates, J.R. and Busing, W.R., *J. Chem. Phys.*, (1974), 60, 2414.
20. Momany, F.A., Carruthers, L.M., McGuire, R.F., and Scheraga, H.A., *J. Phys. Chem.*, (1974), 78, 1595. (And previous papers from H.A. Scheraga.)
21. Allinger, N.L. and Sprague, J.T., *JACS*, (1973), 95, 3893. (and previous papers from N.L. Allinger).
22. Williams, D.E., *Trans. Amer. Cryst. Assoc.*, (1970), 6, 21.
23. Stenkamp, R.E. and Jensen, L.H., *Acta Cryst.*, (1973), B29, 2872.
24. Cheney, B.V., Duchamp, D.J., and Christoffersen, R.E., in: "Quantitative Structure Activity Relationships of Analgesics, Narcotic Antagonists, and Hallucinogens", Barnett, G., Trsic, M., and Willett, R.E., ed., NIDA Research Monograph 22, Washington, (1978), p. 218-249.
25. Von Koningsveld, H., *Rec. Trav. Chim.*, (1970), 89, 375.
26. Beckett, A. and Casey, A., *Prog. Med. Chem.*, (1965), 4, 171.

RECEIVED June 29, 1979.

Studies of Chemical Structure-Biological Activity Relations Using Pattern Recognition

P. C. JURIS, J. T. CHOU, and M. YUAN¹

Department of Chemistry, 152 Davey Laboratory, The Pennsylvania State University, University Park, PA 16802

The attempt to rationalize the connection between the molecular structures of organic compounds and their biological activities comprises the field of structure-activity relations (SAR) studies. Correlations between structure and activity are important for the development of pharmacological agents, herbicides, pesticides, and chemical communicants (olfactory and gustatory stimulants) and for the investigation of chemical toxicity and mutagenic and carcinogenic potential. Practical importance attaches to these studies because the results can be used to predict the activity of untested compounds. In addition SAR studies can direct the researcher's attention to molecular features that correlate highly with biological activity, thus confirming or suggesting mechanisms or further experiments. SAR studies have been used to some extent in the pharmaceutical and agricultural industries. The methods are beginning to be applied to the important problems of chemical toxicity and chemical mutagenesis and carcinogenesis.

The superior way to develop predictive capability is to understand, at the molecular level, the mechanisms that lead to the biological activity of interest. Unfortunately, this knowledge is not yet available for most classes of biologically active compounds. Furthermore, the progress made through a living system by an active compound or its precursors is not usually known. Thus, two choices are presented: study the mechanisms for a very few compounds to develop fundamental information for those few compounds, or use empirical methods to study larger sets of compounds with correlative methods. The latter method comprises an SAR approach to the problem. Thus, one has available a set of compounds that have been tested in a standard bioassay and the observations that resulted from the tests. One can then search for correlations between the structures of the compounds tested and the biological observations reported. One is actually

¹ Current address: Stauffer Chemical Co., 1200 S. 47th Street, Richmond, CA 94804.

0-8412-0521-3/79/47-112-103\$06.75/0

© 1979 American Chemical Society

modelling the entire process of uptake, transport, distribution, metabolism, cell penetration, binding, excretion, etc.

The discovery and design of biologically active compounds (drug design) is a field that has been subject to widespread and well-documented (1-8) changes in the past decade. A host of new techniques and perspectives have evolved. While these techniques have been used largely for the development of pharmaceuticals, they can also be applied to the rationalization of structure-activity relations among sets of toxic, mutagenic, or carcinogenic compounds.

Several approaches to SAR have been reported: the semi-empirical linear free energy (LFER) or extrathermodynamic model proposed by Hansch and coworkers (9,10,11), the additivity or Free-Wilson model (12); quantum mechanically based models (13,14) and pattern recognition methods (8,15). Reviews are cited that describe the progress made using each of the approaches.

Chemical and SAR Applications of Pattern Recognition. Pattern recognition is a subfield of artificial intelligence comprise of a set of nonparametric techniques used to study data sets that may not conform to well-characterized probability density functions. A voluminous literature describes the field (16,17,18,19).

Most of the pattern recognition methods share a set of common properties. The data to be analyzed, here molecular structures of carcinogens and noncarcinogens of interest, are represented by points in a high dimensional space. For a given compound, which is represented by a given point, the value of each coordinate is just the numerical value for one of the molecular structure descriptors comprising the representation. The expectation is that the points representing carcinogenic compounds will cluster in one limited region of the space, while the points representing the noncarcinogenic compounds will cluster elsewhere. Pattern recognition consists of a set of methods for investigating data represented in this manner to assess the degree of clustering and general structure of the data space.

Parametric methods of pattern recognition attempt to find classification surfaces or clustering definitions based on statistical properties of the members of one or both classes of points. For example, Bayesian classification surfaces are developed using the mean vectors for the members of the classes and the covariance matrices for the classes. If the statistical properties cannot be calculated or estimated, then nonparametric methods are used. Nonparametric methods attempt to find clustering definitions or classification surfaces by using the data themselves directly, without computing mean vectors, covariance matrices, etc. Examples of nonparametric methods would include error-correction feedback linear learning machines (threshold logic units or perceptrons) and simplex optimization methods of searching for separating classification surfaces. Complete descriptions of these methods can be found in standard texts (19).

Applications of pattern recognition methodology to chemical problems were first reported in the 1960's (20,21) with studies of mass spectra. Since then papers have described work in a variety of areas (22,23) including mass spectrometry, infrared spectroscopy, NMR spectroscopy, electrochemistry, materials science and mixture analysis, and the modeling of chemical experiments. Diagnosis of pathological conditions from sets of measurements made on complex biological mixtures, e.g., serum, have been reported (24). The successes in these areas have led to the belief that these methods should prove useful in the development of structure-activity relations.

A number of studies of the application of pattern recognition to the problem of searching for correlations between molecular structure and biological activity have been reported. A large fraction of the effort in this area must be devoted to the generation of appropriate descriptors from the molecular structures available. Areas of study include drug structure-activity relations, studies of chemical communicants, etc. Applications of pattern recognition to drug design have been reviewed by Kirschner and Kowalski (15).

Recently, papers have begun to appear reporting the applications of SAR methods to sets of chemical carcinogens. One study reports correlations between the number of carbon atoms per compound and carcinogenic activity for a set of forty-seven carcinogenic nitrosamines (25). Correlations of carcinogenic activity and liposolubility for small sets of cyclic nitrosamines have been reported (26). Carcinogenic activity was correlated with the theoretical reactivity indices for twenty-five representative polycyclic aromatic hydrocarbons (27). A quantitative relationship was reported that correlated carcinogenic activity with the hydrophobic parameter π and the Taft polar substituent constant for twenty-one nitrosamines (28). Correlations were reported between nitrosamine carcinogenic activity and molecular connectivity indices (29). Another study of cyclic nitrosamines found correlations between carcinogenic potency and aqueous and octanol solvation free energies (30). A paper has appeared describing the application of the SIMCA method of pattern recognition to the SAR study of 4-nitroquinoline 1-oxides (31). Hansch and Fujita derived an equation relating the logarithm of carcinogenicity of 47 polycyclic aromatic hydrocarbons with the logarithm of the 1-octanol/water partition coefficient, the total charge of the K-region (32). Franke correlated carcinogenic activity, characterized by the hydrophobic binding p_f of PAH's to human serum albumin, to a chemical reactivity index for 34 compounds (33). Jurš and coworkers used pattern recognition methods to study a set of 209 diverse chemical carcinogens (34).

Experimental

The SAR studies reported here were done using the ADAPT computer software system (35,36). ADAPT has been implemented on the Department of Chemistry MODCOMP II/25 16-bit digital computer in FORTRAN. The fundamental steps involved in performing an SAR study using this system are shown in Figure 1. The individual steps are implemented in a series of independent programs that can be executed individually. They perform the mutually supportive tasks necessary for an SAR study as follows:

- (1) Entry and storage of molecular structures and substructures.
- (2) Entry and storage of a list comprising those structures to be considered as the current data set of interest.
- (3) Generation of three dimensional molecular models of the structures using molecular mechanics.
- (4) Generation and storage of molecular structure descriptors from the topological or geometrical representations of the structures.
- (5) Gathering together of a common set of descriptors for each compound under investigation to prepare a data set for pattern recognition analysis.
- (6) Analysis of the data for correlation, separability, etc. using techniques drawn from statistics, nonparametric statistics, and pattern recognition. Testing of discriminants for predictive ability.
- (7) Identification of which of the currently active descriptors are contributing to solution of the problem and which are unimportant enough to discard; that is, feature selection.

Each of these tasks will be discussed in more detail in the following paragraphs.

Entry of Structures and Substructures. The main data input to the ADAPT system is chemical structures along with the relevant biological or physicochemical property. Structures are entered by sketching on the CRT screen of a graphics display terminal. The information contained in the sketch is converted into a compressed connection table representing the atom types, the bond types, and the bonding pattern, i.e., the topology. Rings are automatically perceived and labelled. The user supplies pertinent stereochemical information during sketching. All the topological information, along with additional information such as the name of the compound, associated numerical information, etc. are saved on ADAPT disc files cataloged under an accession number for each compound.

Routines allow the retrieval for review of any structure, the deletion of any structures, the alteration of any structure. The system is configured on the MODCOMP II/25 so that one cartridge disc will hold one thousand structures in addition to all other ADAPT disc files for other storage.

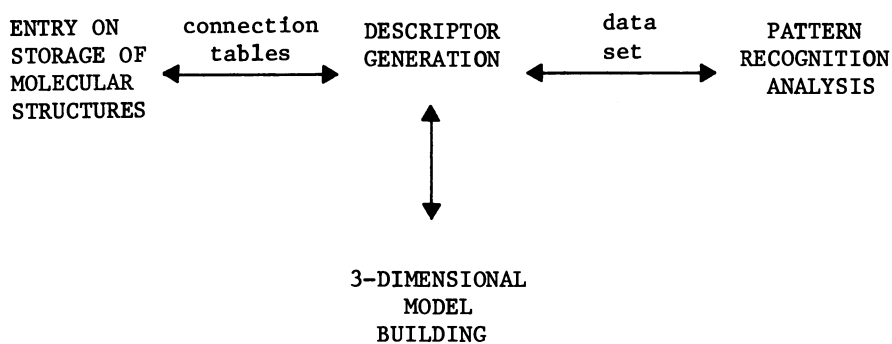


Figure 1. Flow chart of steps involved in structure-activity studies using chemical structure information handling and pattern recognition techniques

Substructures are entered into the system using the same procedures, but they are stored on special, separate disc files. They are used by the substructure searching routine during descriptor generation.

Entry and Storage of List of Structures for Study. The ADAPT disc files can store more structures than can be studied at once, so each SAR study must have its own list of structures of interest. This list is called the worklist. It comprises the set of structures for which descriptors are to be generated and stored.

Molecular Mechanics Model Building. A three-dimensional molecular model builder routine interfaced to the ADAPT system (MOLMEC) is used to derive information on the spacial conformation of molecules. The molecules are viewed as collections of particles held together by simple harmonic or elastic forces. A strain energy can be defined for a molecule which is the sum of contributions for each of several items:

$$E_{\text{strain}} = E_{\text{bond}} + E_{\text{angle}} + E_{\text{torsion}} + E_{\text{nonbond}} + E_{\text{stereo}}$$

The first four terms of the function are commonly found in molecular mechanics strain energy functions, and they are modified Hooke's law functions. The last term has been added to insure the proper stereochemistry about asymmetric atoms. A model is refined by minimizing the highly nonlinear strain energy function with respect to the atomic coordinates. An adaptive pattern search routine is used for the strain energy minimization because it does not require analytical derivatives. The time necessary to obtain good molecular models depends on the number of atoms in the molecule, the flexibility of the structure, and the quality of the starting model.

The molecular mechanics routine is highly interactive and relies on graphical input and output. A molecule being modelled can be displayed in any desired orientation for inspection by the user. If such inspection shows that the model is trapped in a local minimum of the strain energy, then interactive routines allow the user to alter the model to bring it out of the minimum.

An automatic version of MOLMEC has been implemented so that large sets of molecules can be modelled without continuous supervision. The routine is set up so that it can utilize computer time that would otherwise go unused, and so that it can be interrupted and restarted in order to allow modelling to continue for long periods.

The purpose of the molecular modelling is to generate models of sufficient quality to allow the computation of geometric descriptors, allow realistic perspective drawings of molecular

structures, and to provide geometries for input to molecular orbital calculations.

Descriptor Generation. Topological descriptors are generated from the connection tables of the structures. Each descriptor development routine is independent, fetching the structures of the compounds being coded from the disc files, generating the particular descriptor, and storing the results on other disc files.

Fragment Descriptors: Fragment descriptors include the atom type counts, bond type counts, the molecular weight of the compound, the number of ring atoms, and the number of basis rings in the compound. These simple descriptors provide fundamental information regarding the structures being represented.

Substructure Descriptors: This routine generates descriptors using a generalized substructure search algorithm. The substructure(s) to be searched are entered into the substructure storage area using the sketching input described above. Their accession numbers are supplied to the substructure descriptor routine by the user at execution time. The descriptor value for a compound is the number of times the substructure is found imbedded within that structure. A variety of options may be invoked by the user, such as requiring ring atoms to match ring atoms, using relaxed criteria of connectivity for matching, or finding all substructure occurrences or only unique, nonoverlapping occurrences.

Molecular Connectivity Descriptors: The molecular connectivity index of a molecule is a measure of the branching of the structure. The path one molecular connectivity index is formed by summing contributions for each bond in the structure, where the contribution for each bond is determined by the connectivity of the atoms that are joined by that bond. Higher order molecular connectivities can also be computed by considering all paths of length two, three, etc. These descriptors have been shown in several published reports to be correlated with a number of physicochemical parameters, such as partition coefficients and steric parameters (37,38,39).

Environment Descriptors: The immediate surroundings of a substructure, as it is imbedded within a structure, is coded by the environment descriptor. Once a substructure is found within the structure being coded, then a pseudomolecule is formed from the substructure match atoms of the structure and their nearest neighbor atoms. Then a path one molecular connectivity calculation is done over this pseudomolecule to form the environment descriptor. Thus, a value is obtained that indicates not only the presence of the substructure, but also its surroundings in the molecule.

Path and Path Environment Descriptors: Molecular structures can be viewed in the abstract as graphs, in the mathematical sense. Graphs have a number of properties that can be readily

computed, among them paths. A path is a sequence of connected atoms in which no atom appears more than once. The tracing of paths in highly connected graphs, e.g., polycyclic aromatic hydrocarbons, is tedious but can be done. A method for finding all the paths contained in a graph has been published, with a discussion of the chemical utility of this measure of bonding complexity (40). A descriptor generation routine has been implemented and interfaced to ADAPT using this concept. The routine calculates the total number of paths of all lengths contained within the structure being coded.

The path environment descriptor is similar to the molecular connectivity environment descriptor. A substructure is matched against the structure being coded. If it is present, then a pseudomolecule is formed from the substructure atoms and their first nearest neighbor atoms. Then the number of paths originating from the set of atoms forming the pseudomolecule is found. The maximum path length to be considered is input by the user. The descriptor for a particular structure consists of the number of paths originating from the substructure atoms and their first nearest neighbors.

This environment descriptor can be made sensitive to the immediate surroundings of the substructure by making the maximum path length to be considered very small, e.g., two or three. Alternatively, it can be made sensitive to the overall structure of the molecule by making the maximum path length larger.

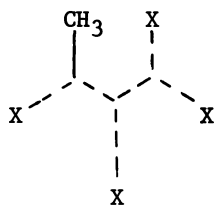
Sigma Charge Calculation: As a first attempt to provide the system with an electronic descriptor, a sigma charge calculation routine has been implemented following the discussion of the Del Re method by Hopfinger (41). The Del Re method (42) is a simple calculation whereby bond charges are calculated in a Hückel formalism and are summed about each atom to give the total sigma charge for each atom. Each bond orbital is described as a linear combination of atomic orbitals.

This sigma charge calculation program has been interfaced with the substructure searching routine to develop the sigma charge descriptor. The routine searches for a specified substructure in a molecule, and, if present, the sigma charge calculation is done for the entire molecule. The descriptor is then generated in one of two ways. In order to investigate the sigma charge of a particular region of a set of molecules, the descriptor can be the average sigma charge of the atoms comprising the substructure used for searching. If the substructure occurs more than once within the structure, the descriptor value is the average sigma charge over all occurrences of the substructure. In its alternate forms the descriptor can code the sigma charge on the most positive atom within the substructure or the sigma charge on a specified substructure atom.

In the event that the Del Re sigma charge parameters are not available for a particular bond in the molecule under consideration, or, if the molecule under consideration is too large for

the program to handle, the weighted average of the active and nonactive classes is assigned to that particular compound and the user is notified by a message.

A specific example will clarify the procedure. When the sigma charge descriptor is generated using the following sub-structure with four differently substituted 12-methyl benz(a)-anthracenes, the differences in electron withdrawing power of the 7 position substituents are reflected in the descriptor value. The dashed lines represent aromatic bonds, and the terminal X's represent atoms of indetermined type and indetermine connectivity.



The four compounds and their average sigma charges on the sub-structure are as follows: (a) 7,12-dimethylbenz(a)anthracene 0.00433; (b) 7-formyl-12-methylbenz(a)anthracene 0.00604; (c) 7-cyano-12-methylbenz(a)anthracene 0.0176; and (d) 7-fluoro-12-methylbenz(a)anthracene 0.0754.

Despite the fact that this semi-empirical technique deals only with the sigma electron contribution and was developed for the study of saturated organic molecules, the descriptors developed by the routine have found utility in the study of polycyclic aromatic hydrocarbons, as will be shown.

Partition Coefficient Estimation: The log P, the logarithm and a model lipid phase, is highly correlated with various types of biological activities. Its inclusion as a descriptor allows the system to have access to information regarding transport and partitioning that may not be as clearly represented in the other descriptors. The calculation of log P can be done using a "constructionist approach" due to Leo, Hansch and coworkers (43, 44). This formulation starts with a set of fundamental fragments whose values are summed with appropriate weightings. Corrections are added to refine the calculation if necessary. The equation used is as follows:

$$\log P = \sum_{i=1}^n a_i f_i + \sum_{j=1}^m c_j$$

where f_i is the fragment constant for the i th fragment, a_i is the number of occurrences of the i th fragment, and c_j is the j th correction factor. The fundamental assumption of this approach is that hydrophobicity is an additive-constitutive property of molecules.

Recently, this fragment constant method has been implemented in a routine and interfaced to ADAPT (45,46). Thus, the estimated log P value for a molecule can be used as a descriptor along with all the other descriptor types.

Geometric Descriptors: Once the set of molecules being studied have been modelled, then descriptors can be generated from these models. Geometrical descriptors that have been used in many structure-activity studies are the three principal moments of inertia of the molecules, the three ratios of these three principal moments, and the molecular volume. Many other geometrical descriptors can be envisioned, and the development of new descriptors is an area of active interest.

Descriptor Manipulation Routines: In addition to the individual descriptor generation routine, the ADAPT system contains several support routines. There is a general maintenance and interrogation routine for the descriptor files that allows the user to review descriptor values by listing them, to delete unwanted descriptors, etc. Another routine allows the user to input descriptors obtained from any source from cards so that they can be studied along with the computer generated descriptors. A manipulation routine allows the user to alter previously generated or stored descriptors to form new descriptors. Addition, subtraction, multiplication, division, exponentiation, logarithmic transformation, autoscaling, inversion, etc. are allowed.

Gathering of Common Set of Descriptors. The descriptor files of ADAPT typically hold many more descriptors for a set of compounds of interest than can be studied at one time. Therefore, a routine allows the user to select which of the available descriptors to form into a data set for analysis. This allows the user to conveniently study different sets of descriptors for the same set of compounds. When a data set is generated, the descriptors are automatically autoscaled in order to give each descriptor equal weight in the subsequent analysis.

Pattern Recognition Analysis. A set of interactive pattern recognition programs have been interfaced to the ADAPT system. The same data set can be studied using a variety of techniques for comparative purposes. The analysis routines are all interactive.

The set of transformed, scaled, and normalized data for the i th compound is expressed as a pattern vector of the form:

$$X_i = (x_1, x_2, \dots, x_j, \dots, x_n, x_{n+1})$$

where n is the number of descriptors per compound. The $(n+1)$ th component of the pattern vector is a constant added for computational convenience. The set of observations belong to two classes, and the fundamental purpose of each of the analysis routines is to find a discriminant that will separate the two classes. For linear discriminants this is equivalent to searching for a weight vector, W , such that the dot product of X_i with W

$$s_i = W \cdot X_i$$

will be positive for all members of one class and negative for all members of the other class.

Logically, the analysis part of the ADAPT system breaks down into several subsections: utility routines, parametric pattern recognition methods, nonparametric pattern recognition methods, and feature selection routines. Each is discussed in the following paragraphs.

A training set/prediction set generation routine allows the user to enter into disc files lists of which patterns in the current data set are to comprise a training set and which patterns are to comprise a prediction set. A number of options exist whereby the user can generate training set/prediction set combinations using random selection of members of the data set, random selection with some of the patterns excluded from consideration, training sets consisting of all the patterns but k where k is a user supplied number, etc. The system is configured to be able to store up to 100 training set/prediction set combinations. The resulting lists of training set members and prediction set members are accessed by the pattern recognition routines if directed to do so.

A weight vector utility routine allows the user to enter a weight vector into the appropriate disc storage area so that it can be used to initialize any of that pattern recognition programs. The routine can also output any stored weight vector. Any of the linear discriminant development routines can store a trained weight vector in the same weight vector storage area. Thus, the weight vector resulting from execution of one pattern recognition program can be used to initialize another of the programs. This feature has proved to be very useful in many studies.

Parametric Pattern Recognition Programs: The parametric pattern recognition programs use the mean vectors and covariance matrices (or other saturated measures) of the two classes into which the patterns fall as their basis for development of discriminants.

One parametric routine implements a quadratic discriminant function using the Bayesian theorem. The equation for the discriminants are as follows:

$$d_k(X) = \ln p_k - \frac{1}{2} \ln |C_k| - \frac{1}{2} [(X-m_k)^T C_k^{-1} (X-m_k)] \quad k=1,2$$

where p_k is the a priori probability for class k , C_k is the covariance matrix for class k , and m_k is the mean vector for class k (19). This discriminant assumes a multivariate normal distribution of the data. A pattern, X_1 , is put in the class for which $d_k(X_1)$ is greatest.

If the assumption is made that the covariance matrices for the two classes are the same, then the discriminant function developed using the Bayes theorem and the multivariate normal assumption simplifies from that above to the following:

$$s = d_1(X) - d_2(X) = \ln p_1 - \ln p_2 + X'X^{-1}(m_1 - m_2) \\ - \frac{1}{2} m_1' C^{-1} m_1 + \frac{1}{2} m_2' C^{-1} m_2$$

where s is greater than zero for one class and less than zero for the other class. This routine implements this linear discriminant function. The covariance matrix, C , can be calculated using the members of either of the two classes of patterns or using the entire training set, at the user's option.

Another parametric routine implements a discriminant function by the method commonly called linear discriminant function analysis. It is nearly identical to the linear Bayesian discriminant, except that instead of using the covariance matrix, the sum of cross-products matrix is used. Results obtained with the routine are ordinarily very similar to those obtained using the linear Bayes routine. The routine implemented as LDFA is a highly modified version of program BMD04M taken from the Biomedical Computer Programs Package (47).

Nonparametric Pattern Recognition Programs: The nonparametric pattern recognition programs develop their discriminants using the training set of patterns to be classified rather than statistical measures of their distributions.

The k nearest neighbor routine is conceptually the simplest. An unknown pattern is assigned to the class to which the majority of its nearest neighbors belong. The metric that is used to determine proximity is ordinarily the Euclidean metric, but any metric can be used.

The system contains a routine that implements the familiar linear learning machine or perceptron (19,22). The algorithm is heuristic. A decision surface is initialized either arbitrarily or is input from a disc file that was previously filled using the weight vector utility routine or one of the other linear discriminant development routines. Then the learning machine classifies one member of the training set at a time. When the current discriminant function correctly classifies a pattern, the discriminant is left unchanged. However, whenever an incorrect classification is made, then the discriminant is altered in such a way that the error just committed is eliminated. The learning

machine continues to classify the members of the training set repeatedly until no errors are committed or until the user terminates the routine. This algorithm for development of linear discriminant functions has the desirable property that if a training set of patterns is separable into the two classes, i.e.; if a solution exists, then this method will find a solution. This routine is the one ordinarily used to train a set of slightly different weight vectors to be used for the variance method of nonparametric feature selection.

Another nonparametric routine develops a linear discriminant function through an iterative least squares approach (22). The function is minimized:

$$Q = \sum_{i=1}^m [Y_i - F(s_i)]^2$$

where m is the total number of patterns in the training set, Y_i equals +1 for one pattern class and -1 for the other class, s_i is the dot product of the weight vector with the i th pattern vector, and $F(s_i)$ is the hyperbolic tangent function of s_i . The hyperbolic tangent function has the values of positive one for all positive values of s_i and negative one for all negative values of s_i . Therefore, minimizing Q is equivalent to minimizing the number of incorrect classifications. The function is non-linear in the independent variables and thus can not be solved directly, so an iterative algorithm is employed. This routine has been found to be particularly useful for dealing with data that are not separable into the two classes.

A linear discriminant function can be found using a linear programming approach (48,49). The objective function to be optimized consists of the fraction of the training set correctly classified. If two vertices have the same classification ability, then the vertex with the smaller sum of distances to misclassified points is taken as better.

Another routine develops a decision tree of binary choices which, taken as a whole, can classify the members of the training set. The decision tree generated implements a piecewise linear discriminant function. The decision tree is developed by splitting the data set into two parts in an optimal way at each node. A node is considered to be terminal when no advantageous split can be made; a terminal node is labelled to associate with the pattern class most represented among the patterns present at the node.

Feature Selection. Once a data set has been established to be linearly separable, then variance feature selection (50) can be used to discard the least useful descriptors and thereby focus on the more useful descriptors. Several routines are implemented

in ADAPT to allow the application of variance feature selection to a data set in a convenient, interactive way.

Polycyclic Aromatic Hydrocarbon Study (51). The data were taken from papers by Dipple (52) and McCann *et al.* (53). In the Dipple data set, several hundred polycyclic aromatic hydrocarbons are arranged according to their carcinogenic potency into four classes: inactive, slight, moderate, and high. These classifications are based upon the percentage of treated animals which developed tumors when exposed to the compounds. Slight carcinogens are those compounds which caused tumors in up to 33% of the test animals. Moderate carcinogens are those compounds which caused tumors in between 33% and 66% of the test animals. High carcinogens are those compounds which caused tumors in 66% or more of the test animals. In the McCann data set, 209 compounds in ten different classes (e.g., aromatic amines, N-nitroamines, polycyclic aromatic hydrocarbons) are compiled. From this group of compounds a number of polycyclic aromatic hydrocarbons not already present in the Dipple data set were selected. These compounds were only classified as carcinogens or noncarcinogens.

In most cases, these compounds were tested in either mice or rats, by subcutaneous injection, giving rise to sarcomas, or by application to the skin, giving rise to benign papillomas or malignant ephtheliomas.

In the Dipple data set, in some cases, compounds were classified into several categories because different results were obtained with different species or with different models. For example, 5-methyl Benz(a)anthracene is classified as an inactive compound when it is injected intramuscularly into rats. It is classified as a slight carcinogen when it is either painted on mouse skin, or subcutaneously injected into rats. Finally, it is classified as a moderate carcinogen when it is subcutaneously injected into mice.

Because most of the pattern recognition algorithms used in this study are binary pattern classifiers, the four categories used by Dipple were reduced to two, namely carcinogens and noncarcinogens. A compound was classified as a noncarcinogen if and only if it was classified by Dipple as either inactive or inactive and slight. A compound was classified as a carcinogen if and only if it was classified by Dipple as either high, moderate, high and moderate, high and slight, moderate and slight, or high and moderate and slight.

A set of 200 PAH compounds were randomly selected from those stored for study. They fall into sixteen structural classes as shown in Table I. There are 73 carcinogens and 127 noncarcinogens in the set.

Each compound was modelled to obtain a three dimensional geometry for use in geometrical descriptor development. For each of the 200 compounds more than 50 descriptors were generated. The 28 descriptors finally found to be the best set are shown in

TABLE I.
CHARACTERIZATION OF THE 200 COMPOUND PAH DATA SET

	<u>Carcinogens</u>	<u>Noncarcinogens</u>
Anthracenes	2	10
Benz(a)acridines	0	17
Benz(c)acridines	3	10
Benz(a)anthracenes	23	41
Benz(c)phenanthrenes	0	4
Benzo(a)pyrenes	13	5
Carbazoles	5	0
Cholanthrenes	5	5
Chrysenes	0	3
Cyclopenta(a)phenanthrenes	2	14
Dibenz(a,h)anthracenes	4	1
Miscellaneous Heterocycles	6	0
Miscellaneous PAH	10	10
Phenanthrenes	0	3
Pyrenes	0	2
Triphenylenes	0	2
Totals	73	127

Table II. The substructures used for the generation of the environment descriptors and the sigma charge descriptors are shown in Figure 2. In this figure a terminal X indicates an atom of undetermined type and undetermined connectivity. An X found imbedded in a substructure, such as the X found in substructure 4, is constrained to have the connectivity found in the substructure to obtain a match, but any atom type. For example, the substituent on the benzo ring in substructure 5 can be a methyl group, a chlorine, or an ethyl group, among others. The mean and standard deviation for each descriptor and the values for 7,12-dimethyl benz(a)anthracene are shown in Table III.

This set of 28 descriptors supported linear discriminants that could completely separate 191 of the PAH's into the active and inactive classes. This is a recognition rate of 95.5 percent. There are 66 carcinogens and 125 noncarcinogens in this set of compounds.

Predictive ability tests were performed with the 28 descriptors. In a series of randomized trials 400 predictions were performed. Actual carcinogens were predicted with 87.5 percent accuracy and noncarcinogens with 92.5 percent accuracy for an average of 90.5 percent.

These results show that pattern recognition can be used as an effective tool to characterize polycyclic aromatic hydrocarbon carcinogens. Using a set of only 28 molecular structure descriptors, linear discriminants can be found to correctly dichotomize 191 out of 200 randomly selected PAH's. This same set of 28 descriptors supports a linear discriminant function that has an average predictive ability of over ninety percent when subjected to randomized predictive ability tests.

The 28 descriptors selected in this study are a notably diverse set. All types of descriptors available have been found useful in the analysis. One can interpret this selection of such a diverse set of descriptors as an indication that PAH carcinogenesis is a complex process involving many factors, and that one might expect the most success at developing structure-activity relationships with a multifaceted approach to the problem.

Among the descriptors chosen is the logarithm of the calculated 1-octanol/water partition coefficient, indicative of the important influence of transport properties. Substructural descriptors, involving both connectivity environment calculations and average sigma charge calculations are perhaps indicative of particular regions of the molecules that might be involved in PAH carcinogenesis. Both bay-region substructures and K-region substructures were retained after feature selection. The influence of overall molecular size is well represented by several geometric descriptors, the fragment descriptor representing the number of rings in the compound, and the descriptor representing the square of the area of the molecule.

Of particular interest are the geometrical descriptors and other descriptors representing some aspect of molecular size or

TABLE II.
 TWENTY-EIGHT DESCRIPTORS USED FOR PAH STUDY

<u>Descriptor</u>	<u>The Number of Non-Zero Occurrences</u>
1. The number of oxygen atoms	52
2. The number of fluorine atoms	14
3. The number of double bonds	31
4. The number of rings	200
5. The molecular volume	200
6. Path 1 molecular connectivity, using heteratom valences, ring corrected	200
7. Path 2 molecular connectivity	200
8. The X principal radius	200
9. The Z principal radius	200
10. The X principal radius/ The Y principal radius	200
11. The X principal radius/ The Z principal radius	200
12. The Y principal radius/ The Z principal radius	200
13. The molecular connectivity of substructure 1	164
14. The mole. connect. of substr. 2	26
15. (X prin. radius x Y prin. radius) ²	200
16. The mole. connect. of substr. 3	102
17. The mole. connect. of substr. 4	17
18. The mole. connect. of substr. 5	44
19. The mole. connect. of substr. 6	51
20. The mole. connect. of substr. 7	16
21. The mole. connect. of substr. 8	26
22. The logarithm of the 1-octanol/water partition coefficient	196
23. The average sigma charge of substructure 9	19
24. The average sigma charge of substructure 1	164
25. The average sigma charge of substructure 8	26
26. The average sigma charge of substructure 10	66
27. The average sigma charge of substructure 6	51
28. The average sigma charge of substructure 5	44

Note: Substructures are numbered according to Figure 2.

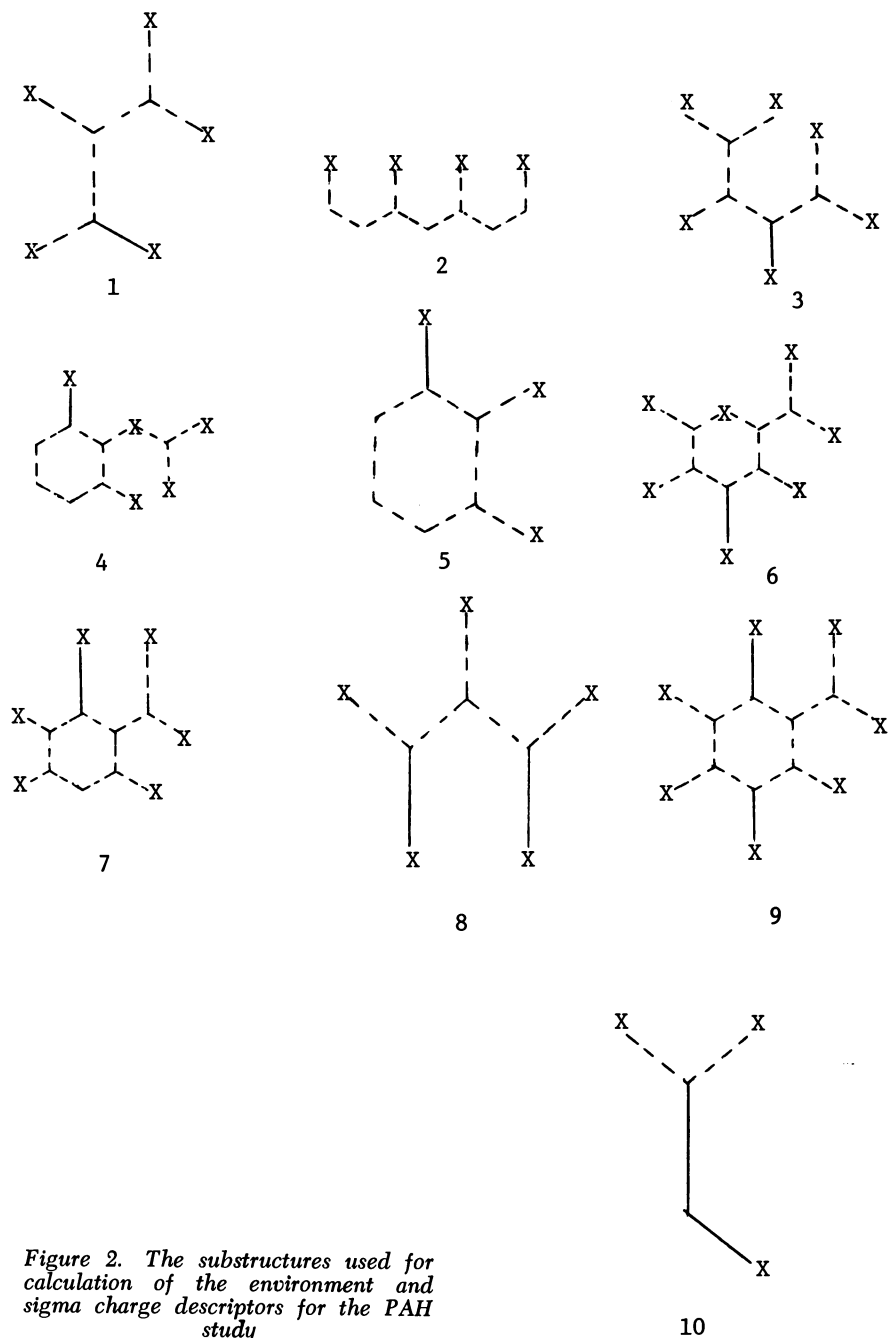


Figure 2. The substructures used for calculation of the environment and sigma charge descriptors for the PAH study

TABLE III.
DESCRIPTOR STATISTICS AND SAMPLE DESCRIPTOR VALUES

<u>Descriptor</u>	<u>Mean</u>	<u>Standard Deviation</u>	<u>Value for 7,12-Dimethyl Benz(a)anthracene</u>
1.	0.365	0.703	0
2.	0.080	0.323	0
3.	0.190	0.485	0
4.	4.390	0.735	4
5.	165.950	20.677	159.001
6.	6.044	0.802	5.721
7.	5.432	0.693	5.106
8.	8.832	2.344	7.285
9.	0.072	0.258	0.012
10.	3.787	1.719	3.185
11.	653.340	403.620	604.796
12.	504.960	435.070	189.889
13.	3.652	1.748	4.486
14.	0.638	1.655	0
15.	585.420	547.861	277.588
16.	2.633	2.600	5.251
17.	0.440	1.451	0
18.	0.910	1.726	0
19.	1.618	2.777	0
20.	0.506	1.724	0
21.	0.592	1.542	0
22.	6.360	1.448	6.940
23.	0.000556	0.002205	0.00388
24.	0.012536	0.015061	0.00433
25.	0.001902	0.009210	0
26.	-0.009188	0.017044	0
27.	0.000006	0.001883	0
28.	-0.001157	0.005645	0

shape. Of the 28 descriptors surviving feature selection there are five of the six geometrical descriptors generated involving principal radii, the molecular volume descriptor, the fragment descriptor representing the number of rings in the compound, and the descriptor related to the area of the planar or near planar polycyclic aromatic hydrocarbons. The abundance of descriptors which relate size of the hydrocarbons to carcinogenic activity supports arguments made by Arcos (54,55) relating carcinogenic activity to the incumbrance area of the polycyclic aromatic hydrocarbon. The incumbrance area is the area of the smallest rectangle that can encompass the molecule's van der Waals envelope (54). Their conclusion was that molecules with too great an incumbrance area would be too large to fit in a receptor site or too large to intercalate into DNA. Those molecules possessing too small an incumbrance area would be bound too loosely to the critical cellular constituent. This size criteria has been supported by other studies involving DNA-hydrocarbon binding (54).

The importance of molecular size has also been emphasized by Hansch and Fujita (32). However, he suggested that the small hydrocarbons were too hydrophilic and were excreted from the body before they could cause damage. The large hydrocarbons were thought to be insufficiently hydrophilic and easily isolated from critical cellular sites by a small aqueous barrier. Support for the size criteria in this study is given not only by the fact that a large proportion of the geometric descriptors were retained after feature selection, but also the single most powerful descriptor was the fragment descriptor representing the number of rings in the compound.

While the molecular size criteria is no doubt important, the carcinogens cannot be discriminated from the noncarcinogens on the basis of size alone. This, again, suggests that polycyclic aromatic hydrocarbons carcinogenesis is a complex process involving many different factors, with molecular size being just one of perhaps many important factors.

N-Nitroso Compound Study (56). The data used were taken from several sources: Dr. William Lijinsky of the Frederick Cancer Research Center provided most of the data; the remainder came from a published compilation (57) and from papers (53,58). A total of 171 compounds were entered into the ADAPT disc files. The distribution of the compounds among structural classes is shown in Table IV. A subset of the total data set comprising 118 carcinogens (labelled + or w in Table IV) and 35 noncarcinogens (labelled - in Table IV) was used for analysis.

For each of the 153 N-nitroso compounds many descriptors were developed and tested. After many trials of different subsets of descriptors, a final set of 15 descriptors was selected. This set of 15 descriptors is shown in Table V. They provided the best overall performance of the pattern recognition routines of

TABLE IV.
DISTRIBUTION OF THE N-NITROSO COMPOUND DATA SET

<u>Structural Class</u>	<u>Number of Compounds</u>			
	<u>+¹</u>	<u>w¹</u>	<u>-¹</u>	<u>?¹</u>
Dialkyl, diaryl, or mixed	47	6	18	3
Piperidines	10	2	7	6
Pyrrolidines	3	1	4	-
Morpholines	2	1	1	-
Piperazines	6	-	3	1
Hydrazine, hydroxylamines	3	-	1	1
Acyl alkyl	29	2	1	7
Misc.	5	1	0	-
	<u>105</u>	<u>13</u>	<u>35</u>	<u>18</u>

¹ + : carcinogen

w : weak carcinogen

- : non-carcinogen

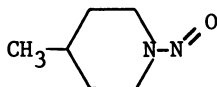
? : controversial or have not been tested

TABLE V.
 DESCRIPTORS USED TO REPRESENT
 153 COMPOUND N-NITROSO DATA SET

<u>Index</u>	<u>Descriptors</u>
1	Number of Carbon Atoms
2	Number of Nitrogen Atoms
3	Number of Bonds
4	Number of Double Bonds
5	Number of Aromatic Bonds
6	Number of Basis Rings
7	Path 1 Molecular Connectivity (Ring Corrected)
8	Path 1 Molecular Connectivity (Ring and Heteratom Corrected)
9	Path 2 Molecular Connectivity
10	Path 3 Molecular Connectivity
11	Path 4 Molecular Connectivity
12	
13	Environment of Substructure
	$\begin{array}{c} \text{X} \\ \diagdown \\ \text{N-N=O} \\ \diagup \\ \text{X} \end{array}$
14	Average Sigma Charge Substructure
	$\begin{array}{c} \text{X} \\ \diagdown \\ \text{N-N=O} \\ \diagup \\ \text{X-X} \end{array}$
15	Average Sigma Charge Substructure
	$\begin{array}{c} \text{X} \\ \diagdown \\ \text{N-N=O} \\ \diagup \\ \text{R} \end{array}$
	where R = $-\text{CH}_3$, $-\text{CH}_2\text{-X}$, $-\underset{\text{X}}{\text{CH}}\text{-X}$, $-\underset{\text{X}}{\overset{\text{X}}{\text{C}}}\text{-X}$

any set of descriptors tested. Included are six fragments, five molecular connectivities, one geometric, one molecular connectivity environment, and two sigma charge descriptors.

Table VI provides the means and standard deviations for the 15 descriptors. The final column of the table shows the descriptor values for 4-methyl-N-nitrosopiperidine:



Among the 153 compounds are included two pairs of diastereoisomers. Because the descriptors cannot distinguish between pairs of diastereoisomers, these four compounds were excluded from further study. In a lengthy series of pattern recognition studies, five additional compounds were identified that had to be excluded from consideration. This leaves a data set of 143 N-nitroso compounds with 116 carcinogens and 28 noncarcinogens. This data set could be completely dichotomized by linear discriminants generated with several different pattern recognition routines.

A series of predictive ability tests were performed with the 115 descriptors. In all, 650 predictions were done. The accuracy of prediction was 91.4 percent overall, 93.4 percent for carcinogens and 84.7 for noncarcinogens.

The fact that linear discriminants could be developed to completely separate 116 carcinogenic N-nitroso compounds from 28 noncarcinogenic ones indicates that information concerning the carcinogenic potential of these compounds is contained in their structures. A diversity of descriptors are shown to be important in relating structure to carcinogenic potential. Simple fragment descriptors are useful predictors in this study. Molecular connectivity indices have been shown to be correlated with a variety of biological activities (37) including mutagenicity of nitrosamines (29). It has been reported that the partition coefficients of nitrosamines are correlated with their carcinogenicity (25,26). The molecular connectivity indices are also known to be correlated with diverse physico-chemical properties, such as partition coefficients (37). The retention of molecular connectivity descriptors may show this type of relationship. The retention of one shape parameter may indicate the importance of geometry. The electronic properties of nitrosamines have been shown to be related to carcinogenic potential as well (25,59,60,61). The retention of two sigma charge parameters in this study supports the hypothesis of the importance of alpha hydroxylation (62,63) in the bioactivation of nitrosamines.

TABLE VI.

MEAN AND STANDARD DEVIATIONS FOR THE 15 DESCRIPTORS AND THE
DESCRIPTOR VALUES FOR 4-METHYL-N-NITROSOPIPERIDINE

<u>Descriptor Number</u>	<u>Mean Value</u>	<u>Standard Deviation</u>	<u>Value for 4-Methyl- N-Nitroso- Piperidine</u>
1	6.28	2.97	6
2	2.54	0.94	2
3	10.50	3.46	9
4	1.63	0.77	1
5	1.07	2.96	0
6	0.58	0.67	1
7	4.21	1.27	2.51
8	3.35	1.16	3.00
9	2.48	1.01	2.72
10	1.54	0.78	1.88
11	3.39	2.27	3.14
12	2.33	0.40	2.43
13	4.09	3.01	6.27
14	-0.032	0.02	-0.044
15	-0.047	0.05	-0.044

Acknowledgements

This research was sponsored by the National Cancer Institute through contract number N01 CP 75926. The computer used for this work was purchased with partial financial support of the National Science Foundation.

Literature Cited

1. E.J. Ariens, "Drug Design", Vol. 1-8, Academic Press, New York, 1971-1978.
2. A. Burger, "Medicinal Chemistry", Part I, Wiley-Interscience, New York, 1970.
3. B. Bloom and G.E. Ulliyot (Eds.), "Drug Discovery", American Chemical Society, Washington, D.C. 1971.
4. W. Van Valkenburg (Ed.), "Biological Correlations -- The Hansch Approach", American Chemical Society, Washington, D.C. 1972.
5. W.P. Purcell, G.E. Bass, J.M. Clayton, "Strategy of Drug Design", Wiley-Interscience, New York, 1973.
6. Y.C. Martin, "Quantitative Drug Design", Marcel Dekker, Inc., New York, 1978.
7. V.E. Golender and A.B. Rozenblit, "Computer-Assisted Drug Design", Zinathe Press, Riga, U.S.S.R. 1978 (in Russian).
8. A.J. Stuper, W.E. Brugger, and P.C. Jurš, "Computer Assisted Studies of Chemical Structure and Biological Function", Wiley-Interscience, New York, 1979.
9. W.J. Dunn, in "Annual Reports in Medicinal Chemistry", Vol. 8, R.V. Heinzelman (Ed.), Academic Press, New York, 1973.
10. R.D. Cramer, in "Annual Reports in Medicinal Chemistry", Vol. 11, F.H. Clarke (Ed.), Academic Press, New York, 1976.
11. C. Hansch, in "Advances in Linear Free Energy Relationships", Vol. 2, N.R. Chapman and J. Shorter (Eds.), Plenum Press, New York, 1979.
12. P.N. Craig, in "Biological Correlations -- The Hansch Approach", W. Van Valkenburg (Ed.), American Chemical Society, Washington, D.C. 1972.
13. W.G. Richards and M.E. Black, in "Progress in Medicinal Chemistry", Vol. 11, G.P. Ellis and G.B. West (Eds.), American Elsevier, New York, 1975.
14. R.E. Christoffersen, in "Quantum Mechanics of Molecular Conformations", B. Pullman (Ed.), Wiley, New York, 1976.
15. G.L. Kirschner and B.R. Kowalski, in "Drug Design", Vol. VIII, E.J. Ariens (Ed.), Academic Press, New York, 1978.
16. N.J. Nilsson, "Learning Machines", McGraw-Hill, New York, 1965.
17. E.A. Patrick, "Fundamentals of Pattern Recognition", Prentice-Hall, Englewood Cliffs, N.J., 1972.
18. H.C. Andrews, "Introduction to Mathematical Techniques in Pattern Recognition", Wiley-Interscience, New York, 1972.
19. J.T. Tou and R.C. Gonzalez, "Pattern Recognition Principles", Addison-Wesley, Reading, Mass., 1974.
20. V.L. Tal'roze, V.V. Rasnikov, and G.D. Tantsyrev, Doklady Akademii Nauk SSSR, 159(1), 182 (1964).
21. P.C. Jurš, B.R. Kowalski, and T.L. Isenhour, Anal. Chem., 41, 21 (1969).

22. P.C. Jurs and T.L. Isenhour, "Chemical Application of Pattern Recognition", Wiley-Interscience, New York, 1975.
23. B.R. Kowalski, Anal. Chem., 47, 1152A (1975).
24. M.L. McConnell, G. Rhodes, U. Watson, and M. Novotny, Jour. Chrom. , in press.
25. J.S. Wishok and M.C. Archer, Brit. Jour. Cancer , 33, 307 (1976).
26. G.M. Singer, H.W. Taylor, W. Lijinky, Chem.-Biol. Interactions , 19, 133 (1977).
27. I.A. Smith, G.D. Berger, P.G. Seybold, and M.P. Serve, Cancer Res. , 38, 2968 (1978).
28. J.S. Wishnok, M.C. Archer, A.S. Edelman, and W.M. Rand, Chem.-Biol. Interactions , 20, 43 (1978).
29. L.B. Kier, J.R. Simons, L.H. Hall, Jour. Pharm. Sci. , 67, 725 (1978).
30. A.J. Hopfinger and G. Klopman, Chem.-Biol. Interactions , in press.
31. W.J. Dunn III and S. Wold, Jour. Med. Chem. , 21, 1001 (1978).
32. C. Hansch and T. Fujita, Jour. Amer. Chem. Soc. , 86, 1616 (1964).
33. R. Franke, Mol. Pharmacol. , 5, 640 (1969).
34. P.C. Jurs, J.T. Chou, and M. Yuan, Jour. Med. Chem. , in press.
35. A.J. Stuper and P.C. Jurs, Jour. Chem. Infor. Comp. Sci. , 16, 99 (1976).
36. A.J. Stuper, W.E. Brugger, and P.C. Jurs, in "Chemometrics: Theory and Practice", B.R. Kowalski (Ed.), American Chemical Society, Washington, D.C., 1977.
37. L.B. Kier and L.H. Hall, "Molecular Connectivity in Chemistry and Drug Research", Academic Press, New York, 1976.
38. W.J. Murray, Jour. Pharm. Sci. , 66, 1352 (1977).
39. G.R. Parker, Jour. Pharm. Sci. , 67, 513 (1978).
40. M. Randic, G.M. Brissey, R.B. Spencer, and C.L. Wilkins, "Computers in Chemistry", submitted.
41. A.J. Hopfinger, "Conformational Properties of Macromolecules", Academic Press, New York, 1973.
42. G. Del Re, Jour. Chem. Soc. , 4031 (1958).
43. A. Leo, P.Y.C. Jow, C. Silipo, and C. Hansch, Jour. Med. Chem. , 18, 865 (1975).
44. C. Hansch and A. Leo, in "Substituent Constants for Correlation Analysis in Chemistry and Biology", John Wiley and Sons, Inc., New York, in press.
45. J.T. Chou and P.C. Jurs, Abstract, 176th American Chemical Society National Meeting, Miami, Fla., September 1978.
46. J.T. Chou and P.C. Jurs, Jour. Chem. Infor. Comp. Sci. , submitted for publication.
47. W.J. Dixon (Ed.), "Biomedical Computer Programs", University of California Press, Berkeley, California, 1973.

48. S.L. Morgan and S.N. Deming, Anal. Chem. , 46, 1170 (1974).
49. G.L. Ritter, S.R. Lowry, C.L. Wilkins, and T.L. Isenhour, Anal. Chem. , 47, 1951 (1975).
50. G.S. Zander, A.J. Stuper, and P.C. Jurs, Anal. Chem. , 47, 1085 (1975).
51. Mark Yuan and P.C. Jurs, in preparation.
52. A. Dipple, in "Chemical Carcinogens", C. Searle (Ed.), American Chemical Society, Washington, D.C. 1976.
53. J. McCann, E. Choi, E. Yamasaki, and B.N. Ames, Proc. Nat. Acad. Sci. USA , 72, 5135 (1975).
54. J.C. Arcos and M.F. Argus, "Chemical Induction of Cancer", Vol. IIA, Academic Press, New York, 1974.
55. J.C. Arcos and M. Arcos, Prog. Drug Res. , 4, 407 (1962).
56. J.T. Chou and P.C. Jurs, Jour. Med. Chem. , submitted.
57. P.N. Magee, R. Montesano, and R. Preussmann, in "Chemical Carcinogens", C.E. Searle (Ed.), American Chemical Society, Washington, D.C., 1976.
58. K. Lee, B. Gold, and S.S. Mirvish, Mutation Res. , 48, 131 (1977).
59. C. Nagata and A. Imamura, Gann , 61, 169 (1970).
60. B. Testa, J.C. Bunzil, and W.P. Purcell, Jour. Theo. Biol. , 70, 339 (1978).
61. B. Testa and D. Mihailova, Jour. Med. Chem. , 21, 683 (1978).
62. S.S. Hecht, C.B. Chen, and D. Hoffman, Cancer Res. , 38, 215 (1978).
63. A.M. Camus, M. Wiessler, C. Malaveille, H. Bartsch, Mutation Res. , 49, 187 (1978).

RECEIVED June 8, 1979.

Chance Factors in QSAR Studies

JOHN G. TOPLISS and ROBERT P. EDWARDS

Schering-Plough Research Division, Bloomfield, NJ 07003

During the past decade quantitative structure-activity relationships (QSAR) have emerged as an important tool in drug design (1). The role of the computer has been essential in allowing practical use of statistical methods such as multiple regression analysis (2). However, there is a risk of arriving at fortuitous correlations when too many variables are screened relative to the number of available observations. In this regard a critical distinction must be made between the number of variables screened for possible correlation and the number which actually appear in the regression equation.

By way of illustration take a case where there are ten observations and corresponding activity values A_1 to A_{10} and eight variables V_1 to V_8 are considered. Multiple regression analysis yields equation (1) where $r^2 = 0.85$, $F_{2,7} = 19.8$ ($p < 0.005$) and

$$A = aV_1 + bV_2 + c \quad (1)$$

V_1 and V_2 are each significant at the level $p < 0.01$. However, the statistical test for significance of the equation only relates to the variables included in the actual correlation equation and does not take account of the fact that eight variables were screened for possible correlation. It is clear that the greater the number of variables which are screened for possible inclusion the greater will be the probability that a chance correlation will occur. The p value of < 0.005 for the equation shown overstates to some extent the probability that the relationship given by the equation is real rather than occurring by chance. The question is how misleading is this p value? The studies to be described will attempt to provide some answer to this question.

0-8412-0521-3/79/47-112-131\$05.00/0

© 1979 American Chemical Society

Earlier work on this problem was reported some years ago (3). However, these studies were too limited in scope to provide more than a demonstration that indeed there was a problem from chance correlations under certain conditions and to provide very general guidelines. Accordingly, in view of the interest in this problem following publication of the earlier study, plans were made to generate more extensive and detailed data.

Method

The basic approach was to set up a large number and wide range of simulated correlation studies using random numbers for both observations and screened variables and assess the incidence and degree of any resulting correlations.

Computer System. The program used was developed by modifying an existing FORTRAN Multiple Regression Analysis Program (BMD 02R), the essence of the modifications being to generate the data to be analyzed using a pseudo-random number generator and to accumulate statistics from each of the individual simulations and produce output reports, frequency distributions and summary tables. The program was run on an IBM 360/158 computer performing under OS/VS2.

Random Number Generator. The pseudo-random numbers were generated using the IBM RANDU subroutine (4) which employs the Multiplicative Congruential Method (5) and returns a vector of random numbers, each uniformly distributed in the range 0 to 1. The random numbers were scaled by a factor of 1000. Starting from a user-defined seed value, RANDU generates a stream of pseudo-random numbers and will produce 2^{29} terms before repeating. By comparison of the observed as against the expected frequency distribution (uniform) of numbers, RANDU is an adequate pseudo-random number generator. To better approximate pure randomness each computer run had a different externally imposed initial seed value. In addition only every fourth pseudo-random number was sampled from the stream.

Computer Program Methodology. The first step in each simulation is to generate the data matrix with r rows (one for each observation) and $n + 1$ columns (n independent variables and one dependent). The values of r and n are specified when the program is

run. The data are then analyzed using a stepwise Multiple Regression Technique (2). At each stage of this procedure the independent variable most highly correlated with the dependent variable is brought into the model provided that the partial F-test value for that variable is significant at the 10% level. Each independent variable in the regression model is then re-examined to see if it is still making a significant contribution. Any variable whose partial F-test value is not significant at the 10% level is dropped from the model. This process continues until no more variables enter the model and none are rejected.

From each simulation, the number of variables entering the regression (zero to n) is stored as well as the r^2 statistic in the case of at least one variable entering the regression analysis. This whole process is repeated for each simulation, the number of simulations to be carried out being specified when the program is run. The results of all the simulations are then consolidated to give the following statistics:

- (1) The number and relative frequency for 0, 1, 2, 3 etc. variables entering the regression analysis.
- (2) The number and relative frequency for at least one variable entering the regression analysis.
- (3) For each significant correlation observed (a) mean r^2 statistic (b) median r^2 statistic (c) minimum r^2 statistic (d) maximum r^2 statistic.
- (4) The number and relative frequency of observed correlations by r^2 interval: 1.0-0.9; 0.9-0.8; 0.8-0.7 etc.

Scope of Study. The number of variables screened was 3-15, 20, 30, 40 and the number of observations was 5-30, 50, 100, 200, 400, 600. Altogether some 300 combinations of variables and observations were examined. The number of runs for each combination ranged from 120-2190. It should be noted that the study was originally designed for far fewer combinations of variables and observations and runs per combination. However, due to the unexpected availability, for a short period, of a computer which was being phased out, the scope of the study was greatly expanded.

Results

Due to space limitations it is not feasible to present in tabular form all of the data obtained in

the study. However, Table I provides an illustrative sample. The columns in the table give the number of independent variables, the number of observations, the number of runs, the frequency of a chance correlation, the average number of variables entered (for those runs for which at least one variable is entered), the maximum, minimum and mean r^2 values, and the frequency of those runs where $r^2 \geq 0.5$ and ≥ 0.8 .

For the situation with ten independent variables screened and ten observations, some two-thirds of the runs gave a correlation and for these an average of 2.20 variables entered. R^2 ranged from 0.28 to 1.00 with a mean value of 0.64. Almost half of the total runs made gave correlations with r^2 values of 0.5 or higher while about one quarter of the total runs made gave correlations with r^2 values of 0.8 or higher. As the number of observations increased there was a downward trend in r^2 values so that for 15 observations only 4% of the total runs gave correlations with $r^2 \geq 0.8$ and for 30 observations no correlations with $r^2 \geq 0.8$ were found.

For 20 variables screened and only 15 observations, practically all of the runs resulted in some level of chance correlation with about 44% of the total runs resulting in a correlation with $r^2 \geq 0.8$. As the number of observations increased to 30 the number of runs yielding correlations with $r^2 \geq 0.8$ declined to about 1%.

P-values for the observed correlations ranged from a minimum value of 0.10 up to 0.00.

At this point it must be noted that the degree of chance correlation found in this study was substantially less than that reported in an earlier study (3). Mean r^2 values averaged about one-half those reported in the earlier study, ranging from three-quarters down to one-quarter. The cause of this discrepancy could not be determined since the basic records from the earlier study were not retained. However, it must be concluded that the earlier study was in error. The present study was very carefully checked and selected combinations set-up and performed manually showed similar results.

From data generated in the present study it is now possible to answer the question raised earlier concerning the p-value of equation 1. With ten observations and eight screened variables the expected frequency of chance correlations comes out to be about 12%. Thus, the validity of the correlation equation is far less certain than the reported p-value indicates.

Table 1
Data from Simulated Correlations using Random Numbers

NO. IND. VAR.	NO. OBS.	NO. RUNS	FREQ. CHANCE CORR.	AV. NO. VAR. ENTERED	MAX R ²	MIN R ²	MEAN R ²	FREQ. 0.5	R ² > 0.8
3	5	2190	0.26	1.26	1.00	0.65	0.83	0.26	0.15
3	10	2190	0.27	1.15	0.95	0.30	0.47	0.09	0.01
3	20	2190	0.28	1.14	0.63	0.14	0.24	0.01	0.00
5	5	120	0.44	1.87	1.00	0.61	0.80	0.44	0.28
5	7	398	0.39	1.54	1.00	0.45	0.73	0.34	0.15
5	10	2190	0.40	1.41	0.99	0.30	0.54	0.21	0.04
5	20	2190	0.41	1.30	0.79	0.14	0.28	0.03	0.00
10	10	120	0.67	2.20	1.00	0.28	0.64	0.46	0.23
10	12	398	0.71	2.09	1.00	0.23	0.57	0.41	0.14
10	15	398	0.71	1.92	0.98	0.18	0.46	0.28	0.04
10	30	2190	0.74	1.84	0.74	0.08	0.23	0.03	0.00
15	10	120	0.88	4.42	1.00	0.26	0.79	0.74	0.54
15	15	120	0.87	3.33	1.00	0.19	0.60	0.56	0.28
15	17	398	0.86	2.77	0.98	0.16	0.53	0.46	0.14
15	20	2190	0.85	2.61	0.97	0.13	0.45	0.34	0.06
15	30	2190	0.85	2.48	0.86	0.08	0.30	0.11	0.00
20	15	120	0.94	5.00	1.00	0.17	0.73	0.76	0.44
20	20	120	0.97	3.53	0.98	0.13	0.56	0.58	0.17
20	30	2190	0.95	3.53	0.94	0.08	0.39	0.29	0.01
20	50	1991	0.95	3.41	0.73	0.04	0.24	0.03	0.00
30	50	1991	0.99	5.59	0.76	0.04	0.37	0.19	0.00
30	100	1991	0.99	5.17	0.52	0.02	0.18	0.00	0.00
40	50	199	1.00	7.14	0.90	0.04	0.49	0.44	0.02
40	100	199	1.00	7.04	0.47	0.03	0.24	0.01	0.00

The results are further illustrated in Figures 1-7. The relationship between r^2 and the number of variables screened for varying numbers of observations is illustrated in Figure 1. It can be seen that for a constant number of observations r^2 increases as the number of variables screened increases. Also, for a given number of variables screened, r^2 increases as the number of observations decreases.

In Figure 2 the relationship between the number of observations and the probability of a chance correlation with ten screened variables for various r^2 values is shown. From the graph corresponding to $r^2 \geq 0.8$ it may be seen that the probability of encountering a chance correlation at this level is about 22% for ten observations, reaching zero for 23 observations.

The graph shown in Figure 3 gives the relationship between the number of observations and the probability of a chance correlation with $r^2 \geq 0.8$ for five, ten and fifteen variables. Thus, for five screened variables the probability of encountering a chance correlation is about 22% for six observations, 10% for eight observations, 3% for ten observations and 1% for twelve observations.

A guideline often stated is that a certain number of observations are required in order to have much confidence in a given correlation equation. This rule of thumb is deficient on two counts; first, variables screened rather than variables included in the equation should be considered and second, the number of observations per variable is not a linear function of the number of variables as it pertains to a specified level of chance correlation. In Figure 4 are plotted the number of observations per variable against the number of variables for chance correlation levels of 1% for r^2 values of ≥ 0.5 , 0.8 and 0.9 . Considering the $r^2 \geq 0.8$ curve, the number of observations per variable is 3.7 for three variables dropping to 1.75 for thirteen variables.

Figure 5 shows the relationship between observations and variables for a chance correlation probability of 1% or less for various r^2 levels. The linear relationships shown are each statistically significant at the $p < 0.0001$ level. This graph permits the determination of the number of observations required to screen, for example, ten variables while keeping the probability of encountering a chance correlation with $r^2 \geq 0.8$ at the 1% level or less. From the graph it can be estimated that this number of observations is about 20. For $r^2 \geq 0.9$, the number required is less,

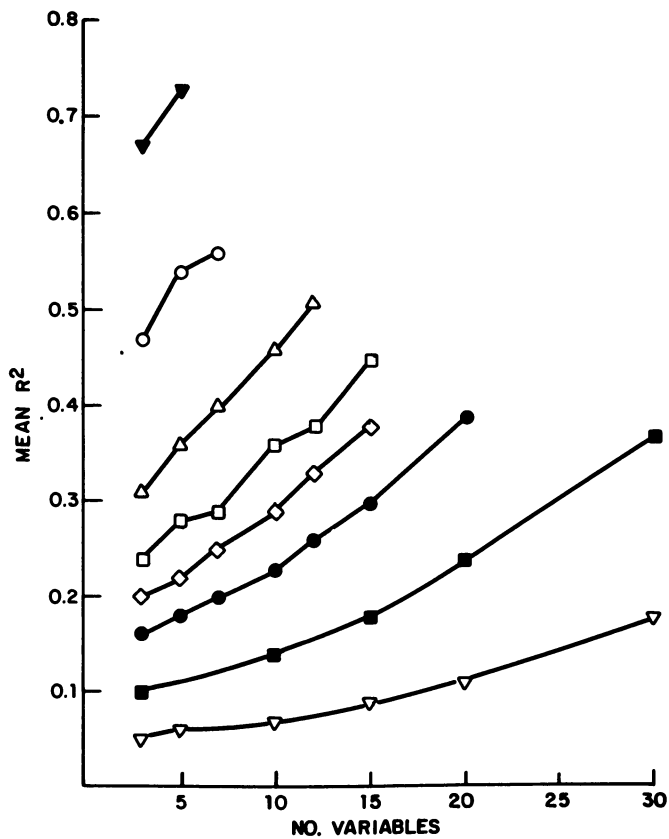


Figure 1. Relationship between mean r^2 values and number of variables screened (number of observations: (\blacktriangledown) 7; (\circ) 10; (\triangle) 15; (\square) 20; (\diamond) 25; (\bullet) 30; (\blacksquare) 50; (∇) 100)

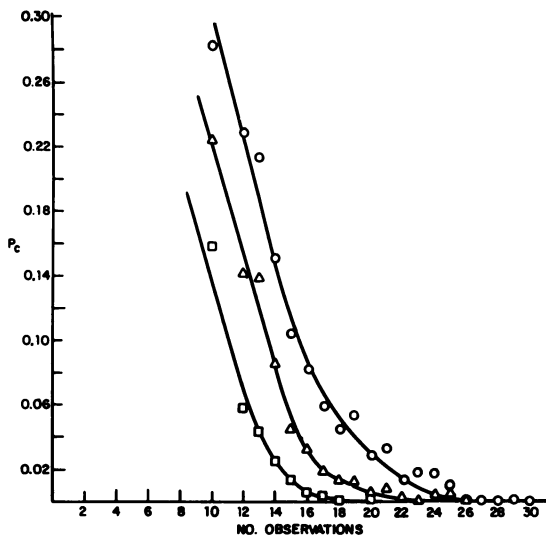


Figure 2. Relationship between number of observations and probability of a chance correlation (P_c) for 10 screened variables with $r^2 \geq 0.7$ (\circ), 0.8 (Δ), and 0.9 (\square)

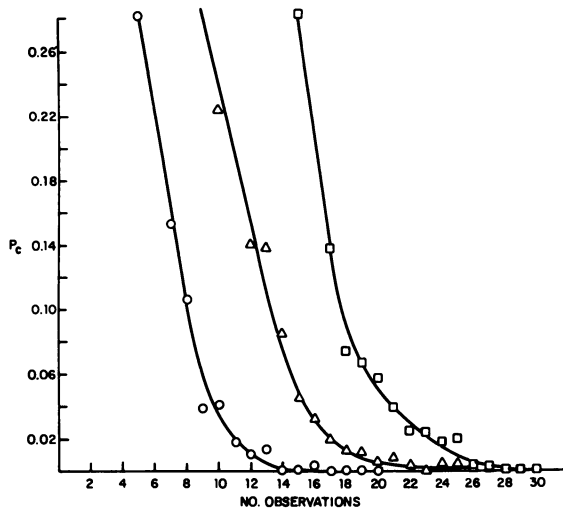


Figure 3. Relationship between number of observations and probability of a chance correlation for 5 (\circ), 10 (Δ), and 15 (\square) screened variables with $r^2 \geq 0.8$

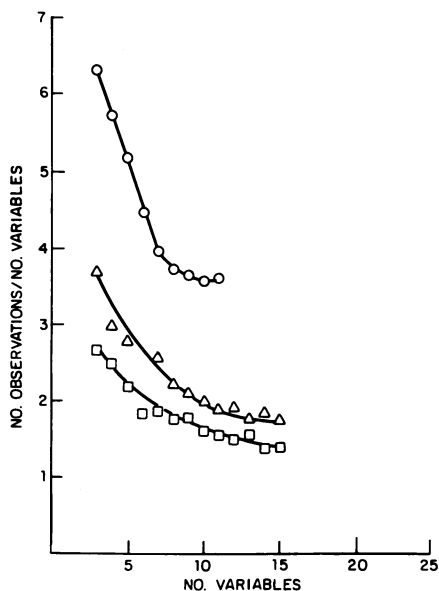


Figure 4. Number of observations/number of variables as a function of number of variables for specified levels of chance correlation ($P_c \leq 0.01$; $R^2 \geq 0.5$ (○); $R^2 \geq 0.8$ (△); $R^2 \geq 0.9$ (□))

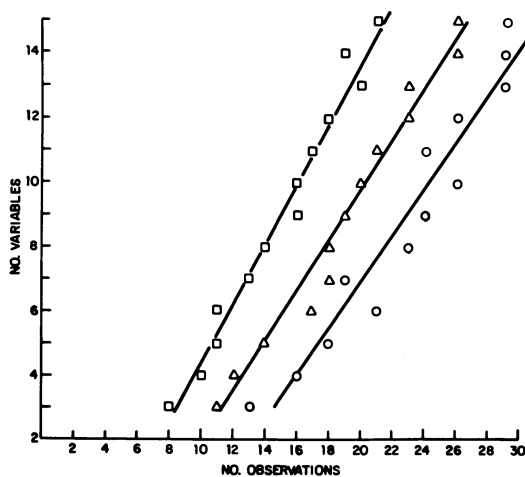


Figure 5. Relationship between number of observations and number of variables for chance correlation level $P_c \leq 0.01$ with $r^2 \geq 0.7$ (○), 0.8 (△), and 0.9 (□)

about 16.

Extrapolation of the previous graph allows one (Figure 6) to make similar estimations for larger numbers of screened variables. Thus, some 56 observations would be required to screen 40 variables while keeping the probability of encountering a chance correlation with $r^2 \geq 0.8$ at the 1% level or less.

The preceding graphs make possible the estimation of the number of observations needed to screen a specified number of variables while maintaining the probability of encountering a chance correlation of defined r^2 level at 1% or less. In Figure 7 the graph shows the effect on the number of required observations as the tolerated chance correlation probability increases to 5 or 10%. Again the depicted linear relationships are each statistically significant at the $p < 0.0001$ level. Thus, for 10 variables, 20 observations are needed at a chance correlation level of 1% which reduces to 15 at the 5% level and 13-14 at the 10% level.

Some limitations must be pointed out to the approach described in this study for estimating probable chance correlation levels. In actual practice screened variables have varying degrees of collinearity and not infrequently some variables are highly collinear. Generally, collinearity will be more in evidence in actual practice among a group of screened variables than among random number simulated screened variables. To the extent that this occurs it has the effect of reducing the number of independent variables actually operative as such. To reference a group of screened variables used in an actual problem with the random number data it is therefore suggested that when the correlation coefficient between two screened variables is 0.8 or higher, then only one of these variables is counted in arriving at the effective number of independent variables screened. This approximate correction device should avoid serious overestimation of chance correlation effects.

A second limitation has to do with use of the stepwise multiple regression method. This method probably misses some correlations and as such leads to some underestimation of the incidence of chance correlations.

Overall, the residual bias after correction for the collinearity effect tends to counterbalance the bias from use of the stepwise regression method so the net estimation error may not be that large.

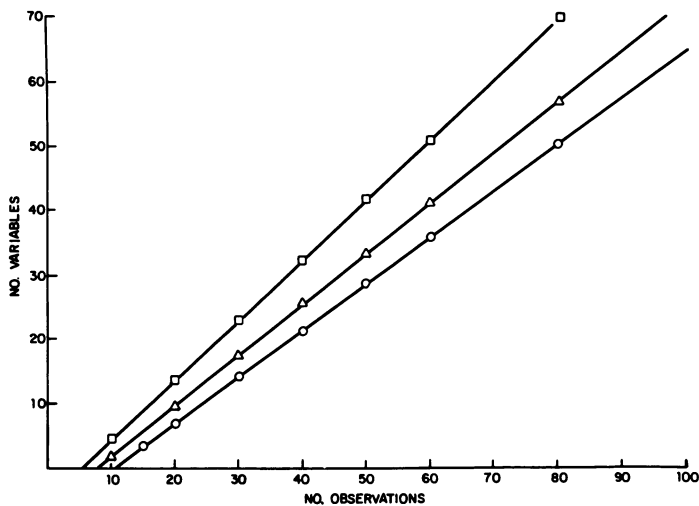


Figure 6. Relationship between number of observations and number of variables for chance correlation level $P_c \leq 0.01$ with $r^2 \geq 0.7$ (\circ), 0.8 (\triangle), and 0.9 (\square) (extrapolated)

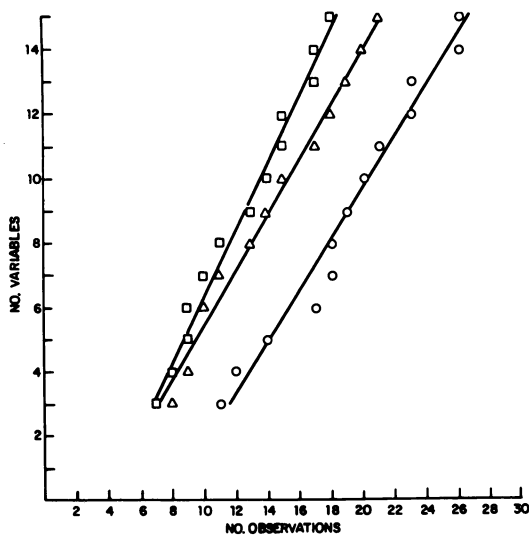


Figure 7. Relationship between number of observations and number of variables with $r^2 \geq 0.80$ at specified probability levels: $P_c \leq 0.01$ (\circ); $P_c \leq 0.05$ (\triangle); $P_c \leq 0.10$ (\square)

Applications to Reported Correlation Studies

In the light of the chance correlation phenomenon described, some comments will be made on a number of correlation studies reported in the literature which involved large numbers of screened variables.

The first of these reported by Peradejordi *et al.* (6) concerned a quantum chemical approach to structure-activity relationships of tetracycline antibiotics. The correlation equation (2) contained seven independent variables each significant at the 1% level

$$\begin{aligned} \log K' = & 18.3975 + 56.1733Q_{010} + 16.9155E_{010} + 48.7956Q_{04} \\ & - 1.1063E_{011} + 71.3254Q_{012} + 18.3655E_{012} + 3.3880Q_{C6} \end{aligned} \quad (2)$$

with the overall equation significant at the 0.1% level. The r^2 value given was 0.986 so that the equation accounts for almost all of the variance in the data. There were 20 observations and some 18 possible independent variables were screened. The random number simulated correlation data generated in the current study did not include the combination of 20 observations and 18 variables. The closest reference points were 20 observations and 20 variables for which the incidence of chance correlations with $r^2 \geq 0.9$ was about 8%, and 17 observations and 15 variables with a 5% chance correlation incidence. The approximate risk of a chance correlation is therefore 5-8%. The reported correlation equation, although it still may be valid, thus appears far less secure than the reported p-value of < 0.001 would indicate. It should be noted that we are not saying that there is a 5-8% probability that the reported equation could arise by chance. We are saying that from 20 observations and 18 variables there is a 5-8% probability that some correlation with $r^2 \geq 0.9$ could emerge from chance factors alone.

Gibbons *et al.* (7) reported on Quantitative Structure-Activity Relationships among selected Pyrimidinones and Hill Reaction Inhibition. The correlation equation (3) arrived at was based on 17 observa-

$$\begin{aligned} \text{RpI}_{50} = & -0.46(\pm 0.11)\pi \text{ ring} + 8.92(\pm 1.85)\sigma_{R3} \\ & + 0.045(\pm 0.008)\Sigma \text{MR} - 0.15 \end{aligned} \quad (3)$$

tions, contained three independent variables, had an

$r^2 = 0.83$ and a p-value of < 0.01 . Some 20 variables were screened for possible inclusion which may be equivalent to about 18 variables allowing for collinearity factors. This gives the combination 17 observations and 18 variables. The nearest reference points available for chance correlations are the combinations 20 observations and 20 variables and 15 observations and 15 variables for which the correlation frequencies for $r^2 \geq 0.8$ are 17% and 28% respectively. It must, therefore, be concluded that there is a substantial risk that the reported equation is spurious in sharp contrast with the reported p-value of < 0.01 .

A study by Timmermans and van Zwieten (8) on QSAR in Centrally Acting Imidazolines Structurally Related to Clonidine led to the development of equation (4). There were 27 observations and effectively

$$\begin{aligned} \log 1/ED_{30} = & -0.00032(\pm 0.0008)(\Sigma \text{Par})^2 + 0.105(\pm 0.03)\Sigma \text{Par} \\ & - 0.695(\pm 0.17)\Delta pK_a^0 + 5.333(\pm 1.89)\text{HOMO}(P) \\ & + 6.752(\pm 2.25)\text{EE}(P) + 2.494 \quad (4) \end{aligned}$$

$$n = 27, \quad r^2 = 0.91, \quad p < 0.001$$

perhaps 30-35 possible independent variables screened. For 27 observations and 35 variables using random numbers the frequency of occurrence of correlations with $r^2 \geq 0.9$ was approximately 45%, with approximately eleven variables entering on average for all significant correlations. For 27 observations and 30 variables there is about a 15% frequency with, on average, seven variables entering overall. It should be noted that in the correlations obtained with random numbers the average number of variables involved exceeds the five in the equation of Timmermans and van Zwieten. Therefore, the probability that the actual equation reported by them is spurious is less than the 15-45% chance correlation incidence suggests since r^2 increases as the number of independent variables in the equation increases. Nevertheless, it can be seen that under the conditions of the Timmermans and van Zwieten study chance factors could be important and have to be addressed. Certainly the stated p-value for the equation of < 0.001 is misleading.

Having established that a particular correlation may have a meaningful risk associated with it from chance factors, the correlation should then be further examined to more precisely define this influence. In the case of a correlation equation (5) which has been

$$Y = aX_1 + bX_2 + cX_3 + dX_4 + e \quad (5)$$

generated from screening variables X_1 to X_{20} , Y values can be replaced by sets (100 or more) of random numbers and the correlation analysis repeated with variables X_1 to X_{20} . Such a procedure has been used by Kier and Hall (9, 10). They also searched for correlations where all X values were replaced by random numbers and real Y values were retained. Since this does not preserve the pattern of collinearity present in the real X variable set this latter procedure may not provide as good an estimate of chance effects. It is also possible to replace selected X variables only, with random numbers. This would be useful in situations where it is suspected that in a correlation equation certain terms, eg. π , π^2 were real correlations while others were spurious. In this case, real π and π^2 numbers would be retained and the other X variables replaced by random numbers.

In evaluating correlations for chance effects attention should also be paid to whether some observations have been dropped from the set in order to obtain a better data fit. This is a fairly common occurrence and usually gives much higher r^2 values. However, sometimes the justification given for dropping the poorly fit observations are questionable. The correlation equation obtained from utilizing all of the observations (ie. after adding back in the dropped observations) should therefore be checked for possible chance correlations to give a better overall evaluation of the reliability of the result.

Summary

The results of the studies described show that chance correlations are a real phenomenon occurring when the number of variables screened for possible correlation is large compared to the number of observations. For this reason some correlations are less significant than their standard p -values indicate as has been demonstrated by reference to some reported correlations.

The present study provides guidelines for the approximate incidence of chance correlations at specified r^2 values for various combinations of observations and screened variables. Specific correlations obtained under conditions where the guidelines indicate an appreciable risk of chance correlations should be individually checked as described.

Finally, it should be pointed out that there is no intention of advocating correlation studies only under conditions where there is a miniscule risk of chance correlations. However, the importance of having a true idea of the reliability of a correlation equation is obvious. It is one thing to develop and use a correlation equation which is known to be somewhat tenuous but quite another to believe that it has a solid foundation when in fact it has not.

Acknowledgements

The authors are indebted to L. Weber for assistance in data tabulation and to M. Miller and A. Saltzman for helpful discussions.

Literature Cited

1. Martin, Y. C., "Quantitative Drug Design", Marcel Dekker, Inc., New York, 1978.
2. Draper, N. R., and Smith, H., "Applied Regression Analysis", Wiley, New York, 1966.
3. Topliss, J. G., and Costello, R. J., *J. Med. Chem.* (1972), 15, 1066.
4. International Business Machines Corporation, "System/360 Scientific Subroutine Package, Version III, Programmer's Manual, Program Number 360A-CM-03X" Manual GH20-0205-4 (Fifth Edition), IBM Corporation, White Plains, New York, August 1970.
5. International Business Machines Corporation, "Random Number Generation and Testing,, Manual G20-8011, IBM Corporation, White Plains, New York.
6. Peradejordi, F., Martin, A. N., and Cammarata, A., *J. Pharm. Sci.* (1971), 60, 576.
7. Gibbons, L. K., Koldenhoven, E. F., Nethery, R. E., Montgomery, R. E., and Purcell, W. P., *J. Agric. Food Chem.* (1976), 24, 203.
8. Timmermans, B. M. W. M., and van Zwieten, P. A., *J. Med. Chem.* (1977), 20, 1636.
9. Kier, L. B., and Hall, L. H., *J. Med. Chem.* (1977), 20, 1631.
10. Kier, L. B., and Hall, L. H., *J. Pharm. Sci.* (1978), 67, 1409.

RECEIVED June 8, 1979.

The Design of Transition State Analogs

P. R. ANDREWS

Department of Physical Biochemistry, The John Curtin School of Medical Research,
Australian National University, Canberra, A.C.T. 2600, Australia

The transition state of an enzymically catalyzed reaction is bound to the enzyme much more tightly than its substrate(s), and may therefore be used as a template for the design of potent and specific enzyme inhibitors (1,2,3,4). Such inhibitors are known as transition state analogues, and have considerable potential as highly selective drugs. They may, for example, be specifically designed to inhibit enzymes which are vital to abnormal or invading cells, but of no importance to the host. Alternatively they may take advantage of quantitative biochemical differences between normal and abnormal tissue.

Two major phases may be distinguished in the design of transition state analogues. They are (i) determination of the transition state structure, and (ii) specification of stable molecules which mimic the geometric and electronic properties of the transition state.

Determination of Transition State Structure

Unlike the substrate(s) and product(s) of a reaction, transition state structures cannot be studied directly, and the structures employed in analogue studies are commonly derived from intuitive interpolations between reactant(s) and product(s) on the basis of an assumed reaction mechanism. The intermediates obtained in this way may differ radically from the actual transition states.

Efforts to deduce transition state structures theoretically have until recently been retarded by the failure of even the more sophisticated molecular orbital treatments to predict accurate activation energies, and the need to avoid geometric and mechanistic assumptions has made the calculation of reaction pathways prohibitively expensive. The introduction of efficient gradient methods for minimizing energy with respect to all geometric parameters, coupled with the advent of faster computers, has now virtually overcome the latter problem, and careful parameterization of semiempirical molecular orbital methods has led to more

0-8412-0521-3/79/47-112-149\$05.00/0

© 1979 American Chemical Society

accurate calculated energies. The use of computers to determine the structures of transition states for enzyme catalyzed reactions has thus become a realistic possibility.

The Reaction Coordinate. To calculate the reaction pathway, an internal coordinate which varies monotonically from reactant(s) to product(s) is chosen as the reaction coordinate, e.g. the length of a making or breaking bond may be used. The energies at various points along the reaction pathway are then obtained by constraining the reaction coordinate to appropriate fixed values while minimizing the energy with respect to all other geometric variables. The high point on the reaction pathway thus obtained provides an estimate of the location and structure of the transition state, which may then be precisely located and identified by minimizing the energy gradient with respect to all geometric variables, including the reaction coordinate (5). Because of the complexity of many reaction surfaces it may sometimes be necessary to calculate reaction pathways for several starting geometries and reaction coordinates; a number of alternative transition state structures may thus be located.

The *lowest* energy transition state obtained in this way is appropriate to the gas phase reaction, and does not necessarily correspond to the transient species for the reaction in solution. Similarly, there is no valid justification for assuming that the pathway catalyzed by the enzyme is the calculated minimum energy reaction pathway. There is, however, good reason to suppose that one of the calculated low energy reaction pathways will be that catalyzed by the enzyme. The calculations may thus provide two or more alternative reference reactions with their corresponding transition states, one of which will be selectively stabilized at the active site of the enzyme.

Calculation Methods. Of the alternative molecular orbital methods available, the MINDO/3 method (6) has proved especially successful for calculating reaction pathways (7). Dewar estimates that the relative energies of alternative transition states are predicted to within 5 kcal/mol, and that the structures of transition states are probably reproduced to a few hundredths of an Angstrom in bond lengths, and a few degrees in angles (7). Unresolved differences between the energies and geometries of the transition state structures obtained for bimolecular reactions using *ab initio*, MINDO/3 and other techniques (8) suggest that any molecular orbital method must still be applied with caution, but provided that alternative reaction pathways within 5-10 kcal/mol of the minimum energy pathway are considered, it seems likely that an accurate representation of the transition state structure for an enzyme catalyzed reaction will be obtained using MINDO/3. Versions of MINDO/3 suitable for various computers are available from the Quantum Chemistry Program Exchange (9,10,11).

For reactions involving large molecules, the time required to compute a reaction pathway with MINDO/3 remains prohibitive. The development of reliable empirical methods for preliminary studies of reaction surfaces will be necessary to alleviate this problem.

Design of Stable Analogues

The long partial bonds and abnormal hybridization of transition state structures preclude precise correspondence between the functional groups of an analogue and those of the transition state, but a range of plausible analogues can usually be derived by inspection of the transition state structure; additional alternatives may be obtained from chemical or crystal structure files. Choice of the best analogues for synthesis is then based on a comparison of their molecular structures with that of the transition state. For this purpose, computed transition state structures and other points on the reaction pathway may conveniently be stored in a molecule library file, either as cartesian coordinates or as geometric variables; it may also be convenient to store a connectivity matrix, since the presence of partial bonds in the transition state structure may not be self-evident from the interatomic distances.

Molecular Manipulation and Superimposition. To facilitate molecular comparisons, a variety of computer graphic techniques are available for three-dimensional manipulation and display of the stored structures in the library file (12,13). In our laboratory lateral stereo pair views of either single line or ball and stick models are displayed on a Tektronix 4014 graphics terminal linked to a Univac 1100/42 computer.

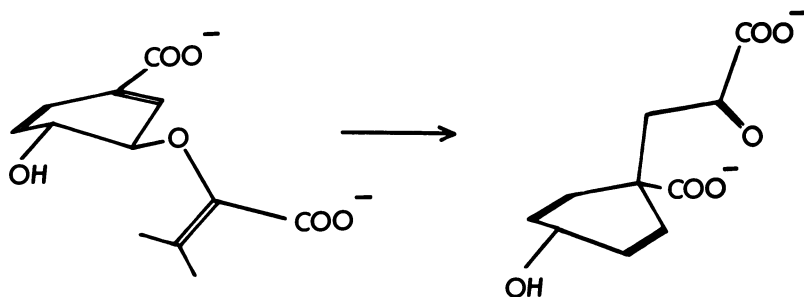
The three-dimensional images of the analogues, derived using standard geometric parameters (14,15), are superimposed on that of the transition state structure using interactive graphics routines to manipulate either molecule. We automate this process by running a stepwise minimization of the extent of misfit between the two molecules. Various functions may be used for the minimization, which is carried out with respect to the six degrees of intermolecular rotational and translational freedom as well as appropriate conformational variables in the analogue molecule. The simplest function appears to be the sum of the squares of the distances between corresponding atoms in the two molecules. In our hands this function has proved most useful when hydrogen atoms are excluded, isoelectric groups are equated, and atoms separated by more than a given distance (e.g. 1 Å) are arbitrarily assigned a separation of that distance.

Analogues which optimally match the transition state structure are chosen for synthesis, which may itself be computer assisted (Wipke, this Symposium). Inhibitory activities against the enzyme may then be measured *in vitro* and/or *in vivo*, and

structure-activity relationships determined. Design of further analogues may then prove desirable on the basis of these data.

Example : Inhibition of Chorismate Mutase

The Claisen rearrangement of chorismate to prephenate, catalyzed by chorismate mutase, is illustrated below. It



is the first specific step in the biosynthetic pathways leading to tyrosine and phenylalanine in bacteria and other organisms (16). The enzyme is absent in man and thus provides a potential target for bacteriostatic action, although this potential is limited by the ability of bacteria to scavenge aromatic amino acids in the bloodstream. The enzyme provides a clear example of the design principles described above, illustrating both advantages and limitations of most of the techniques discussed.

Transition State Calculation. MINDO/3 calculations have been used to describe the reaction surface for the isomerization of chorismate in both the neutral and dianionic forms (17). Two alternative structures were obtained in each case (Figure 1). One, which results from clockwise rotation of the sidechain around the breaking C-O bond, adopts a chair-like configuration; the other results from anticlockwise rotation, which leads to a boat-like structure. Both are asymmetric structures in which the breaking C-O bond (ca. 1.45 Å) is significantly shorter than the making C-C bond (ca. 1.95 Å).

The enthalpy difference between these alternative reaction pathways (<2 kcal/mol) is insufficient to define a preferred transition state structure. Thus, no firm inference regarding the conformation of the intermediate in the enzyme catalyzed reaction may be drawn on the basis of the molecular orbital calculations alone. In principle, the two pathways should differ

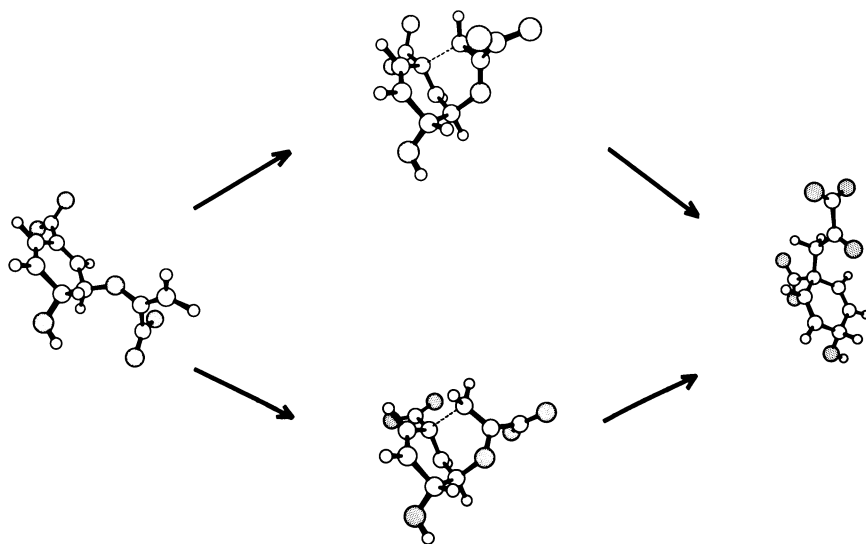


Figure 1. The isomerization of chorismate to prephenate via the chair-like transition state (upper pathway) or the boat-like transition state (lower pathway). The molecules are shown in the lowest energy conformations obtained from MINDO/3 molecular orbital calculations (17). Oxygen atoms are shaded.

in the stereochemistry of the products obtained from stereospecifically labelled chorismate, but efforts to prepare this starting material have not been successful (18). Accordingly, we set out to solve the problem by designing a structural analogue for each of the alternative transition state conformations, and testing these as inhibitors of the enzyme (19). The rationale behind this approach is that the inhibitor resembling the true transition state should produce the strong inhibition characteristic of transition state analogues, whereas that resembling the alternative transition state should not.

Analogues of Alternative Transition States. The chair- and boat-like transition state structures are shown diagrammatically in Figure 2. They illustrate two problems of transition state analogue design. First, the partial bonds in transition state structures give rise to bond lengths and hybridization states which cannot be reproduced in stable molecules. Second, the number and variety of chemical features in transition state structures may be incompatible with synthetic feasibility. Bearing these limitations in mind, we chose *exo*-6-hydroxybicyclo (3.3.1) nonane-1,*exo*-3-dicarboxylic acid (Figure 2c) and *exo*-6-hydroxybicyclo (3.3.1) nonane-1,*endo*-3-dicarboxylic acid (Figure 2d) as the closest practicable analogues of the chair- and boat-like transition state structures, respectively. Although differing from each other only in their stereochemistry around a single carbon atom, they adopt the chair-chair and chair-boat conformations respectively (19), and thus closely reproduce the topographical differences between the alternative transition state structures. When synthesized and tested as inhibitors of chorismate mutase-prephenate dehydrogenase from *Escherichia coli* K12, the analogue of the chair-like transition state was considerably more potent (IC₅₀ 0.4 mM) than that of the boat-like intermediate (no inhibition at 2.5 mM). The inhibition due to the *exo*-COOH nonane is not, however, as strong as expected for an ideal transition state analogue, and it could be argued that an energetically less favoured, and consequently less populated, conformation of the nonane may produce the observed inhibition by substantially tighter binding. To test this possibility an additional analogue, 6-hydroxyadamantane-1,3-dicarboxylic acid (Figure 2e), in which the chair-chair conformation is fixed by the addition of a bridging methylene group, was prepared (19). Inhibition by this compound (IC₅₀ 0.65 mM) was very similar to that of the *exo*-COOH nonane, indicating that the stable chair-chair conformation of the nonane is responsible for the observed inhibition. These data support the view that the chair-chair backbone provides a good topographical fit to the active site, and that the rearrangement of chorismate to prephenate catalyzed by chorismate mutase-prephenate dehydrogenase proceeds via the chair-like intermediate.

Structure-Activity Relations. The rate of chorismate isomerization is enhanced approximately two million-fold by chorismate mutase (20). In the absence of substrate destabilization effects (21,22), an ideal transition state analogue might therefore bind the active site some 10^6 times more strongly than chorismate. The inhibitors above, although more potent than any previously available, do not approach these levels of inhibition. The probable explanation for this is apparent in the superimposition of the *exo*-COOH nonane and the chair-like transition state structure, given in Figure 3. The overlap of the backbone atoms and the substituent groups in the two molecules is excellent for all except the side-chain carboxyl group (arrowed), for which the abnormal hybridization in the transition state structure results in a significantly different orientation to that of the analogue.

This explanation is supported by the inhibitory activities of the series of adamantanes shown in Figure 4. The simplest compound, adamantane-1-carboxylic acid, is a potent inhibitor, but introduction of a second carboxyl group results in substantially weaker inhibition. Activity is partially restored by the replacement of either or both carboxyl groups with acetic acid functions, in which the additional methylene groups may serve to reduce repulsive interactions. The inhibitory activity of these analogues thus emphasizes the importance of correct substituent orientation in the design of transition state analogues.

Unsolved Problems

Literature Retrieval of Likely Analogues. The preceding example also serves to illustrate a problem which may apply generally to computer-based searches in the field of *ab initio* drug design. Early in this study, it was hoped to minimize synthetic effort by locating possible analogues for the transition state structures in the recent literature. For this purpose, a simple chemical family or substructure search is not sufficient; rather, several substructures connected by geometrically (rather than chemically) defined functional groups are required. The problem may thus be posed as shown in Figure 5, where each pair of parentheses encloses a series of optional groupings in order of preference. Some 1300 different chemicals, and a rough order of priority, are defined by this figure, which cannot apparently be handled by standard computer-based chemical searches. Development of search techniques for this class of problem would clearly be valuable.

Quantitative Measures of Steric Correspondence. The comparison of alternative analogues with a transition state structure, prior to synthesis, would be further facilitated if a quantitative method were available for estimating the steric and electronic correspondence between the analogues and the calculated transition state structure. The functions discussed above for optimising

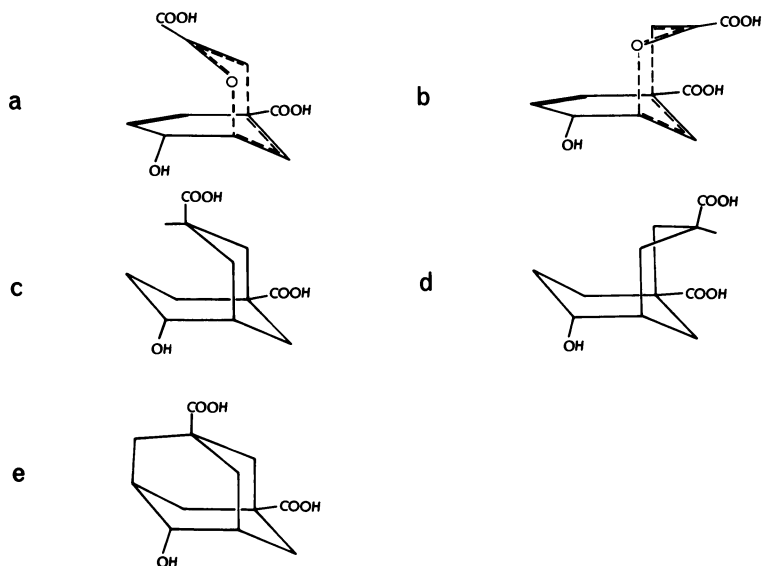


Figure 2. Diagrammatic representations of the alternative transition state structures and some proposed analogs: (a) chair-like transition state structure; (b) boat-like transition state structure; (c) *exo*-6-hydroxybicyclo (3.3.1) nonane-1,*exo*-3-dicarboxylic acid; (d) *exo*-6-hydroxybicyclo (3.3.1) nonane-1,*endo*-3-dicarboxylic acid; (e) 6-hydroxyadamantane-1,3-dicarboxylic acid.

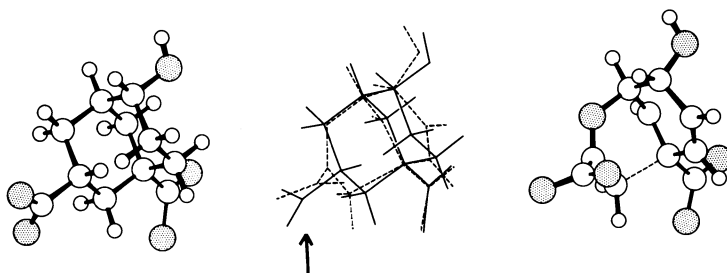


Figure 3. Molecular superimposition of the chair-like transition state (left, ---) and its analog, *exo*-6-hydroxybicyclo (3.3.1) nonane-1,*exo*-3-dicarboxylic acid (right, —). Oxygen atoms are shaded.

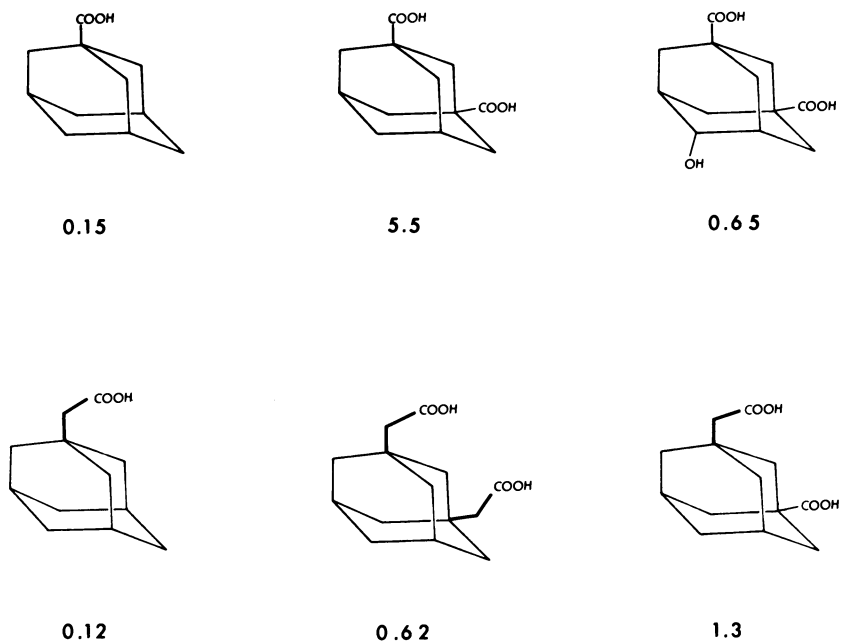


Figure 4. Structures and inhibitory activities of some adamantane analogs of the chair-like transition state

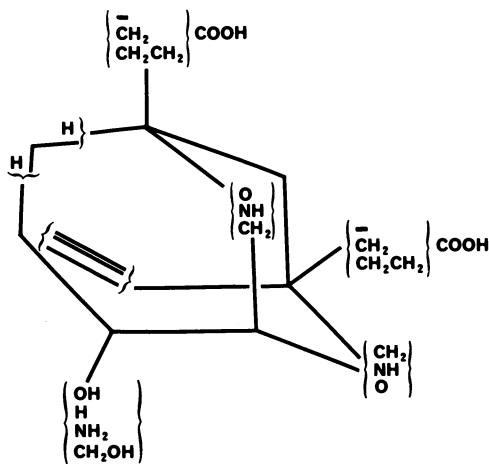


Figure 5. Structural description of 1296 possible analogs of the chair-like transition state

molecular superimposition, possibly supplemented with an allowance for missing or superfluous atoms, could conceivably be used for this purpose. Difficulties arise, however, in determining the relative importance of shape, charge distribution and substituent moieties, and the presence or absence of necessary or superfluous atoms or groups. Several alternative functions were applied to the series of inhibitors treated here, but none reflected the observed inhibitory activities, although the basis for these was self-evident on inspection of the molecular superimpositions.

Conclusions

The present data illustrate the importance of correct substituent orientation in transition state analogue design. In many cases, however, correct alignment will be precluded in stable molecules by the nature of the partial bonds in the transition state structure. For this reason, it may be wise to favour multisubstrate enzymes in analogue design; the additional substituent groups thus provided may permit some leeway in orientation without prejudicing tight binding. In these cases the computer will again provide a means for determining the transition state structure, as well as for rapidly testing a range of possible analogues for steric and electronic correspondence with the transition state.

Literature Cited

1. Pauling, L., Amer. Scientist (1948) 36, 51.
2. Jencks, W.P., "Current Aspects of Biochemical Energetics". p.273, ed. N.O. Kaplan, E.P. Kennedy, Academic Press, New York, 1966.
3. Wolfenden, R., Acc. Chem. Res. (1972) 5, 10.
4. Lienhard, G.E., Science (1973) 180, 149.
5. McIver, J.W. and Komornicki, A., J. Am. Chem. Soc. (1972) 94, 2625.
6. Bingham, R.C., Dewar, M.J.S. and Lo, D.H., J. Am. Chem. Soc. (1975) 97, 1285.
7. Dewar, M.J.S., Science (1975) 187, 1037.
8. Caramella, P., Houk, K.N. and Domelsmith, L.N., J. Am. Chem. Soc. (1977) 99, 4511.
9. Dewar, M.J.S., Metiu, H., Student, P.J., Brown, A., Bingham, R.C., Lo, D.H., Ramsden, C.A., Kollmar, H., Weiner, P. and Bischof, P.K., QCPE (1975), 279.
10. Dewar, M.J.S., Metiu, H., Student, P.J., Brown, A., Bingham, R.C., Lo, D.H., Ramsden, C.A., Kollmar, H., Weiner, P., Bischof, P.K. and Parr, W.J.E., QCPE (1976), 308.
11. Dewar, M.J.S., Metiu, H., Student, P.J., Brown, A., Bingham, R.C., Lo, D.H., Ramsden, C.A., Kollmar, H., Weiner, P., Bischof, P.K., Olson, M.L. and Chiang, J.F., QCPE (1976), 309.

12. Barry, C.D., Ellis, R.A., Graesser, S.M. and Marshall, G.R., "Pertinent Concepts in Computer Graphics", p.104, ed. M. Fairman, J. Nievergeld, Univ. of Illinois Press, Urban, 1969.
13. Perkins, W.J., Piper, E.A. and Thornton, J., Comput Biol. Med. (1976) 6, 23.
14. Sutton, L.E., "Tables of Interatomic Distances", Chemical Society, London, 1965.
15. Gordon, M.S. and Pople, J.A., QCPE (1969), 135.
16. Gibson, F. and Pittard, J., Bacteriol. Rev. (1968) 32, 465.
17. Andrews, P.R. and Haddon, R.C., submitted.
18. Ife, R.J., Ball, L.F., Lowe, P. and Haslam, E., J. Chem. Soc., Perkin Trans. 1 (1976), 1776.
19. Andrews, P.R., Cain, E.N., Rizzardo, E. and Smith, G.D., Biochemistry (1977) 16, 4848.
20. Andrews, P.R., Smith, G.D. and Young, I.G., Biochemistry (1973) 12, 3492.
21. Wolfenden, R., Ann. Rev. Biophys. Eng. (1976) 5, 271.
22. Jencks, W.P., Advances in Enzymology (1975) 43, 219.

RECEIVED June 8, 1979.

The Molecular Basis of Structure-Activity Relationships: Quantum Chemical Recognition Mechanisms in Drug-Receptor Interactions

HAREL WEINSTEIN, ROMAN OSMAN, and JACK PETER GREEN

Department of Pharmacology, Mount Sinai School of Medicine of
The City University of New York, New York, NY 10029

The analysis of the relation between the molecular structure of drugs and their biological effects focuses on the structural characteristics of the drug that are recognized by receptors and on the possible molecular mechanisms that can lead to the tissue response. The basic assumption is that the interaction between drugs and receptors is dependent on the same molecular parameters that determine chemical interactions and reactions (1). On that basis, a description of drug-receptor interactions and mechanisms should follow the formal patterns of chemistry. Clearly the special parameters of biomolecular interactions, such as size and complexity of environment, determine a special class for these phenomena. Theoretical considerations as well as documented success in the study of biological interactions indicate, however, that these complex phenomena can be decomposed into physically meaningful steps that are amenable to analysis by methods of theoretical and quantum chemistry. A schematic representation of one possible form of the causal link between molecular structure and an observable biological response is given in Figure 1. From the physical and chemical basis of molecular interactions it is possible to recognize the elements that would lead to the important step of RECEPTOR RECOGNITION (defined arbitrarily as an intermediate step before the actual drug-receptor interaction). Also, from the vast literature dealing with the relation between the molecular structure of drugs and their biological actions, it is possible to identify alternative ways for the identification of molecular reactivity characteristics that are necessary for the recognition step, e.g., Quantitative Structure Activity Relationships (QSAR, see article by Osman et al. in this volume (2)). These characteristics are not necessarily determinant of the drug-receptor binding energy. For example, it has been shown that in the complex phenomenon of protein-protein interaction, hydrophobicity is the major factor stabilizing protein association mainly due to the very large number of protein-water interactions replaced by the formation of a protein-protein interface (3). Nevertheless, it was shown that the specificity

0-8412-0521-3/79/47-112-161\$06.75/0

© 1979 American Chemical Society

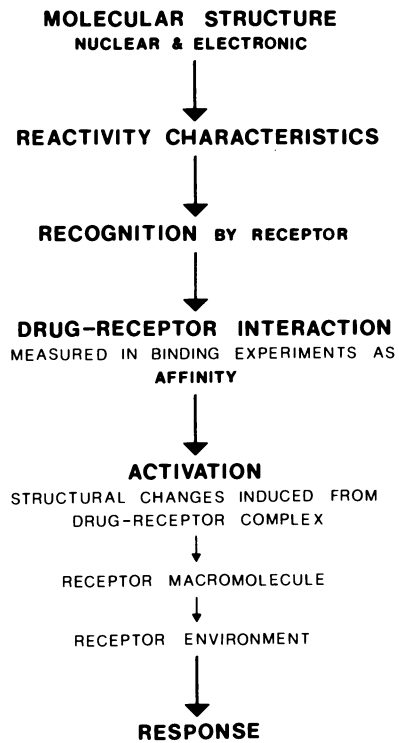


Figure 1. Postulated link between the molecular structure of a drug and its biological activity

of the interaction is entirely dependent on the complementarity of the interfacing protein surfaces (3). The onset of the stabilizing bulk effect of the solvent is thus dependent on the discrete interactions between properly positioned residues. That crucial part of the interactions is obviously dependent on the same factors that govern interactions of small molecules, and the selectivity of the protein-protein interaction could therefore be described by the formalism of chemical reactivity. Studies in which the overall interactions are treated as a superposition of discrete physicochemical components have been documented (4), and specifically for enzyme-substrate interactions (e.g., lysozyme (5)).

In the description of chemical reactivity characteristics, it is important to obtain a representation of the molecular properties as a function of the structure of the entire molecule. Yet, much of the work on structure activity relationships in biological systems focused almost exclusively on the independent contribution of certain structural groups ("functional groups") to the ability of the molecules to interact with putative receptors (2). Very often, this approach carried the tacit assumption that functional groups necessarily affect specific molecular sites that interact directly with the receptor, that the modification of one group does not affect the reactivity of another, and consequently that geometrical and stereometrical factors such as the distance and spatial relationship between functional groups is all important (2). This assumption is especially engrained in the philosophy of drug design. In fact even a simple substitution such as a hydroxyl at the 5-position of tryptamine, markedly alters the electronic properties and thereby the reactivity pattern of the molecule so that, for example, the magnetic circular dichroism spectrum of 5-hydroxytryptamine (5-HT) in solution is markedly different from that of tryptamine (6). Moreover, a nucleotide-serotonin complex in solution has a geometry different from that of the same nucleotide with either tryptamine or 6-hydroxytryptamine (6-HT) and the geometry of self-association also differs for the tryptamine congeners (7, 8, 9). It is noteworthy that no direct interaction with the hydroxyl is observed in these studies. These experimental findings are in agreement with early theoretical structure-activity studies (that were mainly based on semi-empirical quantum mechanical calculations) showing that differences in the pharmacological potency of tryptamine derivatives are attributable to changes produced by substitution at different sites in the electronic structure of the entire indole ring (10, 11, 12). More recent theoretical work on these systems explains the relation between the quantum chemical reactivity characteristics and the preferred orientations of the various tryptamine derivatives in polarization ("charge transfer") complexes (13, 14).

This example demonstrates the need for quantum chemical methods that can describe the fundamental forces that control

molecular mechanisms on the basis of interactions between the electrons and nuclei of the reactants. During the very recent past, 5 to 7 years, quantum chemical calculation methods have developed to the point that they can produce good relative values of physicochemical measurements that can be determined by experiment either with great difficulty or only by inference. These include multipole moments, molecular polarizabilities, ionization potentials, electron affinities, charge distributions, scattering potentials, various spectra, transition state geometries and energies, and the relative likelihood of the existence of various conformations of molecules.

1. The prediction of physicochemical properties

The success of the application of quantum mechanical methods to the description of physical and chemical properties of molecules is evident in every facet of modern chemical research and is well documented. It may be of special interest to point out some of the very recent uses of quantum chemical approaches applied as a tool in the study of molecular structure and interactions. The examples below were not selected for their relative importance in the field of quantum chemistry nor for the use of the most advanced methodology. Each represents one of many illustrations of a specific application.

One of the main advantages of theoretical studies with quantum mechanical methods is the ability to predict processes and mechanisms of interaction with the use of model systems that cannot readily be obtained experimentally but are amenable to experimentation once the most probable mechanism is established. Pople, Schleyer and their collaborators have recently predicted from quantum chemical calculations the existence of planar methane derivatives, and of perpendicular olefins (15, 16). The far reaching implications of this work to organic chemistry in general, and to the quantum chemical approach have been subsequently discussed (17) and the predictions have been reconsidered with more advanced and more accurate methods (18). Others have shown in a comparison of results from quantum chemical calculations with rotational and vibrational spectroscopy that "recent progress in molecular orbital methods promises the potential not only for improving our understanding of spectral features and their relationship to molecular conformation but for providing a closer examination of the role of various intramolecular forces governing conformational stability" (19). The work by Borden, Davidson and Hart (20) supports such conclusions, and indicates the importance of the insight gained from calculations into the complex mechanism of structural transitions in excited states.

2. Theoretical models for biological mechanisms

The ability of the quantum chemical approaches to analyze, interpret and rationalize all experimental observations on molecular processes in a unified language based on the primary principles of physics, represents an important asset in the study of biological systems where experimental data has to be obtained in a variety of forms and from many different sources. Taking advantage of this formal and analytic power of the quantum chemical approaches, Lipscomb and his coworkers studied models of enzymatic mechanisms in chymotrypsin (21) and carboxypeptidase (22). These models reveal the contribution of the various groups in the active site to substrate binding and to the catalytic activity (see also (23)). For example, Scheiner and Lipscomb showed that the presence of an electrophile in the vicinity of the carbonyl group enhances hydrolysis by eliminating the energy barrier calculated in the absence of the electrophile (22). The theoretical results also showed how the electrophile stabilizes the tetrahedral configuration around the carbon prior to dissociation, and indicated that the reaction is thermoneutral and proceeds with little or no activation energy. Such details contribute to the critical evaluation of existing hypotheses about the mechanism of action of the enzyme.

Also on the basis of quantum mechanical considerations, Goddard and collaborators very recently proposed a mechanism for the function of flavin coenzymes (24). In this model, the flavin molecule interacts with the triplet (ground) state of molecular oxygen forming a biradical. In this intermediate the chemical reactivity of the bound oxygen molecule resembles that of singlet oxygen, which unlike triplet oxygen is very reactive towards double bonds. Goddard and his coworkers showed that the formation of the intermediate requires less energy than would be expected, due to the electronic structure of flavins. This first step represents an attractive mechanism for the formation of the "activated oxygen" that has been postulated repeatedly for certain biochemical reactions. Although the mechanism remains to be verified experimentally, theory has already indicated some interesting directions of research such as the prediction of an intermediate "supermolecule" composed of the flavin and the substrate (e.g., phenol) linked through an oxygen molecule. This is another example of the direct advantage offered by the quantum chemical calculations by which hypothetical systems are studied and their properties predicted for further experimental verification.

In fact, theoretical studies on large biological systems have already generated important predictions that were specifically tested and confirmed by experiments. For example, Kirchner and Loew had predicted from quantum chemical considerations that the oxyheme geometry is bent and totally spin paired (25). Experimental studies had been unable to reach a firm conclusion regarding the geometry of the oxygen with respect to the heme due, in

part, to seemingly conflicting data obtained by different spectroscopical methods. Recently, Churg and Makinen (26) stated that they were able to reach useful conclusions from the optical spectrum of oxyheme (specifically polarized single crystal electronic absorption) primarily by following the features indicated by the theoretical analysis in (25). Moreover, the authors point out that the final conclusion from the experiment is reached on the basis of the correlation of the observed transition energies and polarizations of the charge-transfer states with predictions based on the quantum chemical calculations of Kirchner and Loew (25).

In light of these impressive achievements in specific areas of study, and in view of the fact that results in almost every field of chemistry are described and analyzed in terms of concepts introduced by quantum chemistry, it is not surprising to find that quantum chemical methods are being increasingly applied to the study of properties and mechanisms of action of drugs and related substances (for some recent reviews see - (27-32)). In what may seem today an unfortunate choice, many theoretical studies on drug molecules have concentrated almost exclusively on the calculation of conformation in an attempt to predict the most probable geometry of the drug at the receptor and the relation between conformation and activity. With very few exceptions, notably the calculation of acetylcholine by a classical-quantum chemical method (33), the total energy of the molecules is calculated in these approaches for a small number of variable parameters, with most bond lengths and bond angles fixed at their initial values. Unfortunately, freezing most of the molecule in its initial geometry may influence the outcome of the calculation. For example, an *ab-initio* calculation of acetylcholine (ACh) using the crystal geometry of AChCl showed a *gauche* form to be 3.2 Kcal/mole more stable than the *trans* form, whereas with a standard geometry the *trans* form was 3.4 Kcal/mole more stable than the *gauche* (34). Analogously, PCILO calculations of ACh based on crystal geometries of AChCl and AChBr produced different preferred conformers (35). Such ambiguities in the elucidation of the most stable conformation are not peculiar to the theoretical methods. Crystal structures, for example, are influenced by bulk factors (such as the free energy of packing) and by intermolecular forces that prefer certain arrangements in the unit cell. This situation is not necessarily discouraging if it is recognized that the question of the absolute minimal energy conformer is not the most important one in studying biological interactions. The interactions of the drug with its environment may stabilize other allowed conformations that are only a few Kcal less stable than the optimal one. Moreover, a certain degree of flexibility may even be a requirement for agonist activity (36, 37, 39). On the other hand, theoretical methods for the calculation of molecular conformation have the advantage of being able to describe the entire function of energy dependence on conformation. This reveals the more

important information on the relative likelihood of occurrence of all possible conformers. It must be stressed that these theoretical methods depend on the experimental data for calibration and final validation.

In our own theoretical studies (e.g., 1, 12, 13, 14, 38), we have regarded the conformational data obtained from quantum mechanical calculations as a starting point rather than an end. By indicating which conformations should be considered for further work, they simplify the problem of defining the reactive regions of the molecule and the sites of likely intermolecular interaction. Whatever ambiguities evolve from theoretical studies on conformation can often be eliminated or at least reduced by studying a series of similar compounds and by the judicious use of experimental data from x-ray and NMR studies as well as pharmacological data from studies of rigid analogs (e.g., 39). It has to be stressed that good theoretical calculations on molecular structure provide more than conformational data: the electronic structure in each conformation of the molecule can be obtained from the corresponding wave function and the chemical behavior of the drugs can be studied from the changes in this electronic structure upon interaction with model reagents such as models for receptor sites. On this basis, the wealth of methods and formulations developed in quantum chemistry, which were shown to provide reliable predictions and understanding of chemical reactivity, can be directly used to study drug receptor interaction (1).

3. Theoretical models and the effect of the environment

A common criticism of quantum chemical approaches is that however successful quantum mechanics may be in explaining and predicting elements of chemical reactivity, the calculations for biological systems still contain the assumption that molecules are in a vacuum, whereas in reality they are in solution. One can, however, list many chemical reactions that also take place in solvents and that quantum chemistry has been able to treat and predict successfully. This approach is successful especially in cases in which theory is called upon to compare the molecular determinants for the interactions of various groups of compounds and conclude from a comparison of the calculated properties of the molecules and from their comparative behavior what the nature of the interaction is, and what the most probable consequences of these interactions will be. Such structural determinants of the molecular events should be relatively insensitive to the effects of the medium, much more so than are precise energetic balances. In fact, every textbook mechanism offered to describe the binding of substrates to enzymes (e.g., chymotrypsin, cholinesterase, etc.) has no bulk solvent interposed between the interacting species. Almost without exception, molecular mechanisms offered as a result of biochemical studies depend on information directly derived from

the chemical reactivity of the reactants (albeit often from "classical" chemistry considerations). The reaction of drugs with tissue components must follow the same rules.

This is not to say that the specific effect of the environment (water, protein, phospholipid) cannot be considered by theoretical methods if the goal is to obtain quantitatively correct estimates of energy changes. For example, Warshel and Levitt have presented a method by which the full enzyme-substrate complex can be considered with the surrounding solvent (5). The method combines quantum mechanical calculations of energies for bond cleavage and charge redistribution with classical calculations of the "solvation energy" resulting from the electrostatic polarization of the enzyme in the field of the substrate and from the reorientation of the dipoles of the surrounding water in this field (see also (40) for the microscopic model of dipole reorientation and the effect on chemical processes). Others have attempted to learn about specific solvent effects on molecular conformation and on reactivity by using the "supermolecule" approach, i.e., by calculating the system composed of the molecule of interest in the presence of a number of water molecules (for some reviews see (41, 42)). More intriguing, however, are the new approaches by Clementi and coworkers (43, 44, 45), and by Beveridge and coworkers (46, 47) - see also (48, 49) - in which a statistical (Monte Carlo) approach is used to determine the structure of the solvent interacting with a solute on a potential surface that has been obtained from good quality quantum chemical calculations. So far, these methods have been used mainly to study the distribution of water around discrete solutes. Interesting results were obtained for the solvation of formaldehyde (46), and glycine (44) and even for the interaction of water with lysozyme (42), at temperatures other than $T=0$. Such approaches, and other statistical procedures based on potential surfaces calculated mainly by quantum mechanical methods (42, 43, 46, 47, 50), or on electrostatic models of effective solvent fields (51, 52, 53), supplement the earlier methods. Those were developed for the treatment of intermolecular forces in liquids (54) and for the description of the properties of small clusters of liquid water (55, 56, 57). The new theoretical approaches have become increasingly powerful and are able now to address both the microscopic and the macroscopic solvent effects (58, 59). The development of methods has not yet reached the stage where the thermodynamics of solvent displacement related to drug-receptor interactions can be modelled accurately, but the inherent problems are well recognized and most of them are understood. Meanwhile, well defined approximations are used to study solvent effects on drug conformation (60, 61, 62, 63, 64). For example, we have described a "differential solvation" approach based on supermolecule calculations that yielded excellent results and predicted correctly the effects of solvent on the comparative pK_a 's of structurally related drugs such as atropine and scopolamine (64) or hydroxymorphone and

oxymorphone (63). This enabled us to understand structural determinants for the comparative effectiveness of these drugs (64).

Once an understanding of the interaction components is attained, the specific drug-receptor recognition step can be singled out in the theoretical approach and monitored at the level of chemical reactivity. Parameters related to the reactivity of biologically active molecules may be sought from the molecular wave functions and from the calculated properties, providing insight into the likely molecular mechanisms of direct drug receptor interactions and to the biological implications of such interactions. As with experimental studies, theoretical studies carry the risk that model results may be extrapolated too far and the text of resulting hypotheses may be hampered by the inherent ambiguity of crude biological correlations. Nevertheless, considerable evidence is at hand proving the relative success of theoretical approaches in elucidating specific mechanisms of drug action, in providing a basis for rational drug design, and in leading to conclusions about the nature of drug-receptor complexes and about the consequences of drug-receptor interactions (1, 12, 13, 14, 28, 29, 30, 31).

Yet more recently, new theoretical methods have been developed and applied to biologically active substances. This continuous development, combined with more rigorous quantum chemical methods for calculating the wave function, has provided new insight into the basic molecular events determining drug activity, and, at the same time, furnished a useful tool in drug design. In what follows, mainly our own work is offered to support these assertions, although other persuasive examples exist. In the following sections we describe and exemplify our theoretical approaches to the study of drug activity, and their relation to the experimental results from our studies on the action of hallucinogens, mainly LSD and tryptamine derivatives.

4. Methodological framework for quantum chemical studies on drug-receptor interactions.

Our theoretical studies on the molecular determinants for drug action are based on formal representations of the reactivity of drug molecules and the consequences of their interaction with other species. These characteristics are used to describe the elements of drug recognition at the receptor, and the molecular mechanism of the drug-receptor interaction that triggers the response. The identification of the structural and electronic determinants for the specific action of a drug can be divided into several interrelated stages:

- (i) the definition of the reactive sites of the drug and of their mutual relationship in space;

(ii) the identification of the possible and most probable orientation of these reactive sites towards reactants that represent molecular models of receptor sites; and

(iii) the elucidation of the response (at the molecular and electronic level) of the drug molecule and of the receptor models, to the perturbation caused by the environment and by their interaction in a simulated drug-receptor complex.

Clearly, the elucidation of the molecular determinants described in stage (i) and partly in stage (ii) is most intimately related to the first order level of understanding drug action and to the design of drugs. The more elaborate elements of stage (ii) and the simulations in stage (iii) are related to the elucidation of the molecular mechanism of drug action at the most fundamental level.

The study of stage (i) defined above required a method that could identify the most reactive sites in the molecule for a specific type of reactions, and could describe the nature of that reactivity in terms that are independent of the actual chemical identity (i.e., atomic composition) of the reactive site. The latter characteristic is particularly important in comparison of drugs that are structurally unrelated but that can act on the same receptors. A trivial example of a method for stage (i) is the identification of "hydrogen bonding sites" in a molecule, or even more specifically, its "proton donor" and "proton acceptor" sites.

We have developed a formal quantum mechanical approach to this problem by defining an "interaction pharmacophore" (1, 38) that is based essentially on the electric field created by a molecule in its surroundings (represented by the molecular electrostatic potential map (65, 66)). Since molecular systems represent a spatial distribution of electronic and nuclear charges (an averaged time independent distribution within the workable approximations of quantum mechanics), the approach was to seek a description of the molecular reactive sites and interaction ability based on the long-range intermolecular electrostatic forces. By describing the electrostatic potential field generated by the distribution of charge in the molecule, a "picture" could be expected representing the forces that the molecule can apply on another charge or on the continuous electron density distribution of another molecule. Regions of space in which the molecule generates negative potentials (i.e., where the contribution from all the electrons in the molecules is felt stronger than that from all its nuclei) should attract positively charged species or fragments. A positive region (where the nuclear contribution is stronger) would repulse positive, and attract negatively charged fragments. A molecular species approaching the investigated molecule should adopt an orientation that will bring its charged fragments into matching attractive interaction with the positive and

negative potential pattern. On this basis the calculation of molecular electrostatic potentials has been successful in first order predictions of the relative reactivity of functional groups in ionic-type reactions by the characterization of the shape and localization of regions representing attractive and repulsive zones of varying potency. The spatial disposition and relative strength of these regions form a pattern characteristic of a given molecule, in a certain geometric conformation, and can be analyzed to predict with some accuracy the reactivity and orientation of the whole molecule towards reagents of varying complexity. This feature, when added to the current reactivity criteria based on such quantities as net atomic charges, bond orders and bond polarities, bond polarizabilities and the elements of "frontier orbital" reactivity theory, constitute a potent theoretical means of defining stage (i) of the interaction in which we are interested (1, 12, 14).

An important advantage of the electrostatic potential is that unlike the common reactivity criteria, which are artificially constructed indices, the electric field generated by the molecule is an experimentally observable quantity (from high energy electron scattering). Notably, it has been shown that theoretical calculations generate electrostatic potentials that agree with the "static potential" component measured in scattering experiments (67, 68, 69). We have used the interaction pharmacophore for the study of drug action, e.g., phencyclidine (38, 70) narcotic analgesics (63, 71), hallucinogens e.g., (12, 13, 14).

Hopfinger has suggested that an "obvious reservation to the interaction pharmacophore concept is the change in the electrostatic potential about a drug because of the 'bound' water on the drug" (see p. 346-349 in (4)). Since it is difficult to assume that molecules could interact specifically "through" their respective hydration shell, Hopfinger's consideration would seem to negate any kind of specific interaction in solution - whether described by the electrostatic potential or any other descriptor. One therefore concludes that the interaction pharmacophore, like any other descriptor of "complementarity", or "matching site interaction", or "matching hydrogen bonding site" (such as between DNA base pairs), refers to the situation in which the outcome of the interaction is decided for (possibly instantaneously) "naked" sites. This does not assume total desolvation of the molecule but rather the formation of a common solvent-cage for the molecules interacting through specific sites. Even at this stage, the distances between the centers of interacting sites would still be large enough (e.g., $\geq 4\text{\AA}$) for the electrostatic approximation to remain valid, thus justifying the use of the electrostatic potential as a reactivity criterion. This would be the final and probably the decisive step in the interesting "path of unique interaction pharmacophores" suggested by Hopfinger (p. 348 of (4)). The attractive features of the electrostatic potential as a reactivity criterion have also been recognized by

others who have recently begun to use it in studies of drug-receptor and enzyme-substrate interactions (see (23, 72-76) for some examples).

It is obvious even from the most simple theoretical considerations that only the initial steps of the interaction can be represented by the interaction pharmacophore approximation. Other theoretical methods, designed to deal with stages (ii) and (iii) of the theoretical approach defined above, are continually being developed and applied selectively to appropriate problems of drug-receptor interactions: We have constructed a formalism to investigate the consequences of molecular interactions occurring simultaneously at several sites, by studying the perturbation of the reactivity and of the electronic structure following such interactions (77, 78). Applied to the study of multimolecular complexes, such as the simultaneous interaction of adenine with several water molecules (1, 79) it revealed the nature of a possible mechanism for "the preferential activation of (certain) secondary sites when a large molecule is already interacting at a primary active site" (1). We have discussed the importance of such reactivity induction mechanisms in drug-receptor interactions for muscarinic cholinergic agonists (36), and antagonists (70), and also for the mechanism of activation of the histamine H₂-receptor (80). As a generalization of the physical assumptions on which we based this multiple perturbation procedure, we also developed the Interaction Field Modified Hamiltonian (IFMH) method (81). In the IFMH approach, the charge distribution and reactivity of a molecule (e.g., the receptor model or the drug) are recalculated under the effect of the electrostatic field generated by another molecule (e.g., the drug or the receptor model, respectively).

It is important to test continuously the generality of the methodological framework outlined above, and especially the ability of the theoretical approaches to identify the interaction characteristics towards an unknown receptor site from the reactivity properties of the drugs and from simulations of drug receptor interactions. To this end we study the mechanism of action of carboxypeptidase: the active site of this enzyme and its mode of interaction with a substrate is known in detail from crystallographic studies (82). We have therefore modeled the active site of carboxypeptidase by $[\text{Zn}(\text{NH}_3)_2\text{OH}]^+$ with coordinates taken from the crystal structure of the enzyme, and the enzyme-substrate complex (82). The structure of the active site model is shown in Figure 2. (The fourth ligand shown is a water molecule; it is not part of the active site model but indicates the model substrate-see below). The electrostatic potential generated by the active site model in a plane positioned halfway between the zinc in the active site and the oxygen of the substrate in the crystal structure of the complex (Figure 3) exhibits a very localized maximum (i.e., attractive to a nucleophilic ligand) directed towards the zinc atom. The nature of the reactivity of

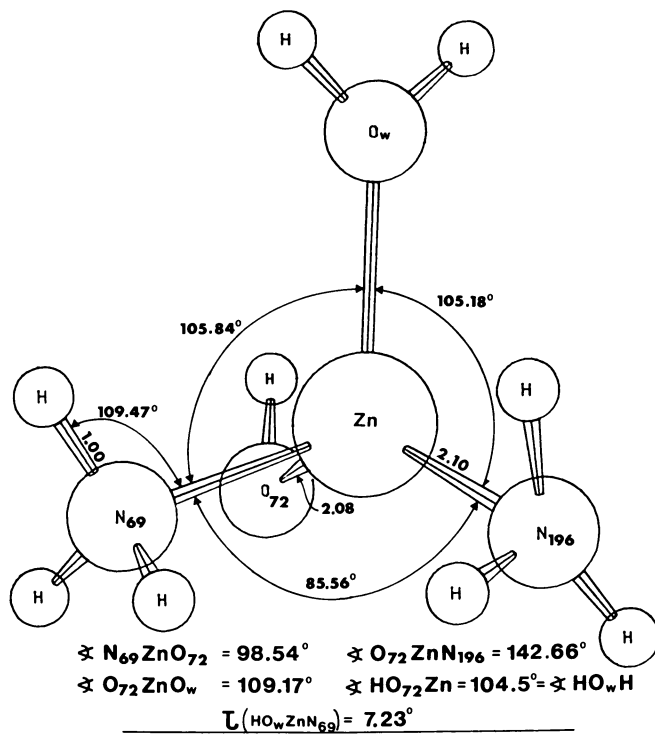


Figure 2. Geometry of the $[Zn(NH_3)_2OH]^+$ model for the active site of carboxypeptidase based on the crystal structure of the native enzyme (82). The interaction geometry with a water molecule corresponds to the coordinates from (82).

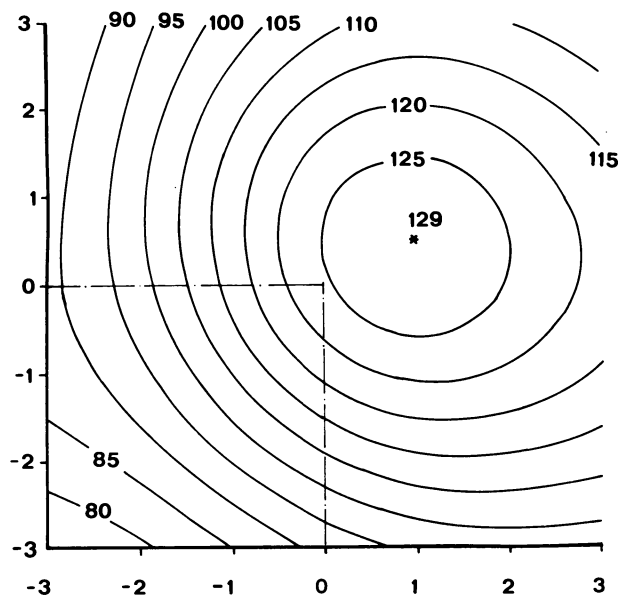


Figure 3. The molecular electrostatic potential generated by the $[\text{Zn}(\text{NH}_3)_2\text{OH}]^+$ model of the active site of the carboxypeptidase in a plane perpendicular to the Z-axis at 4 a.u. from the Zn.

The Z-axis forms an angle of 17° with the Zn-Ow bond shown in Figure 1. The position of Zn in the plane below [i.e., (0.0, 0.0, 0.0)] is indicated by dotted lines. The molecular wavefunctions were calculated with an ab-initio LCAO-SCF method using minimal STO-4G atom optimized basis sets. The basis set for Zn was augmented by additional diffused 3d and 4p functions obtained from an STO-4G expansion of Slater orbitals with exponents of 1.6575 and 1.45, respectively. The method used for the calculation of the electrostatic potentials is described in Ref. 65. Units are Kcal/mol.

the active site model as well as the main geometrical feature of the interaction were therefore well characterized by the "interaction pharmacophore" of the active site in that plane.

Quantum chemical methods for the study of stages (ii) and (iii) of the conceptual framework (see above) were applied to this problem by carrying out a simulation of the interaction between the active site model and a model substrate, formamide. It is one of the tenets of proposed mechanisms for carboxypeptidase activity that, as a consequence of the interaction with the active site, the carbonyl carbon of the substrate is activated for a nucleophilic attack (82). The geometry of the interaction was chosen according to the crystallographic structure of the enzyme-substrate complex (82). From the simulation we observed a large change in the electronic density in the vicinity of the carbon atom of formamide induced by the interaction with the active site model. This redistribution of the electron charge density causes a change in the pattern of the electrostatic potential surrounding the formamide molecule. Thus, in the simulated "substrate-active site complex", the formamide molecule is surrounded by a positive electrostatic potential with a strong maximum localized above the carbon atom (see Figure 4). This finding illustrates the ability of this theoretical approach to identify the creation of new reactive sites as a result of the response of the system to the interaction (see description of stages (ii) and (iii) above).

It thus appears that quantum chemical methods can successfully predict and characterize electronic mechanisms in substrate - enzyme interactions on the basis of the molecular reactivities of the separated entities, and from results of simulated interactions between molecular models - as shown by the following conclusions:

- a. the reactive sites were identified [stage (i)] from the "interaction pharmacophore" (i.e., the pattern of the electrostatic potential maps);
- b. the most probable mutual orientation of reactants [stage (ii)] was predicted by the electrostatic potential to have the geometrical orientation observed in the crystal structure of the "drug-receptor" (enzyme-substrate) complex;
- c. the "response of the substrate molecule to the interaction with the "receptor" [stage (iii)] was shown by the electrostatic potential of the perturbed molecule to match the experimental evidence for the activation of the carbonyl carbon for nucleophilic attack.

Similarly, these methods can be used to gain insight into the molecular basis for recognition and activation in drug-receptor interactions. In the example below (section 6) we discuss the

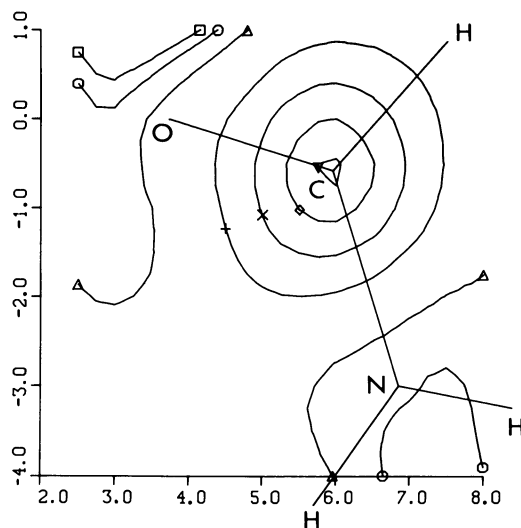


Figure 4. The molecular electrostatic potential generated by the $[\text{Zn}(\text{NH}_3)_2\text{OH} \cdot \text{HCONH}_2]^+$ complex, simulating the interaction of the active site of carboxypeptidase with a peptide substrate; geometries are from (82). The plane is parallel to the molecular plane of formamide at a distance of 1 Å. See legend to Figure 3 for computational details. (\square) 95.0; (\circ) 100; (\triangle) 110; (+) 140; (\times) 170; (\diamond) 200; (∇) 220.

identification of discriminant molecular determinants for the actions of LSD on two different receptors.

5. A link between QSAR and quantum chemical criteria for molecular reactivity.

Excellent correlations were observed in the early QSAR studies by Green and collaborators (10, 11, 84, 85, 86) between frontier densities (i.e., net charges on atoms, calculated from the highest occupied molecular orbital-HOMO) and the ability of tryptamines to contract the rat fundus strip. On this basis it was suggested that the tryptamines might interact with the receptor through a complex involving electron polarization from certain sites on the drugs toward the receptor. Our more recent findings indicated a linear correlation between the potencies of the tryptamines on the fundus and their ability to inhibit LSD binding to brain membranes (87). If correct, the mechanistic conclusion from the QSAR studies should hold also for that particular receptor in brain. It is therefore important to understand the molecular basis for such successful predictions of a defined biological activity.

The general importance of polarizability and electrostatic forces in complexes with indole derivatives had been emphasized by Green and Malrieu (88, 89). The QSAR results of Johnson and Green (11), indicated that the main sites of interaction (besides the side chain amino group, which is essential) would be the indole nitrogen, a region including the C-4 and C-5 bond, and the C-3 atom (see numbering scheme in (12)). Strong indications for the importance of these sites for complexation are also evident from experimental studies in different media. Thus, C-3 and C-5 are sites of close contact in crystalline complexes of 5-HT with both picrate and creatinine sulfate (90). NMR studies on complexes of methylated indoles and trinitrobenzene in solution show that the 3 position of the indole (especially in 3-methylindole) makes the greatest contribution to complexation as explicitly predicted from molecular orbital calculations by Green and Malrieu (88). Proton NMR studies (91, 92) also indicate that the proton at C-4 position is the first proton (on a carbon) to be exchanged in 5-HT and 5-methoxytryptamine. This changes with the site of hydroxyl substitution on the indole ring and the sequence of fastest exchanging protons is different in tryptamine, 6-HT and 4-HT.

The link between these considerations and quantum chemical studies of drug reactivity emerged from our efforts to identify the reactive sites in LSD and the tryptamine congeners, and to learn how their distribution can be described by more rigorous quantum mechanical reactivity criteria than the ones used in QSAR. We found (12, 87) that the positions at which a high frontier electron density was correlated with high biological potency correspond to the sites at which the density of the highest occupied

molecular orbital is localized in 5-HT (12, 87). These correlations are explained in terms of a multiple perturbation expansion (77, 78) of the interaction energies of 5-HT with positively charged reagents: the highest contribution to the polarization term in the interaction energy comes from the HOMO (12, 87). Since the electronic density is localized on specific centers in the HOMO of the various 5-HT congeners, this finding explained the relation between high frontier density at the same sites at which HOMO is localized in 5-HT and more potent 5-HT-like activity. Similarly, it became clear why earlier studies found a negative correlation between potency and the frontier density on certain atoms (11): the localization of HOMO and the orbital Next to HOMO (NHOMO) differs, and a requirement for high frontier density on the atoms on which the HOMO of 5-HT is localized is tantamount to a requirement for low frontier density on the atoms on which the NHOMO of 5-HT is localized. We have shown (87) that the localization pattern in the 5-HT NHOMO is precisely on those sites on which Johnson and Green (11) found a negative correlation between frontier density and biological activity. The relation between the different localization patterns of HOMO in the various congeners and the biological activity also indicate some of the reasons for the apparent failure to find a direct correlation between the energy of HOMO (or the molecular ionization potential) and 5-HT-like biological activity (87).

6. Discriminant molecular determinants for the actions of LSD on the 5-HT and histamine receptors.

The localization pattern of electron density from the HOMO (and NHOMO) of LSD is very similar to that of 5-HT (Figure 5). Furthermore, the patterns of the electrostatic potential maps of LSD and 5-HT also show great similarities, indicating a common interaction pharmacophore (Figure 6). On the basis of our analysis of the molecular determinants for drug-receptor interactions in the tryptamine series (12, 13, 14), it follows that the LSD molecule should be able to assume towards the 5-HT receptor an orientation that is electrostatically equivalent to that of 5-HT (14). This similarity of reactivity characteristics (i.e., electrostatic potential, HOMO localization pattern) between 5-HT and LSD is all the more striking as LSD could be considered to be much closer structurally to the unsubstituted tryptamine (e.g., see (84)). From Figure 6 it is clear that the interaction pharmacophore of LSD is close to that of 5-HT because the C12-C13 double bond in LSD generates a local minimum that mimics the one generated by the hydroxyl in 5-HT.

This commonality of reactivity characteristics of LSD and 5-HT is consonant with the well known experimental findings that LSD and 5-HT can act on a common receptor in brain as well as in peripheral tissue (for a recent review see (93)). However, LSD exhibits a complex pharmacology (83) that includes interactions

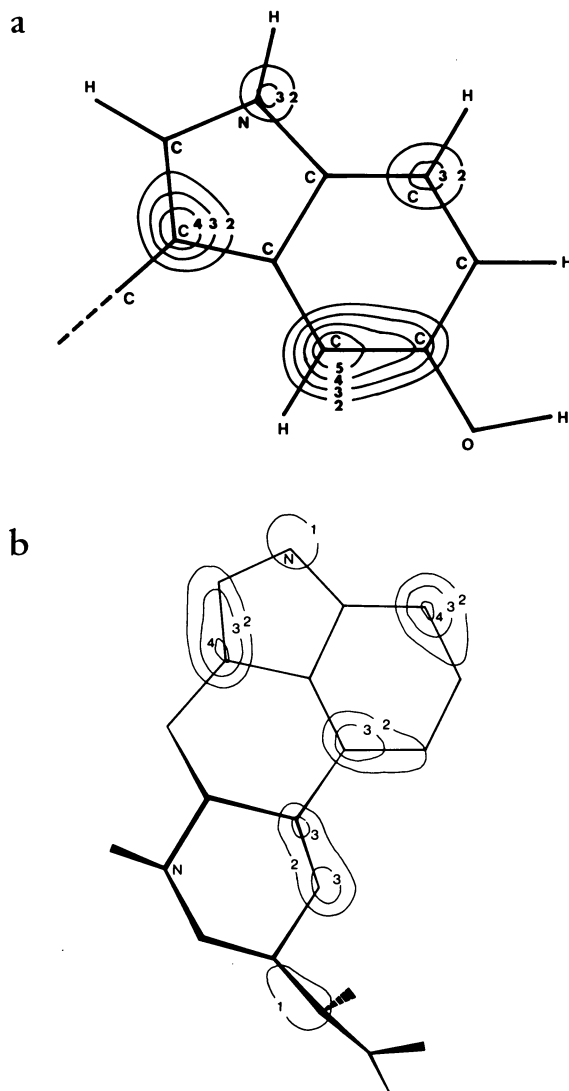


Figure 5. Comparison of the electron density distribution from the HOMO of 5-HT (a) and of D-LSD (b) in a parallel plane 1 Å above the indole portions of the molecules. Molecular wavefunctions for 5-HT and D-LSD were calculated by a pseudopotential method (13). Units are 10^{-3} electron/a.u.³.

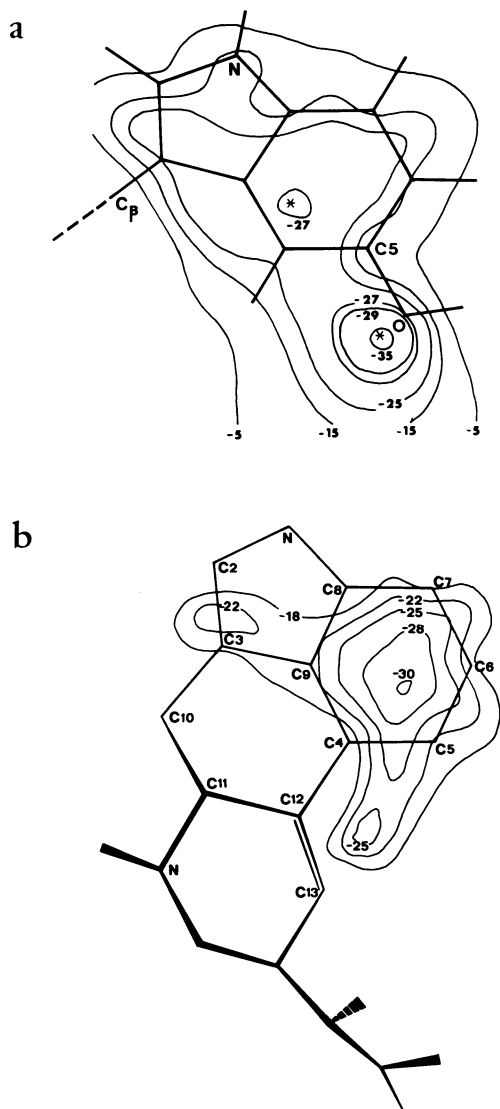


Figure 6. Comparison of molecular electrostatic potentials of 5-HT (a) and D-LSD (b) in a parallel plane 1.5 Å above their indole portions (units are kcal/mol). The electrostatic potentials were calculated (65) from molecular wavefunctions obtained as described in the legend to Figure 5.

with many receptors (e.g., dopamine, histamine, norepinephrine, halopericol).

We have recently shown (94) that the molecular structural factors for antagonist properties of LSD on the histamine H₂-receptor are related to the congruency of a particular portion of the molecule with the novel H₂-antagonists cimetidine and metiamide (83): "the residue in LSD that is congruent with metiamide is hexanoic acid-4-aza-4-methyl-6-(3-pyrrolyl)-N,N-diethylamide." We therefore suggested (94) that "the indolyl and imidazolyl substituted analogs (of that residue) and their derivatives may have H₂-antagonist activity". When chemists at SK&F were made aware of this prediction, Dr. Carl Kaiser provided us with hexanoic acid-4-aza-4,5-dimethyl-6-(3-indolyl)-N,N-diethylamide (SKF 10856 10856), the structure of which is shown in Figure 7. We recently obtained full dose response curves and Schild plots proving the compound to be a competitive antagonist of histamine and dimaprit (a highly specific H₂-agonist) at the H₂-receptor linked to cy-clase (94, 95). The pA₂ value, 6.10, (K_D = 7.9 x 10⁻⁷ M) was not different from that of LSD.

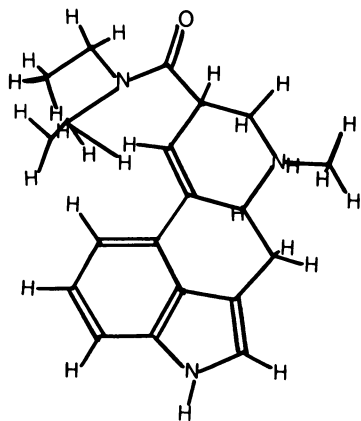
Comparison of the molecular structures of LSD and SKF 10856 (Figures 6 and 7, respectively) shows that the C12-C13 double bond is missing in the latter. Because the similarity of the interaction pharmacophores of 5-HT and LSD is based on the effect of this bond on the electronic properties of LSD, the SKF compound should not resemble 5-HT in its reactivity characteristics. Rather, the resemblance should be with tryptamine. The experimental findings from binding experiments to the P₂ fraction of homogenates from guinea pig cortex confirm this theoretical prediction: The affinity of SKF 10856 for the LSD/5-HT receptor (K_D ≈ .1μM) is 100 times lower than that of either LSD or 5-HT to their receptor (K_D = 5nM) and is comparable to that of tryptamine (K_D = .1μM).

7. Conclusion

This work and that from other laboratories (e.g., 5, 21-25, 28-32) show that quantum chemical methods have become tools to reveal biological mechanisms and, in rigorously relating molecular structure to biological interactions, these tools are ready for use in drug design. Some of the reluctance to use quantum chemistry for this purpose rests on early attempts to use quantum chemical indices e.g., Mulliken charges, superdelocalizability, HOMO energies, as QSAR parameters without mind to their shortcomings and physical meaning (see (2)). Aside from such substantive objections, some people have dismissed the quantum chemical approach out of simple prejudice against "theory". Karl Popper (96) said that it

"is mistaken...to distinguish between theoretical terms...and non-theoretical or empirical or observational or factual or ordinary terms...It is

a



b

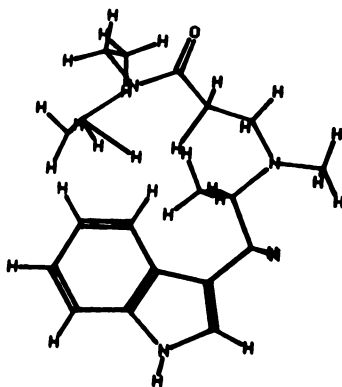


Figure 7. Comparison of the molecular structures of D-LSD (a) and SKF 10856 (b)

perhaps somewhat like this. What people have learnt before reaching a certain critical age, they are inclined to regard as factual, 'or ordinary', and what they hear later, as theoretical... (The critical age seems to depend on the psychological type)."

Acknowledgements

This work was supported by the National Institute on Drug Abuse (DA 01875). The computations were supported in part by a grant of computer time from the University Computer Center of the City University of New York. Analysis of results in this work was carried out in part on the PROPHET System, a national computer resource sponsored by the NIH through the Chemical/Biological Information - Handling Program, Division of Research. H.W. is recipient of a Career Scientist Award from the Irma T. Hirsch Trust.

Literature Cited

1. Weinstein, H., Int. J. Quantum Chem. (1975) QBS2, 59.
2. Osman, R., Weinstein, H. and Green, J.P., this symposium.
3. Chotia, C. and Janin, J., Nature (1975) 256, 705.
4. Hopfinger, A.J., "Intermolecular Interaction and Biomolecular Organization", Wiley-Interscience, J. Wiley & Sons, Inc., New York, 1977.
5. Warshel, A. and Levitt, M., J. Mol. Biol. (1976) 103, 227.
6. Sprinkel, F.M., Shillady, D.D. and Strickland, R.W., J. Amer. Chem. Soc. (1976) 97, 6653.
7. Helene, C., Dimicoli, J.-L. and Brun, F., Biochem. (1971) 10, 3802.
8. Nogrady, T., Hrdina, P.D. and Ling, G.M., Mol. Pharmacol. (1972) 8, 565.
9. Sapper, H. and Lohmann, W., Mol. Pharmacol. (1976) 12, 605.
10. Green, J.P. and Kang, S. in "Molecular Orbital Studies in Chemical Pharmacology", pp. 105-120, L.B. Kier, Ed., Springer Verlag, New York, 1970.
11. Johnson, C.L. and Green, J.P., Int. J. Quantum Chem. (1974) QBS1, 159.
12. Weinstein, H., Chou, D., Kang, S., Johnson, C.L. and Green, J.P., Int. J. Quantum Chem. (1976) QBS3, 134.
13. Weinstein, H. and Osman, R., Int. J. Quantum Chem. (1977) QBS4, 253.
14. Weinstein, H., Osman, R., Edwards, W.D. and Green, J.P., Int. J. Quantum Chem. (1978) QBS5, 449.
15. Collins, J.B., Dill, J.D., Jemmis, E.D., Apeloig, Y., Schleyer, P.v.R., Seeger, R. and Pople, J.A., J. Amer. Chem. Soc. (1976) 98, 5419.

16. Apeloig, Y., Schleyer, P.v.R., Binkley, J.S. and Pople, J.A., J. Amer. Chem. Soc. (1976) 98, 4332.
17. Maugh, T.H., Science (1976) 194, 413.
18. Laidig, W.D. and Schaeffer III, H.F., J. Amer. Chem. Soc. (1978) 100, 5972.
19. Truax, D.R. and Wieser, H., Chem. Soc. Rev. (1976) 5, 411.
20. Borden, R.T., Davidson, E.R. and Hart, P., J. Amer. Chem. Soc. (1978) 100, 388.
21. Scheiner, S. and Lipscomb, W.N., Proc. Nat. Acad. Sci. USA (1976) 73, 432.
22. Scheiner, S. and Lipscomb, W.N., J. Amer. Chem. Soc. (1977) 99, 3466.
23. Hayes, D.M. and Kollman, P.A., J. Amer. Chem. Soc. (1976) 98, 3335.
24. Goddard, W.A., in "Enzymic Oxygenation - Binding, Activation and Reaction Mechanisms", P.A. Leermakers Symposium on Energetics in Chemistry and Biochemistry, Wesleyan University, 1978.
25. Kirchner, R.F. and Loew, G.H., J. Amer. Chem. Soc. (1977) 99, 4639.
26. Churg, A.K. and Makinen, M.W., J. Chem. Phys. (1978) 68, 1913.
27. Kier, L.B., "Molecular Orbital Theory in Drug Research", Academic Press, 1971.
28. "Molecular and Quantum Pharmacology", E.D. Bergmann and B. Pullman, Eds., D. Reidel Publ. Co., Dordrecht-Holland, 1974.
29. Green, J.P., Johnson, C.L. and Kang, S., Ann. Rev. Pharmacol. (1974) 14, 319.
30. Christoffersen, R.E., in "Quantum Mechanics of Molecular Conformations", Ch. 3, B. Pullman, Ed., J. Wiley & Sons, New York, 1976.
31. Christoffersen, R.E. and Angeli, R.P., in "The World of Quantum Chemistry", Ch. 3 B. Pullman and R.G. Parr, Eds., D. Reidel Publ. Co., Dordrech-Holland, 1976.
32. Kaufman, J.J., Int. J. Quantum Chem. (1977) QBS4, 375.
33. Gelin, B.R. and Karplus, M., J. Amer. Chem. Soc. (1975) 97, 6996.
34. Port, G.N.J. and Pullman, A., J. Amer. Chem. Soc. (1973) 95, 4059.
35. Pullman, B. and Courriere, P., Mol. Pharmacol. (1972) 8, 612.
36. Maayani, S., Weinstein, H., Cohen, S. and Sokolovsky, M., Proc. Nat. Acad. Sci. USA (1973) 70, 3103.
37. Richards, W.D., Aschman, D.G. and Hammond, J., J. Theor. Biol. (1975) 52, 223.
38. Weinstein, H., Srebenik, S., Pauncz, R., Maayani, S., Cohen, S. and Sokolovsky, M., in "Chemical and Biochemical Reactivity", pp. 493-512, E.D. Bergmann and B. Pullman, Eds. D. Reidel Publ. Co., Dordrecht-Holland, 1974.

39. Weinstein, H., Maayani, S., Srebrenik, S., Cohen, S. and Sokolovsky, M., Mol. Pharmacol. (1975) 11, 671.
40. Warshel, A., Chem. Phys. Lett. (1978) 55, 454.
41. Pullman, A. and Pullman, B., Quant. Rev. Biophys. (1975) 7, 505.
42. Schuster, O., Jakubetz, W. and Marius, W., in "Topics in Current Chemistry", vol. 60, pp. 1-107, Springer Verlag, Berlin, 1975.
43. Clementi, E., "Determination of Liquid Water Structures. Coordination Numbers for Ions and Solvation for Biological Molecules", Lecture Notes in Chemistry, vol. 2, Springer Verlag, Berlin, 1976.
44. Romano, S. and Clementi, E., Gazz. Chim. Ital. (1978) 108, 319.
45. Bolis, G., Ragazzi, M., Salvaderi, D., Ferro, D.R. and Clementi, E., Gazz. Chim. Ital. (1978) 108, 425.
46. Swaminathan, S., Whitehead, R.J., Guth, E. and Beveridge, D.L., J. Amer. Chem. Soc. (1977) 99, 7817.
47. Swaminathan, S. and Beveridge, D.L., J. Amer. Chem. Soc. (1977) 99, 8392.
48. Huron, M.-J. and Claverie, P., J. Phys. Chem. (1972) 76, 2123.
49. Huron, M.-J. and Claverie, P., J. Phys. Chem. (1974) 78, 1853, 1862.
50. Hagler, A.T. and Moulton, J., Nature (1978) 272, 222.
51. Germer, H.A., Theor. Chim. Acta (1974) 34, 145.
52. Cremaschi, P., Gamba, A. and Simonetta, M., Theor. Chim. Acta (1975) 40, 303.
53. Spanget-Larsen, J., Theor. Chim. Acta (1978) 47, 315.
54. Sinanoglu, O., in "Intermolecular Forces" pp. 283-325, J.O. Hirschfelder, Ed., Wiley-Interscience, New York, 1976.
55. Del Bene, J.E. and Pople, J.A., J. Chem. Phys. (1973) 58, 3605.
56. Lentz, B.R. and Scheraga, H.A., J. Chem. Phys. (1973) 58, 5296.
57. Owicki, J.C., Shipman, L.L. and Scheraga, H.A., J. Phys. Chem. (1975) 79, 1794.
58. Hylton, J., Christoffersen, R.E. and Hall, G.G., Chem. Phys. Lett. (1974) 24, 501.
59. McCreery, J.H., Christoffersen, R.E. and Hall, G.G., J. Amer. Chem. Soc. (1976) 98, 7191, 7196.
60. Germer, H.A., J. Pharm. Pharmac. (1974) 26, 467.
61. Weintraub, H. and Hopfinger, A.J., in "Molecular and Quantum Pharmacology", pp. 131-152, E.D. Bergmann and B. Pullman, Eds., D. Reidel Publ. Co., Dordrecht-Holland, 1974.
62. Beveridge, D.L., Radna, R.J., Schnuelle, G.W., and Kelly, M.M., in "Molecular and Quantum Pharmacology", pp. 153-178, E.D. Bergmann and B. Pullman, Eds., D. Reidel Publ. Co., Dordrecht-Holland, 1974.

63. Loew, G., Weinstein, H. and Berkowitz, D., in "Environmental Effects on Molecular Structure and Properties", pp. 239-258, B. Pullman, Ed., D. Reidel Publ. Co., Dordrecht-Holland, 1976.
64. Weinstein, H., Srebrenik, S., Maayani, S. and Sokolovsky, M., J. Theor. Biol. (1977) 64, 295.
65. Srebrenik, S., Weinstein, H. and Pauncz, R., Chem. Phys. Lett. (1973) 20, 419.
66. Scrocco, E. and Tomasi, J., in "Topics in Current Chemistry, New Concepts II, No. 42, p. 95, Springer Verlag, Berlin, 1973.
67. Truhlar, D.G., Van-Catledge, F.A. and Dunning, T.H., J. Chem. Phys. (1972) 57, 4788.
68. Truhlar, D.G. and Van-Catledge, F.A., J. Chem. Phys. (1973) 59, 3207.
69. Truhlar, D.G. and Van-Catledge, F.A., J. Chem. Phys. (1976) 65, 5536.
70. Weinstein, H., Maayani, S., Srebrenik, S., Cohen, S. and Sokolovsky, M., Mol. Pharmacol. (1973) 9, 820.
71. Loew, G., Berkowitz, D., Weinstein, H. and Srebrenik, S., in "Molecular and Quantum Pharmacology", pp. 355-389, E.D. Bergmann and B. Pullman, Eds., D. Reidel Publ. Co., Dordrecht-Holland, 1974.
72. Petrongolo, C., Tomasi, J., Macchia, B. and Macchia, F., J. Med. Chem. (1974) 17, 501.
73. Petrongolo, C. and Tomasi, J., Int. J. Quantum Chem. (1975) 9, 181.
74. Petrongolo, C., Maachia, B., Macchia, F. and Martinelli, A., J. Med. Chem. (1977) 20, 1645.
75. Politzer, P., Daiker, K.C. and Donnelly, R.A., Cancer Lett. (1976) 2, 17.
76. Kaufman, J.J., in "QuaSAR. Quantitative Structure Activity Relationships, of Narcotic Analgesics, Narcotic Antagonists, and Hallucinogens", p. 250, NIDA Monograph No. 22, G. Barnett, M. Trsic and R.E. Willette, Eds., 1978.
77. Bartlett, R.J. and Weinstein, H., Chem. Phys. Lett. (1975) 30, 441.
78. Chang, S.Y., Weinstein, H. and Chou, D., Chem. Phys. Lett. (1976) 42, 145.
79. Chang, S.Y. and Weinstein, H., Int. J. Quantum Chem. (1978) 14, 801.
80. Weinstein, H., Chou, D., Johnson, C.L. and Green, J.P., Mol. Pharmacol. (1976) 12, 738.
81. Weinstein, H., Eilers, J.E. and Chang, S.Y., Chem. Phys. Lett. (1977) 51, 534.
82. Lipscomb, W.N., Hartsuck, J.A., Reeke, G.N., Quioco, F.A., Bethge, P.H., Ludwig, M.H., Steitz, T.A., Muirhead, H. and Coppola, J.C., Brookhaven Symp. Biol. (1968) 21, 24.

83. Green, J.P., Weinstein, H. and Maayani, S., in "QuaSAR. Quantitative Structure Activity Relationships of Analgesics, Narcotic Antagonists, and Hallucinogens", p. 38, NIDA Research Monograph No. 22, G. Barnett, M. Trsic, and R.E. Willette, Eds., 1978.
84. Kang, S., Johnson, C.L. and Green, J.P., Mol. Pharmacol. (1973) 9, 640.
85. Green, J.P., Dressler, K.P. and Khazan, N., Life Sci. (1973) 12, 475.
86. Kang, S. and Green, J.P., Proc. Nat. Acad. Sci. USA, (1970) 67, 62.
87. Green, J.P., Johnson, C.L., Weinstein, H., Kang, S. and Chou, D., in "The Psychopharmacology of Hallucinogens", p. 28, R.E. Willette and R.C. Stillman, Eds., Pergamon Press, 1978.
88. Green, J.P. and Malrieu, J.P., Proc. Nat. Acad. Sci. USA, (1965) 54, 659.
89. Millie, P., Malrieu, J.P., Benaïm, J., Lallmand, J.Y. and Julia, M., J. Med. Chem. (1968) 11, 207.
90. Thewalt, U. and Bugg, C.E., Acta Cryst. (1972) B28, 82.
91. Daly, J.W. and Witkop, B., J. Amer. Chem. Soc. (1967) 89, 1032.
92. Kang, S., Witherup, T.H. and Gross, S., J. Org. Chem. (1977) 42, 3769.
93. Friedman, D.X. and Halaris, A.E., in "Psychopharmacology: A Generation of Progress", p. 347-359, M.A. Lipton, A. DiMascio, and K.F. Killam, Raven Press, New York, 1978.
94. Green, J.P., Johnson, C.L., Weinstein, H. and Maayani, S., Proc. Nat. Acad. Sci. USA (1977) 74, 5697.
95. Green, J.P., Johnson, C.L. and Weinstein, H., in "Histamine Receptors", p. 185-210, T.O. Yellin, Ed., Spectrum Publ., New York, 1978.
96. Popper, K.R., "The Logic of Scientific Discovery", p. 425, Hutchison, London, 1975.

RECEIVED June 8, 1979.

Modeling Receptor and Substrate Interactions

GREGORY M. COLE, EDGAR F. MEYER, JR.,
STANLEY M. SWANSON, and W. GERALD WHITE

Biographics Laboratory, Department of Biochemistry and Biophysics,
Texas A&M University, College Station, TX 77843

Modelling describes the human perception of physical concepts in a way easily communicated to others. In an empirical process experimentation lets us evaluate intermediary steps, as described by individual models. Thus a model may have predictive powers. Although biological macromolecules containing thousands of atoms have been studied by X-ray crystallography and have been shown to relate the structural and functional aspects of modern biochemistry, functional models of these molecules are exceedingly difficult to study without the aid of a computer. The ability of a digital computer to store, recall, and depict structural information opens the door to interactive molecular modelling. The numerical nature of the model elements are, in principle, in a convenient form for subsequent quantitative evaluation.

Starting data for these modelling studies are derived from measured kinetic parameters of natural or synthetic inhibitors or substrates of enzymic reactions and the three-dimensional structures of the host enzymes as studied by X-ray diffraction methods. Interactive computer graphics lets us express the structural information in graphical terms and generate stereochemical hypotheses. Currently, our progress in using these modelling methods has lead us to propose a potential inhibitor for one of the systems under study. The approach is quite general and so this paper is essentially a progress report of early efforts.

Several speakers have prefaced their discussion of drug design methodology by lamenting the lack of specific geometric

0-8412-0521-3/79/47-112-189\$05.00/0

© 1979 American Chemical Society

information regarding the drug receptor site. While sufficient information is now available for a critical evaluation of molecular modelling methods, the receptor and guest molecules are more closely related to enzymic than pharmaceutical interactions. We propose that these modelling studies are a necessary early step in molecular modelling of site-specific pharmacophores.

These initial efforts encourage us to press forward, fully aware that, when put to the test, some appealing models will exhibit blemishes when considered in the harsh light of chemical/biological/physical testing while others may serve as our "blessed Beatrice" in the search for new routes through uncharted regions.

Hardware/Software

A group may boast of unique computer hardware until it begins to age or show deficiencies. The Biographics' PDP 11/40 + 3-D Vector General Graphics, is, by current standards, venerable, "black and white", and silent, ca. pre "Gone with the Wind". However, a stable and growing software package presents possibilities yet to be plumbed, examples of which are presented herewith. As the potential application of computer modelling methods to drug design have already been discussed (1), here we shall present several areas of recent and current work at the intersection of crystallography, information retrieval, force field calculations, and modelling.

Macromolecular Modelling

The dimensions and complexities of macromolecular models not only overwhelm the eye, they also challenge conventional computer techniques for time and space. We were fortunate that Carl Morimoto's original version of Program FIT (2), used for modelling a nascent polypeptide chain (foreground) into electron density maps (background) on the Biographics display, could be used directly to model the interactions of a small molecule (foreground) with a receptor or active site of a macromolecule (background), as obtained from crystallographic atomic coordinates derived from the Protein Data Bank (3). To this end, program PDB80 reads an entry of the Protein Data Bank and prepares a compressed binary file with an effective 9-fold compression factor. The compressed binary file is the common link between retrieval programs SEARCH (4), PROBAK, and PROSID (5), the electron density contouring and tracing programs (6), and the interactive fitting program, FIT (2). Most of the above as well as other auxiliary programs produce binary picture files unique to the Vector General 3-D display. Program PIXMOV can display up to ten picture files under operator control

and has the option of recording operator command sequences for subsequent playback, leading to photographs and motion picture films as illustrated here. Although most of the methods employed here have elements common to those used by most other active computer graphics groups, one should not underestimate the creative efforts involved in designing and developing such a system, knowing full well that not until the system gains a minimal level of facility can it be possible to attempt useful molecular modelling without undue frustrations.

Two classes of macromolecules have received special attention: 1) horse liver alcohol dehydrogenase, ADH (7), because of its reactivity with a broad range of substrates and 2) the class of serine proteases because of the overlapping body of chemical + structural information and especially because of the high resolution (1.5Å) of the trypsin: trypsin inhibitor complex (8). While Mr. Gerry White's treatment of the former system is primarily descriptive, it may be of interest to the steroid chemist and endocrinologist because it illustrates Theorell's kinetic studies of ADH with steroid substrates (9, 10). In the latter study, Mr. Greg Cole has used the Biographics display to predict a model of a potential inhibitor as a first step in probing the structure + activity relationships of this series leading to the targeted synthesis of promising substrates, inhibitors, or transition state analogs.

Alcohol Dehydrogenase

Using a PDP 11/40 minicomputer with an interactive vector-drawing display, the 3-D coordinates of all atoms (except H) in the active site region of an enzyme and those of the proposed substrate may be stored, displayed, and manipulated relative to one another, so that the substrate can be steered or "fitted" by the operator into an assumed productive orientation in the active site. A series of potentiometer dials interfaced to the display allows for rotation and translation of the substrate, modification of bond lengths and atom positions, and other manipulations to help the user orient the substrate. Fitting of substrates in this way can be correlated with kinetic studies of the enzyme-substrate system to see how various substituent groups in the substrate may alter reaction rates by steric and chemical interaction with amino acid residues and cofactors in the active site. A program package is currently being developed by Dr. Gerard Pepe which will give the user a quantitative means of determining goodness-of-fit in conjunction with the 3-D fitting program. Van der Waals' (VDW) and Coulombic interactions as well as possible hydrogen bond arrangements are considered in determining how well the substrate may fit into the active site.

There are limitations, of course. The coordinates for the atoms of both active site and substrate are obtained from X-ray crystallographic studies and represent the experimentally determined conformation of the molecule in the crystalline form. This may not be the same as the conformation in solution (enzymes may flex and bend slightly to accommodate the substrate in some systems) and although it is relatively easy to alter atom positions of the substrate (foreground) to mimic *in vivo* conformations, the residues in the active site are currently left unchanged. A rigid active site (background) is therefore assumed. Secondly, the sum of the derived van der Waals' and Coulombic interactions and hydrogen bonds may not completely represent all the interactions taking place between active site and substrate. The seriousness of these limitations is unclear, but they probably have at least some effect on the accuracy of the calculations; further study is in progress.

Two examples follow which serve to illustrate the method. The substrate models concerned were fitted into the model of the active site of alcohol dehydrogenase-nicotinamide adenine dinucleotide (ADH-NAD) with VDW contacts, etc. not considered explicitly.

Dutler and Brändén (11, 12) have studied the interaction of the ADH-NAD complex with alkyl-substituted cyclohexanols to determine productive substrate orientations in the active site. Some of their findings were that 1) cyclohexanol fits best in a chair conformation with axial-reacting hydroxyl oxygen and equatorial-reacting hydrogen atoms on C1, 2) that substitutions at C4 (para) have little effect on the reaction rate due to hydrophobic bonding with the residues in the hydrophobic "barrel" region of the active site, 3) substitutions at C2, C3, C6 and C5 axial positions and at C2, C3 and C6 equatorial positions strongly reduce the reaction rate due to steric interactions with active site residues or the nicotinamide ring of NAD. These findings were determined by correlation of reaction rate studies with model-building using Kendrew models. Dutler and White have used the 3-D graphical technique previously described to simulate productive substrate orientations for this system. Using only steric considerations and hydrophobic bonding assumptions, they were able to verify Dutler and Brändén's earlier findings. For example, addition of a methyl group at the C2 axial position of cyclohexanol interferes sterically with the carboxamide moiety of NAD, in agreement with the profoundly reduced reaction rate (cf. Fig 1-2).

In a separate study, Brändén (11) modelled the active site+ substrate interactions of ADH and several 3 β - and 3 α -hydroxy steroids, based on the kinetic work of others (4,5). He showed that 3 β -hydroxy-5 β -cholanoic acid fits in the active site rather

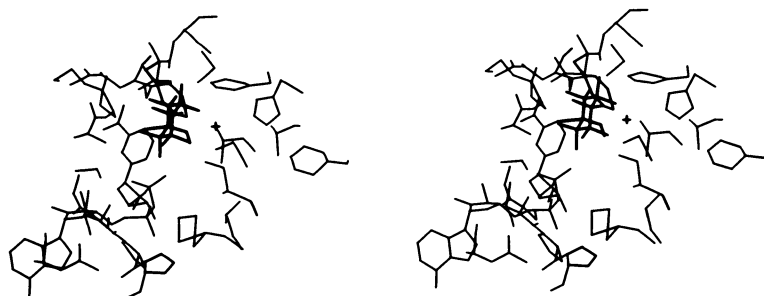


Figure 1. Stereoview of the active site of ADH looking through the protein and out through the entrance cavity of the receptor site. The zinc atom (+) is tetrahedrally coordinated to a his and two cys residues. The fourth position is occupied by the O atom of the substrate, here, cyclohexanol. The cofactor, NAD⁺, is shown from front left to center, where the nicotinamide moiety is placed in bonding proximity with the substrate.

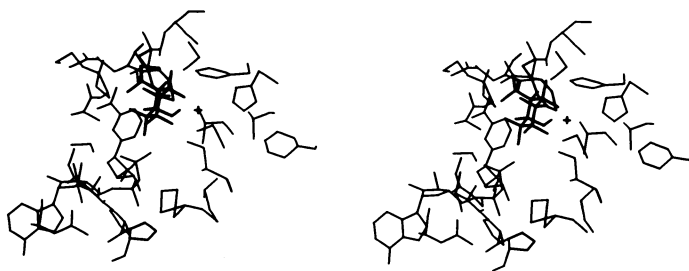


Figure 2. Stereoview of axial orthomethylcyclohexanol in the identical position as cyclohexanol (Figure 1). Close contact between the nicotinamide moiety of NAD⁺ and the axial methyl group make this analog a less favorable substrate.

well except for one interaction between the methyl group at position 18 of the steroid and leu 116 of the enzyme (E-subunit). (The S sub-unit of ADH differs slightly in conformation around leu 116 and is thought to bind the steroid more productively). 3α -hydroxy- 5β -cholanoic acid fits very poorly in the E (or S) type subunit, offering a reason why this steroid is not oxidized by the enzyme. Computer graphic studies (Fig. 3,4) confirm Brändén's findings (this would be expected since both Brändén and White used the X-ray structure of the horse E-type subunit in modelling the active site).

That interactive computer graphic modelling techniques can be useful in structure-activity studies of enzymes is shown by the agreement between computer-generated active site-substrate complexes and the kinetic and non-computer modelling studies already carried out. Once all the parameters governing the positioning of a substrate in the active site can be accurately represented using computer graphical techniques, computer-aided design of substrates and inhibitors can effectively be performed.

Serine Proteases

It is hoped that by studying the peptide binding regions of the serine proteases it will be possible to design specific inhibitors. One begins by studying the conformation of known inhibitors bound to the enzyme determined by X-ray crystallographic means.

Trypsin + PTI Complex. The structure of the complex of the pancreatic trypsin inhibitor (PTI) with trypsin has been determined at 1.5Å resolution by Huber, *et al.* (8). The first step in the hydrolysis of natural trypsin substrates is believed to be a nucleophilic attack on the carbonyl carbon atom of the scissile peptide bond by the oxyanion derived from the hydroxyl group of ser 195. During the course of this attack, the carboxyl carbon atom must change from sp^2 to sp^3 orbital hybridization, eventually forming a tetrahedral adduct to the serine. In the PTI complex the carbonyl group of lys 16 of the inhibitor (which would correspond to the carbonyl group of the scissile bond of a natural substrate) is seen to be partially tetrahedral (Fig. 5). If hydrolysis were to proceed as with a normal substrate, one would expect the carbonyl group to translate towards ser 195 (in order to form the covalent bond), while rotations about other bonds of the bound substrate chain would occur in order to allow this translation and the concomitant change to sp^3 hybridization and yet preserve other binding interactions.

It has been found that when the carbonyl carbon of lys 16 of the model of the PTI complex is replaced by a tetrahedral atom which is covalently attached to the oxygen of 195 (also tetrahedral), the backbone of the inhibitor can be super-

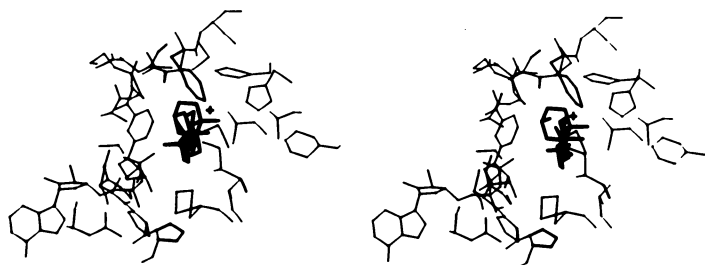


Figure 3. Stereoview of the model of β 30H-5 β cholanoic acid in orientations similar to those in Figures 1, 2, 4. The steroid extends back into the yawning entrance cavity of the enzyme.

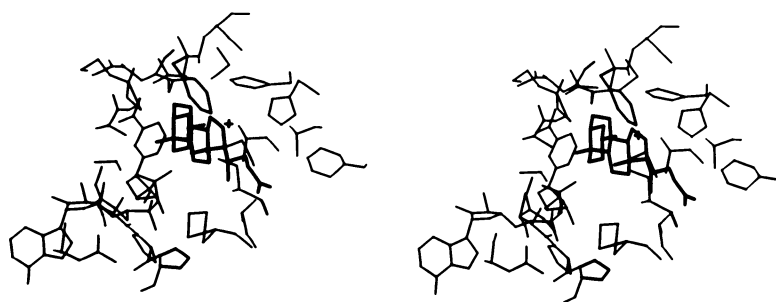


Figure 4. Stereoview of a model of α 30H-5 β cholanoic acid. Placing the α 30H oxygen atom in bonding position causes the rest of the steroid to be shifted in the entrance cavity such that serious overlap near the D ring of the steroid cannot be avoided.

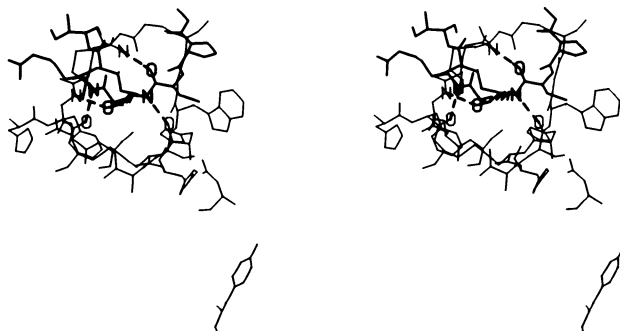


Figure 5. Stereoview of the active site region of trypsin bound to the trypsin inhibitor complex. The lys 16 sidechain extends back into the receptor cavity (cf. Figure 7 for a better view). The proton relay chain, asp 102-his 57-ser 195 arches from right to center with ser 195 appearing in the center. The partially tetrahedral geometry of the carbonyl carbon atom of lys 16 may be seen with some difficulty. The hexapeptide from the trypsin inhibitor protein (Figure 8) is drawn intensely for emphasis; central H bridges are drawn as dashed vectors.

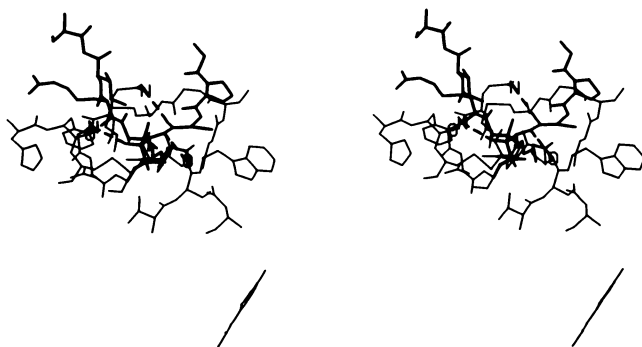


Figure 6. Stereoview of a model of how lys 16 of trypsin inhibitor might appear if it were to form a covalent bond to ser 195 of trypsin. Thus this is a proposed model of the transition state of the enzyme + substrate complex. The orientation is rotated slightly from those of Figures 5 and 7.

imposed back to its position in the PTI crystal structure by appropriate rotations about PHI and PSI torsion angles within three amino acids on either side of lys 16 (Fig. 6). If one studies the resultant changes in binding interactions upon transition from partially to fully tetrahedral carbonyl carbon, one finds that most hydrogen bonds are shortened (strengthened) and that the terminal amino group of the lysine side-chain is pushed closer to the acid residue in the trypsin side-chain-binding-pocket (8). For these reasons it is believed that the functional peptide of the pancreatic trypsin inhibitor as seen in the crystal structure of its complex with trypsin closely mimics the conformation of a natural trypsin substrate bound to the surface of the enzyme, about to be cleaved.

Recall now that the carbonyl group of the lysine of the inhibitor is seen to be semi-tetrahedral (Fig. 5) in the crystal structure. In order to define binding interactions which could be built into a synthetic inhibitor, it could be instructive to determine how the binding interactions of the pancreatic inhibitor might change if the carbonyl of lys 16 were completely planar. If the semi-tetrahedral carbonyl group of the inhibitor model is replaced by an atom with completely planar geometry, it is again possible to superimpose the inhibitor backbone upon its position in the crystal structure by rotations about PHI and PSI angles within three amino acid residues on either side of the lysine (Fig. 7). In this case, however, two H-bonds are lost and two others are lengthened. It was first suggested that various interactions within the inhibitor itself constrained the functional peptide in such a manner as to force the carbonyl group of lysine into a semi-tetrahedral arrangement. A study of the interactions within the inhibitor shows no evidence of this...indeed the functional peptide chain is relatively free from interactions with the remainder of the inhibitor, so much so as to suggest that it might be flexible in solution (Fig. 8). This fact leads to three likely possibilities: 1) Either the model of the bound inhibitor bearing a flat carbonyl group at lys 16 is not a good representation of the conformation of the functional peptide of the inhibitor at the moment of binding to trypsin, or 2) the functional peptide does indeed bind as seen in the planar model but quickly snaps into the semi-tetrahedral conformation with concomitant formation of two additional hydrogen bonds, or 3) perhaps initial binding occurs at some distance from lysine and proceeds towards lysine and passes that point only after the partial nucleophilic attack by ser 195 forces the carbonyl group into the semi-tetrahedral arrangement.

If the third possibility is indeed the case, the binding would have to proceed from the amino terminal end of the peptide since the basic residue for which the enzyme is specific lies in this direction from the scissile bond. If the opposite were true, serine would be able to attack peptide bonds of amino

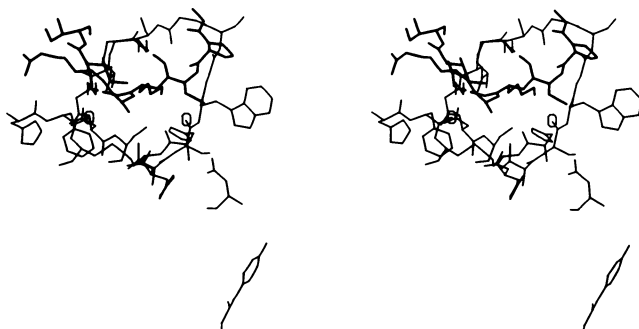


Figure 7. Stereoview of a model of a hexapeptide as it might appear just before attack by ser 195 on the carbonyl carbon atom of lys 16. The lys side chain may be seen extending back into the receptor cavity which helps give specificity to trypsin catalyzed reactions. Transitions in the conformations of Figures 5 and 7 were generated by visual manipulation of angles within the hexapeptide (drawn more intensely).



Figure 8. Stereoview of trypsin + trypsin inhibitor showing backbone atoms. This model shows the separated enzyme + inhibitor prior to formation of the complex.

acids bearing any amino acid in the critical position. In either case, the enzyme would spend a lot of time bound to peptide chains for which it is not specific. For this reason, we favor the second possibility mentioned above i.e. the peptide first binds in a non-optimal manner and then snaps into the semi-tetrahedral arrangement which is stabilized by the formation of additional hydrogen bonds and optimization of preexisting ones.

Subtilisin. Compounds of the form: $R-CO-CH_2-C\ell$ have been shown to inhibit the serine proteases irreversibly by alkylation of the active site histidine. The rate at which inhibition occurs has been found to be dependent on the constitution of the R group. Peptide chains terminating with $-NH-CHR-CO-CH_2-C\ell$ exhibit excellent inhibition. Robertus *et al.* (13) have determined the arrangement of the complex of the chloromethyl ketone (amino to carboxyl) $PH-CH_2-O-CO-ALA-GYL-PHE-CH_2-C\ell$ with subtilisin by difference Fourier analysis. The compound was shown to alkylate the enzyme by covalent attachment to the N-epsilon atom of his 64. The phenylalanine in the first position was seen to be surrounded by hydrophyllic groups. The sidechains of gly and ala had little interactions with their surroundings, and the phenyl ring of the terminal benzyl ester was seen to participate in what seemed to be a stacking interaction with trp 104. The whole peptide chain of the inhibitor was shown to form a quasi-beta structure with residues 125 to 128 of the enzyme.

Powers *et al.* (14) have investigated the kinetics of the reaction of a series of peptide chloromethylketones with subtilisin BPN. Of the various peptide chloromethylketones assayed, they found those with hydrophobic groups adjacent to the chloromethyl moiety to be the fastest inhibitors. The fastest of all being $AC-PHE-GLY-ALA-LEU-CH_2-C\ell$. Considering the stacking interaction seen in Robertus' structure, it is interesting to note that the above was the only peptide assayed by Powers carrying an aromatic group in the number four position.

A model of Powers' inhibitor has been fitted into a model of the active site region of subtilisin (15) to form the quasi-beta structure shown by Robertus. (Fig. 9,10). The phenyl ring of the phenylalanine is seen to be in a good position to stack with tyr 104. Using this peptide as a starting point, one might be able to design an inhibitor of subtilisin using a different method to inactivate the catalytic mechanism and to continue the chain in the other (carboxyl) direction from the bound residues in order to create additional binding interactions. Lindquist has shown (16) boronate derivatives to be good inhibitors of serine proteases, and has suggested that the reason for this is that the boron is able to form a stable tetrahedral adduct to the oxygen of the serine. We propose (17) that substitution of a boronate for the chloromethylketone

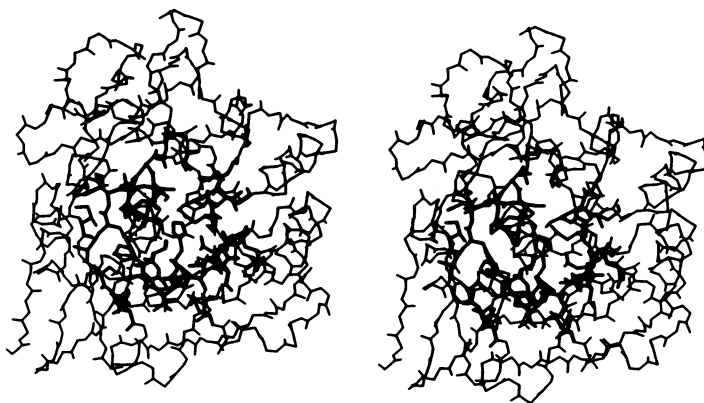


Figure 9. Stereoview of the enzyme subtilisin. Residues near the active site region are drawn more intensely.

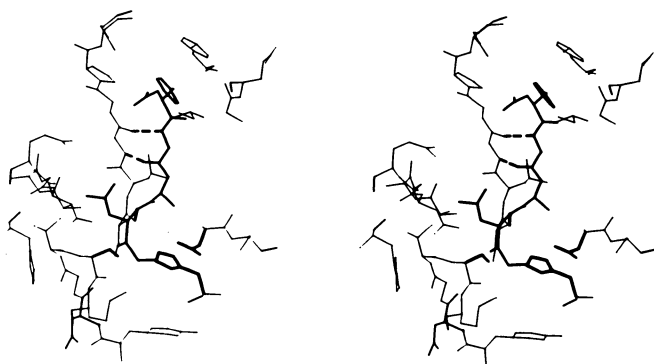


Figure 10. Stereoview of an inhibitor covalently attached to his 64 of subtilisin; to the left is ser 221, to the right of the complex is asp 32. The inhibitor is drawn more intensely for emphasis; hydrogen bridges are drawn as dashed vectors.

moiety of Powers' chlormethylketone + peptide would result in an inhibitor of subtilisin which could be extended in the carboxyl direction from the site of catalysis in order to enhance inhibition through development of additional interactions. Specifically, a model of the compound AC-PHE-GLY-ALA-LEU-BOH-NH-CH₂-(*)CH(CH₂-PH-OH)-(CH₂)₂-CONH₂, with the "S" configuration at the chiral (*) carbon has been fitted into a model of the active site of subtilisin by formation of a tetrahedral boronate adduct to ser 221 and formation of the quasi-beta structure seen in the chloromethylketones (Fig. 11). One is able to place the terminal amide group into a position to form two good hydrogen bonds with the amide and carboxyl groups of asn 218 quite easily. This then leaves the phenol group in a fair position to interact with tyr 217. We feel that the replacement of the unusual C-terminal moiety proposed here by a chain of common amino acids would give better binding interactions and an inhibitor more easily synthesized.

Conclusion

Computer programs to facilitate the synthesis of models for binding studies illustrated here are being developed or extended. They are designed to work on any protein-substrate combination, and thus the results of these modelling effects may be immediately applied to other problems under consideration.

A critical, systematic study of the expanding macromolecular data base may be predicted to yield insights into the mode of action of small and large molecules. The complexity of even these simplified models is such that the uninitiated viewer requires time and a variety of stereo cues in order to perceive the spatial relationships discussed. For this reason a short motion picture film will be shown.

An holistic view would then include non-crystallographic quantitative structure-activity parameters as well as imaging force-field calculation techniques. So long as the structure of the molecular species is a critical factor in a study, interactive computer graphics must be a focal point where various techniques converge because the relative complexity of the molecules and their interactions leave the investigator virtually blind to the ebb and flow of quantitative data (numbers) without visual cues, but especially because the interactive powers would permit the investigator to orchestrate and modulate the assembled battery of techniques. In a few years these predictions will appear obvious but now people and groups must be willing to invest critical skills to explore, model, synthesize, and test. Meanwhile, hardware and software advances will continue to make graphical techniques not only an information-rich diet for perceptive investigators but an aesthetically pleasing medium as well.

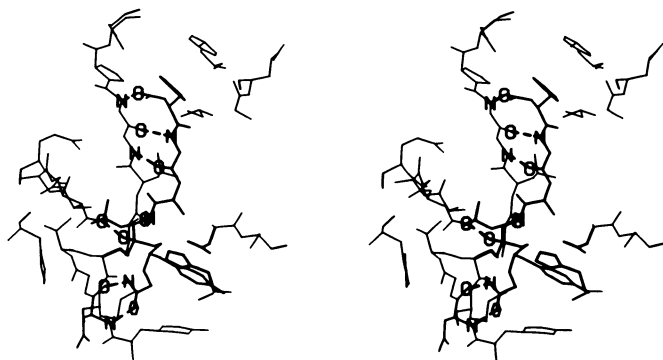


Figure 11. Stereoview of a proposed boronate inhibitor to subtilisin showing favorable stacking interactions and hydrogen bridges

Acknowledgment

We thank Drs. Dutler, Van Opdenbosch, and Pèpe for contributing helpful comments and useful computer techniques. We acknowledge the support of the R.A. Welch Foundation, N.S.F., N.I.H., and the Texas Agricultural Experiment Station.

All stereoviews were photographed directly from the cathode ray tube display. Even in stereo it is difficult to present (and perceive) all the important steric interactions. These views thus must serve as single snapshots of a multifaceted modelling technique. Naturally, the interactive manipulation of groups cannot be illustrated satisfactorily with snapshots; direct access and control of the display is necessary. Never-the-less, we hope that these views will help illustrate both the message and the method.

Literature Cited

- ¹Meyer, E.F., Jr. "Interactive Graphics in Medicinal Chemistry", in "Drug Design", Vol. IX, E.J. Ariens, Ed., in press.
- ²Morimoto, C.N., and Meyer, E.F., Jr., "Crystallographic Computing Techniques", 488-496, F.R. Ahmed, Ed., Munksgaard, Copenhagen, 1976.
- ³Bernstein, F.C., et al., *J. Mol. Biol.* (1977) **112**, 535-542.
- ⁴Meyer, E.F., Jr., *Biopolymers* (1974) **13**, 419-422.
- ⁵Legg, M.J., "Protein Crystallography: New Approaches to X-ray Data Collection, Direct Space Refinement and Model Building", Dissertation, Texas A&M University, 1978.
- ⁶Swanson, S.M., *J. Mol. Biol.*, in press.
- ⁷Eklund, H., Nordstrom, B., Zeppezauer, E., Soderlund, G., Ohlsson, I., Bowie, T., Soderlund, B.O., Tapia, O., Brändén, C.I., and Akeson, A., *J. Mol. Biol.* (1976) **102**: 27-60.
- ⁸Huber, R. and Bode, W., *Acc. Chem. Res.* (1978) **113**: 114-121.
- ⁹Reynier, M., Theorell, H., Sjoval, J.: Studies on the Stereospecificity of Liver Alcohol Dehydrogenase for 3- β hydroxy-5 β steroids. *Acta Chem. Scand.* (1969) **23**: 1130-1136.
- ¹⁰Cronholm, T., Larsen, C., Sjoval, J., Theorell, H., Akeson, A.: Steroid Oxidoreductase Activity of Alcohol Dehydrogenases from Horse, Rat and Human Liver. *Acta Chem. Scand.* (1975) **B29**, 571-576.
- ¹¹Brandén, Carl-Ivar, "Functional Significance of the Structure of Liver Alcohol Dehydrogenase", in "Pyridine Nucleotide-Dependent Dehydrogenases", H. Sund, Ed. W.G. Groeter, Berlin, pp. 325-338 (1977).
- ¹²Dutler, Hans, "Substrate Orientation in the Active Site of Liver Alcohol Dehydrogenase", *ibid*, pp. 339-350.
- ¹³Robertus, J.D., Alden, R.A. Birktoft, J.J., Kraut, J., Powers, J.C., and Wilcox, P.E., *Biochem.* (1972) **11**: 4293-4303.

¹⁴Powers, J.C., Lively, M.O., and Tippet, J.T., Biochem. Biophys. Acta (1977) 480: 246-261.

¹⁵Alden, R.A., Birktoft, J.J., Kraut, J., Robertus, J.D., and Wright, C.S., Biochem. Biophys. Res. Comm. 45: 337-344 (1971).

¹⁶Lindquist, R.N., Nguyen, A.C., J. Am. Chem. Soc. 99(19): 2435-2437 (1977).

¹⁷Kraut, J., Personnel Communication.

RECEIVED June 8, 1979.

The Conformational Parameter in Drug Design: The Active Analog Approach

GARLAND R. MARSHALL, C. DAVID BARRY, HEINZ E. BOSSHARD,
RICHARD A. DAMMKOEHLER, and DEBORAH A. DUNN

Departments of Physiology and Biophysics, of Pharmacology and of Computer Science and the Computer Systems Laboratory, Washington University, St. Louis, MO 63110

Molecular recognition plays a central role in biological systems and is the basis of the specificity seen in antigen-antibody, substrate-enzyme, hormone-receptor, and drug-receptor interactions. The importance of this problem had been recognized by scientists like Pasteur and Ehrlich at the end of the nineteenth century. Structure-activity relations in which one probes the basis of recognition by chemically modifying one of the partners form a significant segment of medicinal chemistry. Attempts to correlate activity with physical properties such as solubility, refractive index, etc. are a reasonable approach to defining the characteristic properties associated with activity in a given biological assay (1). In most systems, these correlations reflect the effective concentration of the drug in the critical biophase rather than the critical features necessary for bimolecular recognition. We would like to concentrate on what can be deduced regarding bimolecular recognition from structure-activity data.

A major dichotomy between quantitative structure-activity relations (QSAR) and the active analog approach (AAA) exists on two levels. First, potency is the primary raison d'etre of QSAR and depends on a multitude of parameters such as distribution, metabolism, transport, recognition, affinity, etc. On the other hand, good pharmacological data implying recognition at a common receptor site is the criterion for inclusion in AAA regardless of the relative potency if one can be convinced the activity lies in the compound tested and not some contaminant, etc. The purpose is to deduce the minimal recognition requirements as a basis for understanding how a diverse set of chemical structures can activate the same receptor. Secondly, recognition is critically dependent on the three-dimensional arrangement of electron density in a molecule and the determination of the receptor-bound conformation is a logical goal in attempting to deduce recognition requirements. QSAR has traditionally neglected the conformational parameter although the use of steric parameters has become increasingly more sophisticated

0-8412-0521-3/79/47-112-205\$05.50/0

© 1979 American Chemical Society

(2).

A number of approaches to the conformational parameter have been taken. Crystallographic determinations of the molecular structure have been made, and proposals relevant to the receptor-bound conformation have been based on the crystal structures. Alternatively, spectroscopic determinations of the solution conformation have been made in attempts to correlate this three-dimensional shape with biological activity. Theoretical calculations have also been used to determine the most probable conformer. Each of these approaches makes the tacit assumption that the conformer of minimum energy is directly correlated with the receptor-bound conformer. This assumption deserves to be challenged based on pure logic as well as experimental observation (3). Each method deals with molecules in a symmetrical force field due to other similar molecules, solvent or vacuum. Biological systems are characterized by stereospecificity due to interactions with asymmetric molecules, e.g. enzymes or receptors. The close proximity to an asymmetric force field characteristic of binding of a small molecule to a receptor must express itself in the conformation assumed by the ligand. Allosteric effects in enzymes have shown that small molecules are capable of inducing conformational changes in large proteins; certainly the converse situation is to be expected.

Experimental data relevant to this issue have become abundant. One example (Fig. 1) is ribonuclease S in which subtilisin cleavage hydrolyzes the peptide bond between residues 20 and 21 of the 124 residue enzyme Ribonuclease A. The S-peptide (residues 1-20) can be dissociated and reassociated. The crystal structure of both ribonuclease A and S have been determined and residues 2-13 form two-three turns of α -helix (4). In the dissociated state, S-peptide contains little detectable helix either by circular dichroism (5) or by NMR (6,7). Either the S-protein induces the bound conformation during the association or a conformer too limited in concentration to be detected by spectroscopy is preferentially bound. It has been argued (8) on kinetic grounds that the induced-fit or "zipper" hypothesis is more likely, but no relevant experimental data exists on this or other systems to our knowledge.

Several cases exist where analogs with essentially equivalent biological activity have been shown to have different conformations in solution. Data exist for somatostatin (9), gramicidin S (10) and angiotensin II (11). One would assume that the analogs assume a similar conformation during activation and that the solution conformation of, at least, one of the analogs must be irrelevant to the biological activity, if not both. A similar argument can be made in the case of gamma-aminobutyric acid in which requirements for recognition by the presynaptic uptake system clearly rest in one conformer while

the post-synaptic receptor recognizes another conformer (12). The relevance of solution conformation to biological activity has been questioned in the case of TRH and enkephalin (13). A conformation for LHRH deduced from empirical energy calculation has been proposed to be responsible for the biological activity (14), but an analog [D-Phe²,Aib⁶]-LHRH which cannot assume this conformation has excellent binding activity (15). With regard to crystal data, numerous examples such as acetylcholine exist where many conformers have been found depending on the conditions of crystallization and, certainly, the activity at both muscarinic and cholinergic receptors cannot be explained on the basis of a single conformer.

In summary, methods which have focused on a single conformer of a flexible molecule have been misleading regardless of the methodology. One must examine the multitude of possible conformations available to flexible molecules in order to adequately sample the possible receptor-bound conformation induced on binding to the receptor. The magnitude of this task has prevented most investigators from pursuing this approach. The active analog approach recognizes the necessity to consider all possible conformations as potential activation states.

Active Analog Approach.

The concept of pharmacophore in which the recognition requirements are defined as the three dimensional arrangement of functional groups essential for recognition and activation is a central concept (Fig. 2). By chemical modification, the medicinal chemist has identified chemical groups essential for activity. Substitution of groups by similar chemical functions retaining activity led to the concept of bioisosterism. The definition of the three dimensional arrangement of the groups essential for activity completes the designation of the pharmacophore. Such pharmacophoric patterns have been proposed for a number of systems (16,17). The initial task is, therefore, the deduction of the pharmacophore. Based on the known structure-activity data, a hypothesis concerning the possible pharmacophoric groups is proposed. The set of possible conformations for each compound shown to activate the receptor under examination is calculated. For each allowed conformation, the relative position of the proposed essential groups is determined. This can be expressed as an orientation map in which each point represents one possible arrangement of the essential groups, i.e. a possible pharmacophoric pattern. For three groups, these points can be plotted in three-dimensions where each point represents a specific triangle (Fig. 3). The intersection of all the orientation maps for the active analogs yields the possible pharmacophoric patterns available to the entire set of analogs. A null result would indicate either incorrect identification of essential groups or inclusion of incorrect pharmacological information. Use of sufficient number of analogs of

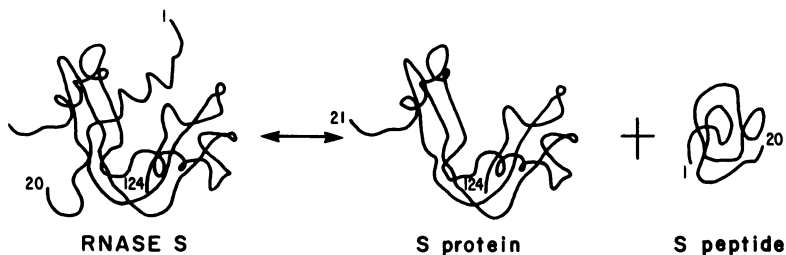


Figure 1. Dissociation of S-peptide from ribonuclease S with associated loss of the helical conformation of a segment of the peptide

Figure 2. The three-dimensional recognition requirements essential for activity—the designated groups A, B, and C and their correct spatial orientation comprise the pharmacophore. The number of groups is not limited to three which is chosen for convenience of presentation.

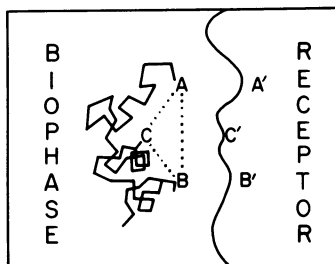
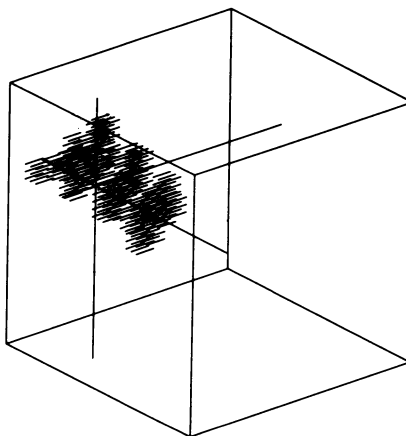


Figure 3. Orientation space map for a molecule with essential groups A, B, and C. Each axis represents the distance between two points.



diverse chemical structures can uniquely determine the pharmacophore. This information can be used as a frame of reference for orienting molecules in the correct receptor-bound conformation as a prelude to mapping the functional volume (excluded volume map) available to drugs which bind. The pharmacophore also serves as a criterion for judging a novel structure for recognition by the receptor. The other criterion is the volume required by the drug when presenting the pharmacophore. Lack of activity in a drug which was capable of presenting the pharmacophore and fitting within the excluded volume space would seriously compromise the proposed pharmacophore unless alternative explanations such as distribution or metabolism could be demonstrated.

A logical extension of mapping the volume available for binding to the receptor is to map the volume which is not available, i.e. receptor essential volume. Certainly, a number of drugs with potential activity may be prevented from interaction by the presence of bulky substituents. Knowledge of the pharmacophore and the excluded volume map allows the unique volume required by the inactive analog to be located in space. Unfortunately, only one segment of this unique volume need be required by the receptor to prevent binding (Fig. 4) and, thus, the total unique volume cannot be used indiscriminately. One scheme which we are investigating normalizes the number of observations of inactivity associated per volume segment by the unique volume associated with each observation. A unique volume requirement of only one volume segment based on an inactive analog would yield a predicted probability of inactivity = 1, if that segment were required by a new test compound. In a similar manner, the pharmacophore can provide a three-dimensional framework for QSAR studies and a basis for correlations across a series of different basic structures.

This discussion should provide the rationale for the following material in which some of the problems which arise in the implementation are presented as well as results illustrated from the various biological systems under exploration.

Systematic Conformational Search.

A major difficulty arises due to the computational complexity of determining all possible conformations at a given grid, i.e. minimal difference between two conformers. Let us compute the number of operations necessary for a molecule of N atoms with M rotatable bonds each of which will be sampled R times, i.e. the angular difference between conformers is $360/R$ degrees. Each atom must be rotated to determine its position in space which takes P_1 operations and then it must be checked against all other atoms for possible Van der Waals collisions which requires P_2 operations for each check and $N(N-1)/2$ pairwise checks. The total number of operations involved in the search is therefore:

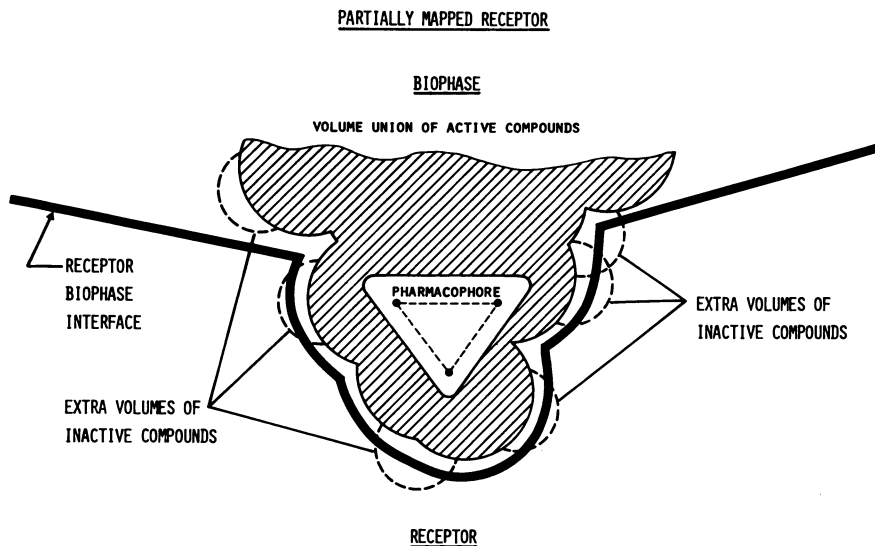


Figure 4. Schematic drug-receptor interaction in which the pharmacophore has been defined and the extra volume requirements of inactive molecules capable of presenting the pharmacophore indicated. If one molecule requires several discrete areas, then one cannot distinguish if all are essential to the receptor, or only a segment of any one, as a single conflict is sufficient to prevent binding.

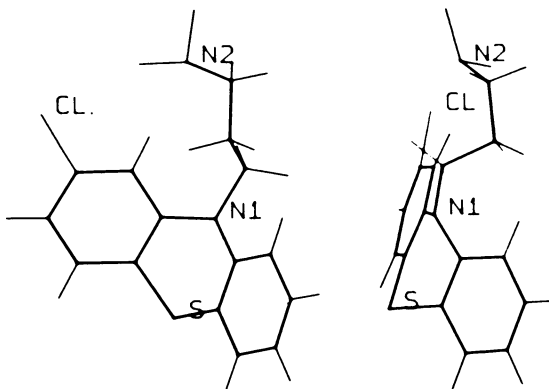


Figure 5. Chlorpromazine which antagonizes a central dopamine receptor has 4 degrees of rotation freedom that position the side chain relative to the phenothiazine ring. Shown is the crystal structure (41).

$$\text{Number of Operations} = (N P_1 + \frac{N(N-1)}{2} P_2) R^M$$

An estimate of the number of operations P_1 required for rotation of a point is given by examining the formulation of Gibbs (18) and others (19).

$$X_i' = T (X_i - X_o) + X_o$$

where T is a matrix (3 x 3) whose elements are functions of the direction cosine and the sine and cosine of the angle of rotation; X_i is the vector (3 x 1) whose elements are the coordinates of atom i , and X_o is a vector of a point along the axis of rotation. P_1 requires 9 multiplications and 12 additions, or 21 operations in our simplified analysis.

P_2 can be derived for the formula for a distance check between two atoms i and j .

$$D_{ij}^2 = [X_i - X_j, X_i - X_j] - VDW_{ij}^2$$

Since the calculation can utilize the squared radius as a criterion, the number of operations required is only 3 multiples and 6 additions without a square root, and P_2 is equal to 9. Our formula can be simplified to:

$$\text{Number of operations} \cong (21 N + \frac{9N(N-1)}{2}) R^M$$

As an example (20) relevant to an application to be discussed later, let us consider the phenothiazine neuroleptic chlorpromazine shown in Fig. 5 which has four rotatable bonds ($M=4$) which determine the orientation of the nitrogen-containing side-chain relative to the phenothiazine ring. If we assume $R=64$ or a grid of approximately 5.6 degrees, then $R^M = 64^4$ or 16,777,216 possible conformations to be examined. If we consider the methyl groups bonded to nitrogen as spheres of rotation, then the number of atoms $N=34$ and the number of operations $\cong 5763 R^M \cong 9.67 \times 10^{10}$.

If one has a computer of moderate capability such as a DEC 20/40 which can perform a floating point multiplication in 2.5 microseconds, then such a calculation should take approximately 241,718 seconds or 67 hours. Over the past several years, numerous improvements in the algorithms used in systematic conformational search have been developed (21) leading to a reduction on the order of $(M+1)^2 2^M$. A prototype system on a DEC 20/40 written in Fortran without any optimization completed this search in under 6 minutes CPU time. Out of the 16,777,216 possible conformations, only 78,670 were sterically allowed. The number capable of presenting the appropriate pharmacophore to the receptor was less than 50. A facility for computer-aided

drug design is being established at Washington University which has an array processor coupled to a 32-bit host computer as the computation engine for the systematic search of conformational space. Estimates indicate that problems of eight to ten rotatable bonds should be approachable utilizing this hardware, enabling consideration of most molecules of pharmacological interest. While the combinatorial nature of a systematic search presents formidable computational demands, both the increased performance and availability of computational hardware as well as the increased understanding resulting in algorithmic improvements make this approach feasible. A number of other problems such as the determination of conformations using lanthanide shift reagents (22) logically require the determination of all possible answers, and avoidance of problems such as local minima associated with minimization is necessary.

Strategy.

Because systematic search is a combinatorial problem, the choice of the order of processing molecules to determine the pharmacophore can dramatically affect the computational demands. Determination of the orientation map for the most sterically constrained molecule with the least number of variables places limits on the possible pharmacophores which translate into distance constraints on subsequent molecules. These distance constraints can be used to effectively truncate the combinatorial search tree eliminating much of the computation. Since the object is to determine if overlap in pharmacophoric patterns between a set of molecules can occur, conformations which cannot satisfy all molecules are not of interest. An extension of this concept to assist in looking for possible pharmacophoric groups derives from the construction of a distance range matrix for each molecule (23). A matrix of the maximum and minimum pairwise distance between atoms is constructed analytically through use of distance geometry (24) as shown in Fig. 6. A minimum condition for selection of a pair of groups as recognition sites is that the ranges in this matrix overlap for all active molecules. In fact, one can limit the search to the range consisting of the smallest maxima and largest minima which contains all possible distances common to the set of molecules.

When faced with a set of flexible molecules with too large a number of variables, one must turn to chemical modification to reduce the number of variables to a manageable set. In the case of peptide hormones, introduction of proline eliminates the torsional degree of freedom ϕ by introducing a cyclic constraint, but adds the necessity of considering both cis and trans isomers of the amide bond. Replacement of the α -proton of an amino acid by a methyl group has been shown to reduce the number of possible combinations of values for the ϕ and ψ torsional rotation to essentially two (25). Introduction of a cyclic constraint in a linear flexible molecule reduces the

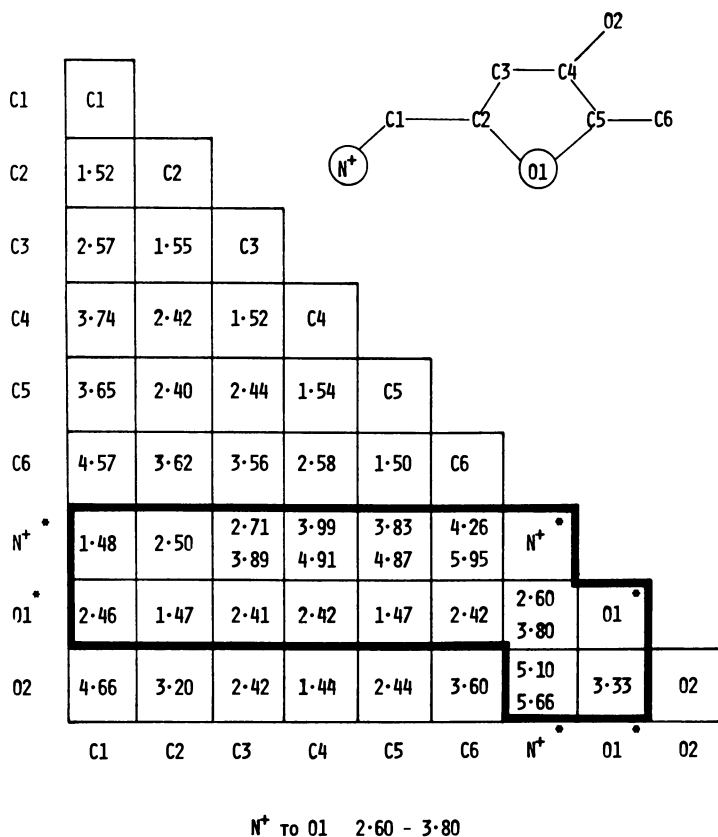


Figure 6. Distance range matrix for sterically allowed conformations of muscimol, a muscarinic agonist. Distances nonvariant under rotation are single entries. Paired entries represent maximum and minimum distances that are sterically allowed.

degrees of freedom by five (26). Recent examples of this approach to determining the receptor-bound conformation of peptide hormones can be found with somatostatin (27) and LHRH (28).

Exploitation of the Pharmacophore.

Once a reasonable candidate for the pharmacophore is available, then the associated distance constraints can be used to examine the remaining analogs active at the receptor under investigation for their ability to assume the pattern. If the question is simply one of the ability to present the pharmacophore, then a constrained minimization can be used. Since there is usually more than one conformer capable of the pharmacophore, however, and volume considerations may dictate which conformer is more likely, a systematic search utilizing the distance constraints from the pharmacophore is preferable. The entire set of active compounds can be screened for consistency with the proposed pharmacophore to test its credibility.

Since the pharmacophore provides a common frame of reference for orienting molecules in their receptor-bound conformations, one can utilize this frame for examining receptor-drug interactions in detail. In particular, the volume available for binding at the receptor can be mapped and accessory binding sites, which contribute to affinity but are not necessary for recognition, defined. In addition, one can correlate affinity or activity across different series of compounds by associating properties with particular volumes of space. Some method of handling volumes and visually indicating them is necessary for manipulation and perception. The MMS-X graphics system developed at Washington University (29) for electron density fitting in protein crystallography has proven essential to these needs. A pseudoelectron density map is constructed for each molecule in its bound conformer utilizing Gaussian functions which have been calibrated in terms of Van der Waals radii (30). A set of programs allows union, intersection, subtraction and addition of these density maps and contouring at levels corresponding to space-filling physical models. One can perform union of the density maps for the set of active compounds in order to determine the excluded volume map, i.e. that volume which is available to drugs interacting with the receptor as evidenced by active analogs requiring that volume. An example with the three stereoisomers of 2-aminonorborene-2-carboxylic acid which competes with methionine for recognition by S-adenosylmethionine synthetase (31) is shown in Fig. 7. The pharmacophore consists of the amino, carboxyl and α -carbon of the amino acids. The fourth stereoisomer of 2-aminonorborene-2-carboxylic acid is not recognized by the enzyme up to the level of solubility. Since it resembles the other three isomers, a reasonable hypothesis would be that it requires some volume when presenting the pharmacophore that is occupied by the

receptor. Subtraction of the volume in common with the excluded volume map (obtained by intersection of the inactive analog and the excluded volume map) from the inactive analog leaves only a small volume which is required by the receptor and designated "receptor essential volume" as shown in Fig. 8.

Applications.

A number of areas are being investigated to determine the applicability of the Active Analog Approach. Emphasis has been placed on systems with minimal conformational degrees of freedom in order to gain insight into the tolerance of receptors for variation in the pharmacophore. This is necessary to establish criteria for acceptable limits on the correctness of fit of a novel structure to the proposed pharmacophore. Use of rigid bicyclic amino acid analogs to probe the methionine binding site of S-adenosylmethionine by the Talalay group (32) provides an opportunity to map an enzyme site prior to structure determination by crystallography. Each analog is positioned uniquely on a frame consisting of amino, carboxyl and α -carbon. The excluded volume map has been constructed from active analogs which compete with methionine (Fig. 7) and used to indicate (31) the novel volumes required by inactive analogs (Fig. 8). Using the bicyclic ring structure as a framework, it should be possible to construct analogs which probe the binding site selectively by replacement of protons with halogen or methyl groups. In view of the increasing importance of S-adenosylmethionine as the key intermediate in biological methylation and the presence of isoenzymes, novel inhibitors which exploit differences in steric requirements could be key pharmacological tools.

Another area of interest has been the recognition of glucose by a proposed glucose receptor in the β -cells of the pancreatic islets which controls insulin release (33). An alternative proposal involving a metabolite of glucose as the primary signal is also advocated, and our study (34) was intended to ascertain the plausibility of the glucose recognition site by examination of compounds active in stimulating insulin release from isolated islets of Langerhans. D-glucose, D-mannose and their anomers initiate insulin release and prevent alloxan (Fig. 9) from irreversibly inhibiting subsequent glucose-induced insulin release. An excluded volume map for the active hexoses was constructed (Fig. 10) and both alloxan and ninhydrin which cause a monophasic burst of insulin release prior to inhibition were shown to fit within this space when presenting the pharmacophore consisting of: an oxygen at position one, either axial or equatorial; an hydroxyl at position two, either axial or equatorial; an equatorial oxygen at position three; and an electron-rich region at position five (35).

Biogenic amines are also under investigation and can be illustrated with the CNS dopamine receptor whose blockade is

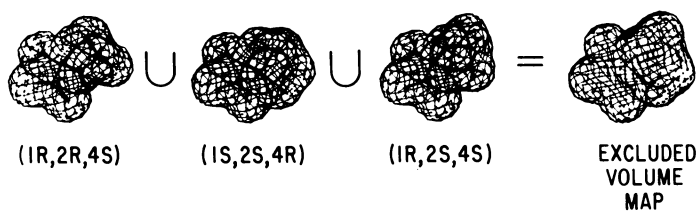


Figure 7. Union of the volume of the active isomers to obtain combined volume available for binding to the methionine site

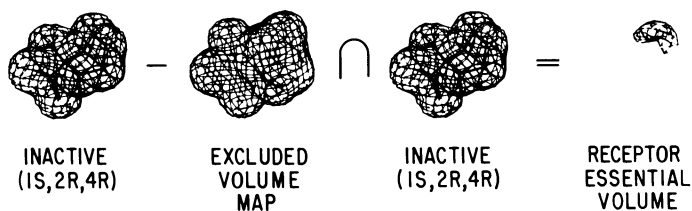


Figure 8. Subtraction of the volume common to the active and inactive isomers from the inactive isomer to obtain the unique volume responsible for negative interaction with the receptor

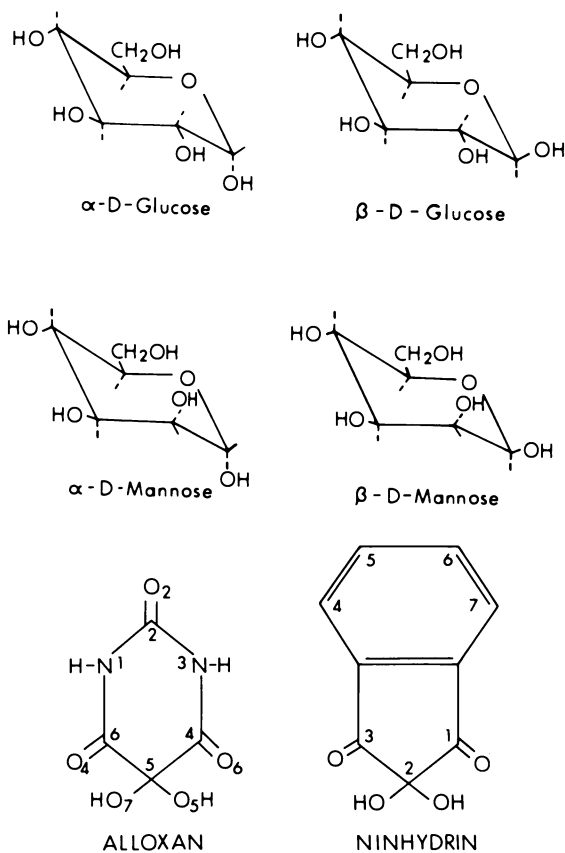


Figure 9. *Compounds that stimulate insulin release from pancreatic islets. Hexoses both stimulate insulin release and prevent alloxan inhibition. Both alloxan and ninhydrin stimulate insulin release prior to irreversible inhibition.*

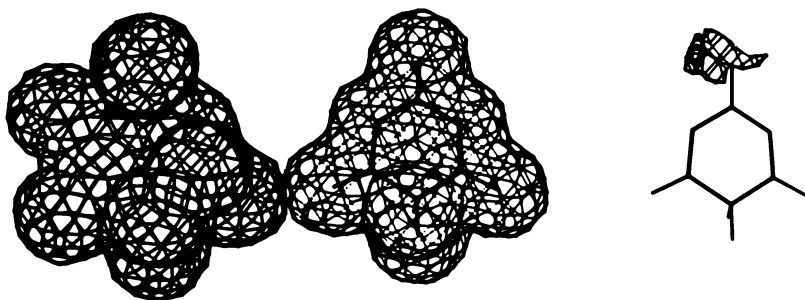


Figure 10. Excluded volume map for active hexoses pictured in Figure 9. Volume map of alloxan and difference map showing small region of alloxan not accommodated by hexose map.

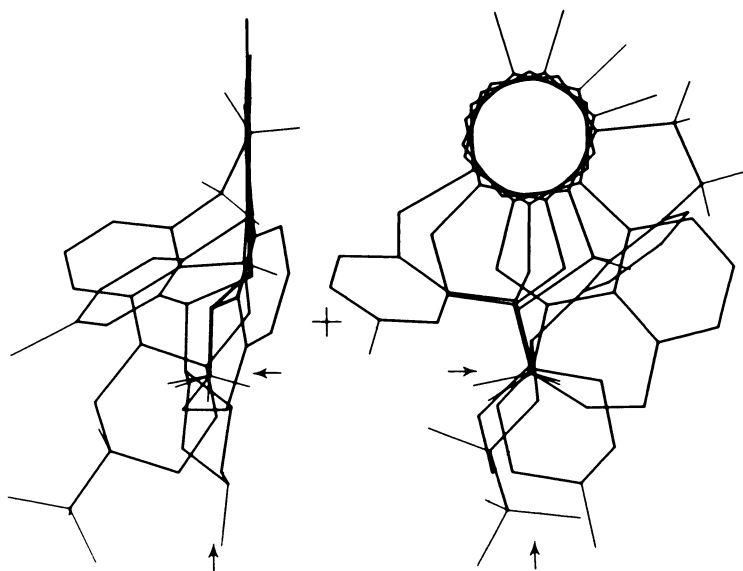


Figure 11. Two orthogonal views of apomorphine, 6,7-ADTN, butaclamol, and octoclothePIN oriented on common frame of reference, the pharmacophore. Nitrogen position indicated by arrows.

associated with neuroleptic activity. Based on the seminal observation of Humber *et al.* (36) of superposition of the dopamine agonist apomorphine with the semirigid neuroleptic butaclamol in which one aromatic ring and the nitrogen overlap, we have extended in collaboration with Dr. Leslie Humber (37) this hypothetical pharmacophore to examine octoclothebin, 2-amino-6,7-dihydroxy-1,2,3,4-tetrahydronaphthalene (6,7-ADTN) as well as flexible neuroleptics such as chlorpromazine. As the pharmacophore for binding, we have required an aromatic ring whose center is approximately 5.1 Å from the projection of a nitrogen lying approximately 1 Å above the plane of the aromatic ring. All of the semirigid agonists and antagonists we have examined can be oriented to present this pharmacophore (Fig. 11). Agonism appears to require a hydroxyl function corresponding to the 5 position of apomorphine in addition to the binding pharmacophore. While one can conceive of multiple modes of binding which would lead to inhibition and it is preferable to concentrate on agonists due to the more rigorous spatial requirements assumed for activation, the correspondence between agonists and antagonists seems compelling in this case. An excluded volume map has been constructed and a consistency check made using the inactive isomers of apomorphine, octoclothebin and butaclamol which show remarkable stereospecificity. Because our pharmacophore has a plane of symmetry represented by a line between the nitrogen and the ring center, all the inactive stereoisomers are also capable of presenting the binding pharmacophore. The fact that they do not bind leads one to the hypothesis that each requires novel volume which is occupied by the receptor. If any of these inactive isomers fit entirely within the excluded volume map when oriented at the pharmacophore, then either the underlying hypotheses would have to be revised or the pharmacological observations questioned. Each of the three, however, requires novel volume as shown in Fig. 12 for (S)-apomorphine.

The procedures have suggested binding activity for a novel class of compound, a substituted *trans*-hexahydrocarbazole, whose crystal structure was available (38) and whose fit to apomorphine is shown in Fig. 13. Activity in both neuroleptic screens and dopamine binding assays (39) confirm this analysis. In addition, suspicions aroused during modeling regarding the optical antipode responsible for the activity of 6,7-ADTN were confirmed by the resolution and characterization of the optical isomers by McDermed *et al.* (40) who have shown the isomer responsible for activity to have an opposite configuration at the carbon bearing the nitrogen from apomorphine. This allows the orientation of 6,7-ADTN as shown in Fig. 14 which allows the 7-hydroxyl to approximate the position of the 5-hydroxyl of apomorphine when presenting the binding pharmacophore thus explaining the agonist activity.

Definitive studies on the thousands of flexible neuroleptics

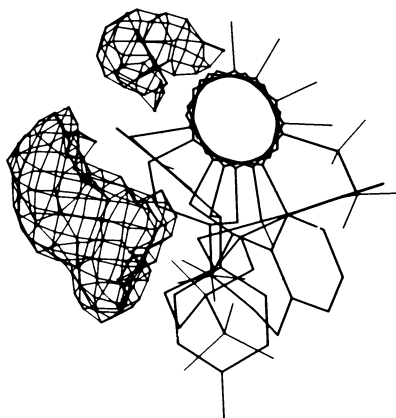


Figure 12. Novel volume required by inactive isomer (S)-apomorphine oriented by pharmacophore

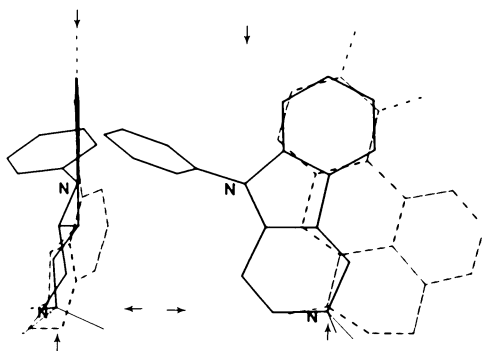


Figure 13. Two orthogonal views of alignment of trans-hexahydrocarbazole analog apomorphine (---). Nitrogen position indicated by arrows.

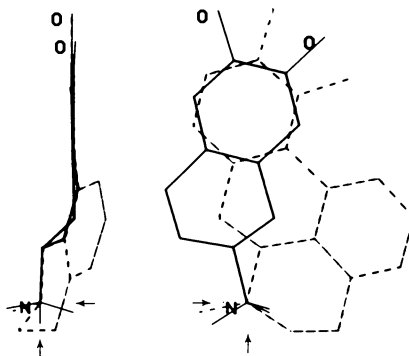


Figure 14. Two orthogonal views of 6,7-ADTN aligned with apomorphine (---). Nitrogen position indicated by arrows.

present more difficulties. Of the 78,000 sterically allowed conformers of chlorpromazine, less than 50 fit the binding pharmacophore and these consisted of very similar families (Fig. 15). Because the phenothiazine ring is capable of inversion (41), the optical isomers must also be considered and thus 4 ring to nitrogen combinations are possible, i.e. crystal configuration nitrogen to either ring or isomerized configuration nitrogen to either ring. While all require some additional volume, with some conformations the novel volume is very small. Additional studies on other flexible neuroleptics which show similar aromatic substituent effects to chlorpromazine should allow assignment of the correct binding mode.

An attempt (42) to use a simplified analgesic pharmacophore to define the receptor-bound conformation of enkephalin has resulted in predictions (43) regarding the activity of enkephalin analogs in which modifications affect backbone conformational possibilities. These predictions have been remarkably consistent with the data (44) considering that the criteria was extremely simple. If an analog was capable of assuming the proposed receptor-bound conformation, then the analog was predicted to be active; otherwise, inactivity was predicted. It is somewhat incredulous that this simplified approach has been so successful as it stresses the backbone conformation and not the pharmacophore itself. Recent synthetic studies (45) have yielded active analogs capable of presenting the pharmacophore, but incapable of assuming the backbone conformation. One interesting observation is the marked similarity in requirements of the analgesic and neuroleptic pharmacophore which implies that the enkephalin receptor may have evolved from a biogenic amine receptor and which led us to discover a whole literature of pharmacological studies indicating crossover between neuroleptic and analgesic compounds. Definition of the pharmacophore and volume tolerance of other receptors should be useful to predict side effects. Toward this end, studies of GABA, TRH, histamine and catecholamine receptors are underway. Preliminary studies of the glutamic acid binding site which inhibits feeding in *Hydra* have been presented (46).

Pitfalls.

There are a multitude of ways in which one can abuse this approach. Of primary importance is the quality of the pharmacological data. Inclusion of a compound with a different mechanism of action should ultimately lead to a null hypothesis, i.e. no consistent pharmacophore, but may require much effort to detect. Alternatively, the procedure may be useful in identifying compounds to be further investigated pharmacologically, similar to the use of outliers in QSAR. One objection often raised is the obvious flexibility associated with proteins. The goal is not to compete with protein crystallography, but rather to derive a functional description of the volume available for

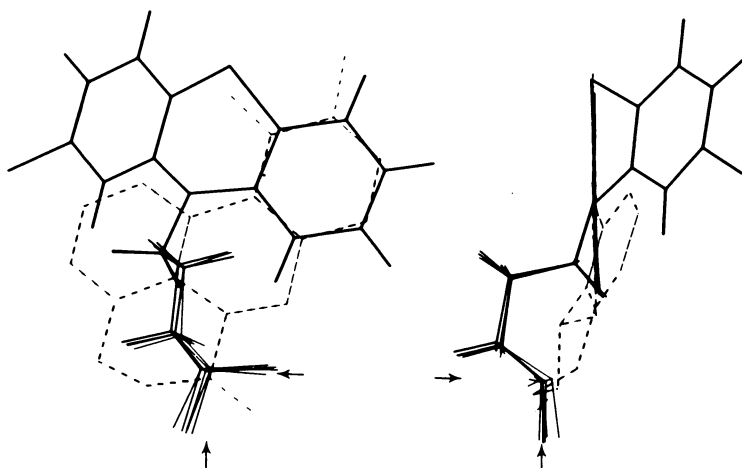


Figure 15. One of a family of conformers of chlorpromazine fit within limits criteria for the neuroleptic receptor. Shown in two orthogonal views aligned with apomorphine (---). Nitrogen position indicated by arrows.

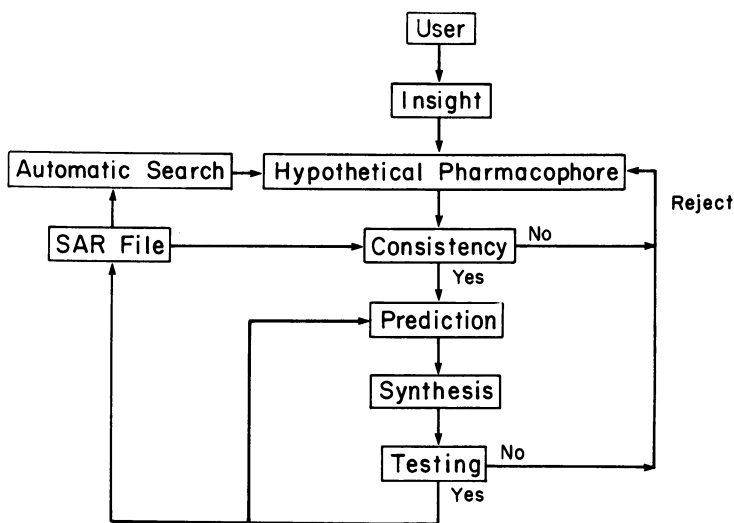


Figure 16. Flow chart of the active analog approach as a tool for the practicing medicinal chemist

analogs at the receptor and an indication of the drug-receptor interface. Changes in the pharmacophore itself are conceptually possible and would present difficulties unless a common spatial arrangement for activation was essential. Otherwise, one is dealing with two mechanisms which are transduced by the same receptor. One interesting mechanism proposed for auxins (47) is a concerted conformational change of both the drug and the receptor. This hypothesis is very attractive in the additional degrees of specificity implied, but the entire pathway would be evident in the orientation map as being common to all agonists. It is only the advent of receptor-binding assays which allows one to critically test for a common recognition event, and this technique provides an essential ingredient in the critical application of the active analog approach.

Conclusions. As an alternative to the methods available for ascertaining the conformational parameter which ignore the influence of the receptor, the active analog approach has been developed in which activity with the receptor itself is the ultimate criterion. Techniques for pharmacophore identification and exploitation including receptor mapping have been developed. Systematic search of conformational space in order to enumerate all possible pharmacophores is an essential element in the active analog approach. What is advocated is a more sophisticated space filling model based on modern computer technology which allows the testing of hypotheses against the structure-activity data base for consistency and advocates prediction of activity as a prelude to synthesis and testing as shown in Fig. 16.

Literature Cited

1. Hansch, C. Acc. Chem. Res. (1969), 2, 232.
2. Verloop, A. and Tipker, J. in "Biological Activity and Chemical Structure", J.A. Keverling Buisman, ed., p. 63-81, Elsevier Scientific Publishing Co., New York, 1977.
3. Marshall, G.R. Proc. 6th Int. Symp. Med. Chem., Brighton, 1978, (in press) and references cited therein.
4. Wycoff, H.W., Hardman, K.O., Allewell, N.M., Inagami, T., Johnson, L.N. and Richards, F.M. J. Biol. Chem. (1967), 242, 3984.
5. Klee, W.A. Biochemistry (1968), 7, 2731.
6. Finn, F.M., Dadok, J. and Bothner-By, A.A. Biochemistry (1972), 11, 455.

7. Chaiken, I.M., Freedman, M.H., Lyerla, Jr., J.R. and Cohen, J.S. J. Biol. Chem. (1973), 248, 884.
8. Burgen, A.S.V., Roberts, G.C.K. and Feeney, J. Nature (1975), 253, 753.
9. Arison, B.H., Hirschmann, R. and Veber, D.F. Bioorg. Chem. (1978), 7, 447.
10. Kumar, N.G., Izumiya, N., Miyoshi, M., Sugano, H. and Urry, D.W. J. Amer. Chem. Soc. (1975), 97, 4105.
11. Turk, J., Needleman, P. and Marshall, G.R. Mol. Pharmacol. (1976), 12, 217.
12. Krogsgaard-Larsen, P. and Johnston, G.A.R. J. Neurochem. (1978), 30, 227.
13. Marshall, G.R., Gorin, F.A. and Moore, M.L. Ann. Rep. Med. Chem. (1978), 13, 227.
14. Momany, F.A. J. Med. Chem. (1978), 21, 63.
15. Beattie, C.W., Corbin, A., Foell, T.J., Garsky, V., McKinley, W.A., Rees, R.W.A., Sarantakis, D. and Yardley, J.P. J. Med. Chem. (1975), 18, 1247.
16. Gund, P. Prog. Mol. Cell. Biol. (1977), 5, 117.
17. Gund, P. Ann. Rep. Med. Chem. (1979), in press.
18. Gibbs, J.W. "Vector Analysis", Yale University Press, New Haven, 1901.
19. Barry, C.D., Ellis, R.A., Graesser, S.M. and Marshall, G.R. in "Pertinent Concepts in Computer Graphics", M. Faiman and J. Nievergelt, eds., p. 104, University of Illinois Press, Chicago, 1969.
20. Marshall, G.R., Bosshard, H.E., Barry, C.D., Dammkoehler, R.A., Dunn, D.A. and Zyda, M., unpublished results.
21. Dammkoehler, R.A., Bosshard, H.E. and Marshall, G.R., unpublished results.
22. Sievers, R.E., ed. "NMR Shift Reagents". Academic Press, New York, 1973.
23. Barry, C.D. and Marshall, G.R. 6th Int. Symp. Med. Chem. (1978), p. 33.

24. Blumenthal, L.M. "Theory and Application of Distance Geometry". Chelsea Publishing Co., New York, 1970.
25. Marshall, G.R. and Bosshard, H.E. Circulation Res. (1972), 31, II-143.
26. Go, N. and Scheraga, H.A. Macromolecules (1970), 3, 178.
27. Veber, D.F., Holly, F.W., Palaveda, W.J., Nutt, R.F., Bergstrand, S.J., Torchiana, M., Glitzer, M.S., Saperstein, R. and Hirschmann, R. Proc. Natl. Acad. Sci. USA (1978), 75, 2636.
28. Seprodi, J., Coy, D.H., Vilchez-Martinez, J.A., Pedroza, E., Huang, W.Y. and Schally, A.V. J. Med. Chem. (1978), 21, 993.
29. Molnar, C.E., Barry, C.D. and Rosenberger, F.U. "MMS-X Molecular Modeling System". Computer Systems Laboratory Technical Memorandum #229, Washington University, St. Louis, 1976.
30. Marshall, G.R. and Barry, C.D. Abst. Amer. Crystallography Assoc. (1979), Honolulu (in press).
31. Sufrin, J.R. and Marshall, G.R. Federation Proc. (1979), 38, 562.
32. Chou, T.C., Coulter, A.W., Lombardini, J.B., Sufrin, J.R. and Talalay, P. in "The Biochemistry of Adenosyl-methionine". F. Salvatore et al., eds., p. 18-36. Columbia University Press, New York. 1977.
33. Weaver, D.C., McDaniel, M.L., Naber, S.P., Barry, C.D. and Lacy, P.E. Diabetes (1978), 27, 1205.
34. Weaver, D.C., Barry, C.D. and Lacy, P.E. Diabetes (1978), 27, 456.
35. Weaver, D.C., Barry, C.D., McDaniel, M.L., Marshall, G.R. and Lacy, P.E. Mol. Pharmacol. (in press).
36. Humber, L.G., Bruderlein, F.T. and Voith, K. Mol. Pharmacol. (1975), 11, 833.
37. Marshall, G.R., Barry, C.D. and Humber, L.G. Abst. Metrochem.'78 Reg. ACS Meeting (1978), 7, and unpublished work.

38. Berger, J.G., Teller, S.R., Adams, C.D. and Guggenberger, L.J. Tetrahedron Lett. (1975), 1807 and personal communication (LJG).
39. Humber, L.G. and Voith, K. personal communication.
40. McDermed, J.D., Freeman, H.S. and Ferris, R.M. in "Catecholamines: Basic and Clinical Frontiers", E. Usdin, ed. p. 568, Pergamon Press, N.Y., 1979.
41. McDowell, J.J.H. in "Molecular and Quantum Pharmacology". E. Bergmann and B. Pullman, eds., p. 269, D. Reidel Publishing Co. Dordrecht, Holland, 1974.
42. Gorin, F.A. and Marshall, G.R. Proc. Natl. Acad. Sci. USA (1977), 74, 5179.
43. Gorin, F.A. and Marshall, G.R. in "Peptides", M. Goodman and J. Meienhofer, eds. p. 277, John Wiley and Sons, Inc., New York, 1977.
44. Gorin, F.A., Balasubramanian, T.M., Barry, C.D. and Marshall, G.R. J. Supramol. Struc. (1978), 9, 27.
45. Gorin, F.A. Dissertation, Washington University, St. Louis, 1978.
46. Cobb, M.H., Marshall, G.R., Heagy, W., Danner, J. and Lenhoff, H.M. Federation Proc. (1978), 37, 1822.

RECEIVED June 8, 1979.

Mapping the Dopamine Receptor: Some Primary and Accessory Binding Sites

L. G. HUMBER, A. H. PHILIPP, F. T. BRUDERLEIN,
M. GÖTZ, and K. VOITH

Ayerst Research Laboratories, Montreal, Canada, H3C 3J1

The concept of biological receptors for neurotransmitters was first enunciated by Langley in 1905 (1). In the intervening decades, particularly in the past years, much progress has been made in the characterization of receptors, especially with the advent of high-affinity, receptor-specific radioligand binding assays (2).

The medicinal chemist faces the task of designing specific, high-affinity receptor agonists and antagonists as potential therapeutic agents. A direct approach to this problem would be to isolate a crystalline complex between the receptor macromolecule and its ligand, and to determine the molecular architecture of the receptor by X-ray crystallographic methods. Such studies with crystalline enzyme-substrate complexes have shed much light on the nature of enzyme active sites (3).

Although the acetylcholine receptor protein, the most extensively studied receptor system, has been isolated in a fair degree of purity (4), the possibility of obtaining crystalline receptor-ligand complexes has not even been broached.

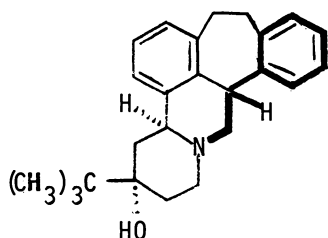
In this circumstance the medicinal chemist must resort to an indirect approach to study the molecular architecture of receptors. This approach, receptor mapping (5), comprises a detailed analysis of the molecular architecture of specific high-affinity, chiral ligands which can undergo minimal, or no conformational change. The results of such analyses can then be interpreted in terms of complementary topographical features of the recognition site on the receptor macromolecule.

The present report describes our studies aimed at mapping topographical features of the receptor(s) for the central neurotransmitter dopamine. These studies evolve from a program which has led to the development of the clinically-active (6-11) anti-psychotic agent butaclamol (12).

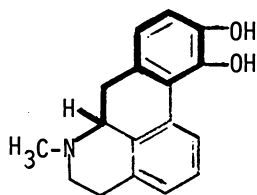
Butaclamol exerts its effects by blockade of central dopamine receptors (DAR), and is unique in that it was the first DAR antagonist wherein all the activity was due to one enantiomer (13, 14). Resolution studies (14) had demonstrated that (+)-butaclamol, I,

0-8412-0521-3/79/47-112-227\$05.00/0

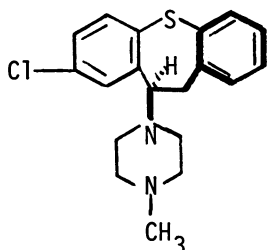
© 1979 American Chemical Society



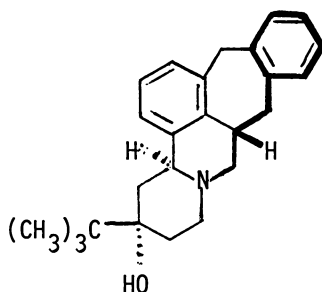
I. (+)-Butaclamol



II. (-)-Apomorphine



III. (+)-Octoclothebin



IV. (+)-Isobutaclamol

was a potent DAR antagonist while the (-)-enantiomer was devoid of activity. This property was utilized by the groups of Seeman (15) and of Snyder (16) to develop a DAR binding assay, since it was now possible to differentiate specific from non-specific binding.

Thus, the enantiomers of butaclamol are differentiated by the DAR with absolute enantiospecificity. Apart from this attribute of butaclamol, another unique feature is the restricted conformational mobility of its nucleus. There exists another DAR ligand, the agonist (-)-apomorphine, II, (17), which also possesses a semi-rigid nucleus and for which the DAR is equally enantio-specific (18).

We have analyzed (14) the molecular structures of these semi-rigid chiral ligands and, in both, we have identified a common chiral pharmacophore, namely a phenyl ring and a nitrogen atom, uniquely juxtapositioned one to the other, which is contained in an extended phenethylamine moiety. We have since shown that the same pharmacophore is present in the DAR antagonist (+)-octoclothebin, III, as well as in several other DAR ligands (19).

We have therefore suggested (14) that the phenyl ring and the nitrogen atom of the pharmacophore are responsible for butaclamol's activity, by interacting with *primary binding sites* on the DAR.

Numerous analogs of butaclamol have been investigated and we

now describe our studies with one of them, isobutacclamol (IV), since, as we will see in the sequel, it provides information on some fundamental properties of the DAR.

Isobutacclamol, is the benzo[5,6]cyclohepta analog of butacclamol. The synthesis and resolution of isobutacclamol have been described recently (20, 21). The racemate and the enantiomers have been studied as DAR antagonists (21), both *in vivo* (antagonism of amphetamine-induced stereotyped behaviour in rats (22)), and *in vitro* (inhibition of tritiated haloperidol binding by homogenized rat caudate nucleus (23)). The results are presented in Table 1. It is apparent that (±)- and (+)-isobutacclamol are about equipotent as DAR antagonists, to (±)- and (+)-butacclamol both *in vivo* and *in vitro*. Furthermore, like (-)-butacclamol, (-)-isobutacclamol is devoid of such activity.

Table 1

DAR Antagonist Properties of Butacclamol and Analogs

Compound	Antagonism of ASB mg/kg, i.p. MED ^a	³ Inhibition of H-Haloperidol binding ^b IC ₅₀ , nM ^b
(±)-butacclamol.HCl	0.62	3.5
(+)-butacclamol.HBr	0.31 ^c	1.4
(-)-butacclamol.HBr	> 50 ^c	> 1000 ^d
(±)-isobutacclamol.HCl	0.62	1.7
(+)-isobutacclamol.HBr	0.31	0.8
(-)-isobutacclamol.HBr	> 25 ^c	> 1000 ^d
(±)-anhydrobutacclamol.HCl	0.62	9.5
(±)-desoxybutacclamol.HCl	0.62	6.3

^aMinimal effective dose antagonizing amphetamine stereotypy. For methodology see reference 22. ^bFor methodology see reference 23. ^cHighest dose tested. ^dHighest concentration tested.

Butacclamol's activity was ascribed to the presence of the neuroleptic pharmacophore within its extended phenethylamine moiety. Inspection of the structure of isobutacclamol reveals no such phenethylamine moiety, but instead, shows the presence of a phenylpropylamine moiety. It thus appears that butacclamol and isobutacclamol are incapable of presenting the same pharmacophore to the DAR. We have therefore studied the molecular structures of butacclamol and isobutacclamol more closely with a view to resolving this apparent anomaly.

The molecular structure and stereochemistry of (+)-isobutacclamol hydrobromide was determined by X-ray crystallography (24) and is shown in Figure 1. It confirms all aspects of the constitution and assigned relative and absolute configurations at

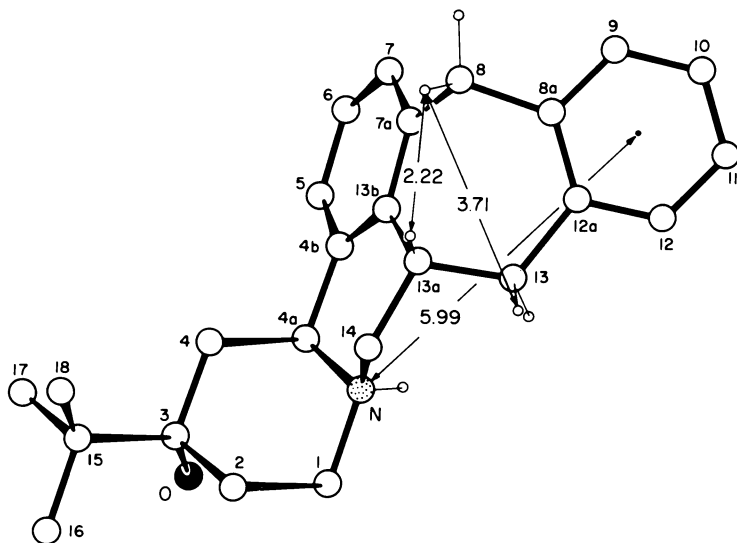


Figure 1. Crystal structure of (+)-isobutacclamol hydrobromide (the distances indicated in A are calculated from the crystallographic data)

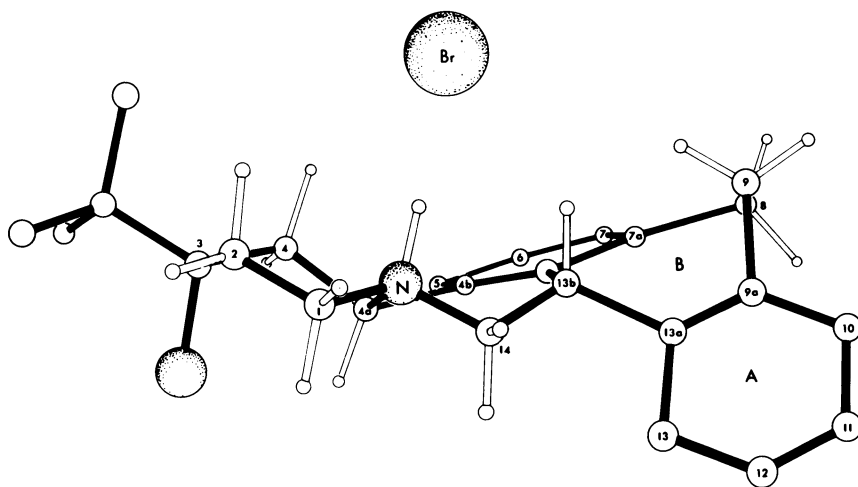


Figure 2. Crystal structure of (±)-butacclamol hydrobromide

positions 3, 4a and 13a (20).

The conformation observed in the crystal structure, however, need not bear any relationship to that adopted by the molecule on interaction with the DAR; it is necessary therefore to consider the nature of that conformation.

In the crystal, (+)-isobutacclamol hydrobromide has a *cisoid* fused rings DE quinolizine system, i.e., it has a 15S absolute configuration. In contrast, in the crystal structure of (\pm)-butacclamol hydrobromide (25), Figure 2, the quinolizine system has a *transoid* fusion. We have suggested previously (20) that at the DAR, (+)-butacclamol and (+)-isobutacclamol exist predominately in the deprotonated forms. In these forms the nitrogen atom is achiral and the *cisoid* and *transoid* quinolizine species are readily interconvertible through nitrogen inversion. We suggest that on interaction with the DAR, (+)-isobutacclamol adopts a conformation with a *transoid* fusion of the quinolizine system. As will become evident in the sequel, only in that conformation can (+)-isobutacclamol present a pharmacophore which is capable of binding to the DAR.

In the crystal structure of butacclamol hydrobromide (25), Figure 2, the cycloheptane ring exists in the conformation with eclipsed hydrogens at positions 9 and 13b (conformer A, Figure 3). In order that (+)-butacclamol would be capable of presenting the same pharmacophore as does (-)-apomorphine, it was postulated that (+)-butacclamol, on interaction with the DAR, adopted the conformation with eclipsed hydrogens at positions 8 and 13b (conformer B, Figure 3), rather than that observed in the crystal structure. In contrast, the conformation of the cycloheptane ring of (+)-isobutacclamol in the crystal structure has eclipsed hydrogens at positions 8 and 13a (conformer B, Figure 4) rather than at positions 8 and 13 (conformer A, Figure 4). We suggest that the cycloheptane ring of (+)-isobutacclamol retains its conformation B on interaction with the DAR.

When models of these deduced receptor-site conformations of (+)-isobutacclamol and (+)-butacclamol, in their conformations B are superimposed, Figure 5, it is seen that the nitrogen atoms of the two ligands are coincident, but the phenyl rings A are not.

Thus, these agents, although closely related structurally, and interacting with a common receptor with affinities in the nanomolar range, do not possess a common pharmacophore.

Closer inspection of Figure 5 reveals however that the rings A of the two ligands lie in the same plane and are immediately adjacent to each other, and suggests that there is a planar binding site on the DAR with the dimensions of at least two adjoining benzene rings, i.e., about 4.8 X 2.4 Å. This binding site can accommodate the phenyl ring A of (+)-butacclamol and the catechol ring of (-)-apomorphine in the α -region, and can accommodate (+)-isobutacclamol's phenyl ring A in the β -region, while the nitrogen atoms of all three ligands are bound to the same site.

The precise relationship of the *aromatic binding site* and

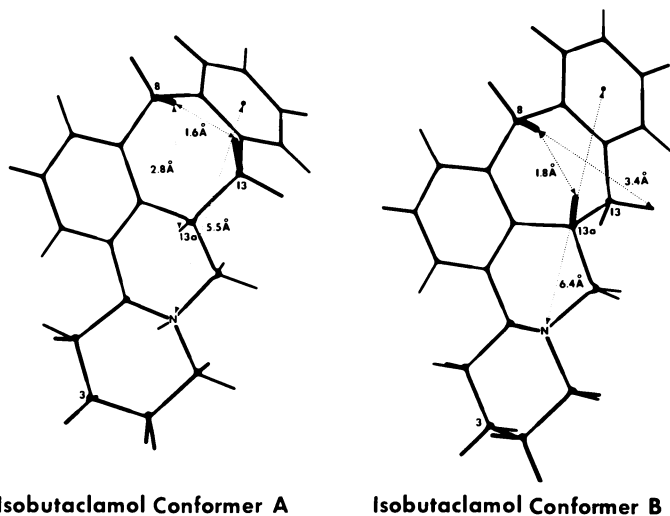


Figure 3. Shadowgraphs of Dreiding models of the nuclei of (+)-butaclamol conformers (the distances shown are from measurements on Dreiding models)

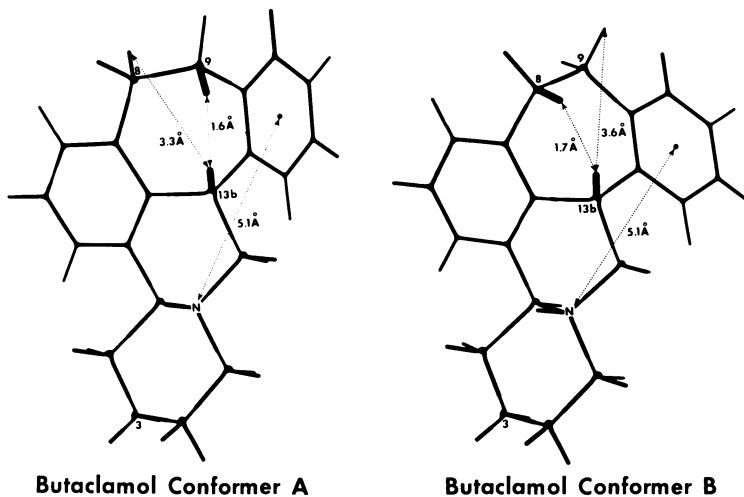


Figure 4. Shadowgraphs of Dreiding models of the nuclei of (+)-isobutacclamol conformers (the distances shown are from measurements on Dreiding models)

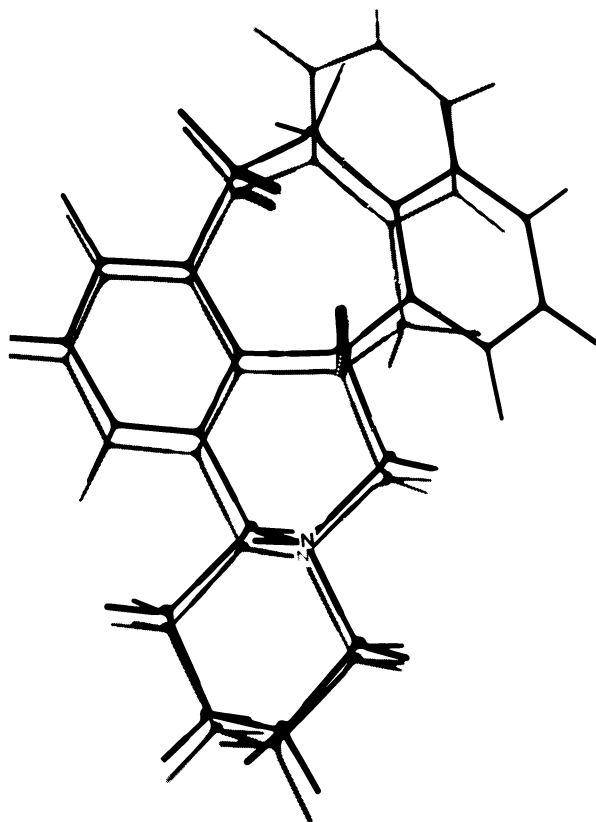


Figure 5. Superimposed shadowgraphs of Dreiding models of the nuclei of B conformers of (+)-butaclamol and (+)-isobutaclamol

the *nitrogen location site* is illustrated, in Figure 6, in terms of their XYZ coordinates. Measurements were obtained from models of conformers B of (+)-butaclamol and (+)-isobutaclamol, oriented in a cartesian coordinate system calibrated in Angstroms, such that the centers of rings A were located at points C and C', respectively, at a distance above the X-axis proportional to 4.7 A, and with the nitrogen atom located at point D in the -XZ plane.

Other significant aspects also emerge from a consideration of this model of the DAR. We have suggested previously (20) that (+)-butaclamol and (+)-isobutaclamol interact with the DAR in their deprotonated forms and that the nitrogen lone pair electrons participate in hydrogen bond formation with another electronegative atom.

Hydrogen bond formation is effective when the distance between the electronegative atoms involved is between 2.5 and 2.7 A, and when the three atoms involved are linearly aligned (26). Consequently, we suggest that an electronegative atom is located at point G, 2.6 A from the nitrogen atom. Because of the rigidity of the butaclamol nucleus, the nitrogen lone pair electrons have a unique directional vector, which is fully defined by the value of 45° for the angle G D D' in Figure 6. The point G has the coordinates -2.0 X and -1.8 Y and is located at the intersection of the -X-Y plane, and a perpendicular -X+Z plane which is 1.8 A (the distance O-F) below the plane in which the nitrogen atom is located. Thus the previously designated (14) nitrogen atom *primary binding site*, must now be viewed as being composed of a *nitrogen location site* D and a complementary *hydrogen bond donor site* G.

This receptor representation also accounts for the chirality of the receptor. Thus, if the pharmacologically inactive (-)-enantiomer of butaclamol is oriented to the receptor with its ring A bound to the β -region of the *phenyl ring binding site*, it requires the existence of a *nitrogen atom location site* at position D' (+3.2 X, +0.9 Z). It would appear that either no such site exists at D', or, that the molecule would have to occupy "disallowed" space in the +Z octants of the coordinate system in which the receptor is oriented.

We have chosen, in Figure 6, to orient the phenyl rings A of (+)-butaclamol and of (+)-isobutaclamol, both in conformation B, on the XY plane such that the nitrogen atoms are situated on the +Z side of the XY plane. Inspection of models of these ligands reveals the remarkable circumstance that with the exception of five hydrogen atoms of the former ligand (those attached to carbon atoms 1, 8, 9, 13b and 14), and the similar hydrogen atoms of the latter (attached to carbon atoms 1, 8, 13, 13a and 14), all the other fifty three atoms of the molecules lie either in the XY plane, or, on the +Z side of that plane. This observation suggests that the XY plane of the receptor represents an essentially flat membrane surface of which the phenyl ring *primary*

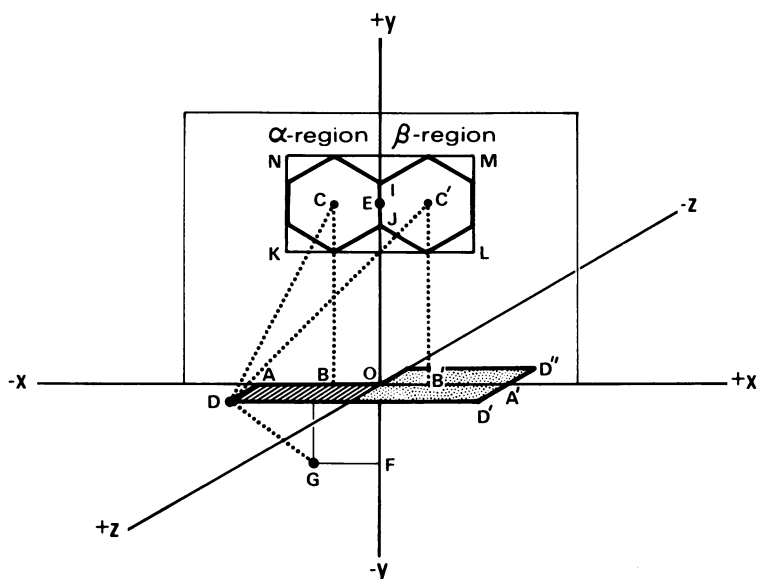


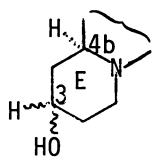
Figure 6. Representation of primary binding sites on the dopamine receptor derived from measurements on Dreiding models of the B conformers of (+)-butaclamol and (+)-isobutclamol (see text). The figure is drawn to a scale of 1 cm = 1 Å. Key distances are: O-E = B-C = B'-C' = 4.7; A-D = A'D' = A''-D'' = 0.9; A-B = A'-B' = 2.0; A-B' = 4.4; D-C = 5.1; D-C' = 6.4; D-G = 2.6; A-O = A'-O' = 3.2; K-L = M-N = 4.8; N-K = M-L = 2.4; O-F = 1.8; F-G = 1.7.

binding site is a part.

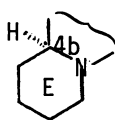
The ligands discussed above could, alternatively, have been oriented with their phenyl rings A bound to a complementary site on the XY plane, but with the nitrogen atoms located at position D' in the -Z side of that plane. This alternative is identical to the first as long as no information is available concerning the nature of the proposed membrane that is represented by the XY plane. However, we consider it improbable that a receptor membrane would be so undifferentiated that specific ligand binding could occur, indiscriminately, on either face.

It was the presence of an extended phenethylamine moiety in (+)-butaclamol that prompted a comparison of its topography with that of the same grouping in (-)-apomorphine, and which led to the detection of common topographical features in the two ligands (14). (+)-Isobutaclamol does not contain such an extended phenethylamine moiety, but rather, a phenylpropylamine group constrained within a semi-rigid ring system. *A priori*, one might not have expected (+)-isobutaclamol to be capable of interacting with the receptor genetically designed for the phenethylamine neurotransmitter dopamine. However, because of the dimensions of the planar *phenyl ring primary binding site*, and because (+)-isobutaclamol and (+)-butaclamol have the same A-D distances they possess the essential features required for recognition by the DAR. This constitutes a unique instance where a phenylpropylamine grouping functions topographically as a phenethylamine moiety.

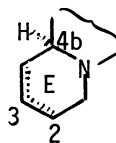
Amongst the numerous butaclamol analogs studied, a set of seven (V-XI), having modifications in ring E were prepared and evaluated as DAR antagonists (27). Their structures are indicated below.



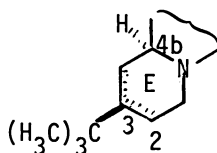
V (3H,4b-H-trans)
VI (3H,4b-H-cis)



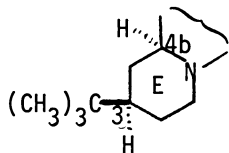
VII



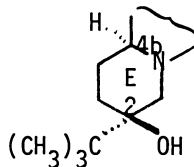
IX



anhidrobutaclamol
X



desoxybutaclamol
XI



VIII

Analog V which retains a 3-hydroxyl group but lacks the *tert*-butyl group, and its epimer VI, are inactive as DAR antagonists when evaluated *in vivo* (27). Similarly, analog VII, with both the hydroxyl and *tert*-butyl substituents absent, is inactive. Analog VIII has both the hydroxyl and *tert*-butyl substituents transposed to position 2 and is inactive, while analog IX, a mixture of unsubstituted Δ^2 and Δ^3 olefins, is also devoid of activity.

Of the series only two compounds, anhydrobutaclamol, X, a mixture of Δ^2 and Δ^3 olefins, and desoxybutaclamol, XI, retain potent DAR antagonist properties. These compounds were evaluated both *in vitro* (28) and *in vivo* as described above, and the results are shown in Table 1. It is apparent that they are qualitatively and quantitatively very similar to (\pm)-butaclamol and (\pm)-isobutaclamol.

Inspection of models of butaclamol and anhydrobutaclamol reveals that the rings E of the olefinic isomers can adopt only a single half-chair conformation. In both of these, the locus in space of the *tert*-butyl groups is very similar to the same group of butaclamol and the groups overlap to a considerable degree. Similarly, the locus of the volume occupied by the *tert*-butyl group of desoxybutaclamol, which was assigned (27) an equatorial orientation on a ring E chair, is identical to that of the same moiety in butaclamol. In contrast, the centers of the volumes occupied by the *tert*-butyl groups of butaclamol and analog VIII, are separated by almost 3 Å.

These findings suggest that DAR antagonist activity in compounds of the butaclamol series is critically dependent on the presence of a *tert*-butyl group attached equatorially at position-3 of the nucleus, and furthermore that a lipophilic *accessory binding site* must exist on the DAR macromolecule, which accommodates the *tert*-butyl groups of butaclamol, isobutaclamol, anhydrobutaclamol, and, desoxybutaclamol.

The dimensions of the lipophilic *accessory binding site* cannot, at this stage, be precisely defined. From our previous studies it is known that an equatorially oriented group as small as ethyl (12), or as large as a substituted phenyl ring (29), will confer neuroleptic activity on the resultant analog. The *tert*-butyl group has dimensions close to those of a benzene ring, however the high potency of the Δ^2 and Δ^3 olefin mixture X, obtained by dehydration of butaclamol, suggests that the locus of the volume in space occupied by the *tert*-butyl group can be displaced, within limits, either above, or below, the plane in which the lipophilic *accessory binding site* is located (Figure 7). Taking these possible displacements into consideration, it would appear that this *accessory binding site* is considerably larger than, simply, the volume occupied by a *tert*-butyl group.

The lipophilic *accessory binding site*, therefore, is viewed as a uniquely shaped cavity on a membrane surface having a minimum diameter of 2.5 Å (Figure 7). Measurements on a model oriented in a cartesian coordinate system calibrated in Angstrom units shows

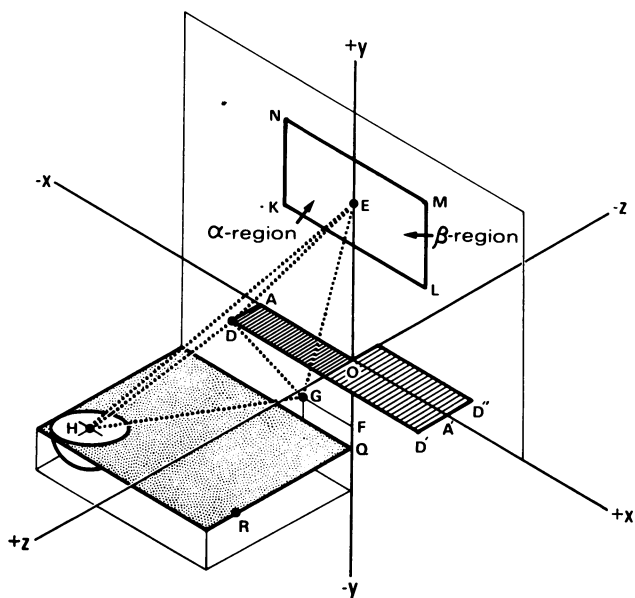
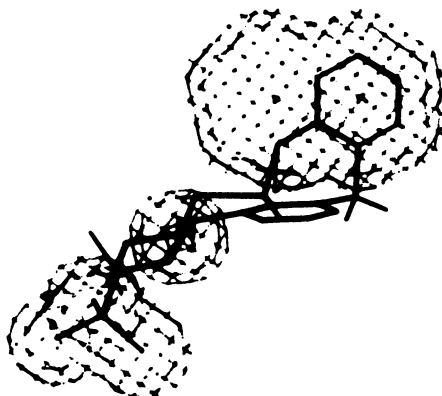


Figure 7. Representation of the lipophilic accessory binding site and the primary binding sites on the dopamine receptor (see text). The figure is drawn to a scale of 1 cm = 1 Å. Key distances are: O-E = 4.7; A-D = 0.9; D-E = 5.7; A-O = 3.2; D-G = 2.6; H-E = 9.6; D-H = 4.5; O-F = 1.8; O-Q = 2.5; H-G = 6.5; Q-R = 5.0; H-R = 4.0; F-Q = 0.7; N-K = M-L = 2.4; K-L = M-N = 4.8.

Figure 8. Model of the dopamine receptor, with the (+)-isobutacramol molecule superimposed, generated by the MMS-X computer-based molecular modeling and graphic display system. (Courtesy of Dr. G. Marshall, Washington University, St. Louis).



that it is centered at the point H (+4.0 Z, -5.0 X, -2.5 Y) in a -X+Z plane parallel to that in which the nitrogen atom is located. The receptor representation in Figure 7 shows that it is 4.5 Å (D-H) from the *nitrogen location site* D, 6.5 Å (H-G) from the *hydrogen bond donor site* G, and 9.6 Å (H-E) from the center of the *phenyl ring binding site* E.

While active compounds of the butaclamol series must utilize this *accessory binding site*, it appears that it is not utilized by any other known DAR antagonists or agonists. It is probable that other *accessory binding sites* exist on the DAR macromolecule. Their identification would require a detailed study of the essential structural and stereochemical requirements for activity in other series of DAR agonists and antagonists.

These key distances and coordinates involving the lipophilic *accessory binding site* H, along with the distances and coordinates related to the *nitrogen location site* D, the *hydrogen bond donor site* G, and the α and β regions of the *phenyl ring binding site* represent a comprehensive preliminary description of the topography of the DAR.

Figure 8 is another representation of the receptor model, generated with the MMS-X computer-based molecular modeling, and graphic display system (30). Input data comprised the XYZ coordinates of those atoms of (+)-butaclamol and (+)-isobutaclamol, in their receptor-site conformations, which served to define the α - and β -regions of the *aromatic binding site*, the *nitrogen atom location site*, and, the *lipophilic accessory binding site*. The coordinates were obtained either from crystallographic results, or, from a program which generates coordinate data from bond lengths and angles (31). The display, Figure 8, is in the form of an electron density map of the atoms referred to above, contoured to their van der Waals radii (light lines). The display also shows the molecule of (+)-isobutaclamol (heavy lines) superimposed on the receptor model.

In conclusion, from the study on the relationship between DAR antagonist activity and detailed molecular structure in certain analogs of butaclamol, it has been possible to confirm the existence of a lipophilic *accessory binding site* on the DAR, to define its probable dimensions, and to specify its locus with respect to the phenyl ring and the nitrogen atom *primary binding sites*.

Acknowledgments. We wish to thank Dr. P. Bird, Concordia University, Montreal for providing Figure 2, and Dr. G. Marshall, Washington University, St. Louis, for providing Figure 8.

Literature Cited

1. Langley, J.N., J. Physiol. (London), (1905), 33, 374.
2. For a review see "Neurotransmitter Receptor Binding", Yamamura, H.I., Enna, S.J., and Kuhar, M.J., eds., Raven Press, N.Y., 1978.
3. Quiocho, F.A., and Lipscombe, W.N., Adv. Protein Chem., (1971), 25, 1.
4. Heidmann, T., and Changeux, J.-P., Ann. Rev. Biochem., (1978), 47, 317.
5. Korolkovas, A, "Essentials of Molecular Pharmacology", p. 187, Wiley-Interscience, N.Y., 1970.
6. Mielke, D.H., Gallant, D.M., Oelsner, T., Kessler, C.M., Tomlinson, W.K., and Cohen, G.H., Dis. Nerv. Syst., (1975), 36, 7.
7. Hollister, L.E., Davis, K.L., and Berger, P.A., Psychopharmacol. Commun., (1975), 1, 493.
8. Gallart-Capdevila, J.M., Arch. Neurobiol., (1975), 38, 545.
9. Imaz, F., Ban, T.A., and Lehmann, H.E., Psychopharmacol. Bull., (1976), 12, 31.
10. Clark, M.L., Paredes, A., Costiloe, J.P., and Wood, F., J. Clin. Pharmacol., (1977), 17, 529.
11. Clark, M.L., Costiloe, J.P., Wood, F., Paredes, A., and Fulkerson, G., Dis. Nerv. Syst., (1977), 38, 943.
12. Bruderlein, F.T., Humber, L.G., and Voith, K., J. Med. Chem., (1975), 18, 185.
13. Voith, K., and Cummings, J.R., Can. J. Physiol. Pharmacol., (1976), 54, 551.
14. Humber, L.G., Bruderlein, F.T., and Voith, K., Mol. Pharmacol., (1975), 11, 833.
15. Seeman, P., Chau-Wong, M., Tedesco, J., and Wong, K., Proc. Nat. Acad. Sci. (U.S.A.), (1975), 72, 4376.
16. Burt, D.R., Enna, S., Creese, I., and Snyder, S.H., Proc. Nat. Acad. Sci. (U.S.A.), (1975), 72, 4655.
17. Colpaert, F.C., Van Bever, W.F.M., and Leyson, J.E.M.F., Int. Rev. Neurobiol., Vol. 19, C.C. Pfeiffer, ed., p. 225, Academic Press, N.Y., 1976.
18. Saari, W.S., King, S.W., and Lotti, V., J. Med. Chem., (1973), 16, 171.
19. Marshall, G.R., Dunn, D.A., Barry, C.D., and Humber, L.G., unpublished.
20. Philipp, A.H., Humber, L.G., and Voith, K., J. Med. Chem., (1979), 22, XXX.
21. Humber, L.G., Philipp, A.H., Voith, K., Pugsley, T., Lippmann, W., Ahmed, F.R., Przybylska, M., J. Med. Chem., submitted for publication.
22. Voith, K., and Herr, F., Psychopharmacologia (Berlin), (1975), 42, 11.
23. Pugsley, T.A., and Lippmann, W., J. Pharm. Pharmacol., (1979), 31, 47.

24. Ahmed, F.R., and Przybylska, M., Acta Cryst., in press.
25. Bird, P.H., Bruderlein, F.T., and Humber, L.G., Can. J. Chem., (1976), 54, 2715.
26. For a review on the characteristics of hydrogen bonds see E.W. Gill in "Progress in Medicinal Chemistry", Vol. 4, G.D. Ellis and G.B. West, Eds., p. 51, Butterworths, London, 1965.
27. Humber, L.G., Bruderlein, F.T., Philipp, A.H., Gotz, M., and Voith, K., J. Med. Chem., (1979), 22, XXX.
28. Lippmann, W., and Pugsley, T.A., personal communication.
29. Voith, K., Bruderlein, F.T., and Humber, L.G., J. Med. Chem., (1978), 21, 694.
30. Barry, C.D., Bosshard, H.E., Ellis, R.A., and Marshall, G.R., Fed. Proc., (1974), 33, 2368.
31. Marshall, G.R., personal communication.

RECEIVED June 8, 1979.

Interaction of Model Opiate Anionic Receptor Sites with Characteristic *N*-Substituents of Rigid Opiates: PCILO and Empirical Potential Energy Calculations

GILDA LOEW, STANLEY BURT, and PAMELA NOMURA

Department of Genetics, Stanford University Medical Center, Stanford, CA 94305

ROBERT MACELROY

NASA Ames Research Center, Moffett Field, CA 94035

Opiate narcotics are thought to act at specific receptors in the brain since they exhibit stereospecific binding (1) and have been shown by fluorescence techniques to be localized at discrete regions in the central nervous system (2). While much effort has been made to isolate and characterize the opiate receptor (3), relatively little detailed information exists about the nature of the opiate binding site. Some investigators describe the receptor as a membrane bound protein or proteo-lipid (4) while others have used nerve cell components such as cerebroside sulfate or phosphatidyl inositol as models for the opiate receptor (5).

In addition to the inherent difficulties in characterizing the opiate receptor, the problem is compounded by the diversity of chemical structures which are active narcotic analgesics. The prototypes of rigid opiates all have fused ring structures which may contain three (benzomorphans), four (morphinans), five (morphine) or six (oripavines/thebaines) fused rings. All active opiates in these classes, and in the so-called flexible opiates, exhibit cross tolerance and are reversibly blocked by the opiate antagonist naloxone.

In the search for a non-addictive analgesic, thousands of analogues in each of these subgroups have been synthesized and tested. The search has extended to flexible classes of opiates among which are 4 ϕ -piperidines, 3 ϕ -piperidines, acyclic compounds such as methadone, and most recently the endogenous peptide opiates, enkephalins and endorphines. As the number of degrees of freedom in these classes increases, the resemblance to rigid opiates decreases, with the endogenous opiates having only a phenethylamine moiety in common with rigid opiates. In previous work we have considered how such diverse classes of opiates can be accommodated at a single, expandable opiate receptor and how this accommodation can account for observed structure-activity profiles (6, 7, 8).

Within each exogenous opiate family there are analogues which are pure agonists, pure antagonists and those which are mixed agonist-antagonists. One clinically promising

0-8412-0521-3/79/47-112-243\$05.00/0

© 1979 American Chemical Society

observation is that in this latter group are compounds with low addiction liability, although many of these have undesirable psychotomimetic side effects (9).

The ratio of agonist/antagonist potency within a given family of opiates appears to be modulated by very small changes in chemical structure. In rigid opiates, the nitrogen is part of a piperidine ring and is a tertiary amine which is thought to act at the receptor site in its protonated form. In fused ring opiates the nature of the substituent on the amine nitrogen plays a key role in such modulation; although other groups such as the C₇ substituent in oripavines (10) and the C₁₄ substituent in oxymorphones are also important (11).

The extent of agonist/antagonist potency in a given analogue must be a receptor related event. Opiates are thought to bind and act at the receptor by weak, reversible, non-covalent bonding, i.e., to form a molecular complex rather than covalent bonds leading to irreversible chemical transformations. The conformations of the opiates (which determine how well they fit at the receptor site) and their electronic structures (which determine their extent of interaction with the receptor) should then be directly relevant to their relative agonist and antagonist potencies.

In previous work we have proposed that N-substituents of rigid opiates which are mixed agonist-antagonists bind and act at the receptor in two distinct conformations corresponding to two different induced receptor site conformations (12). Using this hypothesis, we suggested the synthesis of a series of new morphine analogues predicted to have a wide range of agonist/antagonist potency ratios, among which could be a potentially useful, non-addicting analgesic. These have been synthesized and show promising results in preliminary preclinical tests (13).

Explicit studies of opiate-receptors interactions should be very useful in continuing to explore the factors that modulate the extent of agonism and antagonism in a given opiate and the variation in this ratio among closely related analogues. Such studies are severely hampered on the experimental side by the lack of a detailed description of the opiate binding site. However, every class of opiates, including the endogeneous peptide opiates, have in common an amine group which is almost completely protonated at physiological pH. The interaction of the quaternized amine group with an anionic receptor site is thought to be central to opiate analgesic activity and antagonism. Model receptor studies implicate a sulfate or phosphate moiety as plausible anionic receptor sites (5).

In this study, as a first step in modeling opiate receptor interactions, we have considered the interaction of an ammonium ion and methyl sulfate or phosphate with the series of compounds shown in Figure 1. These compounds, as N-substituents in rigid opiates such as 5,9 dimethyl, 2'hydroxy, 6,7 benzomorphans, exhibit a broad spectrum of pharmacological behavior from

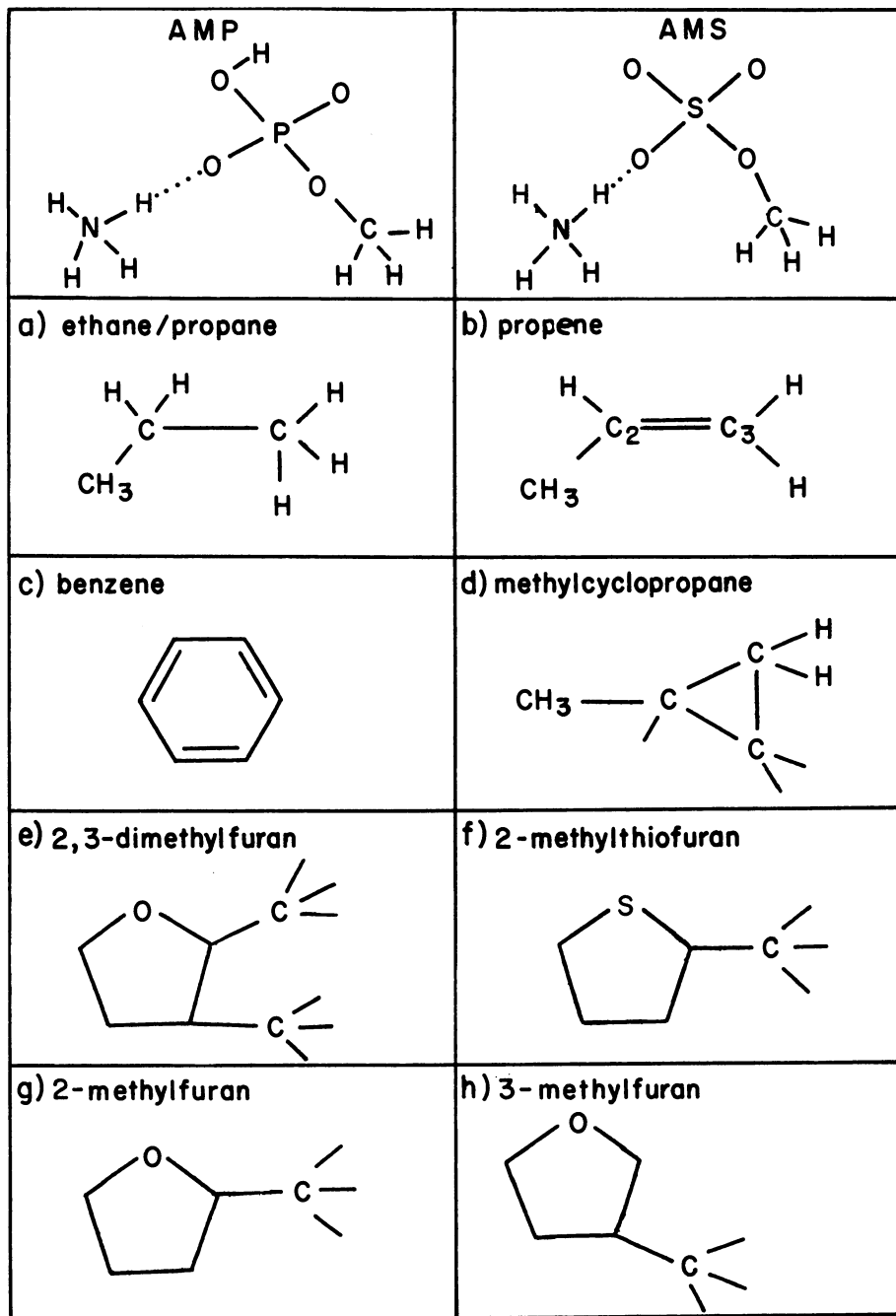


Figure 1. Schematic structures of model anionic receptor sites, ammonium methylphosphate (AMP), ammonium methylsulphate (AMS), and the eight compounds studied

pure potent antagonists to mixed agonist/antagonists, to pure agonists (14,15).

The object of this initial study was to determine whether any net interactions could occur between typical N-substituents and the model anionic receptor site involved in H-bonding with the protonated amine group, without the steric constraint of attachment of the substituent to the nitrogen.

METHOD AND PROCEDURE

The intermolecular energy of complex formation can be characterized in three different ways: Use of 1) a quantum mechanical or 2) an empirical energy method that calculates only the interaction between the molecular systems; or 3) a supermolecule approach in which the molecular system is considered to be a single molecule and the total energy of the molecular system including the intramolecular energy of the separate molecules is calculated. The energy of complex formation can then be obtained by subtracting the energy of each separate molecular system calculated by the same method.

In this study, we have chosen the supermolecule approach and have used the semi-empirical quantum mechanical method called PCILO (Perturbative Configuration Interaction using Localized Orbitals) (16) to calculate intermolecular interactions. This method has recently been used successfully to calculate the intermolecular energies and geometries of hydrogen-bonded dimers of hydrocarbons and water (17,18). H-bonded complexes are particularly well characterized by this method (19).

For other types of complexes, the PCILO method has been shown to give more reasonable energies and geometries than other widely used semi-empirical methods such as CNDO and extended Hückel procedures (19). It should be noted however that in complexes with a dominating dispersion type interaction, the total interaction will be underestimated due to the finiteness of the basis set. Charge transfer interactions, on the other hand, are overestimated due to the basis extension effect and the CNDO approximation for the integrals (20). Equilibrium distance are also underestimated in both types of complexes due to underestimated repulsion energies. To some extent, the underestimated dispersion will be compensated for by the overestimation of the charge-transfer energy and the underestimation of the repulsion energy.

In this study, the NH_4^+ group was used to model the protonated amine in opiates. A hydrogen bond was formed between the ammonium ion and the model anionic sites, methylphosphate anion ($\text{HCH}_2\text{PO}_4^-$) and methyl sulfate anion (CH_3SO_4^-), to mimic the central cation-anion interaction thought to occur at the opiate receptor site. Optimal hydrogen bonding energies for both the ammonium methyl phosphate (AMP) and ammonium methyl sulfate (AMS) were obtained by geometry optimization. These

neutral complexes, were then allowed to interact with each of the eight N-substituents listed in Figure 1. For each of these complexes, both hydrogen bonded and $n \rightarrow \pi$ interactions were investigated and optimized energies and geometries were obtained. In addition, for the $n \rightarrow \pi$ complexes, an empirical energy formulation (21) was used as an alternative method of obtaining interaction energies and complex geometries.

In the empirical energy program used the total interaction energy is written as:

$$E_{\text{tot}} = E_{\text{nb}} + E_{\text{hb}} + E_{\text{es}}$$

where

$$E_{\text{nb}} = \sum_{ij} \left[\frac{A^{ij}}{r_{ij}^{12}} - \frac{B^{ij}}{r_{ij}^6} \right]$$

$$E_{\text{hb}} = \sum_{ij} \left[\frac{A^{ij}}{r_{ij}^{12}} - \frac{B^{ij}}{r_{ij}^{10}} \right]$$

$$E_{\text{el}} = \sum_{ij} \left[\frac{q_i q_j}{D r_{ij}} \right]$$

and A^{ij} and B^{ij} are parameters depending on atom type. A description of the parameters used can be found in the literature (22, 23, 24). This empirical energy program is integrated with a computer graphics system (25) and optimization programs (26) and thus allows a rapid assessment of potential interaction geometries leading to a minimum energy geometry. This method has been tested on nucleic acid base pairing and stacking and interactions of nucleic acid components with amino acids and appears to give reasonable results (27).

For all the calculations reported, the geometry of the methyl phosphate and sulfate was obtained from a composite of X-ray crystal structures of relevant compounds (28).

Geometries for many of the N-substituent compounds were taken from standard bond lengths and bond angles (29). For the furan and methylcyclopropyl compounds, X-ray structures of the isolated compounds and of cyclazocine were used respectively (28, 30).

For hydrogen bonded complexes, one of the phosphate or sulfate oxygens not bonded to either a methyl group or hydrogen atom was considered the electron donor and a hydrogen atom of the substituent compound the electron acceptor, i.e., a linear O-H...R bond was created. The O-H and O-R distances were varied and rotations of the R group about the H-bond axis made in 30° intervals for each O-R distance. At the optimum values of the heavy atom distance and torsion angle,

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

additional local geometry relaxation to a minimum was allowed. The energy of the H-bonded complex, ΔE_{HB} , was then obtained as the energy difference:

$$\Delta E_{HB} = [E_{\text{complex}} - E_{AMP(AMS)} - E_{\text{substituent}}]$$

For the $n \rightarrow \pi$ complexes, an intermolecular coordinate (R_o) was defined from the oxygen lone pair to a specified point of the R substituent placed below it. For benzene, methylcyclopropane, and the four furans the coordinate was the distance from the oxygen lone pair to the center of the ring. For propane and propene the coordinate was the distance from oxygen lone pair to the center of the C_2-C_3 bond.

At each value of R_o , rotation of the R-group by 30° intervals about this coordinate axis was made to find the orientation with minimum energy. For the optimum values of R_o and torsion angle, local geometry relaxation (phosphorous and sulfur atom bond angles) was done to obtain final values of optimized complex energies. Again, the complex energy was obtained from the difference:

$$\Delta E_{n \rightarrow \pi} = [E_{\text{complex}} - E_{AMP(AMS)} - E_{\text{substituent}}]$$

Results and Discussion

Table I presents the optimized energies and heavy atom O-H-R distances obtained by PCILO for H-bonded complexes of AMP and AMS with the eight substituents studied. The equilibrium distances obtained agree to within 0.1Å to those seen in crystals of similar H-bonded systems (31). While no direct comparison with experiment is possible, the calculated energies are all in a reasonable range of values for H-bond complexes. The AMP forms uniformly better H-bonds than AMS consistent with the larger heats of hydration observed for phosphate groups.

Table I also gives the charge transfer associated with the hydrogen bonded complexes in millielectrons transferred from the AMP or AMS to the substituent. Both AMP and AMS donate approximately the same fraction of electrons to each compound. There is no obvious correlation between the extent of charge transfer and stability of the complex formed.

Table II gives the optimized energies of the $n \rightarrow \pi$ complexes of AMP and AMS with the eight compounds studied at the minimum energy value of R_o , the distance from the oxygen atom to the center of the ring or bond in question. While there is no direct experimental data with which to compare these results, the approximations inherent in the method indicate the minimum distance could be underestimated. Thus we have also calculated the interaction energy at an R_o distance of 3.2Å which is closer to experimental distances found in some gas-phase

TABLE I: PCILO CALCULATED ENERGIES AND EXTENT OF CHARGE TRANSFER IN H-BONDED COMPLEXES OF OPIATE N-SUBSTITUENTS WITH MODEL ANIONIC RECEPTOR SITES

Substituent	AMP ^a			AMS ^a		
	$-\Delta E^b$	r_{\min}^c	$\Delta\rho^d$	$-\Delta E^b$	r_{\min}^c	$\Delta\rho^d$
propene 1 ^e	5.8	2.84	15.6	2.9	2.91	15.4
propene 2 ^e	3.8	2.85	22.8	3.2	2.92	27.5
3-methylfuran ^f	5.7	2.77	24.9	3.3	2.77	24.7
2-methylthiofuran ^f	5.5	2.77	24.9	3.0	2.77	24.9
ethane	5.1	2.89	8.1	2.1	2.96	12.2
benzene	5.0	2.85	14.6	2.6	2.92	14.8
2-methylfuran ^f	4.9	2.77	26.4	2.6	2.77	24.8
methylcyclopropane ^g	4.5	2.86	10.6	2.4	2.92	13.5

a) AMP/AMS = ammonium methylphosphate/sulfate

b) $-\Delta E$ = stabilization energy of optimized complex in Kcal/mole.

c) r_{\min} = optimized heavy atom distance 0 - H - C.

d) millielectrons transferred from AMP or AMS to N - substituent

e) propene 1 = H-bonding to allylic H on carbon 3, propene 2 = H bonding to allylic H on carbon 2

f) H bonding to ring H on carbon ortho to ring oxygen or sulfur

g) H bonding to H of ring carbon without CH₃ substituent.

TABLE II: PCILO CALCULATED ENERGIES AND EXTENT OF CHARGE TRANSFER IN $\eta \rightarrow \pi$ COMPLEXES OF OPIATE N-SUBSTITUENTS WITH MODEL ANIONIC RECEPTOR SITES

Substituent	AMP			AMS		
	$-\Delta E_{\min}^a$	R_{\min}^b	$\Delta \rho^b$	$-\Delta E_{\min}^a$	R_{\min}^b	$\Delta \rho^c$
2-methylthiofuran	9.1	(2.6)	-42.2	5.9	(2.9)	6.7
methylcyclopropane	8.2	(2.0)	-89.6	3.4	(2.0)	14.2
benzene	7.3	(2.6)	-14.6	5.3	(2.5)	9.8
3-methylfuran	7.2	(2.2)	-24.9	5.2	(2.2)	12.3
2-methylfuran	6.8	(2.2)	-26.9	4.0	(2.4)	8.6
propene	6.2	(2.2)	-15.6	1.6	(2.2)	17.0
2,3-dimethylfuran	4.9	(2.1)	-	2.0	(2.4)	8.5
propane	3.1	(2.1)	-	1.4	-	-

a) $-\Delta E$ = Stabilization energy of complex expressed in Kcal/mol. Energy at R_{\min} and $R = 3.2\text{\AA}$ obtained by extensive geometry relaxation and torsion angle variation of the model anionic receptor site. R , expressed in angstroms, is the distance from the oxygen atom to the center of the ring plane or bond in question.

b) millielectrons transferred to AMP from the N-substituents

c) millielectrons transferred from AMS to the N-substituents

$\sigma \rightarrow \pi$ complexes such as that between ethylene and Cl_2 (20). All the substituents investigated form stable complexes with both model AMP and AMS anionic sites but with differing stabilities.

The extent of charge transfer involved in these $n \rightarrow \pi$ complexes is also given in Table II. Interestingly, AMS function as an electron donor to the substituent ($n \rightarrow \pi$) while AMP functions as an electron acceptor ($n \leftarrow \pi$). While the extent of charge transfer tends to be greater for the more stable AMP complexes, there is no definitive correlation between extent of charge transfer and stability. This result, similar to the one obtained for the H-bonded complexes, implies that other components such as dispersion and electrostatic interactions are significantly involved in the stabilization of these complexes.

Table III gives the energy of the optimized $n \rightarrow \pi$ complexes obtained from the empirical energy program at optimum values of $R \approx 3.2\text{\AA}$. We see from this table that interaction energies for AMP are comparable to those obtained by the PCILO method at 3.2\AA . For AMS complexes, the empirical method yields somewhat larger energies than those obtained at 3.2\AA with the PCILO method.

For AMP complexes, the major differences in the PCILO and empirical energy results are: 1) the 2-methylthiofuran forms the most stable complex with PCILO and one of the least stable complexes with the empirical energy method, though the absolute difference in energy is only 1.8 Kcal/mole: 2) the PCILO method differentiates the various furan compounds more distinctly than the empirical energy method in which they all form complexes of the same stability.

For the AMS complexes, the major difference in results from the two methods is that methylcyclopropane forms the most stable complex with PCILO and one of the least stable with the empirical energy method. Both methods given relatively small energy variations among the compounds at interaction distances of 3.2\AA .

The major differences in the two methods used are that the empirical method does not contain terms with explicit electron overlap dependence such as a charge transfer term but does include total intermolecular geometry optimization.

Figures 2 and 3 show the optimized geometries obtained from both the PCILO and empirical energy methods for complexes of AMP and AMS with five of the eight compounds studied. The PCILO results give distinct $n-\pi$ complexes involving interaction of one oxygen lone pair of electrons with the π -electron system of the substituent. In all cases, the empirical energy method gives a totally optimized complex which involves mainly electrostatic and dispersion terms. As can be seen from Figures 2 and 3, the interaction of methylcyclopropane with the model anionic receptor site is the one with the greatest difference

TABLE III. ANTAGONIST POTENCY AND BINDING AFFINITY OF BENZOMORPHANS WITH VARYING N-SUBSTITUENTS AND THEIR ENERGIES OF INTERACTION WITH MODEL ANIONIC RECEPTOR SITES.

N-R	ANT ^a	K _e ^c × 10 ⁹	-ΔE _{PCILO} ^d	-ΔE _{EMP} ^d	-ΔE _{PCILO} ^d	-ΔE _{EMP} ^d
methylcyclopropane	3.0	0.8	8.2	5.2	8.0	3.2
2-methylfuran	0.66	2.5	6.8	5.5	2.1	4.2
3-methylfuran	0.66 ^b	3.0	7.2	5.2	2.8	4.0
propene	0.22	-	6.2	3.8	1.0	2.9
ethylbenzene	0.03	10.0	7.3 ^e	4.9 ^e	1.8 ^e	3.8 ^e
2-methylthiofuran	0.024	-	9.1	4.1	1.4	3.8
2,3-dimethylfuran	0.0	-	4.9	5.2	2.1	3.6
propane	-	-	3.1	- ^f	- ^f	- ^f

a) Antagonism relative to nalorphine = 1 in Guinea Pig Ileum (15).

b) The 3-methylfurfuryl and 2-methylfurfuryl benzomorphans are equally potent antagonists in mice and monkeys. No Guinea Pig Ileum Data exists for the 3-methyl analogue

c) Stereospecific binding constants K_e from reference (32).

d) -ΔE = min energy in Kcal/mole

e) Calculated for benzene

f) not calculated

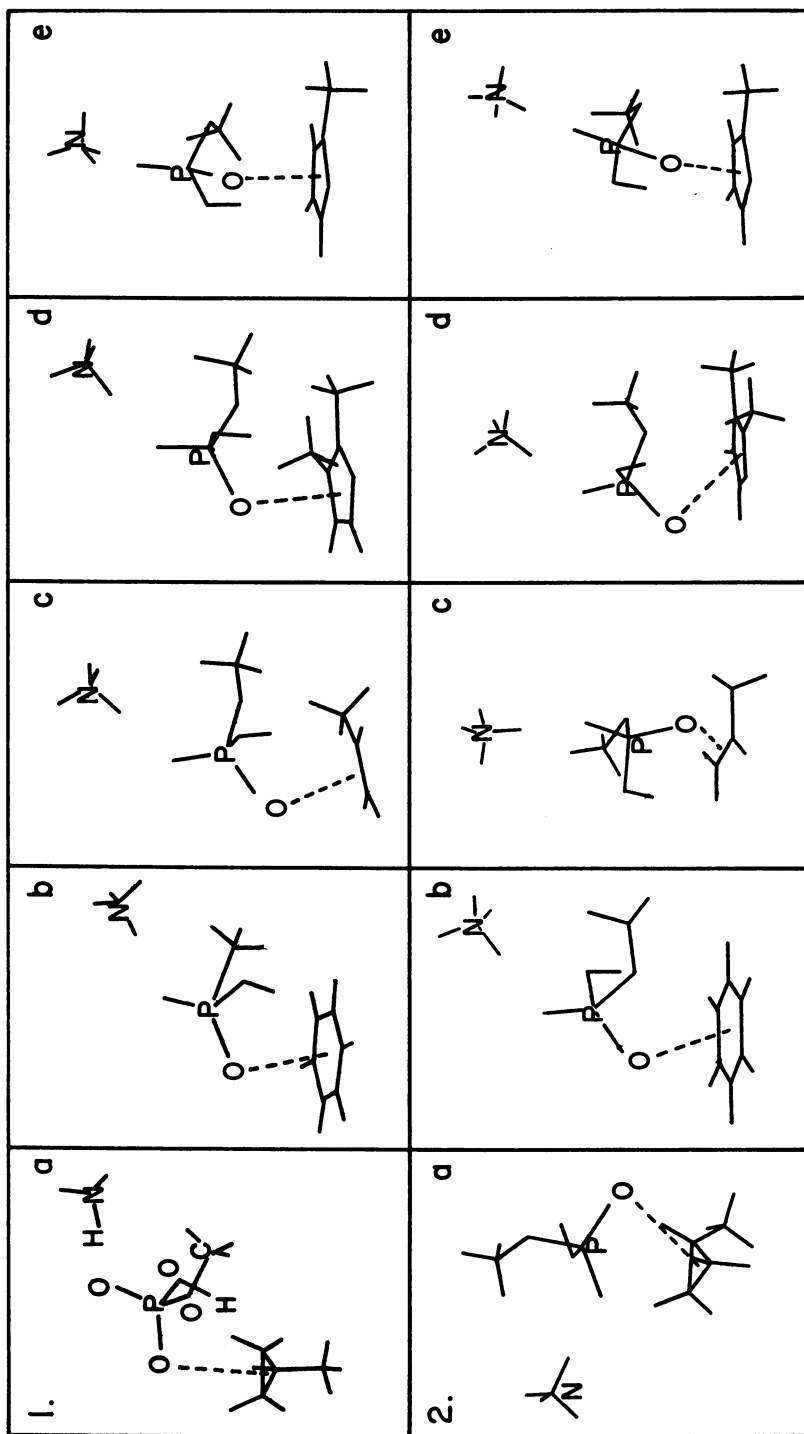


Figure 2. PCIL0 (Row 1) and empirical (Row 2) optimized geometries of complexes of AMP with (a) methylcyclopropane; (b) benzene; (c) propene; (d) 2,3-dimethylfuran; and (e) 2-methylthiofuran

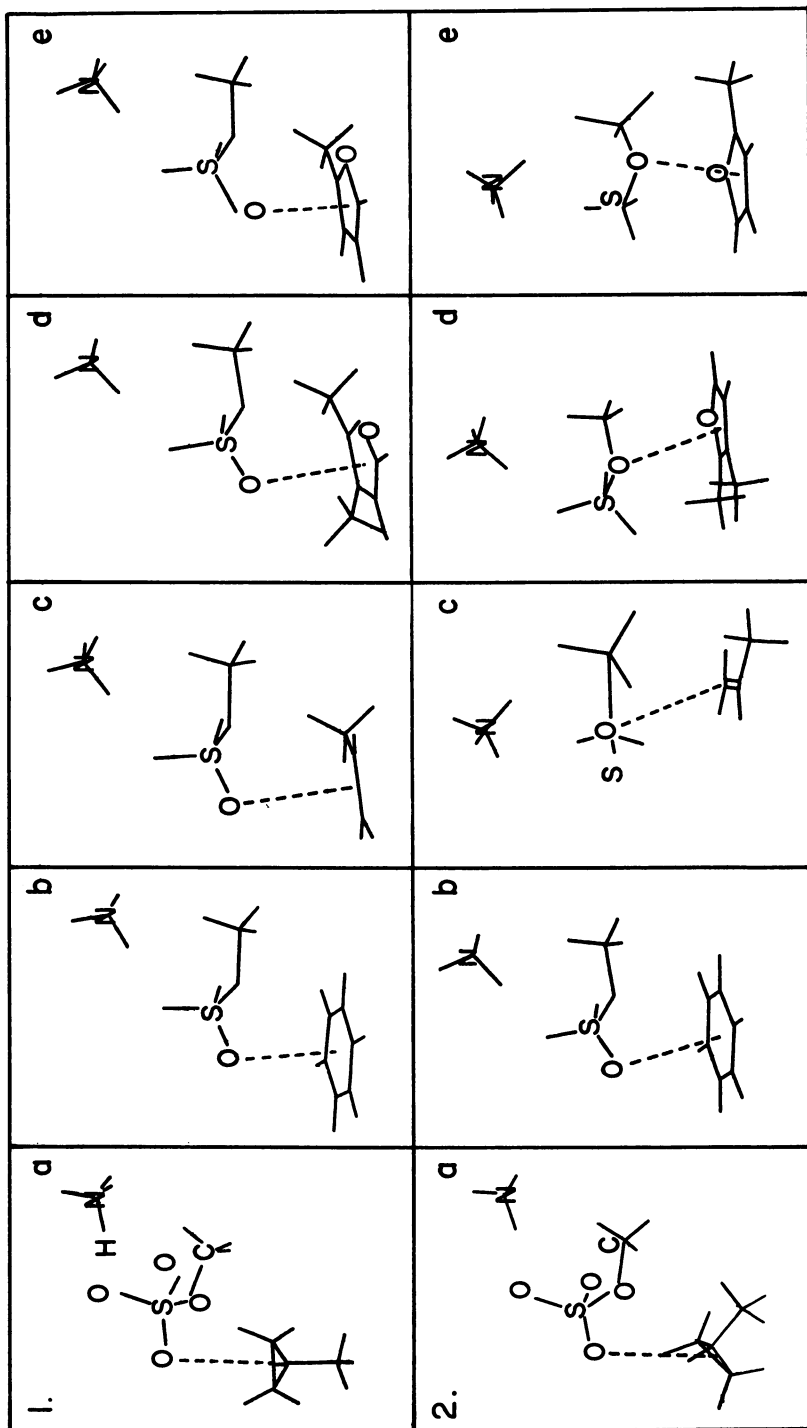


Figure 3. PCILO (Row 1) and empirical (Row 2) optimized geometries of complexes of AMS with (a) methylcyclopropane; (b) benzene; (c) propene; (d) 2,3-dimethylfuran; and (e) 2-methylthiofuran

between the two methods. This geometry difference is reflected in the disparity in the energy of this complex obtained by the two methods.

Relevance to Opiate Activity and Antagonism

Antagonist potencies measured *in vitro* in the guinea pig ileum system or *in vivo* by extent of withdrawal in addicted animals should closely parallel the affinity of opiates for the receptor site. Both antagonist potencies and stereo-specific binding constants have been measured for a number of N-substituted 5,9 dimethyl 2' hydroxy 6,7 benzomorphan analogues (14, 15, 32). Table III gives these values for the compounds studied and shows the extent of correlation between the known affinities and potencies for four analogues. Expanding this list to other N-substituents and other rigid opiate series increases the correlation, though a number of significant deviations occur (32).

In the series shown in Table III only the N-substituent is varying. Thus the observed differences in binding affinities and antagonist potencies could be due directly to differences in the N-substituent interactions with the receptor site. The calculations reported are an initial attempt to determine the extent of interaction, if any, of N-substituents with model anionic receptor sites. Without the constraint of attaching the substituent to the amine group of the opiate, the electronic factors which control optimum interactions could be more fully explored. As already discussed, stable complexes were found for all the substituents investigated but with differing stabilities. Even though these compounds are unconstrained by the steric requirements imposed on them as N-substituents in rigid opiates, it is nevertheless of some interest to compare the calculated value of interaction energy of each compound with the measured affinity and/or antagonist potency for the most closely related benzomorphan analogue.

In Table III, the compounds studied are listed in order of the decreasing antagonist potency they confer as N-substituents of 5,9 dimethyl 2'hydroxy, 6,7 benzomorphans. Calculated stabilities with both anionic sites tend to follow the observed affinities and potencies. The three most potent antagonists are calculated to be the three most stable complexes with AMS and among the four most stable complexes with AMP. On the other hand, the two nearly pure agonist have the smallest interaction energy. For the remaining compounds, the strong complex found for 2-methylthiofuran does not correlate well with its weak antagonism, although there is no measured binding constant for this compound. The binding constant of pentazocine is less than expected from the calculated stabilities of benzene interacting with an anionic receptor site. In phenazocine, with a phenethyl rather than phenyl substituent, the benzene

ring might be too far away from the anionic site for optimum interaction. Similar trends, but with some differences are seen in the stabilities calculated by the empirical method. With the exception of the cyclopropylmethane, the main discrepancy is the relatively large energy calculated for the 2,3 - dimethylfuran. The stabilization energies of the AMS complexes calculated by the empirical method correlate more with the measured affinities and antagonist potencies than those for the AMP complexes. Additionally, more stable complexes with AMS are found by the empirical method than by the PCILO method. Taken together, the results of both methods indicate that either the phosphate or sulfate anion is a reasonable model for the anionic opiate receptor site.

The results obtained in this study, while not definitive, have been encouraging enough to allow us to continue these studies using the same methodology. The next step will be to recalculate the optimum energies and geometries of the methyl phosphate and sulfate anions interacting with these compounds as N-substituents of the piperidine ring of rigid opiates. If the two conformer hypothesis of mixed agonist/antagonist behavior is correct then: a) N-substituent analogues with little or no agonist potency should have one well defined, rather stable complex b) N-substituent analogues which are nearly pure agonists should also have one optimum complex, distinct from that of a pure antagonist and c) N-substituent analogues with mixed agonist/antagonist behavior should have two distinct stable receptor complexes. The more different the stability of each, the more disparate the agonist/antagonist potency ratio for a given opiate should be. The greater the stability of the best complex, the larger the antagonist potency should be. Work is now in progress to test this hypothesis.

Acknowledgement The support of National Institute of Drug Abuse Grant # DA 00770-03 and NASA Ames Interchange #NCA-OR 745-721 for this work is gratefully acknowledged. The authors also wish to acknowledge the use of the AIMS computer graphics system at NASA AMES. Many thanks also to Donald S. Berkowitz and William Maloney for help in the initial stages of this study.

Abstract

In this study both the PCILO and empirical energy methods were used to characterize intermolecular interactions of typical N-substituents of rigid opiates with model anionic receptor sites. Ammonium methylphosphate (AMP) and ammonium methylsulfate (AMS) were used as model anionic receptor sites. Interaction energies of eight compounds which, as N-substituents, modulate different antagonist/agonist potencies

in 5,9-dimethyl 2' hydroxy 6,7 benzomorphans were calculated. Stable H-bonded and $n-\pi$ complexes were obtained. These results suggest that such complexes could be involved in the mode of interaction of the rigid opiates at the receptor site and in their observed affinities and agonist/antagonist potency ratios.

Literature Cited

1. Pert, C. B.; Snyder, S. H. Science, 1973, 179, 1011.
2. Simantov, R.; Snowman, A. M.; Snyder, S. H. Brain Res., 1976, 107, 650.
3. Simon, E. J.; Hiller, J. M.; Edelman, I. Proc. Nat. Acad. Sci. U. S. A., 1973, 70, 1947.
4. Loewney, L. I.; Schilz, K.; Lowery, P. J.; Goldstein, A. Science, 1974, 183, 749.
5. Loh, H. H.; Cho, T. M.; Wu, Y. C.; Way, E. L. Life Sci., 1974, 14, 2233.
6. Loew, G. H.; Jester, R. J. Med. Chem., 1975, 18, 105.
7. Loew, G. H.; Berkowitz, D. S.; Newth, R. C. J. Med. Chem., 1976, 19, 863.
8. Loew, G. H.; Burt, S. K. Proc. Nat. Acad. Sci. U. S. A., 1978, 75, 10.
9. Jasinski, D. R.; Martin, W. R.; Hueldtke, R. Clin. Pharm. Ther., 1971, 12, 613.
10. Lewis, J. W.; Bentley, K. W.; Cowan, A. Annu. Rev. Pharm., 1971, 11, 241.
11. Kosterlitz, H. W.; Waterfield, A. A. Annu. Rev. Pharm., 1975, 15, 29.
12. Loew, G. H.; Berkowitz, D. S. J. Med. Chem., 1975, 18, 656.
13. DeGraw, J. J.; Lawson, J. A.; Crase, J. L.; Johnson, H. L.; Ellis, M.; Uyeno, E. T.; Loew, G. H.; Berkowitz, D. S. J. Med. Chem., 1978, 21, 415.
14. Merz, H.; Langbein, A.; Stockhaus, K.; Walther, G.; Wick, H. in "Narcotic Antagonists", Vol. 8, M. C. Braude, L. S. Harris, E. L. May, J. P. Smith, J. E. Villarreal, Eds., Raven Press: New York, 1974; p.91
15. Kosterlitz, H. W.; Waterfield, A. A.; Berthoud, V. in "Narcotic Antagonists," Vol. 8, M. C. Braude; L. S. Harris, E. L. May, J. P. Smith, J. E. Villarreal, eds., Raven Press: New York, 1974, p.319.
16. Diner, S.; Malrieu, S. P.; Jordan, F.; Gilliert, M. Theor. Chim Acta., 1969, 15, 100.
17. Lochman, R.; Weller, Th. Int. J. Quantum Chem., 1976, 10, 909.
18. Spurling, T. H.; Snook, I. K. Chem. Phys Letts., 1975, 32, 159.
19. Weller, Th. Int. J. Quantum Chem., 1977, 12, 805.
20. Lochman, R.; Hofman, H.J. Int. J. Quantum Chem., 1977, 11, 427.

21. Burt, S.; Egan, J., McElroy, R. D. Compt. in Chem., 1979, submitted.
22. Momany, F. A.; Carruthers, L. M.; McGuire, R. F.; Scheraga, H. A. J. Phys. Chem., 1974, 78, 1595.
23. Millner, O. E.; Anderson, J. A. Biopolymers, 1975, 14 2159.
24. Jordan, F. A. J. Theo. Biol., 1973, 41, 375.
25. Coeckelenberg, Y.; Hart, J.; McElroy, R. D.; Rein, R. Computers and Graphics, 1978, 3, 9.
26. Gill, P. E., Murray, W.; Pitfield, R. A. National Physics Laboratory Division and Numerical Analysis and Computation Depart. No. 11, 1972.
27. Burt, S. K.; McElroy, R. D.; Egan, J. E. Unpublished results.
28. Sutton, L. E. "Tables of Interatomic Distances in Molecules and Ions," The Chemical Society: London, 1965, 11; 1958, 8.
29. Pople, J.; Beveridge, D. "Approximate Molecular Orbital Theory"; McGraw-Hill: New York, 1970; p 111.
30. Karle, I. L.; Gilardi, R. D.; Fratini, A. V.; Karle, J. Acta Cryst, 1969, 85, 469.
31. Zefirov, Y. V.; Zorkii, P. M. Z. H. Strukit, Khim., 1976, 17, 994.
32. Ionescu, I; Klee, W.; Katz, R. in "The Opiate Narcotics," A. Goldstein, Ed., Pergamon Press: New York, 1975, p. 41.

RECEIVED June 8, 1979.

Functional Receptor Mapping for Modified Cardenolides: Use of the PROPHET System

DOUGLAS C. ROHRER—Medical Foundation of Buffalo, Inc., Buffalo, NY 14203

DWIGHT S. FULLERTON and KOUICHI YOSHIOKA—School of Pharmacy,
Oregon State University, Corvallis, OR 97331

ARTHUR H. L. FROM—Cardiovascular Division, Department of Medicine,
Veterans Administration Medical Center, Minneapolis, MN 55417, and
University of Minnesota, Minneapolis, MN 55455

KHALIL AHMED—Toxicology Research Laboratory, Veterans Administration
Medical Center, Minneapolis, MN 55417, and Department of Laboratory
Medicine and Pathology, University of Minnesota, Minneapolis, MN 55455

Digitalis preparations have been used therapeutically in the management of cardiovascular disease for almost 200 years (1). Their ability to increase the contractility of the heart muscle (the inotropic activity) combined with the slowing of the heart rate (the chronotropic activity) resulting in a generally improved heart efficiency keep these drugs among the top ten prescribed even today (2). Unfortunately, these drugs are also extremely toxic, with cardenolide toxicity accounting for up to half of drug induced in-hospital deaths (3). When the dose level required to attain the desired therapeutic response has been administered, 60% of the toxic dosage has also been given (4). Clearly, the demonstrated importance of these types of drugs combined with the need for improvement in the therapeutic-toxic ration provides a strong incentive for their study and for reaching a better understanding of their mode of action.

The pharmacological effects of the digitalis glycosides, [such as digitoxin (Ia) or digoxin (IIa)] and their genins [Ib or IIb] appear to be the result of the inhibition of the membrane-bound Na^+, K^+ -ATPase (5,6,7). Thus it has been suggested that the Na^+, K^+ -ATPase enzyme is the digitalis receptor or certainly very closely related to it (7). This enzyme system is found in the plasma membranes of nearly all mammalian cells and is involved in the active transport of Na^+ and K^+ across the cell membrane. This transport process is particularly important in cardiac muscle cells where the transport process must occur prior to each heart contraction. A number of often conflicting models (8,9) have been proposed to describe the chemical and structural characteristics of a genin which govern its ability to inhibit the Na^+, K^+ -ATPase. Our earlier studies (10-18) have shown that none of these models has been able to consistently explain the activities of a number of the modified genins. A multi-disciplinary approach, including X-ray crystallography,

0-8412-0521-3/79/47-112-259\$05.25/0

© 1979 American Chemical Society

conformational energy calculations, organic synthesis and Na^+, K^+ -ATPase inhibition studies has been used to find the direct relationship between genin structure and activity reported here.

I. Cardenolide Structural Studies

The structural and electronic characteristics of a molecule determine its ability to bind to and/or cause a physiological response with a given receptor. As described in the previous section, the old models for cardiac glycoside action did not account for the activity of the modified genins. To delineate the true nature of the structure-activity relationship of cardenolide genins, a thorough study of cardenolide structure was begun using X-ray crystallography and the graphic analytical features of the NIH PROPHECT computer system.

Digitoxigenin, Ib, is generally considered to be the cardenolide prototype. Its detailed structural features were therefore of particular interest. Figure 1 shows its crystallographically determined structure (19). The curved shape of the steroid backbone caused by *cis* fusions between the A and B rings and the C and D rings and the axial methyl substituents are common features of most genins. Most synthetic modifications of digitoxigenin involve the 17β -substituent. This substituent has generally been recognized to be a major contributor to receptor binding of digitoxigenin analogues. All of the natural genins including digitoxigenin, Ib, have a 17β -lactone substituent (see Table I: I, II, and VIII). The 3β - and 14β -hydroxyl substituents are also common structural features. Until recently, the 14β -hydroxyl substituent has been considered necessary for binding (13). The molecular structures of 11 other cardenolides were determined crystallographically [Table I: IIB (14), III (18), IVa (16), V (18), VI (18), VII (18), VIII (20), IX (17), X (18), XI (18) and XII (21)] and compared with the structure of digitoxigenin, Ib (19), on the NIH PROPHECT computer system. All of the genins in this study were synthesized in our laboratories (10, 11, 13, 16) from digitoxin, Ia, - except for XII, a gift from Dr. Mitsuru Yoshioka, Associate Director of Research, Shionogi Research Laboratories, Osaka, Japan; IVa, a gift from Dr. Romano Deghenghi, Director of Research, Ayerst Laboratories, Montreal, Canada; and VIII and IIB which were obtained commercially. The synthesis of the methyl ester, VI, and the nitrile, VII, follow the procedures reported previously (8). These structures together with the Na^+, K^+ -ATPase inhibition data were used to formulate the structural and activity correlations to be presented.

Major Structural Classification Based on Ring D

All genins examined in this study can be divided into classes with; (1) a fully saturated steroid backbone (I \rightarrow VIII), (2) a 14-ene-steroid backbone (IX \rightarrow XI) or (3) an 8(14)-ene-steroid backbone (XII). Crystallographic analysis of these genins revealed that the A and B rings and most of the C rings are nearly identical. It therefore seems reasonable to assume that this constant portion of the molecule must fit into the

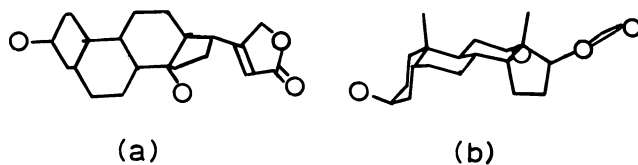


Figure 1. *PROPHET* (a) top and (b) side views of the molecular structure of digitoxigenin Ib from crystallographic coordinates

TABLE I

Ia	Digitoxose ₃	H		
Ib		H		
IIa	Digitoxose ₃	OH		
IIb		H		
III		H		
IVa	β -Glucose	H		
b		H		
V		H		
VI		H		
VII		H		

same location at the receptor site. Any structural (as listed above) or conformational variation in the D-ring portion will have a marked effect on the position of the 17 β -side group relative to the constant portion and also influence the Na⁺,K⁺-ATPase inhibitory potency.

Saturated D-Ring Structures. Normally, saturated cardenolide backbones are thought to be rigid structures with very little conformational variability. Analysis of the 10 crystal structures representing 8 different molecules, however, indicates that there is unexpected conformational freedom in the D rings of these molecules. Furthermore, there are conformational trends which can be linked to the nature of the 17 β -side group. There are three types of 17 β -side groups represented; the normal lactone ring (structures I, II, and VIII), and the modified lactone rings with a substituent on a carbon adjacent to the bond to the backbone (structures III and IV) and the acyclic side chains (structures V, VI and VII).

The major classes of D-ring conformations are a C14 β -envelope, Figure 2a, where C14 is displaced in the β direction from the remaining four coplanar atoms and a C14 β /C15 α -half chair, Figure 2b, where C14 and C15 are displaced in the β and α directions respectively from the three atom plane formed by C13, C16 and C17. Since the symmetry of these conformational forms (a mirror plane in the envelope and a two-fold rotation axis in the half chair) is manifested in the intra ring torsion angles, the torsion angles can be used to evaluate the deviation of a given conformer from the ideal form (22). A plot of the asymmetry parameters representing the deviation from ideal C14 β -envelope symmetry, $\Delta C_s(C14)$, versus the deviation from ideal C14 β /C15 α -half chair symmetry, $\Delta C_2(C17)$, is given in Figure 3. The asymmetry parameter is zero when the conformer has the ideal symmetric conformation. From this plot, it is apparent that the D-ring conformations of the normal lactone structures are clustered in the C14 β -envelope region, while the modified lactone structures have D-ring conformations in the C14 β /C15 α -half chair region. The acyclic structures show a large amount of D-ring conformational freedom which probably results from the decreased strain in the 17 β -side group. The D-ring conformation of V is a C13 α -envelope, VI is a distorted C14 β /C15 α -half chair, and VII is a C13 α /C14 β -half chair. It appears from these results that the D-ring of the modified lactone group has a highly preferred conformation as a result of the location of the substituent on the lactone ring. The normal lactone places some limitations on the D-ring conformation, but not as great as the modified lactones. Finally, the acyclic side groups place few restrictions on the D-ring conformations.

The effect these D-ring conformational differences have on the orientation of the 17 β -side group relative to the constant portion of the molecular are shown in Figure 4. It is apparent

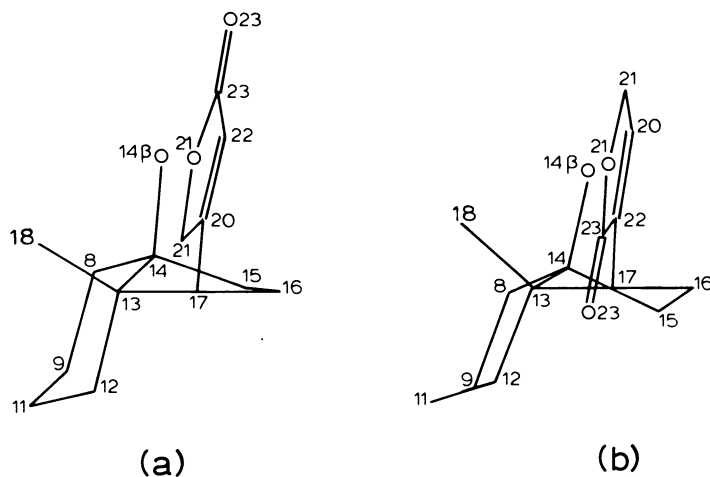


Figure 2. PROPHECT side views of the common saturated D-ring conformations: (a) a 14β -envelope in the structure of Ib; (b) a $14\beta/15\alpha$ -half chair in the structure of IVa

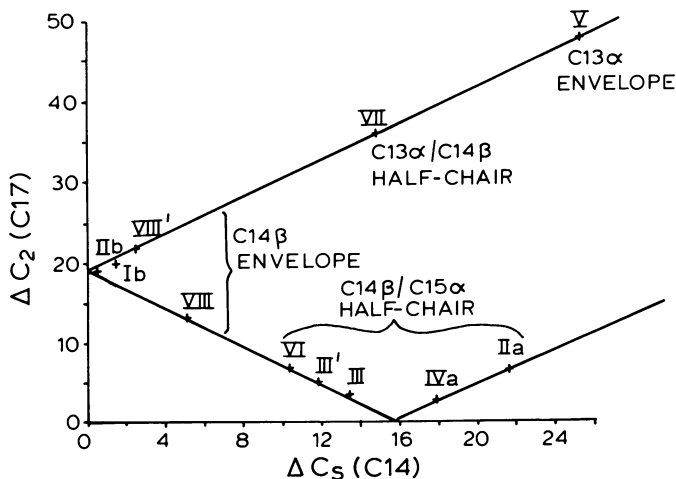


Figure 3. Correlated variation in ring D principle asymmetry parameters. The two-fold rotational asymmetry parameter $\Delta C_2(C_{17})$ is plotted vs. the mirror plane asymmetry parameter $\Delta C_8(C_{13})$.

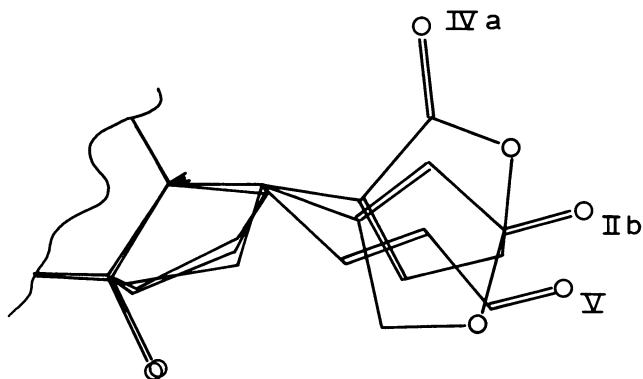


Figure 4. PROPHET/FITMOL overlay of the D-ring regions in structures IIb, IVa, and V showing the shift in location of the 17 β -bond associated with different saturated ring D conformations

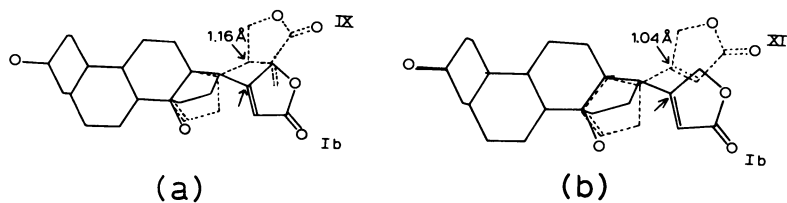


Figure 5. PROPHET/FITMOL overlays of (a) Ib and IX and (b) Ib and XI showing the shift in the location of the 17 β -side group associated with the presence of a C₁₄-C₁₅ double bond in the D ring

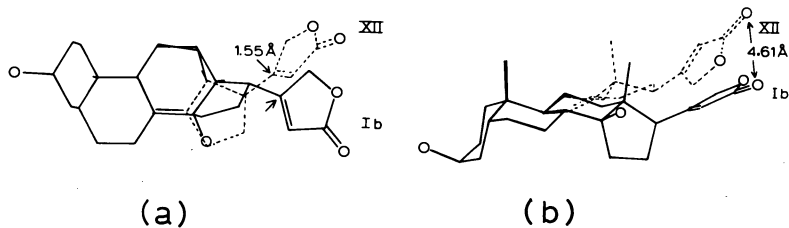


Figure 6. (a) Top and (b) side views of the PROPHET/FITMOL overlays of Ib and XII clearly showing the change in the overall backbone structure in XII associated with the presence of a C₈-C₁₄ double bond

that the exocyclic C17 β bond orientation changes dramatically with change in conformation. This change will be amplified at each successive location out on the chain causing a substantial change in the relative location of the functional oxygen, independent of differences in the nature of the side group.

14-Ene D-Ring Structures. The C14-C15 double bond in these molecules restricts the D-ring conformation to a C17 α -envelope. This observation is consistent with the structures of other non-cardenolide steroid molecules with a C14-C15 double bond (23,24). There are 3 structures with this type of D ring (IX, X, and XI). Figure 5 shows the marked effect of this D-ring conformation on the orientation of the 17 β -side group relative to the constant portion.

8(14)-Ene D-Ring Structure. Here the double bond is in the C ring, but its effect extends into the D ring. There is only one cardenolide structure of this type available for comparison XII. The D-ring conformation is a C17 α -envelope, but the C8-C14 double bond in ring C also affects the directionality of the 17 β -side group relative to the constant portion of the molecule (Figure 6).

17 β -Side Group Orientation

The 17 β -side group also has rotational freedom about the C17-C20 bond. In order to explore this conformational freedom, relative conformational energies were calculated using a version of the CAMSEQ (25) program which was specially modified to be used in conjunction with the NIH PROPHET computer system (26). Starting with the crystallographic coordinates, the energy was evaluated at 10° rotation steps of the 17 β -side group while minimizing the nonbonded interaction with the C13 angular methyl by also allowing it to rotate. No differences were found in the locations of minima or shape of the curve when solvent (i.e. water) interactions were included in these calculations. In every case, the crystallographically observed conformation was at a potential energy minimum.

Saturated D-Ring Structures. These calculations agree with earlier reported observations (17,27) that there are two energy minimum conformations for the normal lactone side group related by a rotation of approximately 180°. Figure 7 shows the resulting energy curves shifted so that the lowest energy is zero kcal/mole. Since the differences in the three energy curves result from relatively small differences in the overall structures of the molecules, a curve representing both structural and conformational flexibility can be generated by using the minimum energy at each side group orientation, Figure 7d. The very large energy barrier at 180° results from the steric interactions be-

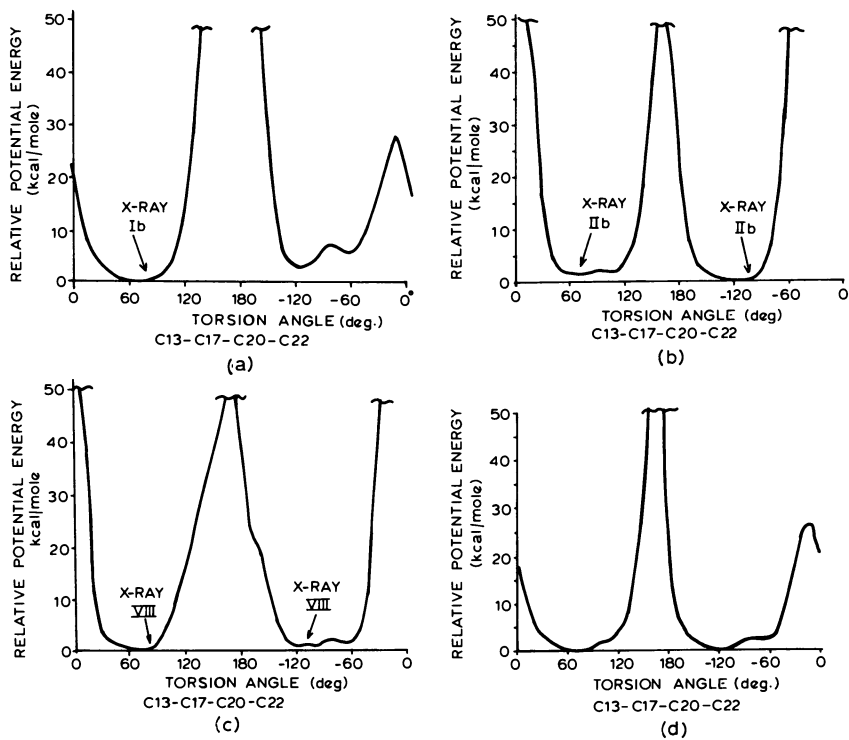


Figure 7. Potential energy diagrams for the rotations of the normal lactone 17 β -side group of (a) Ib, (b) IIb, (c) VIII, and (d) a combined diagram using the minimum energy at each conformational point. The arrows indicate the conformations of the crystal structure.

tween the hydrogens on C18 and C21, while the lower barrier at 0° involves the hydrogens of C22 with those on C18.

The energy diagrams, Figure 8, for the modified lactone ring structures indicate much less rotational freedom than the normal lactone structures. This is a result of the exocyclic substituent adjacent to the rotating bond interacting with the steroid backbone. The tilt of the lactone ring relative to the C17-C20 bond in III caused by C20 being a tetrahedral sp^3 atom rather than planar sp^2 also effects its energy plot. In each case, the conformations in the crystal structure (two independent structures for III and one for IV) are located in the lowest and widest energy minimum region of the plot which should represent the largest conformational populations.

The acyclic structures represent a somewhat more difficult problem since there is an added degree of freedom associated with the C21-C22 bond. However in each of these structures the side group atoms are nearly coplanar. This means that only the cisoid and transoid orientations for the C21-C22 bond need to be considered. The resulting calculations, however, show no effect on the rotation energy of the C17-C20 bond associated with changes in the side group C21-C22 conformation. Figure 9 shows the resulting energy diagrams for V, VI and VII. While the resulting curves have a number of minima, in each case the X-ray conformation represents the preferred orientation - i.e., in the deepest and broadest minimum.

14-Ene D-ring Structures. The energy curves for IX and X are interesting for several reasons. First, C20 is an sp^3 carbon rather than the sp^2 carbon of a normal lactone. This means that the lactone ring is not coplanar with the C17-C20 bond. Second, these C20 epimeric molecules have remarkably different Na^+, K^+ -ATPase inhibition activities which will be discussed later. Figure 10 shows the energy diagram. In these molecules there is only one principal energy minimum conformation, with one or more additional very narrow minima. Again the crystal structure orientations are in the principle energy minima, one structure of IX and two independent structures for X.

The energy plot for XI, Figure 10c, is very similar to the plots obtained for the normal lactone structures (Ib, I Ib, and VIII) discussed above. The C14-C15 double bond in the D ring evidently does not greatly change the rotation of the side groups. This structure, like structures I Ib and VIII, has a disordered 17β -side group. The two side group orientations of the crystal structure represent two separate minima rather than a single wide minimum on the energy plot.

8(14)-Ene D-Ring Structure. The change in the steroid backbone and associated D-ring conformation greatly reduces the steric interactions between the lactone ring hydrogens and the

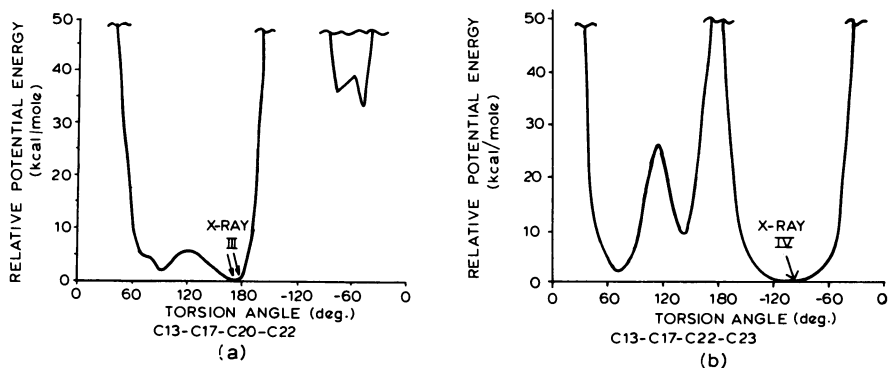


Figure 8. Potential energy diagrams for the rotation of the modified lactone 17β -side group of (a) III and (b) IVa. The arrows indicate the conformations of the crystal structures. Compound III has two independent structural observations.

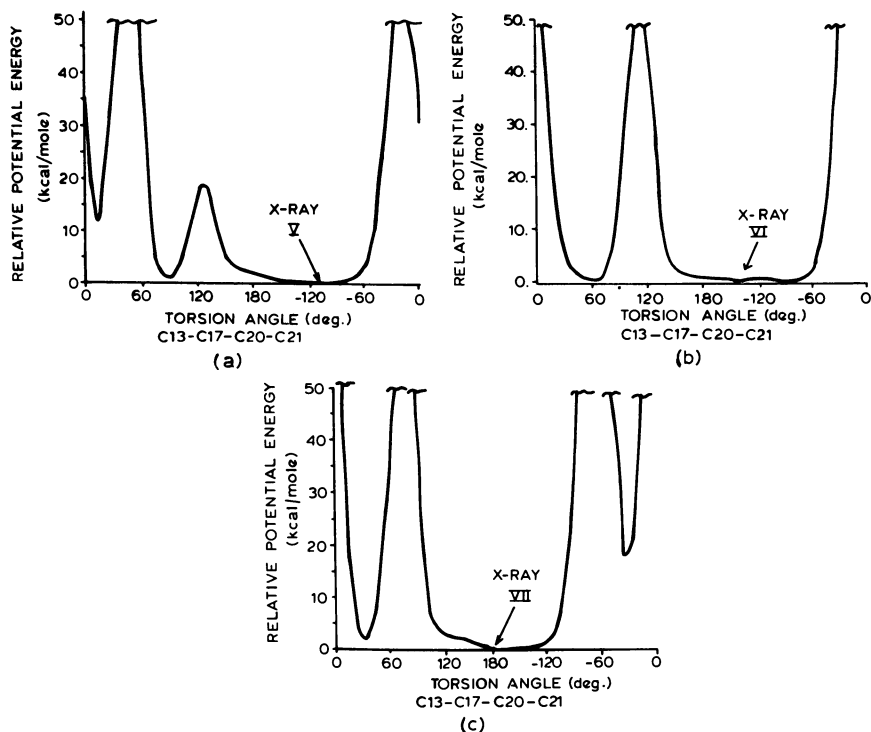


Figure 9. Potential energy diagrams for the rotation of the acyclic 17β -side groups of (a) V, (b) VI, and (c) VII. The arrows indicate the conformations observed in the crystal structure.

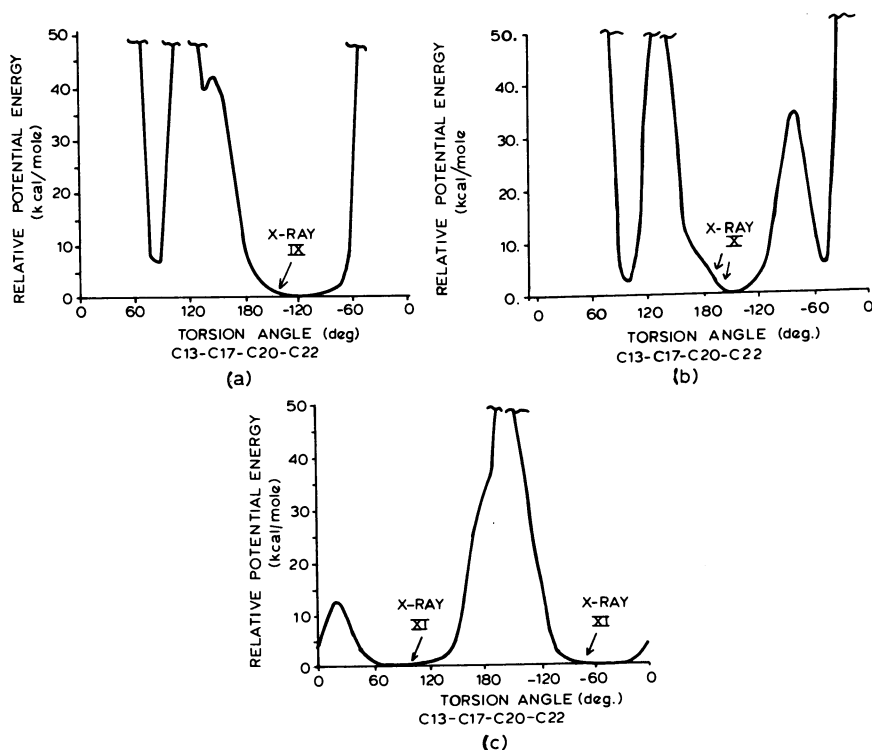


Figure 10. Potential energy diagrams for the rotation of the 17 β -side group of the C₁₄-C₁₅ double bond D-ring structures: (a) IX, (b) X, and (c) XI. The arrows indicate the locations of the crystal structure conformations. Compound X has two independent conformations and XI has two alternate orientations.

angular methyl group, see Figure 11. This means that the lactone ring has much greater orientational freedom than any of the other molecules studied.

II. Na^+, K^+ -ATPase Inhibition Studies

The I_{50} (molar concentration for 50% inhibition) *in vitro* Na^+, K^+ -ATPase inhibitory activities of the genins were determined using rat brain Na^+, K^+ -ATPase (E.C. 3.6.1.3), the preparation and assay of which have been reported previously by our laboratories (28,29). Rat brain enzyme was used in these studies for two reasons: (1) the ease of preparation of high activity enzyme; and (2) the comparable sensitivity of most heart and brain enzyme preparations (29). I_{50} ranges for the less water soluble genins X and XII were extrapolated from their dose-response curves at concentrations where they were completely soluble. Each genin was preincubated with the enzyme ten minutes, i.e. mixing steroid, enzyme, and media lacking K^+ before adding KCl to maximize inhibitory effects. Each genin was also tested without preincubation, however, there was little or no difference in I_{50} with or without preincubation for any of the genins except digoxigenin (I**b**) and strophanthidin (VIII). The added hydroxyl group of these genins decreased lipophilicity with the result that binding to the enzyme is concomitantly slower.

Structure/Activity Correlation

The results of the potential energy calculations show clearly that in every case the crystallographically observed orientation of the 17β -side group is in the energetically preferred conformation or conformations represented by the widest and deepest energy minima when calculated both with and without including a contribution for solvent water. Therefore, the crystallographically determined structure is a suitable model for use in the biological activity and structural comparisons to follow.

The PROPHET procedure FITMOL (30) was used to analytically superimpose the constant portion of structures I**b** through XII upon the corresponding portion of the most active genin in the series digitoxigenin, I**b**, using a least squares method. This permits a quantitative measure of the effects the modification of the 17β -side group structure and variation in the D-ring conformation have upon the functional groups, in particular the position of the carbonyl oxygen.

Comparison of the relative carbonyl oxygen separations obtained from FITMOL with the Na^+, K^+ -ATPase inhibition activities given in Table II revealed a striking correspondence. A simple linear regressor model was used to test the relationship between oxygen separation and the I_{50} data for Na^+, K^+ -ATPase

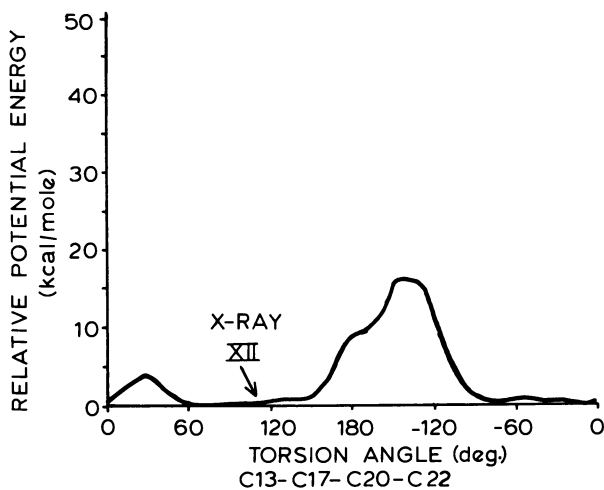


Figure 11. The potential energy diagram for compound XII. The arrow indicates the x-ray crystal structure conformation.

Table II. Na^+ , K^+ -ATPase Inhibition Data (I_{50} , M) and Side Group Structural Data for Compounds I to XII.

<u>Compound</u>	<u>I_{50} (M) with 10 min. Pretincubation.</u>	<u>X-Ray Structure Side Group Torsion Angle ($^\circ$) C13-C17-C20-C22.</u>	<u>Carbonyl Oxygen Separations (\AA) obtained from FITMOL.</u>
<u>Ia</u>	4.0×10^{-8}	-	-
<u>Ib</u>	3.5×10^{-7}	76.3	0.00
<u>IIa</u>	7.0×10^{-8}	-	-
<u>IIb</u>	7.0×10^{-7}	70.4 (alternate: -99.6)	0.42 (2.31)
<u>III</u>	7.0×10^{-5}	180.0	5.06
<u>III'</u>		175.9	4.90
<u>IVa</u>	1.0×10^{-6}	(C13-C17-C22-C23) -106.9	5.22
<u>IVb</u>	7.0×10^{-5}	-	-
<u>V</u>	1.2×10^{-6}	(C13-C17-C20-C21) -107.6	1.30
<u>VI</u>	2.2×10^{-7}	(C13-C17-C20-C21) -129.0	1.43

Table II. (Continued)

<u>Compound</u>	<u>I₅₀(M)</u> with 10 min. <u>Preincubation.</u>	<u>X-Ray Structure</u> <u>Side Group</u> <u>Torsion Angle (°)</u> <u>C13-C17-C20-C22.</u>	<u>Carbonyl Oxygen</u> <u>Separations (Å)</u> obtained from <u>FITMOL.</u>
<u>VII</u>	2.2×10^{-7}	(C13-C17-C20-C21) -129.9	-
<u>VIII</u>	7.5×10^{-7}	85.0 (alternate: -111.0)	0.88 (2.60)
<u>IX</u>	2.0×10^{-5}	-144.2	4.08
<u>X</u>	$(1+3) \times 10^{-4}$	173.5	5.75
<u>X'</u>	[extrapolated from I ₁₈ data]	-178.8	5.67
<u>XI</u>	1.6×10^{-5}	-102.4 (alternate: -76.9)	3.76 (5.51)
<u>XII</u>	$(3+6) \times 10^{-5}$ [extrapolated from I ₂₀ data]	100.3	4.61

with preincubation. Eleven points were found to fit the line shown in Figure 12:

$$\log I_{50} = 0.459 D - 6.471$$

where D is the relative carbonyl separations in Å. The resulting r^2 is 0.993 and for test of the regression relationship, $p = 0.0001$. This indicated that for each 2.2Å the oxygen was displaced from that of digitoxigenin, activity drops by one order of magnitude [An earlier report (15) with nine analogues showed nearly the same relationship, $r^2 = 0.994$.]

In genins IIb, VIII and XI where two 17β -side group orientations were observed in the crystal structure, it is clear that (as one would predict) only one is preferred by the receptor. As shown in Figure 13, the alternate conformations of these genins are more active than can be explained by their carbonyl oxygen separations. However, a nearly perfect correlation Figure 12, exists for their other 17β -side group conformations. The carbonyl oxygen separations for both structures of III and X are nearly equal, Figure 12, so either conformation would be expected to have equivalent ability to inhibit Na^+, K^+ -ATPase.

One structure, ester VI, is more active than its carbonyl oxygen separation would suggest. However, even if this point is included in the linear regression model, the r^2 value is still excellent (0.946). Several possibilities may explain this apparent anomaly and further studies on VI are in progress.

This simple linear relationship provides a quantitative explanation for a number of genins where other models have failed. In particular:

- (A) The predicted activity of aldehyde, V, was up to 123 times higher than digitoxigenin based on previous models, but it was found to be slightly less active (10). Figure 12 clearly shows that V's activity is a direct result of its carbonyl oxygen position. The ring D conformational difference and the structural differences of the acyclic 17β -side group result in this displacement.
- (B) It has been generally believed that the C20-C22 double bond in the lactone ring has an active role in inhibiting the Na^+, K^+ -ATPase (31,32). A lactone ring with an exocyclic double bond would be predicted to be even more active than one with an endocyclic double bond (32) and certainly more active than one with no double bond. However, IX and X, which should therefore be more active than digitoxigenin, are much less active (11). Furthermore, the 20,22-

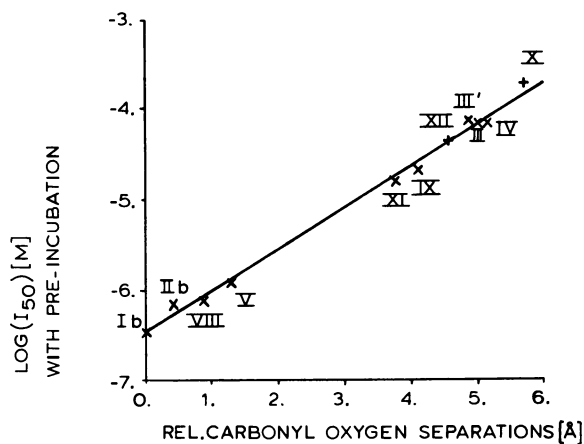


Figure 12. Correlation between the carbonyl oxygen separations relative to digitoxigenin, Ib, and the log of the Na^+, K^+ -ATPase inhibition activity: (x) measured I_{50} data; (+) I_{50} data extrapolated from lower concentrations in which the analog was completely soluble

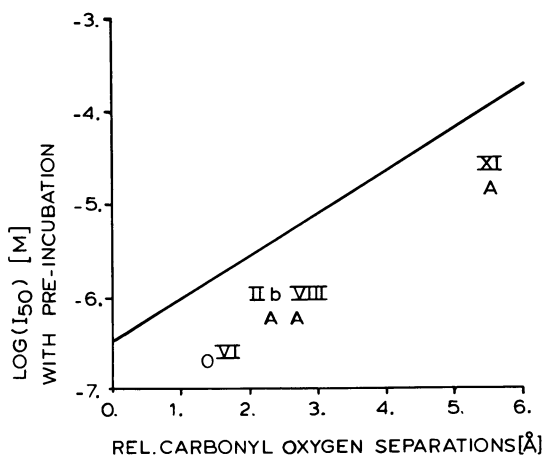


Figure 13. Correlation between relative carbonyl oxygen separations and the log of I_{50} showing the oxygen separations associated with the alternate 17β -lactone orientations (A); carbonyl separation for VI (o)

dihydro analogues with no double bond though 100 times less active than digitoxigenin are more active than IX and X (13). These activity differences are readily explained by the conformational and structural effects on the position of the carbonyl oxygen; that is, the C20-C22 double bond is primarily serving a passive role in Na^+, K^+ -ATPase inhibition - keeping the carbonyl oxygen in the "right" position for maximal inhibitory effect (see Figure 12).

- (C) It has also been generally believed that the 14β -hydroxyl greatly enhances the inhibition of Na^+, K^+ -ATPase (5,6,7,8,9,33). However, IX without the 14β -hydroxy is more active than the corresponding 14β -hydroxy analogues III and its 20R stereoisomer (11). Figure 12 suggests that changes in carbonyl oxygen position may account for these differences in activity.
- (D) The activity of genin IVb is remarkably low (70 times lower than its β -D-glucose analogue (IVa)). The effect of the glucose on the activity was not predicted by a recently proposed binding model (34). The structural correlation shown in Figure 12, however, accurately fits the activity of genin IVb based on its modified lactone structure.

Conclusions

The data presented here are consistent with the following structural correlations of digitalis genin- Na^+, K^+ -ATPase inhibition:

- (A) Rings A and B with a *cis* fusion, most of the C ring, plus the axial methyl group form a structurally rigid CONSTANT PORTION of the molecule.
- (B) The saturated D rings have a great deal of conformational flexibility with preferences influenced by the nature of the 17β -side group.
- (C) There are substantial barriers to rotation of the 17β -side group which provide orientational preferences modeled by the crystal

structures.

- (D) The acyclic α,β -unsaturated 17β -side groups form a coplanar unit and are not as flexible as may have been thought.

The relationship of these structural features to Na^+, K^+ -ATPase inhibition and receptor modeling are:

- (A) The CONSTANT PORTION of the genin molecules must all fit onto a specific position on the receptor surface.
- (B) The 14β -hydroxyl and C20-C22 double bonds do not seem to play as direct a role in the inhibition as was previously thought.
- (C) The minimum energy conformation of the molecule is preferred at the receptor site.
- (D) There is a single carbonyl oxygen position relative to the CONSTANT PORTION which directly controls the activity (a 2.2\AA shift decreases the activity ten fold).

Acknowledgement

The authors are grateful to Miss Melda Tugac, Miss Gloria Del Bel and Mrs. Q. E. Bright for preparation of illustrations and to Miss Kathleen Castiglione, Mrs. Brenda Giacchi and Miss Deanna Hefner for manuscript preparation. This work was supported in part by Grant Number HL-21457 from the National Heart, Lung and Blood Institute and LM-02353 from the National Library of Medicine, the Oregon and Minnesota Heart Associations, and the Veterans Administration Medical Research Fund. The organization and analysis of the data base associated with this investigation were carried out in part using the PROPHET system, a unique national resource sponsored by the NIH. Information about PROPHET including how to apply for access can be obtained from the Director, Chemical/Biological Information-Handling Program, Division of Research Resources, NIH, Bethesda, Maryland 20205.

Literature Cited

1. Withering, W., "An Account for the Fox Glove and Its Medical Uses"; (1785), in "Readings in Pharmacology"; Shuster, L. Ed.; Little Brown: Boston, 1962; p. 109.
2. Pharmacy Times, "1977: Top 200 Drugs", 1978, April, p. 41.

3. Oglivie, R. I.; Ready, J. Can Med. Assoc. J., 1967, 97, 1450.
4. Dreyfus, L. S.; Watanabe, Y. Seminars in Drug Treatment, 1972, 2, 179.
5. Schwartz, A.; Lindenmayer, G. E.; Allen, J. C. Pharmacol. Rev., 1975, 27, 3.
6. Flasch, H.; Heinz, N. Naunyn-Schmiedeberg's Arch. Pharmacol., 1978, 304, 37.
7. Akera, T. Science, 1977, 198, 569.
8. Thomas, R.; Boutagy, J.; Gelbart, J. J. Pharm. Sci., 1974, 63, 1649.
9. Guntert, T. W.; Linde, H. H. Experientia, 1977, 33, 697.
10. Fullerton, D. S.; Pankaskie, M. C.; Ahmed, K.; From, A. H. L. J. Med. Chem., 1976, 19, 1330.
11. Fullerton, D. S.; Gilman, D. M.; Pankaskie, M. C.; Ahmed, K.; From, A. H. L.; Duax, W. L.; Rohrer, D. C. J. Med. Chem., 1977, 20, 841.
12. Yoshioka, K.; Fullerton, D. S.; Rohrer, D. C. Steroids, 1978, 32, 511.
13. Fullerton, D. S.; Yoshioka, K.; Rohrer, D. C.; From, A. H. L.; Ahmed, K. J. Med. Chem., 1979, 22, in press.
14. Rohrer, D. C.; Fullerton, D. S. Acta Crystallogr., Sect. B, submitted.
15. Fullerton, D. S.; Yoshioka, K.; Rohrer, D. C.; From, A. H. L.; Ahmed, K., Science, submitted.
16. Fullerton, D. S.; Yoshioka, K.; Rohrer, D. C.; From, A. H. L.; Ahmed, K. Mol. Pharmacol., submitted.
17. Rohrer, D. C.; Duax, W. L.; Fullerton, D. S. Acta Crystallogr., Sect. B, 1976, 32, 2893.
18. Rohrer, D. C.; Fullerton, D. S. Unpublished results.
19. Karle, I. L.; Karle, J. Acta Crystallogr. Sect. B 1969, 25, 434.
20. Gilardi, R. D.; Flippen, J. L. Acta Crystallogr., Sect. B, 1973, 29, 1842.
21. Gilardi, R. D.; Karle, I. L. Acta Crystallogr., Sect. B, 1970, 26, 207.
22. Duax, W. L.; Weeks, C. M.; Rohrer, D. C. "Topics in Stereochemistry", Vol. 9, Allinger, N. L. and Eliel, E. L., Eds., Interscience: New York, NY, 1976; p. 271.
23. Rohrer, D. C.; Strong, P. D.; Duax, W. L.; Segaloff, A. Acta Crystallogr., Sect. B, 1978, 34, 2913.
24. Rohrer, D. C.; Duax, W. L.; Segaloff, A. Acta Crystallogr., Sect. B, 1978, 34, 2915.
25. Weintraub, H. J. R.; Hopfinger, A. J. Int. J. Quantum Chem., Quantum Biol. Symp., 1975, 2, 203.
26. Rohrer, D. C.; Lavin, M. "EDITMODEL"; in "Public Procedures: A Program Exchange for PROPHET Users"; Wood, J. J., Ed., Bolt, Beranek and Newman, Inc.: Cambridge, MA, 1978.

27. Rohrer, D. C.; Fullerton, D. S. American Crystallographic Association Meeting Abstracts, 1978, Norman, Oklahoma, Abstract: J1.
28. Quarforth, G.; Ahmed, K.; Foster, D. Biochem. Biophys. Acta, 1978, 526, 580.
29. Ahmed, K.; Thomas, B. S. J. Biol. Chem., 1971, 246, 103.
30. Rohrer, D. C.; Perry, H. "FITMOL"; in "Public Procedures: A Program Exchange for PROPHET Users"; Wood, J. J., Ed.; Bolt, Beranek and Newman, Inc.: Cambridge, MA, 1978.
31. Thomas, R.; Boutagy, J.; Gelbart, A. J. Pharmacol. Exptl. Ther., 1974, 191, 219.
32. Kupchan, S. M.; Ognyanov, I.; Moniot, J. L. Bioorg. Chem., 1971, 1, 24.
33. Witty, T. R.; Remers, W. A.; Besch, Jr., H. R. J. Pharm. Sci., 1975, 64, 1248.
34. Thomas, R.; Allen, J.; Pitts, B. J. R.; Schwartz, A. Eur. J. Pharm., 1979, 53, 227.

RECEIVED June 8, 1979.

Thyroid Hormones-Receptor Interactions: Binding Models from Molecular Conformation and Binding Affinity Data

VIVIAN CODY

Medical Foundation of Buffalo, Inc., Buffalo, NY 14203

Thyroid hormones (Figure 1) are iodinated amino acids which are synthesized in the thyroid gland and travel through the general circulation bound to serum proteins. While they have been shown to elicit a multitude of biological responses, the specific nature of their actions still remains unclear. However, as early as 1959 Jorgensen and his co-workers (1) had proposed that specific structural and stereochemical features of the thyroid hormones were responsible for their hormone action. Because the thyroid hormones are transported through the bloodstream bound to the serum proteins thyroxine-binding globulin (TBG), thyroxine-binding prealbumin (TBPA) and serum albumin (SA), and are bound to nuclear proteins which reportedly initiate hormone action (2), the question of conformational preferences becomes increasingly important.

These special stereochemical properties arise from the steric influence of the diortho iodines upon the diphenyl ether conformation. Minimal steric interaction between the 3,5-iodines and the 2',6'-hydrogens is maintained when one ring is coplanar with, and the other perpendicular to, the plane of the two C-O ether bonds. This gives rise to two skewed conformations (Figure 2) which can be described by the torsion angles ϕ (C5-C4-O41-C1') and ϕ' (C4-O41-C1'-C6') of $0^\circ/90^\circ$ and $90^\circ/0^\circ$ for ϕ/ϕ' , respectively. Only the skewed conformer $\phi/\phi' = 90^\circ/0^\circ$ has been observed structurally, although the other has been implicated as an active conformer (3).

This skewed conformation also imparts further stereospecific characteristics to the hormone T_3 which contains only a single outer ring iodine. Because of the restricted rotation about the two diphenyl ether bonds, the chemically equivalent 3' and 5'-iodines are conformationally distinct, giving rise to a distal (away) or proximal (near) conformation (Figure 3). To verify the importance of this conformational feature, numerous structural analogues of triiodothyronine were synthesized (4) in an effort to determine the biologically active conformer of T_3 .

0-8412-0521-3/79/47-112-281\$05.00/0

© 1979 American Chemical Society

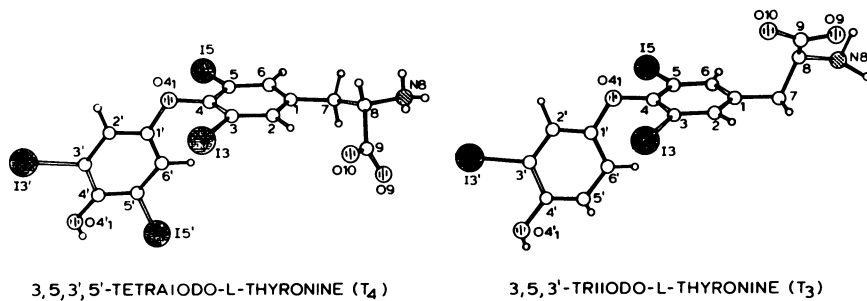


Figure 1. Principal thyroid hormones thyroxine (T₄) and triiodothyronine (T₃) showing their molecular conformation and numbering scheme

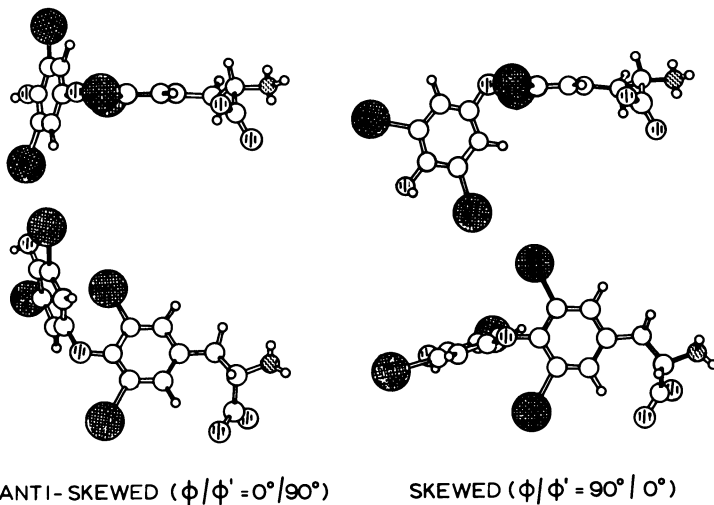


Figure 2. Two examples of the skewed ($\phi/\phi' = 90^\circ/0^\circ$) (right) and antiskewed ($\phi/\phi' = 0^\circ/90^\circ$) (left) diphenyl ether conformation of thyroxine. In each case the molecule is viewed perpendicular and parallel to the inner ring plane.

Distal/Proximal 3'-Substituents

The conformational requirements for activity and protein binding were investigated with 2',3'-dimethyl-3,5-diiodothyronine as the model for a distal conformer of T₃ and 2',5'-dimethyl-3,5-diiodothyronine as the proximal model (1,5,6). These studies showed that the distal analogue is hormonally active, having 50% as much goiter-suppressing activity as T₄ and 13% as much calorigenic effect as T₃ (5). The proximal analogue shows only 1-2% as much activity as the distal analogue in these tests. Similar results were obtained in experiments using pure TBG or whole serum (7), indicating that the requirements for serum protein binding are the same as that required for hormone activity.

The first nuclear magnetic resonance (NMR) studies of T₃ failed to isolate two distinct spectra for the distal and proximal conformers (8). In addition, molecular orbital (MO) calculations predicted a high barrier to internal rotation about the diphenyl ether bonds (9). These data suggested that there was only one conformer present in solution, presumably distal, since this was shown to be the active form of the hormone. However, the first crystal structure of T₃ showed the 3'-I conformation to be proximal (10), inconsistent with available binding and activity data.

This apparent dilemma was resolved when further crystallographic data on T₃ compounds became available (11,12,13,14,15). (See reference 11 for a detailed list of thyroactive crystal structure studies.) These determinations show the 3'-I in the distal conformation and further showed that either the distal or proximal conformer could be isolated by changing the crystallizing media (13,14). These data prompted further NMR and MO studies which showed a nearly equal distal/proximal ratio in solution (16) and a much smaller barrier to internal rotation (17). Thus these data have shown that the distal and proximal conformers are readily accessible in solution and that the energy barrier to free rotation is small enough so that the stereospecific conformer required by a receptor is easily accessible.

Cisoid/Transoid Conformation

In addition to the conformational asymmetry of the outer phenyl ring, the thyronine nucleus possesses other conformational flexibility (Figure 4). The preferred relative orientation (18) of the inner phenyl ring with respect to the amino acid ($\chi^2 \approx 90^\circ$) is such that the alanine side chain and the outer phenyl ring lie either on the same side of the inner phenyl ring plane (cisoid) or on opposite sides of the inner ring plane (transoid) as shown in Figure 5. Although the biological relevance of these conformers is not known, they are observed with equal frequency.

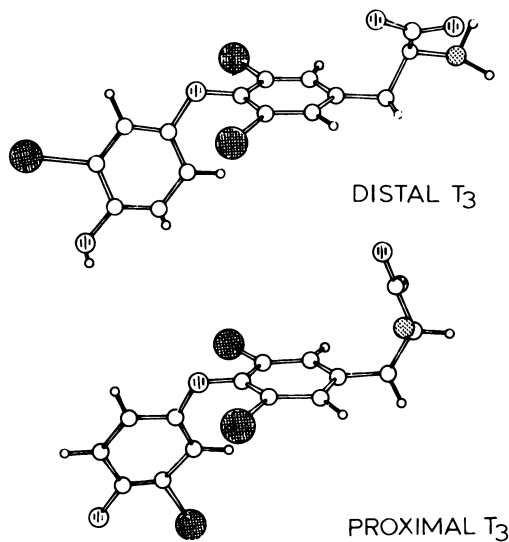


Figure 3. Distal and proximal orientations of the 3'-iodine of T_3 .

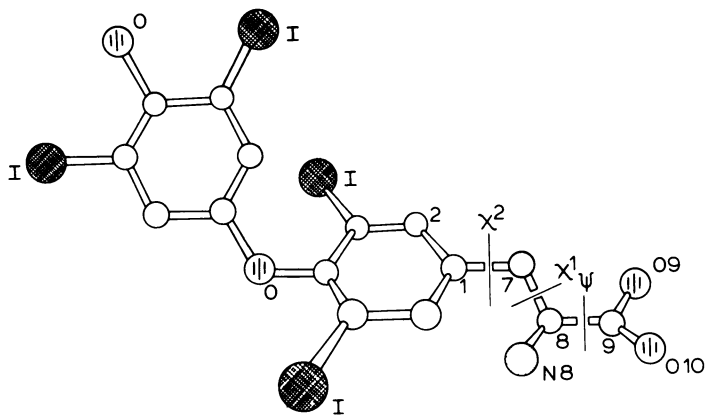


Figure 4. Thyroxine with amino acid conformational parameters defined

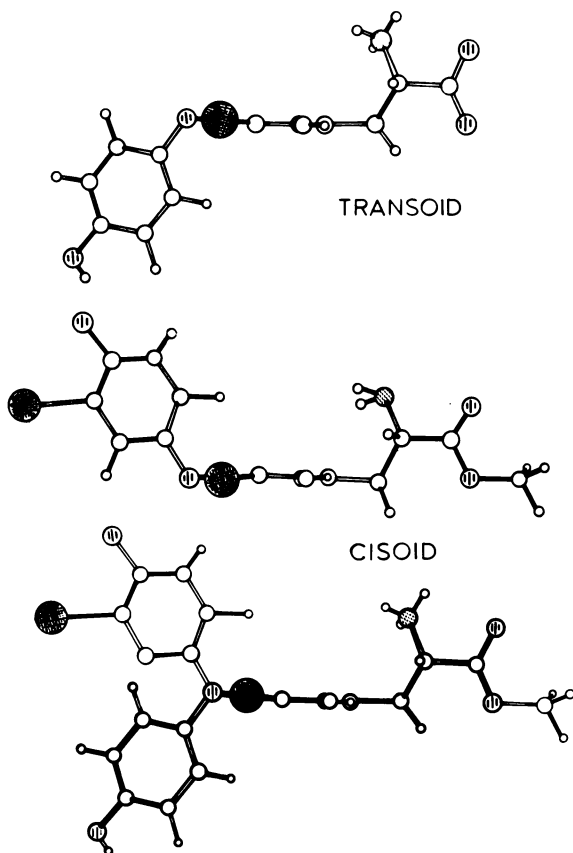


Figure 5. *Thyroid hormones illustrating transoid and cisoid overall conformation*

Diphenyl Ether Conformation

As mentioned, the torsion angles ϕ and ϕ' describe the relative orientations of the two phenyl rings with respect to the C-O-C ether plane. While the bulky 3,5-iodines maintain a perpendicular ring arrangement (6), (the ideal skewed conformation is $\phi/\phi' = 90^\circ/0^\circ$) analysis of these data suggest further correlations in the parameters which influence the diphenyl ether conformation. The plot of ϕ versus ϕ' (Figure 6) shows two lines corresponding to the cisoid and transoid conformers. The graph also shows that the data are grouped into two clusters, one near the skewed conformer ($\phi/\phi' = 90^\circ/0^\circ$) and another, twist-skewed, near $\phi = 108^\circ$, $\phi' = -28^\circ$ or $\phi = -108^\circ$, $\phi' = 28^\circ$. Further inspection of those structures involved shows that the deamino acids (thyroformic, acetic, propionic) tend to be skewed (19) whereas the iodothyronines are twist-skewed. These observations suggest that the α -amino nitrogen is responsible for the long range transmission of conformational effects.

The significant differences between these conformers are reflected in the disposition of the 4'-OH (Figure 7a) when the inner ring is used as the common structural feature or in the displacement of the inner phenyl ring and side chain (Figure 7b) when the outer ring is used as the common feature. Depending on the receptor-site binding requirements, these differences may be directly linked to variations in receptor binding affinity and hormone activity.

Hydrogen Bonding

An analysis of the hydrogen bonding and molecular packing of the thyroactive compounds in their crystal lattice offers an insight into the molecular details of hydrogen bond strength and directionality at the hormone-receptor site. In the thyronine nucleus (1) the amine acts as a hydrogen bond donor, (2) the carboxylic acid group acts as a hydrogen bond acceptor, and (3) the 4'-OH can act both as a hydrogen bond donor and acceptor, depending on its environment. A study of the hydrogen bonding observed in these crystal structures (11) shows that there is a high degree of directional specificity in the location of hydrogen bond donors and acceptors.

In those structures where the 4'-OH acts as a hydrogen bond donor (Table I, Figures 8 and 9), the acceptor atom is a carboxylic oxygen from an adjacent molecule in the lattice. The acceptor atoms approach the 4'-OH from directions nearly coplanar with the phenoxy ring plane (C3'-C4'-O4'1...X) and at an angle (C4'-O4'1...X) of 120° . In addition, for T_3 structures, the direction of approach is *trans* to the 3'-iodine, whether distal or proximal. When the 4'-OH acts as a hydrogen bond acceptor, the donor atoms (Table I) tend to approach the phenoxy ring from directions either above or below the plane and opposing the

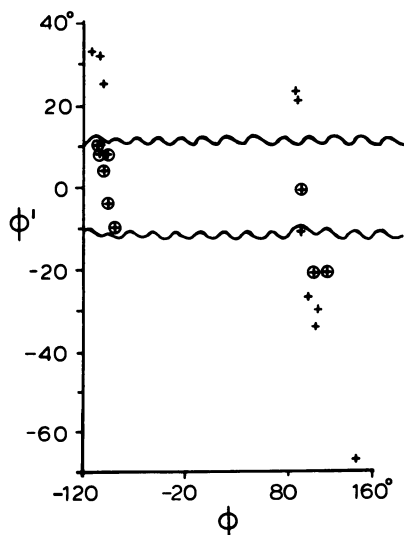


Figure 6. Plot of the torsion angle ϕ (C5-C4-O41-C1') and ϕ' (C4-O41-C1'-C6) for those thyroactive crystal structures listed in Ref. 11. Diphenylether conformation for all thyroid hormones.

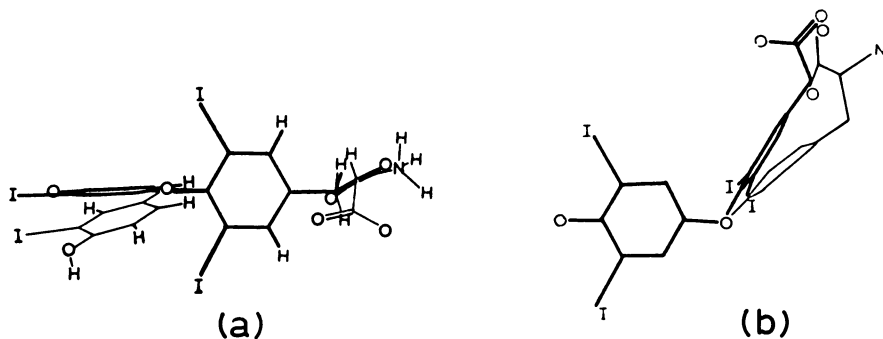


Figure 7. Superposition of a skewed (dark) diphenyl ether conformation on a twist-skewed (light) conformation. (a) Overlap of inner ring as common structural feature and (b) the outer phenyl ring.

Table I

Hydrogen Bond Directionality in Thyroactive Structures

Structure	4'-OH as Donor			4'-OH as Acceptor			
	X	C3'-C4'- O4'1...X	C4'-O4'1 ...X	X	C3'-C4'- O4'1...X	C4'-O4'1 ...X	
difodothyronine (20)	09	-175°	108°	N4 N8	38° -77	124° 116	0...X 2.97Å 2.04
dibromothyroacetic acid (11)	08	162	119		2.59		
difodothyropropionic acid (21)	010	-177	118		2.82		
trifodothyronine methyl ester (14)				N8	168	126	2.69
trifodothyroacetic acid (19)	09	-176	119	07	33	145	3.02
trifodothyronine (13)	09	151	124	N8	-10	151	2.82
trifodothyronine, HC ₂ , H ₂ O (22)				02 03	108 176	127 119	2.68 2.69
trifodothyropropionic acid ethyl ester (23)	09	160	115		2.75		
2',3'-dimethyl-difodothyronine H ₂ O (24)				01 02	-94 49	113 137	3.17 2.51
3'-isopropyl-difodothyronine H ₂ O (25)				01 01	-83 -6	127 127	2.51 2.73

Table I. (Continued)

Structure	4'-OH as Donor			4'-OH as acceptor		
	X	C3'-C4'- O4'1...X	C4'-O4'1 ...X	X	C3'-C4'- O4'1...X	C4'-O4'1 ...X
thyroxine, HCl, H ₂ O (<u>22</u>)	09	118	108	04'1	-25	145
tetraiodothyroacetic acid (<u>19</u>)			2.59	07*	178	150
tetraiodothyroformic acid (<u>11</u>)				N4*	89	132
				N4**	-78	118
thyroxine (<u>11</u>)				N8	73	113
A				N44	-102	134
				N8*	73	113
B				N43	109	130
						0...X
						2.91
						3.08
						2.71
						2.70
						2.65
						2.60
						2.65
						2.70

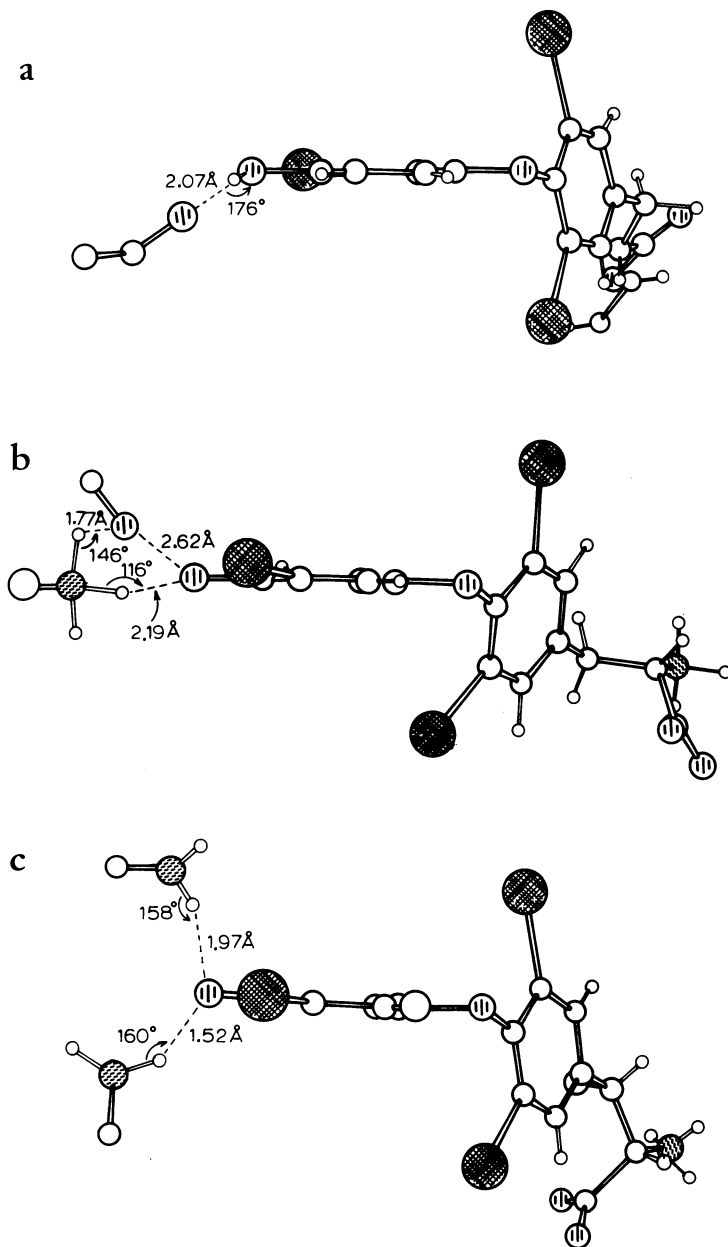


Figure 8. Example of the direction of approach of (a) hydrogen bond acceptor to the 4'-OH of T_3 ; (b) hydrogen bond donor and acceptor to the 4'-hydroxyl of T_3 ; and (c) hydrogen bond donors to the 4'-phenoxide of T_4 .

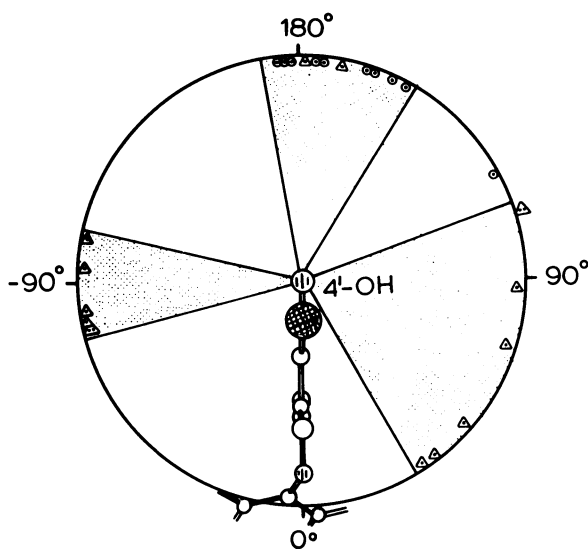


Figure 9. Distribution of hydrogen bond directionality with 4'-OH as acceptor (Δ) and donor (\odot). Plot is of hydrogen bond torsion angle C3'-C4'-O4'1 . . . X.

acceptor atom (Figure 8b).

As seen from Table I, the average value for the hydrogen bond angle C4-O4'1...X is 116° when the 4'-OH acts as a donor and 129° when it is an acceptor. Also, when the 4'-OH is simultaneously involved as a donor and acceptor, the donor atom to 4'-OH tends to have an angle near 150° . In those situations where the hydrogen bonding atoms are water, one adopts the orientation observed for the carboxylic oxygens, suggesting that it is a hydrogen bond acceptor from the 4'-OH.

In the case of tetraiodothyroactive structures where the 4'-oxygen is a phenoxide ion (11) and thus only a hydrogen bond acceptor, the donor atoms approach the 4'-phenoxy ring symmetrically from directions both above and below the plane (Figure 8c). However, where the tetraiodo structures are hydroxyls (T₄A (19)), the outer ring iodines cause the approach of the donor/acceptor atoms to deviate from this pattern.

A study of the 4'-O...X hydrogen bonds in these structures shows that the average O...O distance is 2.74\AA and the O...N distance is 2.76\AA , in general agreement with other studies (26). When the donor and acceptor O...O values are considered separately, these become 2.66\AA and 2.81\AA for 4'-OH as a donor and acceptor, respectively.

The results of theoretical energy calculations of inter- and intramolecular hydrogen bond strengths of ortho substituted phenols and phenoxides (27), as models to investigate likely orientations of thyroid hormones at their protein binding sites, are in general agreement with the hydrogen bonding patterns observed in these crystallographic determinations. These energy calculations predict an average O...O distance of 2.63\AA and a C-O...O angle of 125° , irrespective of whether the 4'-OH is acting as a proton donor or acceptor. Unfortunately, this study (27) did not compute the hydrogen bond geometry when the 4'-OH is simultaneously acting as a hydrogen bond donor and acceptor and consequently the detail observed from the crystallographic data is not apparent.

The observation that in the 3'-substituted structures the hydrogen bonding atom approaches the 4'-OH *trans* to the 3'-substituent is verification of quantitative structure activity relationship (QSAR) data which also suggest that *in vitro* binding of T₃ probably involves hydrogen bond donation of the 4'-OH to the 5'-side of the nuclear receptor (28).

Protein Binding Characteristics

Because thyroid hormones are protein bound throughout the general circulation as well as in the cell membranes and nucleus, protein binding is central to the transport, tissue distribution and metabolism rates of the various thyroid analogues (29). Current theory proposes that the hormone first binds at a specific site on the protein receptor and then an interaction

between the hormone and a complementary receptor component takes place. Steric correspondence between the receptor site and the hormone then permits electronic interaction. Thus, the structural specificity of the hormones can be both steric and electronic in nature.

The importance of specific structural features in binding to the various thyroid binding proteins has been suggested from numerous studies measuring the relative binding affinities of the hormones and their analogues for these serum, nuclear and membrane proteins (30-36). These studies indicate that the binding site requirements are different for each of the proteins.

Table II lists a few of the activity and binding affinity data available for the thyroid hormones and their analogues to membrane, serum and nuclear proteins (11). Structural requirements for potency correlate closely with those for *in vitro* binding to nuclear proteins (37). These correlations, in conjunction with other thyroid hormone studies, provide a general description of the structural features required for binding and hormonal activity (Table III). As illustrated, the major binding requirements differ principally with regard to the 3',4',5'-substituents. The nuclear receptors preferentially bind to distally 3'-oriented compounds, as do the membrane bound proteins, while the serum proteins have maximal binding for 3',5'-disubstituted compounds with electron withdrawing groups. Also, from pH dependency studies of T₃ and T₄ binding, it is apparent that 4'-OH binds to the nuclear receptors in its un-ionized form whereas it binds to the serum protein receptors in its ionized form (28,30). These results correlate well with the crystallographic observations of T₄ structures as 4'-phenoxide ions and T₃ as un-ionized 4'-hydroxyls.

Receptor Models

Recent crystallographic studies of thyroxine-binding prealbumin (38,39,40) show it to be a tetramer with a channel running through the structure. The four units are related by two perpendicular 2-fold axes. Complexes of TBPA with T₄ and T₃ show the hormones are bound inside the channel (Figure 10) with the 4'-phenoxy ring pointed toward the center of the core. However, because the thyroid hormones cannot be accommodated by the symmetry requirements of the protein, their exact orientation in the receptor site cannot be defined. Instead, an average position of two symmetry related orientations is observed. Nevertheless, this description of TBPA structural features, combined with relative binding affinity data (41) of analogues to TBPA, permits a detailed model of the hormone-receptor interaction to be formulated. In this case, the 4'-phenoxy ring is tightly bound, via hydrogen bonds, to a water or peptide functional group (39) suggesting that 4'-OH interactions are a primary requirement for binding. If, on the other hand, the side

Table II
 Relative Binding Affinities and Biological Potencies of Selected Thyroid Hormones
 and Metabolites to Thyroid Binding Proteins

Compound	TBG (33)	TBPA (41)	Nuclear (37)	Membrane (35)	Potency (33)
L-thyroxine (T ₄)	100.0%	39.3%	12.5%	96.6%	12.5%
D-thyroxine	54.0	.95	-	63.0	-
tetraiodothyronopionic acid (T ₄ P)	3.6	76.4	-	23.0	1.9
tetraiodothyroacetic acid (T ₄ A)	1.7	100.0	-	30.3	6.25
tetraiodothyroformic acid (T ₄ F)	-	90.2	-	21.5	-
tetrabromo-DL-thyronine	40.0	-	-	-	-
tetramethylthyronine	0.0	-	0.1	-	0.3
2',3'-dimethyl-3,5-diiodothyronine	0.71	-	1.1	-	6.3
2',5'-dimethyl-3,5-diiodothyronine	0.13	-	0.1	-	0.1
3,5,3'-triflodo-L-thyronine (T ₃)	9.0	1.4	100.0	100.0	100.0
3',5',3-triflodothyronine (RT ₃)	38.0	3.1	0.1	67.5	<0.1
triflodothyroacetic acid (T ₃ A)	0.3	4.7	-	60.6	4.4
trimethylthyronine	-	-	0.1	-	0.4

Table III

Structural Features Required for Optimal Protein Binding

<u>Protein</u>	<u>Structural Requirements</u>
TBG	4'-phenoxide, tetraiodo, L-alanine side chain L-thyroxine (T ₄) best binder
TBPA	4'-phenoxide, tetraiodo, acid side chain tetraiodothyroacetic acid (T ₄ A) best binder
Nuclear	4'-hydroxyl, triiodo (3'-distal), L-alanine side chain distal-L-triiodothyronine (T ₃) best binder
Membrane	appears to be similar to the nuclear proteins

All proteins require a diiododiphenyl ether nucleus. The implications of these data are that TBG and nuclear proteins prefer a twist-skewed diphenyl ether conformation whereas TBPA prefers a skewed diphenyl ether.

chain orientation and composition were of primary importance, as suggested from TBG and nuclear protein data (30,31,32,33), then the consequences of changes in the side chain orientation (χ^1 , χ^2 , Figure 4) would be significant.

The most striking feature of the data listed in Table II is that the acid metabolites bind more strongly to TBPA than thyroxine; just the opposite of the order observed for TBG and the nuclear proteins. A factor which may influence these parameters is the relative importance to binding of the diphenyl ether conformation and the side chain conformation. As was shown (19), the thyroactive acid structures prefer a skewed diphenyl conformation whereas the thyronine structures adopt a twist-skewed conformation. As illustrated (Figure 7), there is significant displacement of the 4'-OH and side chain functional groups between these two conformers. Therefore, if 4'-phenoxy interactions control binding, these differences will affect the position of the side chain groups within the binding channel.

When the molecular structures of the acid metabolites T₄A (dark, Figure 11) is successively superimposed over that of (a) T₄F, (b) T₄P, (c) T₄, and (d) all of them; it can be seen that the acid side chains describe a probably binding volume from which the T₄ amine is excluded. In addition, the T₄ inner ring has the largest deviation from the positional range described by the acid metabolites. Also, the carboxylic oxygens of the various metabolites (except T₄F), irrespective of composition,

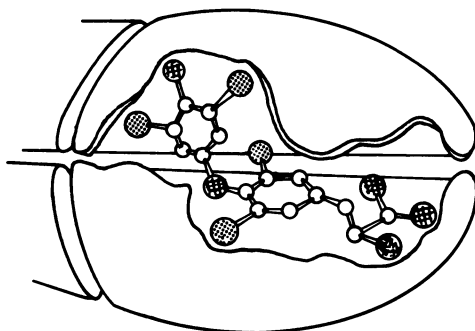


Figure 10. Model of thyroxine binding prealbumin-receptor interaction based on crystallographic protein data (39)

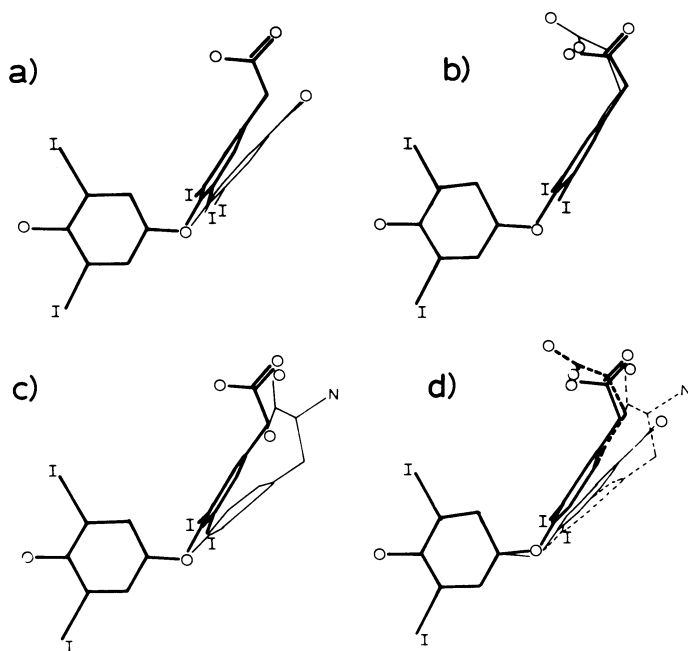


Figure 11. Superposition of molecular structure of T_4A with (a) T_4F ; (b) T_4P ; (c) T_4 ; and (d) all of them, illustrating binding volume of thyroactive acid metabolites, assuming 4'-phenoxy ring fixed.

tend to cluster together, suggesting that this feature may also be of importance in determining relative binding affinity.

Analysis of the TBPA-T₄ complex (39,40) indicates that the binding site for the hormone is located deep inside the channel. The hormone makes extensive interactions with the protein side chains that project into the channel. The 4'-hydroxyl of T₄ interacts with a patch of hydroxy-amino acids of the protein while each of the iodines makes contact with a number of hydrophobic protein residues. The T₄ amino acid side chain functional groups are in appropriate positions to interact with glutamic acid and lysine residues. Thus, this channel provides a favorable environment for each of the characteristic substituents of the thyroid hormone (40). However, because of the T₄ orientation disorder in the protein complex, this structural model is not a sensitive measure of the observed correlations between diphenyl ether conformations and binding affinity data.

The differences in binding orders between TBPA and TBG suggest that different structural features may play a key role in receptor interactions. It has been shown (4,28) that TBG also preferentially binds to a tetraiodo-4'-phenoxide ion, but since T₄ is the strongest binder, this suggests a different side chain stereochemistry. Here we can assume that it is the twist-skewed diphenyl ether conformation which orients the T₄ side chain for optimal receptor-hormone interactions. In the case of the nuclear proteins optimal binding is observed for a distally oriented 3'-I and a 4'-hydroxyl. Side chain requirements appear to be similar to those of TBG (28,31).

Therefore, changes in the relative binding affinities of thyroid hormone structures to receptors will ultimately depend upon the specific steric requirements of the binding site and the ability of the hormones to adopt the required conformation. In addition, the net charge of the hormone could also influence its binding ability. Thus, the observation of conformational patterns among thyroactive structures which correlate with function, activity and binding data provide useful information in describing specific types of hormone-receptor interactions.

Acknowledgements

This research was supported in part by a grant from HEW AM-15051. Graphs and figures were prepared through interrogation of the thyroid data stored in the NIH PROPHET system, an NIH interactive computer network. The secretarial and technical assistance of Mrs. B. Giacchi, Miss M. Tugac, Miss G. Del Bel and Mrs. C. DeVine are gratefully acknowledged.

Literature Cited

1. Zenker, N.; Jorgensen, E. C. J. Amer. Chem. Soc., 1959, 81, 4643.

2. Oppenheimer, J. H.; Schwartz, H. L.; Surks, M. I.; Koerner, D.; Dillman, W. H. Recent Progr. Horm. Res., 1976, 32, 529.
3. Lehmann, P. A. J. Med. Chem., 1972, 15, 404.
4. Jorgensen, E. C. Pharmac. Ther., 1976, B2, 661.
5. Jorgensen, E. C.; Lehmann, P. A.; Greenberg, C.; Zenker, N. J. Biol. Chem., 1962, 237, 3832.
6. Jorgensen, E. C. Mayo Clinic Proc., 1964, 39, 560.
7. Schussler, G. C. Science, 1972, 178, 172.
8. Lehmann, P. A.; Jorgensen, E. C. Tetrahedron, 1965, 21, 363.
9. Kier, L. B.; Hoyland, J. R. J. Med. Chem., 1970, 13, 1182.
10. Camerman, N.; Camerman, A. Science, 1972, 175, 764.
11. Cody, V. Recent Progr. Horm. Res., 1978, 34, 437.
12. Cody, V. Science, 1973, 181, 757.
13. Cody, V. J. Amer. Chem. Soc., 1974, 96, 6720.
14. Cody, V. J. Med. Chem., 1975, 18, 126.
15. Cody, V.; Duax, W.L. Biochem. Biophys. Res. Comm., 1973, 52, 430.
16. Emmett, J. C.; Pepper, E. S. Nature, 1975, 257, 334.
17. Kollman, P. A.; Murray, W. J.; Nuss, M. E.; Jorgensen, E. C.; Rothenberg, S. J. Amer. Chem. Soc., 1973, 95, 8518.
18. Cody, V.; Duax, W. L.; Hauptman, H. A. Int. J. Peptide Protein Res., 1973, 5, 297.
19. Cody, V.; Hazel, J. P.; Langs, D. A.; Duax, W. L. J. Med. Chem., 1977, 20, 1628.
20. Cody, V.; Duax, W. L.; Norton, D. A. Acta Cryst., 1972, B28, 2244.
21. Cody, V.; Erman, M.; DeJarnette, F. E. J. Chem. Res., 1977, S, 126.
22. Camerman, A.; Camerman, N. Acta Cryst., 1974, B30, 1832.
23. Camerman, N.; Camerman, A. Can. J. Chem., 1974, 52, 3048.
24. Fawcett, J. K.; Camerman, N.; Camerman, A. Can. J. Chem., 54, 1317.
25. Fawcett, J. K.; Camerman, N.; Camerman, A. J. Amer. Chem. Soc., 1976, 98, 587.
26. Mitra, J.; Ramakrishnan, C. Int. J. Peptide Protein Res., 1977, 9, 27.
27. Andrea, T. A.; Dietrich, S. W.; Murray, W. J.; Kollman, P. A.; Jorgensen, E. C.; Rothenberg, S. J. Med. Chem., 1979, 22, 221.
28. Dietrich, S. W.; Bolger, M. B.; Kollman, P. A.; Jorgensen, E. C. J. Med. Chem., 1977, 20, 863.
29. Pittman, C. S.; Pittman, J. A. In: "Physiology, Vol. III, Thyroid", Greer, M. A.; Solomon, D. H., Ed.; American Physiological Society: Washington, D.C., 1974.
30. Tabachnick, M.; Korcek, L. Biochimica et Biophysica Acta, 1978, 537, 169.
31. Koerner, D.; Schwartz, H. L.; Surks, M. I.; Oppenheimer, J. H.; Jorgensen, E. C. J. Biol. Chem., 1975, 250, 6417.
32. Goldfine, I. D.; Smith, G. J.; Simons, C. G.; Ingbar, S. H.; Jorgensen, E. C. J. Biol. Chem., 1976, 251, 4233.

33. Snyder, C. M.; Cavalieri, R. R.; Goldfine, I. D.; Ingbar, S. H.; Jorgensen, E. C. J. Biol. Chem., 1976, 251, 6489.
34. Spindler, B. J.; MacLeod, K. M.; Ring, J.; Baxter, J. D. J. Biol. Chem., 1975, 250, 4113.
35. Singh, S. P.; Carter, A. C.; Kydd, D. M.; Costanzo, R. R. Endocrine Res. Comm., 1976, 3, 119.
36. Sterling, K.; Milch, P. O.; Brenner, M. A.; Lazarus, J. H. Science, 1977, 197, 996.
37. Jorgensen, E. C. In: "The Thyroid", Werner, S. C.; Ingbar, S. H., Ed.; Harker & Row, 1978; p. 125.
38. Blake, C. C. F.; Geisow, M. J.; Swan, I. D. A.; Rerat, C.; Rerat, B. J. Mol. Biol., 1974, 88, 1.
39. Blake, C. C. F.; Oatley, S. J. Nature, 1977, 268, 115.
40. Blake, C. C. F. Endeavour, 1978, 2, 137.
41. Andrea, T. A. Ph.D. Thesis, 1977, University of California, San Francisco.

RECEIVED June 8, 1979.

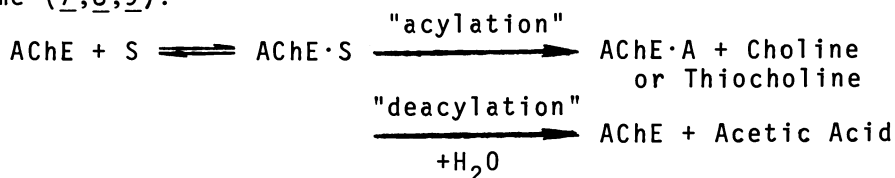
Theoretical Modeling of Enzymic Hydrolysis of Acetylcholine Compared to Acetylthiocholine

JOYCE H. CORRINGTON

Xavier University of Louisiana, New Orleans, LA 70125

Acetylcholine (ACh, 51-84-3) is a molecular ion which functions as a neurohumoral transmitter in the peripheral nervous system. It is released at the end of one cell, diffuses across the synapse, induces permeability changes in the next cell, and is then rapidly removed from the synaptic region by a hydrolytic reaction catalyzed by the enzyme acetylcholinesterase (AChE, EC 3.1.1.7). Replacement of the acetate oxygen of ACh by sulfur yielding the thiol ester analog acetylthiocholine (ASCh, 4468-05-7) profoundly modifies the biological action of this molecular ion (1,2,3). The relative ability of these isosteres to induce polarization in isolated single cell electroplax preparations is 1.0:16.7 (4). However, the relative rate at which these isosteres are hydrolyzed by AChE is very similar, 1.0:0.6 (5).

Proposals made concerning this hydrolysis mechanism suggest that the active site of AChE consists of a nucleophilic serine residue and a histidine residue which probably serves as a proton donor or receiver (6). Hydrolysis is thought to occur through the formation of a Michealis complex between the active site of the AChE and the substrate (S), the conversion of the complex in an "acylation" reaction to an acyl enzyme intermediate (AChE·A) and the product choline (or thiocholine), and the subsequent hydrolysis of the acyl enzyme in a "deacylation" reaction to acetic acid and the regenerated enzyme (7,8,9).



The acylation reaction appears controlled by a functional

0-8412-0521-3/79/47-112-301\$05.00/0

© 1979 American Chemical Society

group of $pK = 5.3$, while the deacylation reaction appears catalyzed by a group of $pK = 6.3$, and for both ACh and ASCh the rate determining step is the deacylation reaction (10).

Semiempirical molecular orbital calculations have been employed to compare the conformation (11) and the electronic properties of ACh and ASCh (12). Two studies have investigated proposed hydrolysis mechanisms for ACh (13,14). This study similarly employs semiempirical calculations to investigate the commonly proposed steps of the mechanism for the AChE catalyzed hydrolysis of ACh and ASCh.

In this study the molecular orbital calculation program employed, ARCANA, utilized only atomic data and iterated to charge self-consistency (see Calculation Note). The substrates, ACh and ASCh, and the products of hydrolysis, choline, thiocholine, and acetic acid, were represented in their entirety, while the AChE enzymic active site was modeled by representing the histidine residue and the serine residue thought to be at the active site by imidazol and methanol respectively. Molecular orbital calculations have been performed for reactants, three possible intermediate acylation complexes, and the acylation products. Since the deacylation mechanism is identical for both ACh and ASCh and has been reported elsewhere (14), those results will not be repeated here. For each of the acylation steps, three cases were calculated: (a) general acid catalyzed, (b) non-catalyzed or neutral, and (c) general base catalyzed. Molecular orbital coefficients and energies, total valence energies, Mulliken net atomic charges, bond and total overlap populations, and bond and total overlap energies were calculated for each run.

The atomic coordinates used for ACh were those employed in a previous theoretical study (15) and were calculated employing idealized hybridization and average bond distances. While crystallographic studies of ACh are available (16,17), the gauche crystal conformation of ACh was not used for this study, but rather an anti-planar conformation similar to the transition state geometry suggested by kinetic data (18-22). The atomic coordinates for ASCh in its anti-planar crystal structure were employed (23). The ASCh crystallographic bond distances and the ACh generalized bond distances were identical within 0.04 Å, the accuracy of the crystallographic data, with the exception of those involving sulfur. Table I presents the bond distances employed in these calculations for the substrate, either ACh or ASCh.

Table I
Interatomic Distances (Å)

	X = O	X = S
$R_3C - XCR_3$	1.42	1.97
$RC(O) - XCR_3$	1.31	1.77
$R_2C = O$	1.23	1.23
$R_3C - CR_3$	1.52	1.52
$RX - H$	0.96	1.33
$R_3C - H$	1.09	1.09
$R_3C - NR_3$	1.48	1.48

The imidazol coordinates were taken from an x-ray diffraction crystallographic study of histamine (24) with a proton being substituted for the ethylamine side chain of histamine. The coordinates of other reactants, products, and intermediate complexes were calculated employing idealized hybridization and average bond distances or are as discussed.

Results

The acylation mechanism was assumed to consist of the four steps following:

Step I. The reactants (ACh or ASCh, methanol, imidazol(s) and/or imidazolium(s) form a "planar" intermediate complex between the carbonyl carbon of the substrate (which is in an sp^2 "planar" hybridization state) and the oxygen of the attacking nucleophile methanol. The nucleophile is assumed to attack perpendicular to the plane of the substrate acetate (or thioacetate) moiety with the electrons from the methanol oxygen filling the lowest empty molecular orbital which is centered on the carbonyl carbon of the substrate.

Step II. The "planar" acetate (or thioacetate moiety of the ACh (or ASCh) forms a "tetrahedral" complex with the attacking nucleophile methanol as the hybridization of the substrate carbonyl carbon changes from sp^2 to sp^3 .

Step III. A "protonated ester" intermediate is formed by the transfer of the hydroxy proton from the

attacking nucleophile methanol to the acetate oxygen (or thioacetate sulfur), except in the base catalyzed pathway where this proton has already been donated to an imidazol, and this step is not assumed.

Step IV. The "protonated ester" intermediate breaks into choline (or thiocholine) and an acyl enzyme intermediate, modeled here by methyl acetate.

Each of the above steps was calculated for three pathways: (a) assuming an imidazolium present at the AChE active site acted as a general acid catalyst and donated a proton to the acyl oxygen of the substrate (thus two imidazols were included in the calculations), (b) assuming there was no acid or base catalyst (thus one imidazol and one imidazolium were included in the calculations), and (c) assuming an imidazol present at the AChE active site acted as a general base catalyst and abstracted a proton from the nucleophile making it a methoxy (thus two imidazoliums were included in the calculations). In all cases the number of atoms present were thus kept constant from run to run. Since no crystallographic data of the active site are available, modeling of the specific interaction of the histidine residue with the proposed reaction intermediate complexes was not attempted. While a few angles and distances were varied during the modeling, in general these were not optimized. The distance between the substrate carbonyl carbon and the nucleophilic oxygen in the "planar" intermediate complex was selected so that the coordinates of the nucleophile did not have to be altered as the substrate changed hybridization and formed the "tetrahedral" intermediate complex with a normal carbon to oxygen bond distance.

Discussion

Total Valence Energy. The calculated total valence energies presented below in Table II are the equivalent of total energies but include only factors due to the valence electrons and the nuclear charges corresponding to them. All runs were converged to a 1 kcal/mole tolerance, but since arbitrary and rigid conformations were employed these energies should be taken only as indicative of the energy changes involved. Also, of course, entropy and solvation effects are not included in such isolated or "gas phase" molecular calculations. But even with these qualifications, the total valence energies indicate that both ACh and ASCh would follow a lower energy acylation pathway when acid catalyzed (when

Table II

Total Valence Energy for Steps in the Enzymic Acylation of ACh and ASCh (g-kcal/mole, relative to the reactants)

Mechanism Step	(A) Acid Catalyzed		(B) Non-Catalyzed		(C) Base Catalyzed	
	ACh	ASCh	ACh	ASCh	ACh	ASCh
I	-12	25	162	153	315	293
II	-63	15	96	110	283	274
III	-47	-58	138	120	-	-
IV			57	79		

the imidazolium assumed present at the catalyst's active site donates a proton to the acyl oxygen of the substrate). The energies of the ACh "planar" and "tetrahedral" intermediate complexes are lower than those for ASCh. Both the ACh and the ASCh "tetrahedral" intermediates are lower in energy than the corresponding "planar" intermediates, as might be expected. A noticeable difference exists in the intermediate formed when a proton is shifted from the attacking nucleophile to the acetate oxygen (or thioacetate sulfur), since the ACh intermediate is not as stable as its "tetrahedral" intermediate, while the ASCh intermediate is considerably more stable than its "tetrahedral" intermediate and indeed roughly equivalent to the ACh "tetrahedral" intermediate. These ACh and ASCh intermediates are both more stable than the proposed acylation products choline (or thiocholine) and the acyl enzyme (modeled here by methylacetate). In both cases the acylation reaction is calculated to be endothermic.

Since the resulting acyl enzyme is identical for both substrates, the deacylation reaction would be identical for both and is not discussed here. However, a similar previous study (14) did indicate that deacylation would have to proceed through a higher energy pathway than either the ACh or the ASCh acylation, and thus would be rate limiting as experimental data suggests (10).

Mulliken Net Atomic Charges. The net atomic charges calculated according to a Mulliken population analysis (25) are presented below in Table III for the low energy or general acid catalyzed pathway. Both ACh and ASCh are molecular ions with positively charged cationic "heads." The calculated total Mulliken net atomic charges on the onium methyls of both ACh and ASCh total approximately 0.98 and stay essentially constant throughout the proposed mechanism steps. Similarly, the net charge on the onium nitrogen is initially a negative -0.31 (in contrast to common notation which denotes the nitrogen as positively charged). The nitrogen net charge becomes somewhat less negative in the proposed mechanism steps, but net charges in the total onium functional groups in ACh, ASCh, choline and thiocholine stay essentially constant and equivalent.

The carbonyl oxygen is the most negative atom in both substrates, and has a Mulliken net charge of -0.698 in ACh and -0.661 in ASCh. Centered on it in both substrates is the highest occupied molecular orbital, thus the carbonyl oxygen is the most subject to electrophilic attack or protonation. In the intermediated formed in

Table III
Mulliken Net Atomic Charges for Acylation Intermediates for Substrates ACh and ASCh

Mechanism Step	(H ₃ C) ₃	N ⁺	-(CH ₂) ₂	X	H	O	C(=====O---H ⁺)-(CH ₃)
Reactants							
ACh, Methanol	0.986	-0.314	0.669	-0.612	0.364	0.306	-0.670 0.877 -0.698
IA	0.966	-0.266	0.644	-0.558	0.396	0.361	-0.492 0.958 -0.529 0.376 0.136
IIA	0.964	-0.269	0.625	-0.606	0.447	0.434	-0.348 0.932 -0.593 0.341 0.074
IIIA	0.981	-0.252	0.679	-0.379	0.439	0.395	-0.625 0.907 -0.593 0.357 0.090
IV	0.972	-0.316	0.594	-0.614	0.363	0.410	-0.662 0.895 -0.712
Reactants							
ASCh, Methanol	0.976	-0.308	0.334	0.118	0.364	0.306	-0.670 0.438 -0.661
IA	0.965	-0.265	0.327	0.177	0.381	0.373	-0.511 0.519 -0.498 0.382 0.149
IIA	0.962	-0.268	0.308	0.067	0.429	0.469	-0.366 0.551 -0.580 0.351 0.077
IIIA	0.979	-0.251	0.394	0.542	0.151	0.370	-0.646 0.597 -0.591 0.368 0.088
IV	0.972	-0.310	0.242	-0.021	0.067	0.410	-0.662 0.895 -0.712

the acid catalyzed pathway, this oxygen is assumed to be protonated by a proton donated by an imidazolium at the AChE's active site. The Mulliken net charge of this oxygen becomes slightly less negative when protonated, but is similar whether the substrate is ACh or ASCh. The proton retains only a fraction of its net positive charge, varying from 0.341 to 0.382 among the proposed intermediates. It thus apparently withdraws considerable electron density, but not from the adjacent acyl oxygen, rather from the oxygen of the attacking nucleophile, thus helping stabilize the developing bond between the substrate and the attacking nucleophile.

The carbonyl carbon is the most positive atom in both substrates, but is considerably more positive in the ACh (0.877) than in the ASCh (0.438). However, the lowest energy empty molecular orbital (LEMO), which is centered on this atom perpendicular to the plane of the acetate (or thioacetate) moiety, has similar orbital energies for both the ACh (-12.00 eV) and the ASCh (-12.35 eV). Both the positive Mulliken net atomic charge and the location of the LEMO on the carbonyl carbon make it subject to nucleophilic attack. In the intermediates formed in the proposed acid catalyzed mechanism steps, the Mulliken net atomic charges of the carbonyl carbon does not change greatly despite the formation of the developing bond between the substrate and the attacking nucleophile.

The acetate oxygen is the second site of net negative charge in the ACh substrate (-0.612), but the corresponding thioacetate sulfur has a net positive charge in the ASCh substrate (+0.118). This difference in net charge has been proposed as the explanation why the crystal structure of ACh and ASCh are different (12). It may also explain why the total valence energies of the proposed ASCh "planar" and "tetrahedral" acylation intermediates are higher than the corresponding ACh intermediates.

This acetate oxygen (and thioacetate sulfur) is also the primary site of the second highest energy occupied molecular orbital, an anti-bonding pi-type orbital shared with the carbonyl oxygen. Because of this, it was proposed that a possible acylation intermediate (IIIA) could result if a proton was transferred from the hydroxy of the attacking nucleophile to the acetate oxygen (or thioacetate sulfur). This intermediate resulted in a moderate positive change in the net atomic charge of the acetate oxygen, but a large positive change in the net charge of the thioacetate sulfur.

In the separated choline molecular ion, the hydroxy oxygen returns to a net atomic charge (-0.614) similar

to what it was in the choline moiety of the ACh, while in the separated thiocholine molecular ion the mercaptan sulfur has a more negative net atomic charge (-0.021) than it had in the thiocholine moiety of ASCh.

Bond Overlap Population. The calculated bond overlap populations for some of the bonds involved in the acylation reaction are presented below in Table IV. The overlap population (OP) between atoms A and B is calculated by:

$$OP_{AB} = 4 \sum_u \sum_i \sum_j C_{iu} C_{ju}$$

where *i* and *j* are summed over all atomic basis orbitals and *u* over all molecular orbitals.

In the ACh substrate, the OP indicates a higher electron density (0.698) between the carbonyl carbon and the acetate oxygen than the electron density (0.579) between the alkyl carbon and the acetate oxygen due to partial pi-bonding between the acetate oxygens. However, the ASCh corresponding electron density (0.980) between the carbonyl carbon and the thioacetate sulfur is even larger due to even more partial pi-bonding between the thioacetate sulfur and the carbonyl oxygen. This enhanced pi-bonding occurs without expanding the atomic basis set for the sulfur atom to include extravalent d orbitals (12). Since this ester (thioester) linkage is the one which commonly fissions in ester hydrolysis, it is obvious that this relatively strong bond must weaken before the choline (thiocholine) moiety can depart to conclude the acylation reaction. It is observed that the formation of the proposed "tetrahedral" intermediate disturbs this partial pi-bonding such that the electron density of the ester linkage falls to 0.637 (and the thioester linkage to 0.734), but not to a lower electron density than that of the adjacent alkyl linkage. Not until a proton is transferred to the acetate oxygen (or to the thioacetate sulfur) does the electron density of the ester linkage fall to 0.485 (and that of the thioester linkage to 0.663), both below the electron density of the adjacent alkyl linkage.

The electron density of the bond forming between the attacking methanol oxygen and the substrate carbonyl carbon increases steadily for all proposed intermediates, but is initially larger for the ACh substrate (0.220) than for the ASCh substrate (0.074).

Also, for all intermediates, the negative anti-bonding electron densities between the attacking methanol oxygen and the thioacetate sulfur and carbonyl oxygen are larger than are the corresponding electron densities

Table IV

Bond Overlap Population for Acylation Intermediates for Substrates ACh and ASCh

Mechanism Step	Alkyl C - Acetate O or Thioacetate S		Carbonyl C - Acetate O or Thioacetate S		Methanol O - Acetate O or Thioacetate S	
	ACh	ASCh	ACh	ASCh	ACh	ASCh
Reactants	0.579	0.670	0.698	0.980	-	-
IA	0.578	0.673	0.765	0.893	-0.130	-0.170
IIA	0.588	0.669	0.637	0.734	-0.101	-0.157
IIIA	0.528	0.665	0.485	0.663	-0.134	-0.155
IV	0.641	0.692	-	-	-	-

Mechanism Step	Methanol O - Carbonyl C		Methanol O - Carbonyl O		Carbonyl C - Carbonyl O	
	ACh	ASCh	ACh	ASCh	ACh	ASCh
Reactants	-	-	-	-	1.020	0.858
IA	0.220	0.074	-0.146	-0.161	0.841	0.981
IIA	0.588	0.499	-0.104	-0.123	0.655	0.623
IIIA	0.653	0.665	-0.118	-0.128	0.664	0.638
IV	0.708	0.708	-0.186	-0.186	0.995	0.995

for the acetate, helping explain why the total valence energies for the proposed ASCh mechanism steps were higher than for the corresponding proposed ACh mechanism steps.

Total Overlap Population. While the above study of the individual bond overlap populations is helpful in following the changes occurring in the proposed mechanism steps, the total overlap population, presented in Table V, does not appear to be a useful value. The total overlap population is the sum over all pairs of atoms A and B of the overlap population OP_{AB} , defined above. All intermediates and products have lower total overlap than the reactants, and the apparently lower energy acid catalyzed pathway does not have the largest total overlap of the three pathways studied.

Total Overlap Energy. However, while total overlap population does not appear to be a useful parameter, the calculated total overlap energy, as presented in Table VI, does. The total overlap energy is the sum over all pairs of atoms A and B of the overlap energy, OE_{AB} , which is defined similarly to the overlap population above, but elements of the H-matrix are used in place of elements of the overlap matrix:

$$OE_{AB} = 4 \sum_u \sum_i \sum_j H_{ij} C_{iu} C_{ju}$$

where i and j are summed over all atomic basis orbitals and u over all molecular orbitals.

In both the ACh and ASCh acylation mechanisms, the total overlap energy is a minimum for the acid catalyzed pathway, thus corresponds to the calculated total valence energy. Total overlap energy also becomes more negative as the mechanism steps proceed. Analysis of the components of the calculated total overlap energy indicates that in the acid catalyzed pathway the individual bond energies of the large ACh or ASCh intermediates were stabilized by the presence of the imidazolium donated proton more than the imidazolium was destabilized, while in the base catalyzed pathway the large ACh or ASCh intermediates were destabilized by the absence of the proton donated to the imidazol by the attacking methoxy more than the imidazol was stabilized. (The opposite was observed to be true in deacylation when the acyl enzyme was modeled as the smaller molecule methylacetate (14).)

Table V

Total Overlap Population (relative to the reactants)

Mechanism Step	(A) Acid Catalyzed		(B) Non-Catalyzed		(C) Base Catalyzed	
	ACh	ASCh	ACh	ASCh	ACh	ASCh
I	-0.423	-0.541	-0.455	-0.498	-0.909	-0.817
II	-0.393	-0.591	-0.377	-0.402	-0.870	-0.802
III	-0.544	-0.329	-0.671	-0.449	-	-
IV			-0.100	-0.223		

Table VI

Total Overlap Energy (g-kcal/mole, relative to reactants)

Mechanism Step	(A) Acid Catalyzed		(B) Non-Catalyzed		(C) Base Catalyzed	
	ACh	ASCh	ACh	ASCh	ACh	ASCh
I	-640	-442	254	236	959	773
II	-806	-579	121	123	881	700
III	-816	-728	27	-29	-	-
IV			266	308		

Conclusions

This study, employing ARCANA, a semiempirical molecular orbital calculation program, to model the acylation step of the enzymic hydrolysis mechanism for ACh and ASCh indicates significant differences in the Mulliken net atomic charges and the overlap densities for the bonds most involved in the acylation for the two substrates and the proposed intermediates. However, these differences do not seem as important as the similarities, such as the location of similar lowest energy pi-type empty molecular orbitals on the positively charged carbonyl carbon subjecting it to nucleophilic attack, and the stabilization of such an attack by the protonation of the negatively charged carbonyl oxygen. Moreover in the case of both substrates, there is significant partial pi-bonding in the acetate (or thioacetate) moiety, and protonation of the acetate oxygen (or thioacetate sulfur) is required to diminish the ester (thioester) linkage and allow the choline (thiocholine) to depart to complete the acylation step. Thus, this study indicates that the acylation step for either the ACh or the ASCh substrate would proceed through a low energy (and thus rapid) acid catalyzed pathway, consistent with experimental data on ACh which suggests that acylation is catalyzed by a functional group of $pK = 5.3$ (10) and with experimental data which show that both ACh and ASCh are hydrolyzed at nearly the same rate with the slow deacylation step being the rate determining step (5).

Calculation Note

The theoretical basis of the molecular orbital program employed in this study, ARCANA, has been developed elsewhere (26, 27) and will not be repeated here at length. Basically, ARCANA is an interactive extended Huckle method which includes neighbor atom interactions and iterates to charge self-consistency. A minimum basis set of invariant but overlap optimized atomic orbitals (28) are employed to calculate overlap integrals. The empirical orbital energy (29) and an effective radius calculated by evaluating $R_i = 1/\langle 1/r \rangle$ for accurate ab initio atomic wave functions (30, 31) are employed to calculate the Hamiltonian elements. Only the values for hydrogen deviate from standard atomic values. Radial moment analyses of accurate molecular orbitals (32) indicated that a contracted, scaled helium-like representation is required for hydrogen in a molecule. Table VII summarizes the atomic orbital parameters employed in the calculations.

Table VII
Orbital Parameters Employed in ARCANA Calculations

Element	Orbital	n_{STO}	z_{STO}	Orbital Energy (eV)	$R_i = 1/\langle 1/r \rangle$ (a.u.)
H	1s	1 ^a	1.45 ^a	-9.0 ^a	0.65 ^a
C	2s	2 ^b	1.57 ^b	-19.5 ^c	1.12 ^d
	2p	1 ^b	0.88 ^b	-9.9 ^c	1.28 ^d
N	2s	2 ^b	1.88 ^b	-25.5 ^c	0.94 ^d
	2p	1 ^b	1.06 ^b	-12.5 ^c	1.04 ^d
O	2s	2 ^b	2.19 ^b	-32.0 ^c	0.80 ^d
	2p	1 ^b	1.22 ^b	-15.3 ^c	0.90 ^d

a Ref. 32, b Ref. 28, c Ref. 29, d Ref. 31

After iterating to charge self-consistency, a total valence energy is calculated. This includes appropriate factors for valence electron repulsion (but excludes the core electrons) and includes nuclear repulsion (but excludes the corresponding core nuclear charges).

A recent comparison (33) of quantities calculated from molecular electronic wave functions for pyrrole and pyrazole indicated that ARCANA values compared more favorably with large-scale ab initio values than those calculated by other non-rigorous methods in the case of orbital energies of occupied molecular orbitals, gross atomic populations of heteroatoms, and total overlap populations including negative overlap populations between nonbonded atoms.

Acknowledgements

The author wishes to thank the following student research assistants for their contributions to this work: Warren Welters, Ramona Calvey, Debra Simmons, Cheryl Mackie, Mary Dean, and Larry Givens. This study was supported by Grant RR-08008 from the General Support Branch, Division of Research Resources, National Institutes of Health.

Literature Cited

1. Scott, K. A. and Moutner, H. G., Biochem. Pharmacol., 13, 907 (1964)
2. Mautner, H. G., Bartels, E. and Webb, G. D., Biochem. Pharmacol., 15, 187 (1966)

3. Mautner, H. G., J. Gen. Physiol., 54, No. 1, Pt. 2, 2715 (1969)
4. Mautner, H. G., Dexter, D. D., and Low, B. W., Nature (London), New Biology, 238, 87 (1972)
5. Scott, K. A. and Mautner, H. G., Biochem. Pharmacol., 13, 907 (1964)
6. Froede, H. C. and Wilson, J. B., Enzymes, 3rd Ed., 5, 87 (1971)
7. Wilson, I. B. and Bergmann, F., J. Biol. Chem., 186, 683 (1950)
8. Krupka, R. M., Biochem., 5, 1983 (1966)
9. Krupka, R. M., Biochem., 5, 1988 (1966)
10. Krupka, R. M., Biochem., 6, 1183 (1967)
11. Pullman, B. and Courriere, P., Mol. Pharmacol., 8, 371 (1972)
12. Aldrich, H. S., Int. J. Quantum Chem., Quantum Biol. Symp., 2, 271 (1975)
13. Farkas, M., Kruglyak, Y. A., Nature (London), 223, 523 (1969)
14. Corrington, J. H. (submitted)
15. Genson, D. W. and Christoffersen, R. F., J. Amer. Chem. Soc., 95, 362 (1973)
16. Camepa, F. G., Pauling, P., and Sorum, H., Nature (London), 210, 907 (1966)
17. Herdklotz, J. K. and Sass, R. L., Biochem. Biophys. Res. Commun., 40, 583 (1970)
18. Clothia, C. and Pauling, P. J., Nature (London), 223, 919 (1969)
19. Smissman, E. E. and Chappell, G. S., J. Med. Chem., 12, 432 (1969)
20. Smissman, E. E., Helson, W. L., LaPridus, J. B. and Day, J. L., J. Med. Chem., 9, 458 (1966)
21. Stephen, W. F., Jr., Smissman, E. E., Schowen, K. B., and Self, G. W., J. Med. Chem., 15, 241 (1972)
22. Schowen, K. B., Smissman, E. E., and Stephen, W. F., Jr., J. Med. Chem., 18, 292 (1975)
23. Shefter, E., Mautner, H. G., and Smissman, E. E., Acta Crystallogr. Sect. A, 25, S201 (1969)
24. Rerat, C., Bull. Soc. Franc. Min. Crist., 85, 153 (1962)
25. Mulliken, R. S., J. Chem. Phys., 23, 1833, 1841 (1955)
26. Corrington, J. H., Aldrich, H. S., McCurdy, C. W., and Cusachs, L. C., Int. J. of Quantum Chem., 5, 307 (1971)
27. Cusachs, L. D. and Miller, D. J., Advan. Chem. Ser. 110, 1 (1972)
28. Cusachs, L. C. and Corrington, J. H., "Atomic Orbitals for Semiempirical Molecular Orbital Calculations," in Sigma Molecular Orbital Theory, Sinanoglu, O.,

- and Wiberg, K. B., eds., (Yale University Press, New Haven, Conn., 1970), p. 256
29. Cusachs, L. C. and Reynolds, J. W., J. Chem. Phys., 43, S160 (1965)
 30. Clementi, E., J. Chem. Phys., 41, 295, 303 (1964); J. Chem. Phys., 38, 996, 1001 (1965), IBM J. of Res. and Dev., 9, 2 (1965)
 31. Froese, C., J. Chem. Phys., 45, 1417 (1966) and supplement and private communication
 32. Cusachs, L. C. and Aldrich, H. S., Int. J. Quantum Chem., 6S, 221 (1972)
 33. Kaufman, J. J., Preston, H. J. T., Kerman, E., and Cusachs, L. C., Int. J. of Quantum Chem., 7S, 249 (1973)

RECEIVED June 8, 1979.

A New Approach to Bioactive Synthesis

PHILIP S. MAGEE

Chevron Chemical Co., 940 Hensley Street, Richmond, CA 94804

The application of QSAR to bioactive synthesis has always suffered from an unfortunate paradox. In order to develop a useful equation, it is necessary to first complete a substantial fraction of the synthesis. Only then can the derived equation assist in extending or optimizing the bioactive series. No help is available for the earliest or intermediate stages of synthesis which have already been passed. Nor is it certain that a useful equation can be gained from the first 10-20 members of a series. Poor selection of structural changes, variable bio-data, differential metabolism of some members and the presence of unknown factors can all lead to poor correlations of little practical use. These problems are common to anyone who has attempted QSAR on novel bioactive series.

There is another problem that is equally vexing and this relates to drug or pesticide modification. In most cases, the complete structural series and biodata were developed in another laboratory and are unavailable. The only available knowledge may be the structures of the commercial and patented bioactives. No equation is forthcoming unless the entire study is repeated in your own laboratories, a project unlikely to gain approval. How then can one apply computer assisted methods based on QSAR to aid synthesis at any stage in either type of problem?

Inventive use of the transport, enzyme association and reaction model in conjunction with a large operational table of physically measured parameters provides a partial solution.

0-8412-0521-3/79/47-112-319\$05.50/0

© 1979 American Chemical Society

Master Data

The key to computer assisted synthesis based on these models lies in the size and format of the data base. In its current version, the Master Data parameter table lists 390 substituents in 780 lines. As shown in Figure 2, each substituent is entered on two lines. The odd lines contain the meta parameters; the even lines contain the para and aliphatic values. Following the case number and a simplified substituent name, the headings are: molecular weight of fragment, molar refraction, pi, pi squared, Hammett's sigma, Brown's sigma plus, Charton's sigma localized and Charton's epsilon values.

<u>CASE</u>	<u>NAME</u>	<u>MWF</u>	<u>MR</u>	<u>PI</u>	<u>PIQ</u>	<u>SIG</u>	<u>SIGP</u>	<u>SIGI</u>	<u>U</u>
17	CN	26.0	6.33	-0.57	0.32	0.56	0.56		
18	CN	26.0	6.33	-0.57	0.32	0.66	0.66	0.60	0.40
701	SOME	63.1	13.70	-1.58	2.50	0.52			
702	SOME	63.1	13.70	-1.58	2.50	0.49	0.39		

Figure 2. Master data

Originally designed as a data base for multiple regression, the main table has several sub-table routines for combining selected lines with kinetic, equilibrium or biodata. One of the routines converts biodata in ppm, a common industry form, to the molar equivalent, log MW/ED50. This is generated by introducing the parent MW into a program that uses MWF from Master Data.

The most important feature of Master Data is the simple data base format. This allows manipulation of the data by computer programs designed to assist synthesis. To facilitate use of these programs, MWF and each of the parameters are assigned a numerical code (1-8) as column identifiers.

In the following sections, the Master Data programs are described with relevant examples.

RANGE Program - Isolipophilic Groups

Many drug and pesticide series are either transport or binding dependent and exhibit optimum behavior in log P or in substituent π values. This is often apparent from the structures of commercial members of a class. Within a factor of two of the optimum (log 1/C-0.3), the width of most parabolic or bilinear plots is 1.5-3.0 log P units(1). Thus, if a class is log P or $\Sigma\pi$ dependant, its best members should cluster about

the optimum with the majority falling in the range, $\log P$ (avg) ± 1.0 . Figures 3 and 4 show four drug series that fit this criterion.

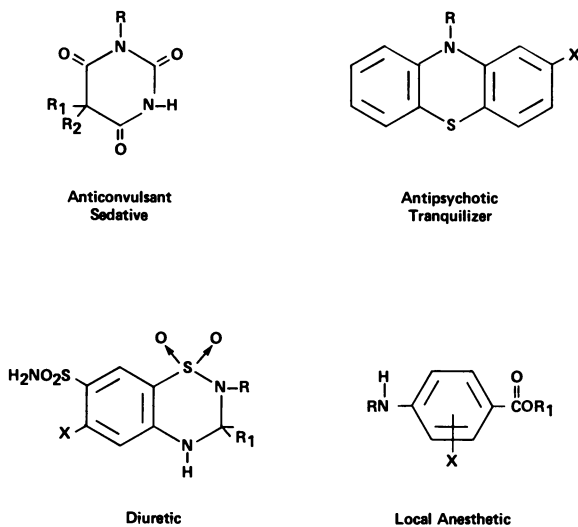


Figure 3. Some drug classes

PROBABLE TRANSPORT OR BINDING DEPENDANCE

Class	Avg. $\Sigma\pi$	Range (± 1.0)	No. in Range	n
Barbiturates	3.23	2.23–4.23	30	34
Antipsychotics	1.43	0.43–2.43	19	22
Diuretics	2.58	1.58–3.58	14	16
Local Anesthetics	2.32	1.32–3.32	15	18

Figure 4. Probable transport or binding dependance

Concentrating on one of these classes, we note in Figure 5 that 27 of 34 commercial barbiturates have ethyl or allyl as one of the gem-dialkyl groups(2). In designing a program to replace ethyl or allyl with novel groups, the most common cause of low activity could well be serious departure

from optimum log P. What is really needed then is a design program for novel isolipophilic groups. Such groups may still fail to excel but not because they possess inadequate π values.

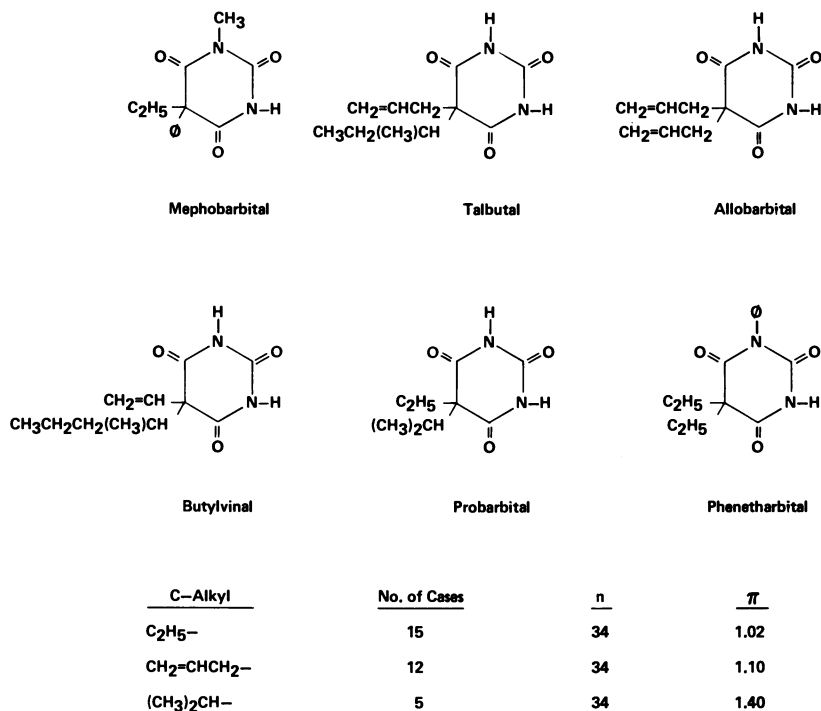


Figure 5. Various barbiturates

In a simplistic approach, Master Data can be rapidly searched by a PL/1 program called RANGE for groups having π values close to ethyl and allyl. RANGE is activated by selecting the parameter code for π ($=2$) and specifying the lower and upper limits of the search. The program segregates the desired data, ranks it in ascending order and prints out within seconds. Response on a CRT terminal is immediate. This is useful, but a more inventive approach uses group fragments called RANGE modifiers.

Shown in Figure 6 are a few of the groups used to modify any selected range of π values by combining these groups with those on the program print-out. In this procedure, we are taking advantage of an approximation, the near additivity of π values. Since these groups contribute to the total π value,

their values must be subtracted from the range being searched.

RANGE MODIFIERS

-S-	-CN	$\overset{\text{O}}{\parallel}\text{-CO-}$
-O-	-OCH ₃	$>\text{NCH}_3$
-CH=CH-	-SCN	-N(CH ₃) ₂
$>\text{C=O}$	-(CH ₂) _n -	-SCH ₃
-Cl, Br	-SO ₂ -	-SO ₂ CH ₃

Figure 6. Range modifiers

As shown in Figure 7, if three -CH₂- groups are added to substituents printed out by the RANGE program, then clearly this contribution must be subtracted to stay in the desired range.

MODIFIER CORRECTION

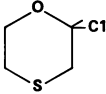

<u>MODIFIER</u>	<u>π</u>	<u>CORRECTION</u>
-(CH ₂) ₃ -	1.55	-1.55
-CH=CH-	0.82	-0.82
-OCH ₃	-0.02	+0.02
$>\text{C=O}$	-1.06	+1.06
-SO ₂ -	-2.14	+2.14

Figure 7. Modifier correction

In actual practice, the RANGE problem of generating groups isolipophilic with ethyl-allyl would be handled as follows. As both groups are close in π value, a single nominal range of 1.02-1.10 can be used. This range is arbitrarily extended to accommodate errors in π and to avoid missing adjacent groups of interest. An extension of 0.3 was selected for this problem, a value which includes two other small groups used in barbiturates, vinyl ($\pi = 0.82$) and isopropyl ($\pi = 1.40$). The extended range (0.72-1.40) is now broadened to cover all modifiers in a single run (search range = -0.83 to 3.54). As the print-out is ranked in π , it is simple to mark and label the top

and bottom of each modifier. Within these individual ranges, one looks for interesting groups to combine with the modifiers.

The near additivity of π values is an approximation, but the width of most log P curves is such that few outliers will be generated(1). It should be remembered that no attempt is being made to duplicate exact numbers but simply to fall within or near a selected range. Because of this scope and "margin for error", the combination process of modifier + print-out groups can be highly inventive, with cyclizations and isomerizations freely allowed. Figure 8 shows a few of the many groups generated as ethyl/allyl replacements. Note that part of the inventive process includes casting the isolipophilic group into a suitable intermediate for the desired reaction, alkylation of malonic ester in this case. Master Data and the RANGE program is merely an assistant or co-inventor at best. Human inventive skills are still required to complete the procedure. The potential for generating novel isolipophilic groups is nearly unlimited even though Master Data is presently far from complete in π values.

Modifier	Print-Out	Isolipophilic Reactant
-SCH ₃	-CH ₂ CH ₂ OCH ₃	
-C≡C-	-CH ₂ CH ₂ OCH ₃	CH ₃ OCH ₂ C≡CCH ₂ C1
-N(CH ₃) ₂	-CH ₂ CH=CH ₂	(CH ₃) ₂ NCH=CHCH ₂ C1
	-CH ₂ C(CH ₃) ₃	(CH ₃) ₃ C-C(=O)CH ₂ Br

Example: CH₃OCH₂C≡CCH₂C1 + \ominus C(COOC₂H₅)₂

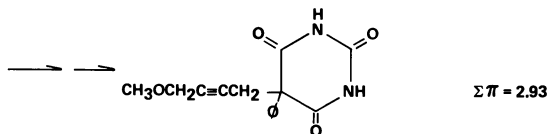


Figure 8. Groups isolipophilic with C₂H₅-C₃H₅

Figures 9 and 10 describe a similar pesticide example based on the Stauffer Chemical series of thiolcarbamate herbicides. The implied optimum in transport or binding is

clear from inspection of Figure 9, these four being the strongest of the series. The problem chosen was replacement of the RS-alkyl group with groups isolipophilic with ethyl/propyl. As seen in Figure 10, the range searched is nearly the same as that for ethyl/allyl. This is coincidental and implies no limitations on the procedure. Any group from most lipophilic to most hydrophilic can be processed. Figure 10 also contains an interesting example of the inventive process. In the third example, the valerolactone generated from the carboxyl modifier and the butyl group is difficult to prepare but the isomeric bromobutyrolactone is available from Aldrich Chemical.

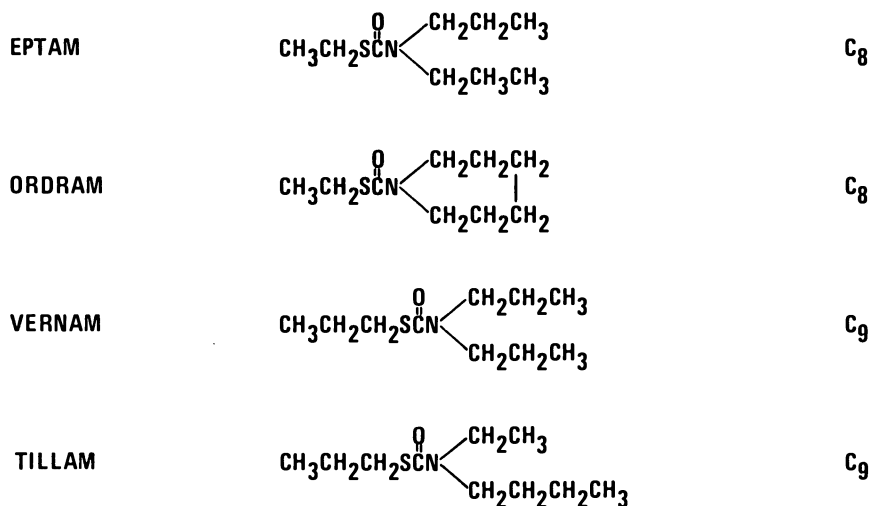


Figure 9. Thiocarbamate herbicides

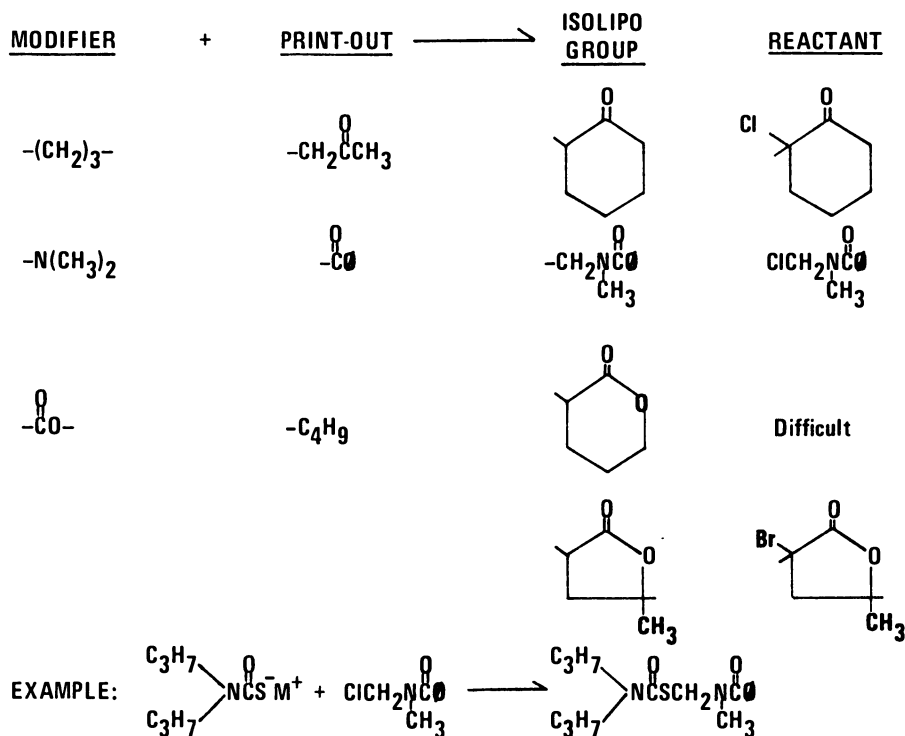


Figure 10. Groups isolipophilic with $C_2H_5-C_3H_7$ ($C_2H_5-C_3H_7$ range: nominal = 1.0 to 1.5; extended = 0.75 to 1.75; search = -0.80 to 3.89)

RANGE Program - Mechanistic Input

When some insight to the mechanism of inhibition is available, RANGE can provide a powerful assist to molecular design at the most fundamental level. This is best illustrated by considering an example. Figure 11 shows the structures of three fungicides, two from agriculture (one systemic, one contact), the other a human systemic. Although mechanistic details are not available, each system is designed to inhibit dehydrogenases bearing a cysteine unit at the active center. In the case of Griseofulvin, QSAR studies of analogs clearly indicate nucleophilic attack at the beta carbon of the cyclohexenone system(3).

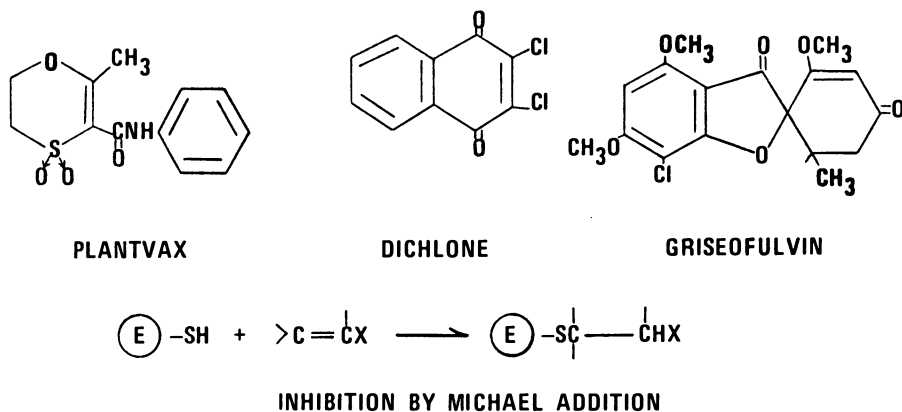


Figure 11. Fungicides inhibiting dehydrogenases

From a pattern recognition point of view, over 50% of commercial fungicides are, in fact, SH inhibitors⁽⁴⁾. These three systems represent a broad general class and have in common an inhibition step related to the Michael addition.

The RANGE program could be used to examine each fungicide individually for the purpose of improving each class. But a much more interesting approach is to treat the Michael addition itself. This can be done by considering the electronic range of those groups known to activate the double bond for nucleophilic addition. Five of the nine groups selected for this study are shown in Figure 12 and these include the two groups at each end of the range (-NO₂ and -CONH₂). Sigma para (σ_p) was selected to represent the electronic range of these groups because of its high resonance component. A much better choice would be one of Charton's sigma delocalized parameters

<u>X</u>	<u>σ_D</u>
-NO ₂	0.78
-SO ₂ CH ₃	0.72
-CN	0.66
-COOCH ₃	0.45
-CONH ₂	0.36
RANGE	0.36 - 0.78
SEARCH	0.3 - 0.9

Figure 12. Michael addition activators ($-\text{CH}=\text{CH}-\text{X}$)

(σ_D or σ_D^-). As shown in Figure 13, the range for these parameters is more tightly defined and descriptive of mechanism than σ_P . However, these parameters are not now included in Master Data. Fortunately, σ_P proved adequate for the study although part of the print-out was irrelevant to the Michael addition.

<u>X</u>	<u>σ_P</u>	<u>σ_P^-</u>	<u>σ_L</u>	<u>σ_D</u>	<u>σ_D^-</u>
-NO ₂	0.78	1.22	0.71	0.07	0.51
-SO ₂ CH ₃	0.72	1.01	0.61	0.11	0.40
-CN	0.66	0.92	0.60	0.06	0.32
-COOCH ₃	0.45	0.68	0.34	0.11	0.34
-CONH ₂	0.36	0.62	0.30	0.06	0.32
-CO \emptyset	0.43	0.88	0.35	0.08	0.53
-COCH ₃	0.50	0.81	0.31	0.19	0.50
RANGE	0.36-0.78	0.62-1.22	0.30-0.71	0.06-0.19	0.32-0.53
SPAN (Δ)	0.42	0.40	0.41	0.13	0.21

Figure 13. Michael addition activators ($-\text{CH}=\text{CH}-\text{X}$)

Considering the mechanism involved, it is somewhat surprising that neither σ_p^- nor σ_D^- were superior to the benzoic acid parameters in describing a tighter range.

The richness of the print-out is shown by the partial listing in Figure 14. Of course, many of these would be anticipated or considered simple extensions of common knowledge, but not all. There is a residual of unexpected suggestions that could easily form the basis for new experimental systems likely to have fungicidal activity. It is interesting, for example, that both 2- and 4-pyridyl appeared on the print-out, the latter group due to a prudent extension of the range. Both vinyl pyridines have long been known to undergo nucleophilic additions analogous to Michael addition⁽⁵⁾. It would appear from this study that these are, in fact, Michael additions. The double bond is identically activated because the pyridyl groups fall in the same sigma range as the more familiar Michael groups. The print-out also suggests that 2, 3, 4, 5, 6-pentafluorostyrene will exhibit Michael addition behavior. This is an interesting case of a false positive as pentafluorostyrene reacts exclusively with nucleophiles by replacement of the 4-F group⁽⁶⁾.

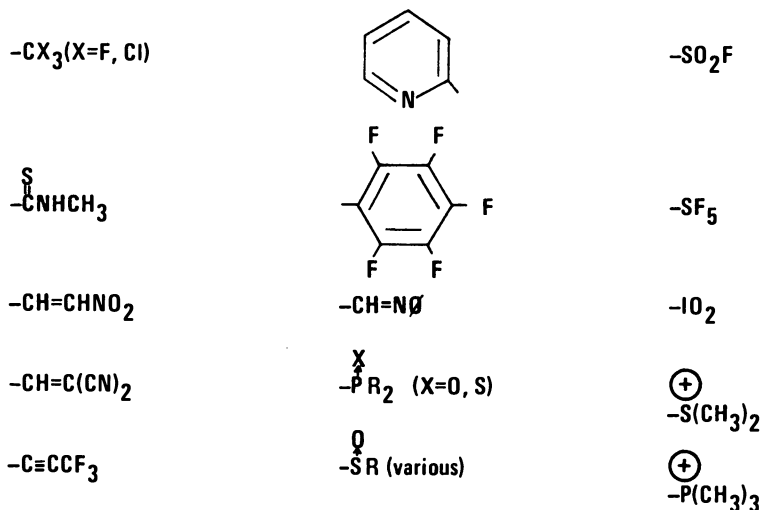


Figure 14. The michael printout

A general word of caution is necessary regarding all RANGE print-outs. There can be false positives for reasons that are not immediately obvious. For example, diethyl vinyl phosphonate does not appear to add thiols under very mild

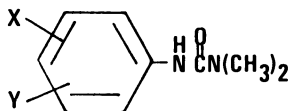
Michael conditions (e. g., triethylamine) that are normally rapid and exothermic with most of the standard Michael olefins. It does, however, react readily with stronger catalysis such as sodium ethoxide (7, 8). Clearly, there is a range of conditions as well as a range of electronic factors. In this case, factors other than electronic may operate to partially block the expected delocalization of the charged intermediate. Because unexpected factors may be present in some cases, it is wise to treat RANGE print-outs as ideas worth trying rather than established fact. In this sense, the programs act as good consultants by suggesting systems that have a high probability of success.

In addition to potential bioactive systems, RANGE can be used to generate new synthetic ideas in general organic chemistry. Every reaction has a scope in terms of certain boundary substituents that cause it to fail or take another course. If the mechanism is known, these groups can be parameterized by a related electronic factor. The derived range in this parameter is a quantitative expression of the reaction scope. Used as input for the RANGE program, the output can be expected to contain all the groups in Master Data that should favor the reaction. Many of these will be novel and may represent new directions in synthesis. We are currently parameterizing the scope of over 30 well-known reactions for RANGE studies.

RANGE2 - A Two-dimensional Program

RANGE can be extended to handle two parameters simultaneously, a program we call RANGE2. This program is activated by specifying the parameter codes with the lower and upper limits of each. As in RANGE, the program scans the data base and segregates values within the selected limits. However, only those groups having values within both ranges are printed out. Thus, whenever two major factors in a bioactive series can be identified, RANGE2 can provide new ideas that have a high probability of success. This is best illustrated by an example.

Regression work by Hansch on a series of urea herbicides inhibiting the Hill reaction indicates that in vitro binding depends on log P and inhibition on electronic factors [Figure 15(9)]. Later work by Giacobbe indicates similar factors in actual field testing of urea herbicides(10). Following this concept, Figure 16 shows some of the commercial and patented ureas used to establish a RANGE2 search based on π and σ . The nominal ranges were extended upward in both parameters but not downward as Fenuron ($\pi = 0.00$, $\sigma = 0.00$) is the weakest of the ureas listed. The output of RANGE2 contains only groups with π and σ parameters that are mutually in the same range as those in Figure 16 or within a reasonable extension thereof.



HILL REACTION INHIBITORS (SPINACH CHLOROPLASTS)

$$\log 1/C = 1.29 \log P + 0.544\sigma + 2.87$$

$$n = 12 \quad r = 0.944$$

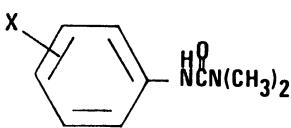
Figure 15. *Herbicalid ureas.*

<u>UREA</u>	<u>X,Y</u>	<u>Π</u>	<u>σ</u>
MONURON	4-Cl	0.71	0.23
DIURON	3,4-diCl	1.42	0.60
COTORAN	3-CF ₃	0.88	0.43
FENURON	H	0.00	0.00
PATENTED	4-SO ₂ CH ₃	0.27	0.70
PATENTED	4-SF ₅	1.23	0.68
RANGE	0.00 - 1.42	0.00 - 0.70	
SEARCH	0.00 - 2.00	0.00 - 0.80	

Figure 16. *Range2-herbicalid ureas*

The herbicalid urea class is very old and much of what was found proved to be patented. This was also true of the isolipophilic thiolcarbamate study (Figs. 9, 10) and clearly supports the procedures. The presence of known bioactives in a RANGE or RANGE2 output is an important indicator of correct input. Figure 17 contains some of the output that we considered to support the RANGE2 study of herbicalid ureas. Excepting the last example, those predicted to be active in both meta and para positions were patented for both. Those predicted to be active only in the meta position were patented only for this position. A close analog of the m-ethoxy example was prepared at Chevron Chemical and found to have excellent herbicalid activity. It is covered generically, but not specifically, by another company(11). The p-carboethoxy group

proved to be a false positive showing little or no activity. In this case, it is probable that soil metabolism is responsible for the degradation of this group within the lengthy test procedures. This is speculation, of course, as it is reasonable that the output will always be a mixture of hot and cold groups. The important point is that the output will contain many more hot groups than could be selected at random. Again, the program functions as a good consultant by suggesting structures that have a high probability of success.



<u>X</u>	<u>META</u>	<u>PARA</u>	<u>STATUS</u>
$\text{O}=\text{C}-$	✓	✓	PATENTED
$\text{O}-$	✓	-	PATENTED
$\text{C}_2\text{H}_5\text{O}-$	✓	-	ACTIVITY VERIFIED
$\text{CF}_3\text{C}(\text{O})\text{NH}-$	✓	-	PATENTED
$\text{CF}_3\text{S}-$	✓	✓	PATENTED
$\text{CF}_3\text{O}-$	✓	✓	PATENTED
$\text{C}_2\text{H}_5\text{OC}(\text{O})-$	✓	✓	PARA INACTIVE

Figure 17. Predicted herbicidal ureas

As a final example, it is instructive to note that RANGE2 carried out around Diuron alone ($\Sigma\pi + 0.4$, $\Sigma\sigma + 0.2$) yields only five groups on the print-out and yet, three of these groups represent patented herbicides. There can be no doubt that RANGE2 in π and σ is a correct procedure for this class.

Applications of RANGE2 as an aid to bioactive synthesis are nearly unlimited. The methodology was illustrated with a class of herbicides because we had some valuable follow-up experience to present. However, applications can be made to drug series with equal ease and reliability. The series shown in Figure 18 are typical of those that could be treated by a RANGE2 procedure. In each case there is prior knowledge of two major factors from the QSAR equations. Moreover, experience and analogy can often lead to an educated guess of the two most important factors thereby allowing one to use RANGE2 intuitively.

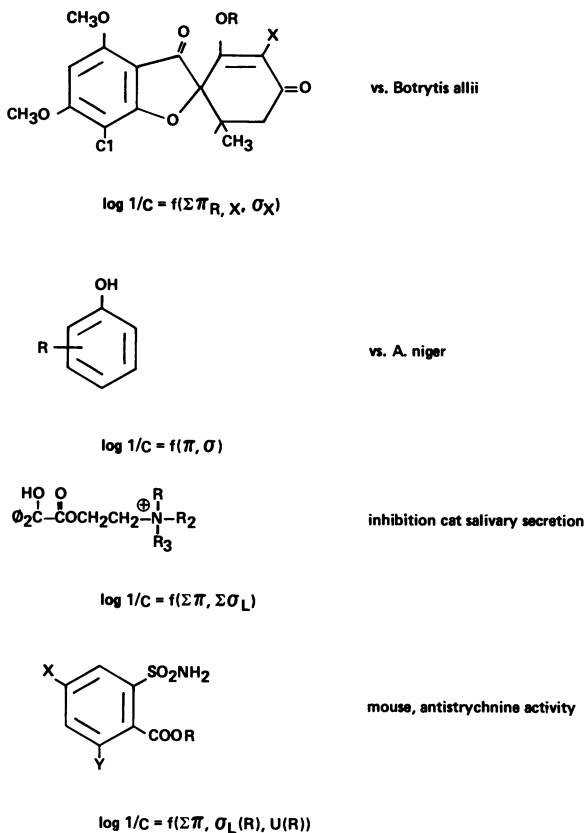


Figure 18. Drug structures—Range2

RANGE2 also has some valuable simplistic uses. For example, all electron-withdrawing lipophilic groups, all electron-donating hydrophilic groups or all highly polarizable groups smaller than *t*-butyl can be printed out or viewed on the CRT in seconds. This versatility can be a powerful aid to the inventive process. In a more specific way, RANGE2 can use any two properties of a group to find its nearest neighbor in these parameters. This is done by simply expanding a range around the parameters of the desired group, e. g., $\pi \pm 0.2$, $\sigma \pm 0.1$.

Figure 19 shows searches conducted around the para-isopropyl and iodo groups. As shown, these groups are best

simulated in π and σ by the meta groups, anilino and ethylthio, both non-intuitive results. The latter is quite interesting as an inexpensive substitute for iodo. This procedure effectively converts Master Data into a library of group similarities. Various combinations of MR, π , the σ 's and U result in a massive data base capable of assisting human judgement in different types of problems.

	<u>π</u>	<u>σ</u>
4-CH(CH ₃) ₂	1.40	-0.15
3-NH \emptyset	1.37	-0.12
4-I	1.12	0.18
3-SC ₂ H ₅	1.07	0.18

Figure 19. Range2-related groups

RANGEX - A Multi-dimensional Program

RANGEX extends the capabilities of RANGE2 by allowing selection of three or more parameters in the data search. Although as many as eight parameters can be called, the most useful version appears to be RANGE3. Most conceptual problems in drug or pesticide synthesis require consideration of three key parameters, π (transport/binding), σ 's (electronic) and U (steric). RANGE3 is in natural accord with this thought process. Other variations are possible, of course, and RANGEX can be used to locate nearest neighbors in any desired combination of properties. This can be useful near the beginning of a problem where only one active compound is in hand or in attempts to equal a desirable commercial compound by a novel substitution. Thus RANGEX can provide valuable direct assistance at a stage in synthesis before regression analysis is possible. Unlike RANGE2, no knowledge of mechanism is required since all likely factors can be included in the search.

Figure 20 shows the results of a RANGE3 search around the chloro and bromo groups. As an uncommon substituent, ethynyl is a most interesting replacement for chloro. It is more costly to use but has the possible advantage of being biodegradable. Perhaps this observation will stimulate more

frequent use of this group along with improved synthetic methods. The similarity of the difluoromethoxy group to bromo in both meta and para-positions is even more striking than shown by RANGE3; MWF, MR, σ_L and U are all remarkably close. These few examples were selected from many that we have run and give some idea of the encyclopedic ability of RANGEX to relate groups. It should be noted that it is just as easy to select ranges that produce dissimilar groups in any desired parameter dimension. This can be a powerful tool in developing the scope of a QSAR study, of much greater utility and speed than cluster analysis⁽¹²⁾.

	π	σ	U
4-C1	0.71	0.23	0.55
4-C \equiv CH	0.40	0.23	0.58
3-Br	0.86	0.39	0.65
3-OCHF ₂	0.90	0.31	0.66
4-Br	0.86	0.23	0.65
4-OCHF ₂	0.90	0.18	0.66

Figure 20. Rangex-related groups

A minor use of RANGEX proves to be convenient on occasion. When the print-out is long it can be nicely sorted by using MWF as an added dimension. Thus, our RANGE2 study of herbicidal ureas was actually run in RANGE3 (π , σ , MWF) using four MWF segments (1-80, 80-120, 120-180, 180-300). Sorting of this type simplifies the print-outs and focuses attention on preferred MWF ranges. In most bioactive problems there is some intuition concerning the size of substituents. Groups for new urea herbicides are of most interest with MWF <120, of lesser interest with MWF = 120-180 and of little interest at higher MWF. Reviewing these individual ranges is much easier than covering the entire study on one print-out.

EXBIOCOR - Extrapolated Biocorrelation Program

Completed or provisional QSAR equations are usually treated in a manner that we consider prejudicial. This consists of searching for interesting groups, assigning parameters and

calculating an estimated bioactivity. Most chemists will limit the search to common groups with which they feel comfortable, usually less than 50 groups.

EXBIOCOR is a non-prejudicial program operating through Master Data. It is activated by entering the coefficients and intercept of the regression equation along with the parameter codes. Also entered is a log 1/C cut-off value below which there is no interest. This value is usually chosen to be 0.5 log units (factor of 3) below the most active member of the regression series. The program does a linear transform for every substituent in Master Data having values in the equation parameters. All calculated log 1/C values are ranked and those exceeding the log cut-off value are printed out. Synthetic ideas are drawn directly from the listing. Although unusual groups may be suggested, our experience to date indicates that few series have been optimized, no matter how old the class. We see EXBIOCOR as a powerful aid to the final stages of bioactive synthesis.

Concern for extrapolation beyond spanned substituent space (SSS) has been logically discussed by Hansch(13). His points are well-taken and the possibility of failure due to unknown factors is very real. Extrapolation with EXBIOCOR of a valid QSAR equation(14) led directly to three inactive Parathion analogs that were predicted to have exceptional activity (m-phenoxy, p-benzoyl, p-benzenesulfonyl). Knowledge of the cavity size of AChE allows the reasonable conclusion that each of these substituents is too large to permit binding. Nevertheless, these represent failures to extrapolate and there will be many more as EXBIOCOR is extended beyond SSS. It is not without reason that we set the log 1/C cut-off well within SSS and one needs to be suspicious of predicted values more than 0.5 log units beyond the best existing compound. However, the excitement lies in synthesizing compounds predicted to be most active and not otherwise. Statistics are favorable for EXBIOCOR to prove its worth whether or not it accurately predicts bioactivity. We do not expect extrapolations of log 2.0 beyond SSS to be real, but if they prove to be even 0.3 beyond, the discovery will be significant.

Work in Progress

Master Data, now in its third edition, is approaching its final length though a few new substituents may be added each year. However, there is plenty of room for new parameters and we expect to add several new sigmas(15), Bondi's volume(16) and some of the Sterimol constants(17) in the next edition. Minor revisions of the RANGE and EXBIOCOR programs will allow for the additional parameters.

In another project, we are nearing the end of a two year study on estimating various parameters. Confidence in these methods comes from many earlier estimates of missing parameters in regression studies. No outliers have yet been generated by these estimates. While not of academic precision, we expect to fill out Master Data with estimated values suitable for approximate RANGE and EXBIOCOR solutions. There are no sharply defined boundaries or exact solutions in these programs. Thus, the inclusion of estimated parameters will give useful insights more often than not. As measured values appear in print, the estimated values will be replaced. We expect the use of estimated values nearly to double the output power of the Master Data programs.

Our most ambitious project is the generation of a multi-substitution parameter table. All meta and para combinations of Master Data would total about 100,000,000 and be highly populated with synthetically inaccessible examples. Such a data base would be too costly to store and impractical to run. Even very narrow ranges would have printouts with thousands of examples most of which are useless. In order to restrict the data base to synthetically possible combinations we will generate the data in a novel way. One of the sub-tables of Master Data is called Comsub Data and contains 50 common groups. Comsub Data will be used in conjunction with Master Data to generate the additive Multi Data table with the following restrictions. For disubstitution, one of the groups will come from Comsub Data, the other from Master Data. If we decide to include tri-substitution, two of the groups will come from Comsub Data. This insures synthetic feasibility for most of these combinations and reduces the data base from 100,000,000 to 100,000, a manageable size. To avoid excessively large print-outs, a query step will be entered in the program with the response showing the number of lines generated by RANGE or EXBIOCOR. At that point, the operator will have the option of narrowing the parameter ranges, raising the log cut-off point or accepting the output.

Despite some obvious limitations, we expect the programs in conjunction with Multi Data to have enormous power to suggest new synthesis. The results that we now obtain with Master Data in treating single substituents could be duplicated manually, though with great difficulty. With di- and tri-substitution from Multi Data, it is not humanly possible to achieve the output manually. Thus we arrive at a valuable area that only the computer can handle. We recognize that parameters of some adjacent groups will not be strictly additive and that the proposed use of estimated parameters will extend the error. However, the purpose of RANGE is to fall within or near boundary limits and that of EXBIOCOR is to exceed a log cut-off value. Granted, there will be more outliers and false positives in the Multi programs than in the simpler Master programs. However, we are

confident that the bulk of the output will meet the stated objectives. The potential is there to accomplish things with two or more substituents that are simply inaccessible with one.

Implementation of these improvements will be the subject of later reports.

Acknowledgements

I acknowledge with thanks considerable assistance from my colleagues, Dr. Edward I. Aoyagi and Dr. Donald E. Pack. Dr. Aoyagi helped in the difficult task of assembling all three editions of Master Data and with many of the concepts involved. Dr. Pack is responsible for the elegant formatting of Master Data and for all sub-routines and PL/1 programs.

Literature Cited

1. The mean width at (log 1/C-0.3) for 167 cases drawn from Hansch, C. and Clayton, J.M., *J. Pharm. Sci.* **62**, 1 (1973) is 2.62 log units with 65% falling in the 1.5-3.0 range.
2. Barbiturates have indeed been shown to have a log P optimum (mean log Po = 1.9 for hypnotic effect on various animals), Hansch, C. et al, *J. Med. Chem.* **11**, 1 (1967). Correcting for the barbituric acid group ($\pi = -1.35$) gives $\Sigma\pi = 3.25$ for the alkyl substituents, nearly identical to the simple average value of Figure 4.
3. Hansch, C. and Lien, E. J., *J. Med. Chem.* **14**, 653 (1971). See discussion of griseofulvin analogs, p. 667.
4. Magee, P.S., unpublished survey of 106 commercial fungicides. Of these, 55 were clearly E-SH inhibitors.
5. Klingsberg, E., "Pyridine and its Derivatives", *Interscience Publishers, Inc.*, New York; Part 2 (1961), pp. 209-210; Part 3 (1962), pp. 74-75, 355, 383; Part 4 (1964), pp. 158-159.
6. Burdon, J. and Westwood, W. T., *J. Chem. Soc.*, C (1970), 1271.
7. Pudovik, A.N. and Denisova, G.M., *Zhur. Obsh. Khim.* **23**, 263, 267 (1953); *Chem. Abst.* **48**, 2572h, 2573c (1954), and references cited therein.
8. Pudovik, A.N. and Zakharova, M.M., *Uchenye Zapiski Kazan, Univ.* **113**, 3 (1955); *Chem. Abst.* **51**, 5687i (1957).

9. Hansch, C., Prog. in Photosyn. Research, Vol. III, 1685 (1969).
10. Giacobbe, T.J. and Blau, G.E., 1974 Pacific Conference on Chemistry and Spectroscopy, 10th ACS Western Regional Meeting.
11. Martin H. et al (Ciba Ltd.), Ger. Pat. 1,905,598; Chem. Abst. 72, 20787b (1970).
12. Hansch, C. and Unger, S.H., J. Med. Chem. 16, 1217 (1973).
13. Hansch, C. in "Biological Activity and Chemical Structure", Ed. by J.A. Keverling Buisman, Elsevier, Amsterdam, 1977, pp. 47-61.
14. The equation is based on fly brain AChE inhibition by Paraoxon analogs: $pI_{50} = 0.433 \pi + 4.67 \sigma + 3.82$ ($n=9$, $r=0.968$), Fukuto, T.R. and Metcalf, R.L., J. Ag. Fd. Chem. 11, 930 (1956), data reworked by P.S. Magee.
15. The new sigmas under consideration are the resonance parameters, σ_D , σ_{D^-} , σ_{D^+} . These will be used in mechanistic RANGE studies.
16. Bondi, A., J. Phys. Chem. 68, 441 (1964).
17. Verloop, A., Hoogenstraaten, W. and Tipker, J. in E. J. Ariens, ed., "Drug Design", Vol. VII, Chapter 4, Academic Press, New York, 1976.

RECEIVED June 8, 1979.

Syntheses of Drugs Proposed by a Computer Program

MALCOLM BERSOHN

University of Toronto, Toronto, Canada M5S 1A1

The Computer and the Synthetic Organic Chemist

There are now three ways in which the organic chemist can use a computer to aid him in devising a synthesis. The most direct and simple mode is to query a data base about the existence of a reaction for transforming a given substructure into another given substructure. The system available commercially from Derwent Publications Ltd. (1) as well as those available inside pharmaceutical companies (2) are examples. By means of such systems, synthetic organic chemists are finally realizing the dream of keeping up to date on the best way to do reactions and the possibilities for transformations. The records in such on line data bases are complete with literature references, yields, solvent, temperature, catalyst and certain interfering groups.

The next mode of operation for the synthetic chemist using a computer is to use a program such as the one written by Wipke (3,4), which, given a product substance, generates the structures of reactants that might reasonably be expected to give rise to it. The ability of such a program to store these reactants and, on command from the user, to use them as products from which to generate reactants, means that the user can devise a synthesis while being on line to the computer.

The third mode of operation for the synthetic chemist using a computer is to use a noninteractive program to derive a synthesis for a given substance, e.g. a drug. The first such program (5,6) was operating early in 1970. There has been another attempt (7) but it appears that the program at the University of Toronto is the only such program still operating.

The chemist using the Toronto program specifies the maximum number of steps acceptable for a synthesis, the minimum overall yield, and, optionally, the starting material. The starting material can be suggested specifically, i.e. by providing the exact structure, or else more generally, by providing the desired starting skeleton, without restricting oneself to particular

0-8412-0521-3/79/47-112-341\$05.00/0

© 1979 American Chemical Society

functional groups. If the chemist does not specify a starting material, the program will decide the starting material for itself. The backward search from the goal molecule to an acceptably simple molecule is guided by the restriction that it must be compatible with the starting structure decided upon. At present, our definition of "compatibility" means that the skeleton of substructures which are present in both the starting material and the goal molecule may not be altered.

At input time, the user also defines to the program what he means by an acceptably simple starting material by providing the maximum number of chiral centers, functional groups and rings that an unknown starting molecule may have. The user may also specify the maximum number of consecutive facilitating reactions which he will accept in the synthesis.

Building Reactions and Facilitating Reactions

Somewhat arbitrarily we may divide all reactions into the two categories of building reactions and facilitating reactions. A building reaction joins two hitherto unattached carbon atoms or completes a heterocyclic ring not used as a protection for some functionality or introduces functionality at a previously unfunctionalized site. All other reactions are facilitating reactions.

The use of facilitating reactions may reflect the inherent limitations of known organic chemistry -- it is not possible, for example to join ethanol to propanol via a carbon carbon bond in a single reaction -- it may also reflect our ignorance or neglect of some efficient process in favor of some more familiar reaction which the molecule at hand does not yet have the desired functionality. Consequently, to ask the program to restrict the number of successive facilitating reactions that may be performed, is to raise the level of efficiency demanded of the synthesis solutions that are suggested. The ideal synthesis would of course consist only of building reactions.

Pruning the Combinatorial Thicket

Those familiar with the numbers involved might even doubt that a program could evergenerate a suitable answer to a nontrivial synthetic problem. The number of possible incomplete pathways is such that no program could consider a significant fraction of the pathways even if it devoted a year to the task. Suppose for example that the synthesis is to take ten steps and there are, on the average 40 conceivable reactions which could give rise to each molecule considered. Then the number of molecules in the problem is 10^{40} . Even with current fast hardware, such as the IBM 3033, the generation of one molecular structure takes a millisecond or so; we would need about 10^{32} years of computer

time to inspect the whole of the synthetic graph, thus discovering all the feasible ten step pathways. The thicket is impenetrable indeed, if we are to do it blindly. The chemist himself in his thinking suggests the solution. He does not attempt to consider all pathways; something familiar strikes his attention and he pursues a fairly straight path backward from the goal substance to a suitable starting material. The Toronto program attempts to do the same. A breadth-first approach, which generates all possible immediate predecessors of each molecule examined, is hopeless; there are too many possibilities.

The "something familiar" recognized by the chemist and by our program can be of two kinds: 1. A substance known to be available looks like a subset or a superset of the molecular structure of the goal molecule. Following this discovery the program will use the available substances as a required starting point and will only attempt to build the parts of the goal molecule that do not exist in the starting material. 2. A substructure of the goal molecule is identical to or derivable from the substructure produced by some key building reaction. If the exact product substructure is present the reaction will be simulated and the reactant(s) generated then present an easier problem of synthesis, closer to a satisfactory starting material. If the substructure present is derivable from the product substructure of the key building reaction then there must be recognition routines in the program which call in the deriving reactions. For example, a secondary alcohol next to a branched carbon atom suggests that the alcohol may have been a ketone which was alkylated. The reaction of reduction of the ketone is thus suggested. Currently a very major task in developing the program is filling the repertory with recognition routines which can "see" building reaction product substructures after one, two and three steps of alteration.

We can, evidently, distinguish three kinds of facilitating reactions. First, there are those concerned with protecting and deprotecting. Secondly, there are those which bring about unreversed changes in order to enable a reaction to take place. These changes are required in order to change a particular building reaction product substructure into the substructure at hand. A third category consists of the reactions which alter substructures so that the molecule at hand will be closer to the structure of a selected starting material. A sophisticated theory of synthesis optimization should be able to place separate limits on the numbers of these three kinds of facilitating reactions used in a synthesis. In any case a limit on the total number of facilitating reactions that can be employed in succession means an avoidance of high priced building reactions, ones that can only be performed with drastic amounts of protection and/or functionality alteration. The user of the program described here indicates at input time the limit on the number of sequential facilitating reactions which he wishes

to impose on any solution to the problem.

Examples of Output from the Program

I. Depoprovera

6- α -methyl-17- α - acetoxyprogesterone, as the only thoroughly tested and widely used injectable contraceptive, seemed to have enormous practical importance. Because of the close structural resemblance to progesterone, the latter was chosen as the recommended starting material. The program suggests an eight step method according to the accompanying scheme. I asked a class of eleven graduate students to produce a short synthesis of the goal substance, starting from progesterone. None of them were successful because the students did not know how to solve the problem of introducing the methyl group at C-6, i.e. the gamma alkylation problem. Evidently the performance of the program was superior to that of the graduate students for this problem but since the repertory is at this writing still only about three hundred reactions the program will fail to find syntheses for many substances with which the graduate students will have no trouble.

The most important reason for the superiority of the program in this case is that it was aware of the reaction developed by Yoshikoshi and his group, of formylation of a dienol ether(8). It appears that the total recall of all reactions in its repertory is responsible for much of the power of the program. A very simple case of this is the synthesis of 1,4-benzenediamine. The program, given this problem and told that the starting material must have no functional groups (other than a benzene ring which is not counted as a functional group in this context) produced an unexpected result. I expected the familiar six step route via the nitration of acetanilide. Instead the program recommended a two step route: oxidation of o-xylene to phthalic acid and then treatment of cupric phthalate with ammonia under pressure to give 1,4-benzenediamine. Obviously the replacement of aromatic carboxyl by hydrogen and the simultaneous amination of the ortho position is an unfamiliar reaction(9), which, however, is just as acceptable to the program as a familiar reaction. With a more ample repertory one may anticipate many such surprising but presumably efficient syntheses(Figure 1).

We observe parenthetically that to a first approximation no long synthesis proceeds as planned. One may ask what is the use of a computer program to generate synthetic plans when any plan from whatever source will have to be revised when exposed to laboratory practice? The answer is that it is more convenient to revise (or in the worst case abandon) an eight step synthesis than to do the same for a twelve step synthesis.

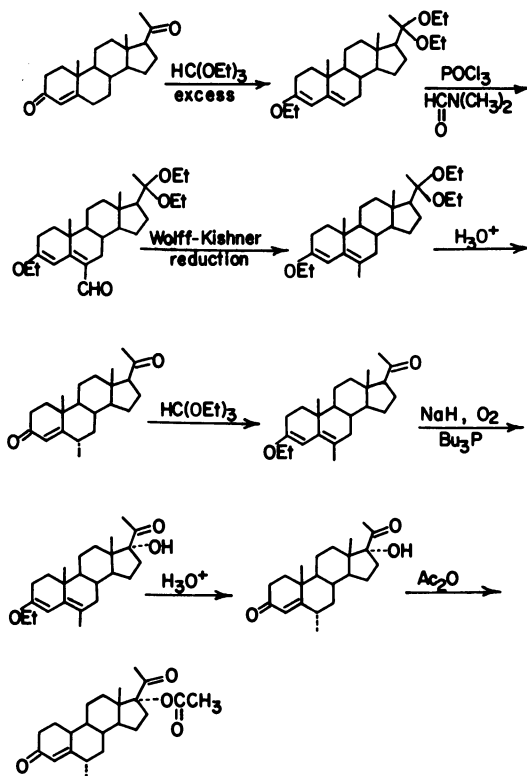


Figure 1. *Synthesis of Depoprovera proposed by the program*

II. PGF_{2α}

The prostaglandin structure is a ring with two side chains. A simple overall strategy is to take the ring and the side chains as already built and try to connect the side chains to the ring. A slight modification of this idea is to introduce a single functional carbon atom at the position of a chain and then to add the rest of the side chain to this functionality. In other words, the side chain minus the atom connecting to the ring may be taken as already synthesized.

This is the strategy successfully used by Stork and Isobe (10) in the briefest of all syntheses of PGF_{2α}. A most ingenious part is the stereoselective conjugate addition of the vinyl cuprate to the cyclopentenone followed by the trapping of the resultant anion with formaldehyde all in one step. The program, taught this unusual reaction, was able to generate the following seven step synthesis (Figure 2).

The synthesis recommended by the program differs in some small respects from the one actually achieved by Stork and Isobe. The most important point of difference is that lithium in ammonia is used to reduce the ketone at C-9. It may certainly be questioned as to whether the Li/NH₃ reduction would be as stereoselective as the method actually used by Stork, i.e. Brown's lithium trisiamylborohydride reagent (11).

A Complete List of the Tree-pruning Devices Used in the Program

1. Avoid duplicate molecules in the tree.

The program is unable to detect all duplicate molecules. This would be possible if we were prepared to store two hundred thousand structures onto a disk in a single run of the program. That would raise the price of a computation to a prohibitive level.

In the author's experience, the most common cause of duplicate generation of the same molecule is the occurrence of pairs of "commuting reactions." Given that A and B are synthetic reactions it frequently happens that the effect of A then B is the same as the effect of B then A. Such duplicate generations of the same molecule are avoided entirely by the following device in the program.

Assume that there exists a substructure X which is transformed into substructure X' by reaction A. Similarly there is a substructure Y which is transformed into substructure Y' by reaction B. We further assume that X and Y have no atoms in common in the molecules of interest. Suppose that a molecule contains both substructure X and Y and reactions A and B are used successively to produce a molecule with substructures X' and Y'. In many cases the order of performance of A and B is immaterial. In other words, neither substructure X nor substructure X' interfere with reaction B and neither Y nor Y'

preclude reaction A. A typical A and B pair are hydrolysis of an ester and hydrogenation of an isolated carbon-carbon double bond. In such a case, as we progress backward from the molecule with substructures X' and Y' to the molecule with substructures X and Y we can reach the latter along two paths, i.e. A then B or B then A. For brevity we can describe the situation as $XY \rightarrow X'Y \rightarrow X'Y'$ and $XY \rightarrow XY' \rightarrow X'Y'$. Before we call in reaction B on product XY' to generate reactant XY we should first inquire whether the pathway from XY to X'Y' through X'Y has previously been considered by the program. Since we do not retain any molecular structure not in the direct line of ascent to the goal molecule from the current molecule XY', this means that we cannot ask the simple question "Has X'Y been generated previously?" Instead we have to get at this information indirectly. We can ask whether the reaction B has been performed to produce the molecule X'Y'. For each molecule whose structure is still stored in the memory, a record is kept of all the reactions simulated which give rise to the molecule as a product. For each of these reactions the particular instance in the molecule of the substructure produced is also noted. Hence we can get a definite answer as to whether B has previously been used as a reaction in which X'Y' is the product. If it was not, we have no duplication problem here. If B was simulated to produce X'Y' then we have to ask whether A is interfered with by Y. If so, then there was no sequence AB. If not, then there must have been a sequence AB; consequently we should avoid reaction B preceding A. In real life, in the program, this means that we have to ask these questions before using any reaction except those that produce the goal molecule directly.

There exist of course, such sequences of "commuting" reactions as ABC, BCA and CAB. We have not programmed the elimination of these triple occurrences yet but are planning to do so.

2. Limitation on the number of steps.

At input time the user specifies the maximum number of steps for an acceptable synthesis. This cuts off the bottom of the tree (problem "trees" grow downward). This device is sine qua non but by itself it is still quite a weak constraint.

3. Minimum requirement for the overall yield.

This is also specified at input time by the user.

4. Minimum requirement for the yield of the last three steps.

As everyone knows who has done syntheses experimentally, the amount of material that has to be carried through the various steps is most important. If there are low yielding steps they should be near the beginning of the synthesis rather than near the end. This factor is not necessarily reflected in the overall yield. Consequently, in the program itself there is the

requirement that the overall yield of the last three steps must be at least 35%. This means that the middle of the tree is pruned rather drastically. The user does not specify the 35% parameter. Obviously it could be altered in minutes but most users have no objection to this value so we do not burden the user with this decision.

5. Limitation on the use of facilitating reactions.

Building reactions can be used wherever appropriate but facilitating reactions can only be used if they were called for specifically. The "call" can come from the failure of a building reaction to proceed on account of interference with the reaction for some functional group in the molecule. In such a case the group must be protected and, since we are proceeding backwards from the goal molecule, the reaction to remove protection from this functional group is called in. If the number of different kinds of groups to be protected exceeds the limit of the successive number of the facilitating reactions imposed by the user then none of the protections are effected.

Sometimes facilitating reactions are called in by reaction product derivation routines. An example is the change of functionality required to obtain a six membered ring at hand from one which is the product of a plausible Diels-Adler reaction.

This single heuristic has resulted in a drastic reduction of the number of reactions simulated by the program in investigating a problem. For example, the Depo-provera synthesis was found after simulating only 280 reactions and the total number of reactions simulated was only 1385. This was obtained using input requirements of 8 for the upper limit on the number of steps, 3 for the upper limit on the number of consecutive facilitating reactions and 5% for the lower bound on the overall yield of the synthesis. Before introducing this heuristic we had been used to generating 20,000 to 200,000 molecular structures in a run.

The exact number of reactants generated by the program in investigating a synthetic problem of any complexity will vary with almost any slight change in the program. It is increased by adding reactions to the program and of course it is decreased by refining the scope and limitations of any reaction in the program.

6. Requirement for compatibility with a starting structure.

This is the requirement that the skeleton of substructures which are present in both the starting material and the goal molecule may not be altered. This pruning device is optional. Ordinarily it is not used when the problem thinks backward from the goal molecule without any particular direction in mind. Any substance which satisfies the input restrictions on the number of chiral centers, functional groups and rings is considered to

be available and the program declares a synthesis to be found if it reaches an available substance in its backward search.

For complex structures, this is an extremely important tree pruning method. When we are thinking in the forward direction about a synthesis the requirement of compatibility with a starting structure is automatically satisfied. As soon as we start to think backward from the goal molecule, the question of the description of the starting material arises. At first thought it seems far more elegant to ignore this question; we simply work backward and as soon as we reach an acceptably simple molecule or an available molecule, we have completed a synthesis plan. However, this often results in unintelligent behavior. To see this, let us consider a string A-B-C-D, where A,B,C and D are structural fragments. Suppose that on the list of available substances is A-B and C-D. Now the program "knows" that A-B and C-D are available but it accesses this knowledge only under a single fixed circumstance, i.e. a new reactant has been generated in the course of simulating a reaction and the program compares the structure of this reactant with those on its list of available substances. But the program does not keep its attention on the availability of A-B and C-D when it is selecting a reaction to simulate. For example, it may generate A-B-C from A-B-C-D. Then it may generate A-B from A-B-C and thus a two step process. But in fact the one step process of joining A-B to C-D may exist and the program will not immediately find it. If the planner in the program decides that the synthesis must proceed from A-B and C-D then the only problem remaining is to see if the bond joining B to C in the goal molecule can be made. If so, then the one step synthesis is immediately uncovered. We encountered analogous advantages of deciding on a starting skeleton in more complex circumstances. If we are seeking a total synthesis of a steroid, for example, and we have the B/C ring system already formed in the starting material, then the problems to be solved are the addition of the A and D rings, not the formation of the B and C rings. However, when thinking backward from the goal molecule, without a definite starting skeleton in view, a program can easily lose its way, now adding pieces of the A ring, then adding pieces of the C ring, etc. There tends to emerge, after much searching and many steps down the synthetic tree, a highly branched structure that has many chiral centers and is as difficult to make as a steroid. Such was the usual experience on asking the program to synthesize a steroid. With a suitable upper limit on the number of chiral centers required for a starting material that is not on the available substance list, the program does not actually propose such a synthesis, but investigation showed that these were the lines along which much of the usually fruitless search was proceeding. To avoid this there was incorporated into the program a planning routine which at the outset of the problem solving process decides on a starting skeleton. When the planner

is given control, no reaction simulated may build this skeleton.

The planning routine uses the starting skeleton for a given number of simulated reactions, e.g. 32,000 and then chooses another starting skeleton and so on until an acceptable synthesis is found or the time allowed for the job is exhausted. The first choice for the starting skeleton is that input by the user. The second choice is that imposed by a required starting substance. The third choice consists of a particular ring of the structure of the goal molecule. Which ring should the starting material contain? The algorithmic answer to this question ultimately must be provided by many experienced synthetic organic chemists. Corey(12) provided a first answer to this question for the case of bridged ring systems. However, subsequently Johnson's very efficient synthesis of longifolene(13) which made two rings at a time, violated Corey's preliminary rules by making a bond that was not "strategic" at the crucial step. Evidently the theory needs refinement. Particularly, there is no theory about an unbridged system. For example, we cannot say whether the most convergent synthesis of a steroid starts with the A and B rings or the B and C rings, etc. Presumably the answer depends on what functionality is around the skeleton, which methyl groups and C-17 substituents are present, etc. It may even depend on what reactions have been invented, e.g. microbial oxidation at C-11: most practicing chemists are highly likely to conclude that general principles are not discoverable because they do not exist. The planning routine described here takes for the time being an ad hoc approach of trying everything, i.e. each ring is tried separately as a starting skeleton and then all possible pairs of rings are used. As the general principles emerge they will be included, to replace the ad hoc approach.

7. Requirement for success ively better syntheses.

After finding an n step synthesis the program changes its requirements so that any further solutions can be no longer than n-1 steps and must have a higher overall yield than that already achieved in the n step synthesis.

Acknowledgement: The author gratefully acknowledges financial support for this work from the National Science and Engineering Council of Canada.

ABSTRACT

A report is given about a program that proposes many-step syntheses for a given organic compound, without interacting with the user after the initial description of the problem. The problem description includes the structure of the desired goal molecule, complete with chirality, the maximum number of steps for an acceptable synthesis and the minimum overall yield required. Optionally, the problem description may also include

the specification of acceptable starting materials. These may be specifically stated, or generally indicated, i.e. with a given skeleton. The maximum number of rings, functional groups and chiral centers for an acceptably simple starting material may also be given. An eight step synthesis of the contraceptive Depo-Provera proposed by the program, which starts from progesterone, is presented. A short prostaglandin synthesis, similar to one reported in the literature, was also proposed by the program.

Various ways of managing the combinatorial explosion problem were described. Basically we want to maximize the number of building reactions and minimize the number of facilitating reactions. The facilitating reactions should not be suggested unless they actually facilitate a specific building reaction. There must also be an upper limit on the number of sequential facilitating reactions that may be performed without doing any building reactions.

Literature Cited

1. Finch, A.F., to be published in J. Chem. Inf. & Comp. Sci., (1979)
2. Fugmann, R., "The IDC System" in "Chemical Information Systems", Ash, J.E. and Hyde, E. editors, John Wiley & Sons, New York, (1975), pp 195 et seq.
3. Corey, E.J. and Wipke, W.T., Science, (1969), 166, 1978
4. Wipke, W.T., Dolata, D., Huber, M. and Buse, C., in this symposium volume, and references therein
5. Bersohn, M., Bull. Chem. Soc. Japan, (1972), 45, 1897
6. Bersohn, M., Esack, A. and Luchini, J., Computers & Chemistry, (1978), 2, 105
7. Gelernter, H.L., Sanders, A.F., Larsen, C.L., Agarwal, K.K., Boivie, R.H., Spritzer, G.A., Searleman, J.E., Science, (1977), 197, 1041
8. Kato, M., Kurihara, H., Kosugi, H., Watanabe, M., Asuka, S. and Yoshikoshi, A., J. Chem. Soc. Perkin I, (1977), 2433
9. Arzoumanidis, G.G. and Rauch, F.C., Chem. Commun., (1973), 666
10. Stork, G. and Isobe, M., JACS, (1975), 97, 6260
11. Brown, H.C. and Krishnamurthy, S., JACS, (1972), 94, 7159
12. Corey, E.J., Q. Rev. Chem. Soc., (1971), 25, 455
13. Volkmann, R.A., Andrews, G.C. and Johnson, W.S., JACS, (1975), 97, 4777

RECEIVED June 8, 1979.

CAMSEQ/M: A Microprocessor-Based Conformational Analysis System

HERSCHEL J. R. WEINTRAUB

Department of Medicinal Chemistry and Pharmacognosy, School of Pharmacy and Pharmacal Sciences, Purdue University, West Lafayette, IN 47907

Conformational analysis is becoming a widely utilized tool in drug design, molecular modeling, and the determination of structure-activity relationships. Of the many techniques presently available for conformational studies, classical, empirical potential energy functions hold great promise in providing relatively inexpensive explorations of conformational hyperspace in various simulated solvent environments. An excellent example of the application of these classical techniques is embodied in the CAMSEQ Software System (1,2,3).

Current computer technology has provided us with very powerful "stand alone", desk top microcomputers at very low cost. The availability of such a system which also provided interactive graphics capability prompted the development of "CAMSEQ/M".

CAMSEQ employs classical-type empirically fitted potential energy functions which have been shown to accurately reproduce a large variety of experimental data (3-10). The conformational energy is determined using sets of pairwise potential energy functions, harmonic torsional correction terms, hydrogen-bonding terms, solvation and other free energy contributions. These terms have been described in detail elsewhere, but will be briefly reviewed at this time (3,11). Specific technical details of certain CAMSEQ and CAMSEQ/M implementations of these potential functions are included in this discussion.

Steric, non-bonded, pairwise interactions are approximated by sets of Lennard-Jones 6-12 potential functions (3,11). Built-in parameters account for the interactions indicated in Table I. The user is free, however, to substitute particular sets of parameters with his own values (at run time) and even to change the nature of the function. For example, often a crystallographer feels more comfortable with a 6-9 type function rather than the built-in 6-12; this change is very easily accommodated by simply specifying the "powers" to be used in the calculation. In this way, the system is very adaptable to the user's needs and can also be used for the development of new parameters.

0-8412-0521-3/79/47-112-353\$05.00/0

© 1979 American Chemical Society

Table I
Atom Types "Built-in" to the CAMSEQ/M Potential Functions

C (sp ²)	C (sp ³)
Cl	F
H (aliphatic)	H (aromatic)
N (sp ²)	N (sp ³)
O (sp ²)	O (sp ³)
P	S

Typically, electrostatic interactions are represented by some form of Coulomb's law-type function. In this type of calculation, the atoms in the molecule are assigned a charge distribution consisting of discrete point charge approximations centered on each atom. Other variations of these simple functions include exponential-type approximations (12) and variable dielectric approximations (3,11). The point charge approximation is often determined by *ab-initio* or semi-empirical molecular orbital calculations. The user may select the particular functional representation used in the calculation as well as the parametric form of the function. The variable dielectric function is represented by an approximation to the rigorously derived function (3,11) for computational efficiency.

Since classical-type calculations cannot intrinsically reproduce such effects as atomic orbital interactions, it is necessary to adjust the conformational potential energy terms to accurately reproduce experimentally observed torsional barrier heights. Harmonic torsional correction functions are incorporated in the repertoire of classical potential energy terms in order to economically satisfy this requirement. These functions may be described in several ways. Sets of parameters may be given to fit a general purpose torsional correction term which includes provisions for describing *cis/trans* barriers. The general form of this equation is

$$E(\text{tor}) = \text{Barrier1} \cdot [A + B \cdot \text{COS}^N(C \cdot \chi + \phi_1)] + \text{Barrier2} \cdot [1 - \text{COS}(D \cdot \chi + \phi_2)]$$

This function has nine adjustable parameters and is quite effective in describing most torsional correction terms.

Alternatively, a set of discrete values may be defined which are then used instead of a harmonic function. Linear interpolation is then used as needed to describe any missing intermediate values. Similarly, a two-dimensional torsional correction term may be used when the one-dimensional function is inadequate. An example of such a case is the glycosidic linkage in disaccharides (13).

A phenomenon which must be accounted for in many conformational problems is that of the hydrogen-bonding effect. The general steric, non-bonded interaction potential functions and the electrostatic potential approximations do not, intrinsically, correctly account for the special properties encountered in

hydrogen-bonded systems. For such cases, CAMSEQ/M utilizes a special, empirically-fitted, hydrogen-bonding potential function (14). This function was developed to accurately reproduce experimentally observed hydrogen-bonding energies for many systems. An angular dependence term accounts for geometrical deviations from the "ideal", while another empirically-fitted function yields the observed hydrogen-bonding energies. The product of these two functions yields reasonable hydrogen bonding energies for a large variety of molecular systems.

Solvation effects are an important aspect of conformational studies involving biologically important molecules. CAMSEQ/M utilizes a modified hydration shell model to account for these environmental effects (3,9,11,15). Several simulated solvent environments may be calculated. These include aqueous, 1-octanol (15), methanol, ethanol, acetic and formic acid (3,9,11). Using this method, the free energy of solvation of the molecular species may be estimated. Because of the statistical nature of the hydration shell model, the dynamics of the solvation effect are accounted for (9).

By using classical-type empirical potential energy functions, accurate conformational calculations which account for experimentally observed phenomena can be performed. Geometry optimization is employed to improve the efficiency of the calculations. In this way, only those atoms whose positions change as a function of conformation are recalculated.

A multi-level (major-minor) bond rotation algorithm has been developed for CAMSEQ/M and is currently being implemented in CAMSEQ. Using this scheme, two levels of importance are assigned to the torsional bond rotations. These may be simply described as "major" rotations, or those involving backbone or other major bonds, and "minor" rotations involving methyl groups, amine groups or other sidechain rotations.

The algorithm was designed to work in either sequential scan, random scan, or multidimensional minimization modes. The technique may be closely compared to a quantitative model building exercise. In molecular model building, the "major" bonds are adjusted to their approximate desired positions, and the minor groups are then "fine-tuned" to positions which minimize unfavorable close contacts. Other conformations are examined until the desired structure is constructed. In a similar manner, CAMSEQ/M adjusts the "major" torsion angles according to the selected search mode (prototype versions are limited to sequential scan modes, but random and multi-dimensional minimization procedures are currently being implemented) and follows this by a limited search for low energy conformers for each of the "minor" torsion angles. Upon setting the latter bonds to their lowest energy positions, the total conformational energy is determined. While this procedure increases the calculation time per conformational state, in general, the total search of conformational hyperspace is considerably more rapid and economical, as fewer conformational

states are (fully) examined.

The savings achieved by this method result from the techniques employed. The calculation of the geometries and energies associated with rotations about the "minor" bonds are very rapid. These calculations involve only the atoms which have moved as a result of the "minor" rotations. The pairwise energy terms involve only order N calculations (where N is the number of atoms in the molecule) as compared to order N^2 for the calculation of one conformational state. Thus the "minor" rotation calculations serve only to describe the optimum positions of the respective atoms.

Since the efficiency of multidimensional minimization schemes is dependent on the number of degrees of rotational freedom, the separation of "major" from "minor" torsional rotations should dramatically enhance the speed of convergence of these routines. Of course, "normal" rotational methods may be applied simply by specifying no "minor" or secondary rotation parameters.

The preceding discussion has briefly described the potential functions and geometry determination algorithms employed in CAMSEQ and particularly in CAMSEQ/M. The "batch" version of CAMSEQ currently in use is totally compatible with CAMSEQ/M. CAMSEQ/M can be interfaced (via a communications link) to a host computer for more lengthy calculations where speed is a requirement (rather than economy).

CAMSEQ/M has been developed as an alternative means for the determination of the conformational potential energy surfaces of molecules. The study of biologically important molecules led to the original development of CAMSEQ (1,3-10). More recent developments have led to an interactive version modified by Hopfinger and Potenzzone (2) for the NIH-EPA-Chemical Information System (CIS) (16). This version (designated as CAMSEQ-II) reflected more current state-of-the-art computer technology and requires the use of a moderately large mini-computer. Some of its features include the availability of large data bases containing many thousands of molecular structures, and the convenience of "conversational mode" input sequences to aid the user in data preparation.

The availability of microprocessor-based interactive computer graphics devices and rapid access disk storage has provided the CAMSEQ/M system with many of the features of the NIH-EPA-CIS (on a smaller scale, of course) and in addition, provides "no-cost", local computational facilities.

In practice, it is rarely necessary to maintain hundreds of diverse molecular structures for immediate access. Conformational studies are generally performed on a small congeneric series of molecules which, of course, should be immediately available. CAMSEQ/M provides for the local storage of several hundred structures. If additional structures are required, the NIH-EPA-CIS can be linked to the local terminal via telephone

communications and the necessary structures located and "sent" to the terminal. These files of new structures may then be saved for future use. Therefore, CAMSEQ/M is designed to supplement the NIH-EPA-CIS by providing local processing power, relieving the latter system for data base manipulation rather than time-consuming conformational energy calculations.

CAMSEQ/M was developed around a Tektronix 4051 Microcomputer Graphics System. The hardware required includes the basic 4051 microcomputer with a minimum of 16k bytes of memory (32k is recommended), a "joystick" graphical input device, and a matrix function package (in a read-only memory firmware pack). A vastly more versatile system requires the addition of a file manager/disk system. In order to communicate with the NIH-EPA-CIS or another "host" computer, a communications interface is necessary. The total hardware cost is approximately \$17,000. Table II outlines the required hardware. The 4051 utilizes a direct view storage display which does not employ a selective erase feature. Therefore, it is necessary to replot the screen frequently in order to remove unwanted information. This handicap is the price one must pay for low cost, but quite sophisticated computer graphics features, and does not pose any major problems in CAMSEQ/M.

Table II
Hardware Item

Tektronix 4051 Microcomputer w/16k memory
Additional 16K memory
Matrix function ROM pack
Joystick input device
File manager/disk system
Communications interface

Upon turning on the system and initiating CAMSEQ/M, the user is shown a "menu" from which to select the particular task to be performed. Figure 1 illustrates the screen display after the "crystallographic coordinate" input feature has been selected. As can be seen, the program prompts the user for the various unit cell parameters. At nearly any point, a "question mark" may be typed, and CAMSEQ/M will highlight the appropriate information on the screen to assist the user in determining what he should do.

CAMSEQ/M provides the user with the capability to input complete molecular structures from the internal (disk) data base, to input cartesian or crystallographic coordinates, or to use a "joystick" controlled model-building system. In addition, a molecule may be constructed from one or more substructures with substituents attached using the joystick model-builder routines. Data input is therefore quite flexible, and the user is guided through every step by the program.

Continuing with the example depicted in Figure 1, the user

CAMSEQ/M MICROPROCESSOR-BASED CONFORMATIONAL ANALYSIS SYSTEM DO YOU HAVE A JOYSTICK ? (Y/N)		RESTART 9
		INPUT COORDS 8
	+	CARTESIAN 7
	+	X-RAY (FRACT) 6
		GRAPHICS MODEL BLD 5
		SAVE ON SUB STR FILE 4
		SAVE ON DATA FILE 3
		DISPLAY MOLECULE 2
		LIST DATA 1
		SELECT OPTIONS 0

SELECT CARTESIAN OR XRAY (FRACTIONAL) COORDINATES	
ENTER UNIT CELL PARAMETERS:	
A:	ALPHA:
B:	BETA:
C:	GAMMA:

Figure 1. This "menu" is provided to allow the user to select the appropriate work to be done by CAMSEQ/M; in this example, crystallographic coordinate input has been selected.

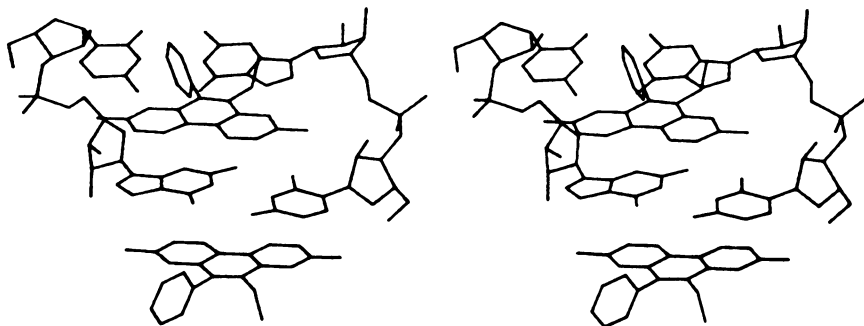


Figure 2. Stereoscopic illustration of the ethidium/dinucleoside monophosphate (CpG) crystalline complex (18) in the stick model mode. There are 130 atoms in this structure.

is prompted to supply the appropriate coordinates, and the atom type (C, O, N, H, etc.). The program determines the exact atom type (sp³, etc.) by analysis of the molecular connectivity data determined in a subsequent step. Following the input of co-ordinates, this connectivity table is calculated (by locating bonds between atoms which are less than 1.6Å apart. A sketch of the molecule can then be produced to verify the data input. Bonds may be added or deleted while viewing this display. The pictorial view thus provided is more convenient for locating incorrect data (coordinates, bonds) than a table of similar information. However, tables of all topological and topographical information are also available. To rotate the molecule, or produce a stereoscopic view, one simply requests a "stereo" display rather than the "quick" display described above. Various molecular parameters including bond lengths, bond angles, and torsion angles may be determined while viewing this display. Publication-quality stick models or "disk and stick" models may be produced as indicated in Figure 2.

If the molecular coordinates are not available, they may be obtained from an external data base such as the NIH-EPA-CIS. A local copy may be made and used for subsequent input. Data stored locally may be treated as entire molecules or as substructures. An index of available structures may be displayed to assist in selecting the appropriate molecule. The attachment of substituents to substructures or of substructures to one another is accomplished using the joystick model builder. Therefore, a discussion of the model builder routines is in order at this time.

Figure 3 shows the menu obtained when requesting the graphics model builder option. A list of 9 "frequently used" substructures is provided to speed the selection of appropriate subunits. Also shown in Figure 3 is the "rough" joystick-assisted construction of a simple, sample compound, phenethylamine. The sequence of steps required to construct this molecule is as follows:

- a. The benzene substructure is selected from the list (by "pointing" to it with the joystick cursor) and placed in the desired location on the screen. The grid is provided for convenience in estimating distances. This is only a "ghost" structure, however, and becomes part of the molecule only after step "c", below.
- b. The "add atom" feature is selected and the number of atoms is defined.
- c. The joystick cursor is moved about the screen and the desired atom letter (C, O, N, H, etc.) is typed. If the cursor is placed near a substructure atom (within 0.2Å or grid units) and a "space" is typed, the substructure atom is added to the molecule. Thus, only the desired portion of the substructure need be incorporated into the molecule.

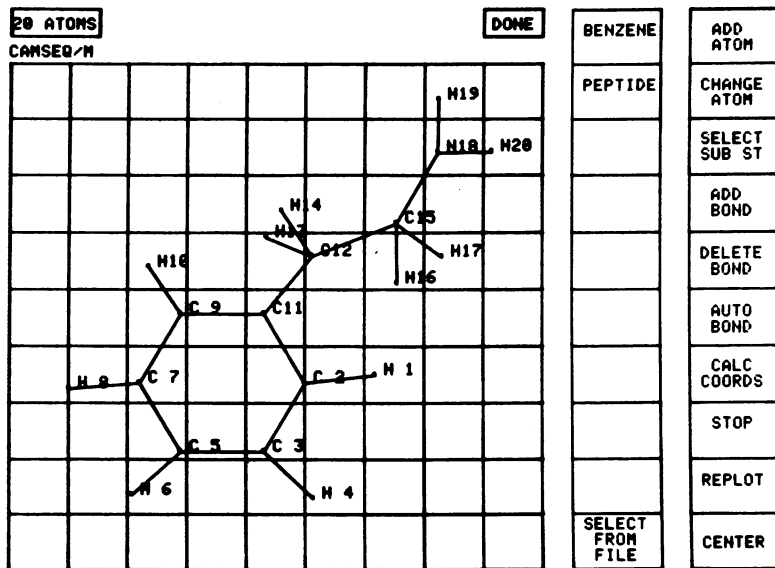


Figure 3. Graphics Model Builder menu. Also depicted is the "rough" drawing of a molecule using the joystick cursor device.

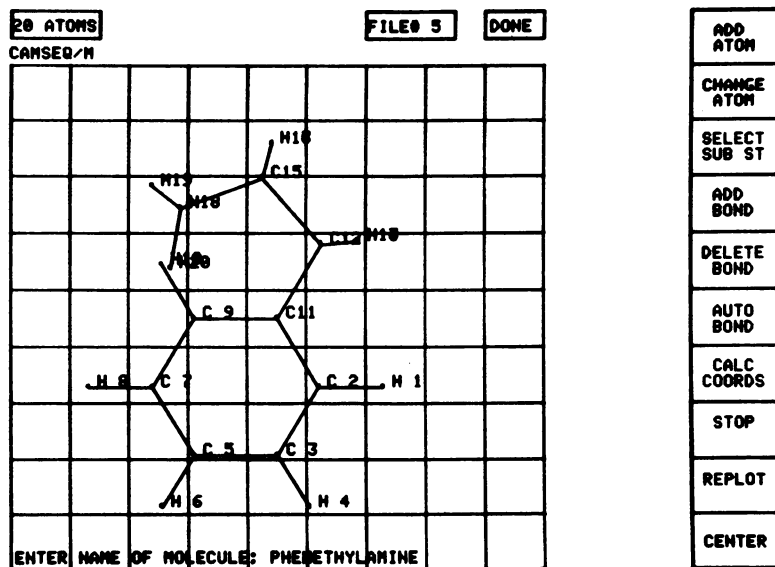


Figure 4. Final structure of phenethylamine generated by CAMSEQ/M from the approximate drawing in Figure 3

- d. After all atoms have been approximately located, bonds are added, either explicitly or implicitly by specifying the "add bonds" or "auto bond" options, respectively. The implicit bonds are determined by calculating the interatomic distances as drawn on the screen, and bonding all atoms within 1.6 Å of one another. A "delete bond" feature allows one to remove extra, unwanted bonds. The "add bond" feature, allows the bonds to be assigned explicitly by indicating the atom pairs to be bonded and depressing a key on the keyboard. The bonding operations automatically determine the connectivity data.
- e. Exact, "standard" bond lengths and bond angles (17) are then used to transform the approximate structure displayed on the screen into an accurate molecular structure. The bond angles and distances to be used in the calculation are displayed and may be modified if desired. This procedure is accomplished by specifying "calculate coordinates" from the menu.

Figure 4 illustrates the "standard" model generated by CAMSEQ/M. This structure may be saved for later use, and/or be used as an input structure for a conformational study. The structure may later be used as a substructure for studying other congeners, by modifying its substituents. For example, amphetamine may easily be constructed by adding the alpha-methyl group to the ethylamine sidechain. The stereochemistry of the molecule is determined by the order of input of the substituents.

Several substructures may be connected together to form more complex molecules. As an example, a polycyclic aromatic hydrocarbon consisting of "standard" bond lengths and angles may be constructed by connecting several benzene rings together. The sequence of operations is as follows:

- a. Select the benzene substructure from the list and place it in position on the screen (as described above for the general graphics model builder example).
- b. Enter the total number of atoms in the complete structure by using the "add atom" feature.
- c. Attach appropriate substituents to the first ring, and assign bonds using either the "add bond" or "auto bond" menu items. Next, select another substructure (by "pointing" to "select substructure" on the menu). Using the joystick, "point" to the "common atoms" of the two structures and CAMSEQ/M will attach the structures to one another.
- d. Repeat this sequence of operations (c) until the entire structure is generated. If desired, all substructures may be connected initially, and the substituents added last. The resultant structure may be saved as either a substructure or a molecule for later use.

In this example, it would be more convenient to maintain the substructure in the substructure file, and to generate smaller bi- or multi-cyclic ring systems by using portions of the larger molecule.

Figure 5 illustrates the generation of the diphenyl methane structure from two benzene rings and other atoms. The first step involves the selection of the first benzene ring, the assignment of 25 total atoms to the molecule, and the location of the first 15 atoms. This is indicated in Figure 5a. The calculation of the "standard" geometry has begun, as indicated by the list of bond lengths. Figure 5b illustrates the "uniting" of two molecular fragments; the 15 atom structure depicted in the first figure and the second benzene ring substructure (shown in the upper right corner). The two vectors or line segments connecting the two substructures define the common bond linkage (i.e. C12-C15 and C-X are to become a single bond). The transformation of the two fragments into one molecular structure is shown in Figure 5c. The remaining 10 atoms have been approximately located in this illustration. The "standard" geometry for the entire structure is illustrated in Figure 5d, and a stereoscopic view is provided in Figure 5e.

At this stage, the molecular coordinates have been determined, and the structure has been verified for accuracy by inspection of the drawings and tables provided by CAMSEQ/M. The next step involves setting up the desired conformational study to be performed. A well documented procedure has been developed for CAMSEQ/M to assist the user in this often difficult task. The difficulty lies in the large number of available options. Table III lists several of these options alphabetically.

Table III

ANGLE:	Lists all bond angles in the molecule
CLOCK:	Performs "clockwise" rather than "counterclockwise" bond rotations
CMAPS:	Produces iso-energy contour maps
DISTANCE:	Lists all interatomic distances in the molecule
ELECTRO:	Selects appropriate electrostatic potential energy function
HBOND:	Determines hydrogen-bonding energies
LINPLOT:	Generates an energy vs. rotation angle plot
LISTCOORD:	Lists all atomic coordinates
LJPAR:	Allows modification of Lennard-Jones, nonbonded potential function parameters
LOCALMIN:	Lists all low energy conformers
MINIMIZE:	Performs multi-dimensional minimization
RANDOM:	Performs random type scan of conformational hyperspace
REFERENCE:	Allows user to specify a "standard" reference conformation

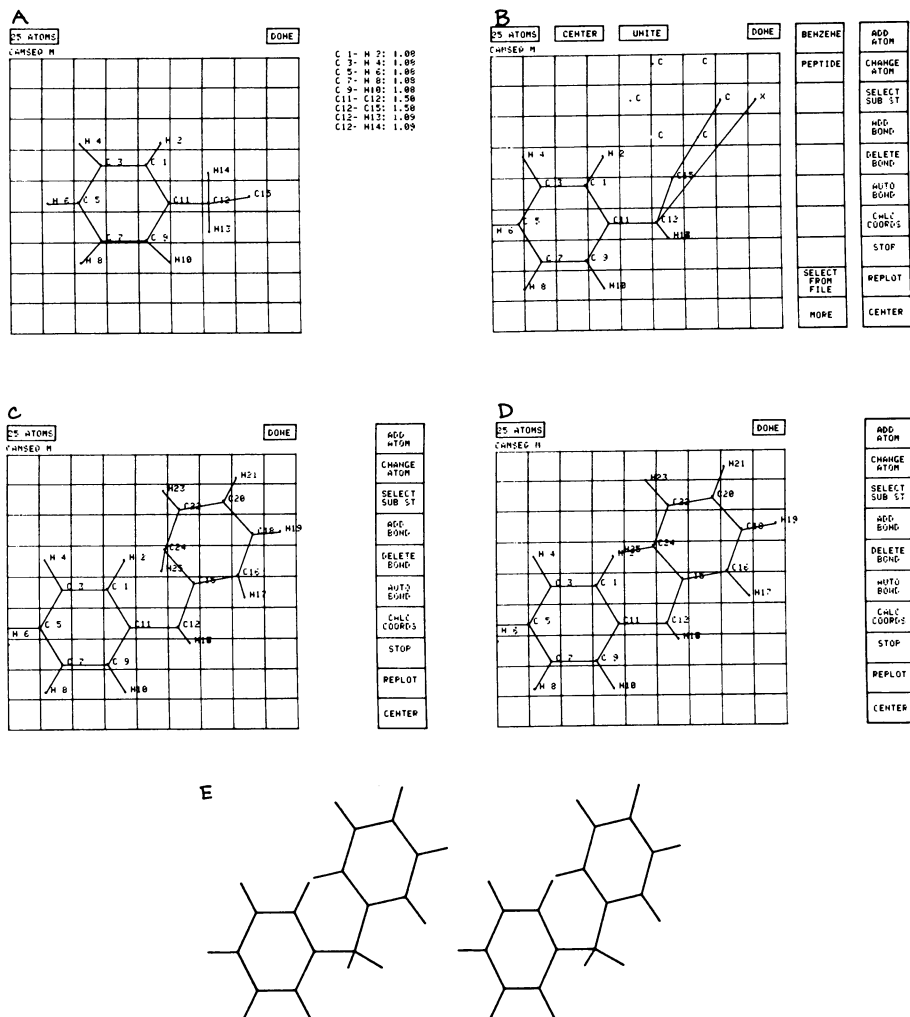


Figure 5. Steps involved in generating the structure of a molecule (diphenyl methane) from several fragments or substructures

SCAN: Performs a uniform sequential scan of conformational hyperspace
THERMO: Lists thermodynamic probabilities
TORSIONAL: Allows user to specify torsional correction functions

Since CAMSEQ/M has been designed to be totally compatible with the "batch" versions of CAMSEQ, additional options are provided which are appropriate for that version of the program. A total of 50 options are available. The advantage of having total compatibility with the batch CAMSEQ is that CAMSEQ/M may be used either as a "stand-alone" conformational analysis device, or it may be used as an interactive, conversational, off-line input device for batch CAMSEQ. Since an available host computer would probably perform the calculations more rapidly than the Tektronix microcomputer, this provides an easy to use batch program interface when results are needed "immediately".

The option selection process is simplified by providing tables of options arranged according to function. A complete description of each option is also available by pressing a single key and specifying the option name. Of course, the user need not concern himself with the internal workings of either CAMSEQ/M or batch CAMSEQ. When an option name is selected, CAMSEQ/M prompts for each input item with a short description. Figure 6 illustrates such an input sequence for two representative options, "ATOMSIZE" and "TORSIONAL". As can be seen, the user need not be familiar with the internal workings of CAMSEQ/M in order to select the various options needed to describe the job to be done.

In the examples depicted in Figure 6, all necessary information is provided to enable the option to be used effectively. Thus, the form of the torsional barrier function is shown indicating the various parameters involved. The "shaping parameters" are shown to be terms used to describe such functions as "cos" as "0+1*cos", or "sin" as "1-1*cos²", etc. It should be pointed out at this time, that although the options are selected in a manner similar to that used in the batch CAMSEQ program, only those options which must be specified in advance (such as those controlling the type of conformational search to be performed, and those specifying the internal parameters to be used in the calculation) need be specified in advance. All options which affect the ultimate data analysis, such as "CMAPS", or "LINPLOT" are specified after completion of the conformational search. If CAMSEQ/M is used as a data preparation device for batch CAMSEQ, then of course, these options must all be specified. However, even data sets produced by the batch versions may be analyzed using CAMSEQ/M.

Specifying the options in this manner allows the rapid description of the conformational analysis task. If desired, the user is prompted for additional options at the appropriate point in the program. For example, prior to setting up the

THE AVAILABLE OPTIONS, IN ALPHABETICAL ORDER, ARE:

ANGLE	ATOMSIZE	CAVITY	CHARGE	CHKBL
CLOCK	CMAPS	CONTACT	COORD	DEBUG
DIFSOLV	DISTANCE	ELECTRO	HBOND	LINPLOT
LISTCOORD	LJPAR	LOCALMIN	MAPAXES	MAPPARAM
MESSAGE	MINIMIZE	MODEL	MODPIC	NEWDEL
NEWSOLV	NEWTOR	NOCONF	NOHEADING	NOMODEL
NOPICT	ORTEP	QUANTUM	RANDOM	REFERENCE
REFPIC	RENUMBER	SCAN	SCREEN	SELECT
SKIPDATA	SOLVDATA	THERMO	TIMELEFT	TORSIONAL
UNITNO	UTLPIC	UACONLY	XRAY	ZMATRIX

ENTER OPTION NAME (OR COMMAND BY PRESSING USER KEY "1"): torsional

ENTER TYPE OF FUNCTION (STANDARD, 2-D, DISCRETE, 2-D AND STANDARD):
: standard

FUNCTION: $E_{\text{tor}} = \text{Barrier1} * \langle \text{SP1} + \text{SP2} * \cos(\text{Mult1} * \text{Bond} + \text{phase1}) \rangle^{\text{Power}} + \text{Barrier2} * \langle 1 - \cos(\text{Mult2} * \text{Bond} + \text{phase2}) \rangle$

BOND # > 12
USE INTERNAL FUNCTION ? no
BARRIER 1 > 2.5
SHAPING PARAMETER (SP1) > 1
SHAPING PARAMETER (SP2) > -1
POWER > 2
MULTIPLIER 1 > 1.5
PHASE ANGLE 1 > 0
CIS/TRANS BARRIER (BARRIER 2) > 0
CIS/TRANS MULTIPLIER (MULT 2) > 0
CIS/TRANS PHASE ANGLE (PHASE 2) > 0

ENTER OPTION NAME (OR COMMAND BY PRESSING USER KEY "1"): atomsize

ENTER NUMBER OF DATA ITEMS: 4

THE SPECIFIED OPTION REQUIRES THAT THE FOLLOWING 4
DATA ITEMS BE SUPPLIED.

RADIUS OF "H" > 1.1
RADIUS OF "C(SP2)" > 1.4
RADIUS OF "C(SP3)" > 1.38
RADIUS OF "N(SP2)" > 1.44

Figure 6. CAMSEQ/M provides a brief description of each input item for the selected options. A full description of the option is available by pressing the "help" key (user key 10 on the Tektronix keyboard) and entering the option name.

bond rotation data, the user is asked if he wishes to describe a reference conformation. If "REFERENCE" was specified previously, the user will be prompted for the appropriate information at this time without the query. In this manner, even if one forgets to select an option in advance, CAMSEQ/M provides a reminder that the option is available. If using CAMSEQ/M as a "stand-alone" computer, it is not even necessary to specify the "select options" feature depicted on the menu in Figure 1. One need merely press the key labeled "perform conformational search", and the appropriate options will be requested.

Individual conformational states may be examined by specifying "single conformation" when prompted for the bond rotation parameters. The coordinates for this conformer may be saved and the resultant structure may be displayed. The user will then be prompted for another single conformer or for an energy calculation. In "single conformation" mode, individual energy terms may be inspected. Thus the total, steric, electrostatic, or other energy terms can be determined, or the interaction energy between a given pair of atoms may be computed. The latter pairwise energy contribution is available only in the "single conformation" mode.

Automatic searches of conformational hyperspace may be requested using the multi-level bond rotation algorithm described above. A selection is made for a uniform sequential scan, a random scan, or a multi-dimensional minimization search of the conformational energy surface. (As indicated above, prototype versions of CAMSEQ/M provide only the sequential scan mode.) The "major" or primary bond rotations are then described, followed by the secondary, or "minor" rotational information.

Based on the options specified, CAMSEQ/M may request additional information regarding the energy parameters to use, the solvent environments to be simulated, and other pertinent information regarding the conformational analysis to be performed. The user is then given the opportunity to assemble a batch CAMSEQ data set to be sent to an alternate computer. This feature assembles all data accumulated into the form of a CAMSEQ run deck and saves the information on a disk file. Alternatively, CAMSEQ/M begins the conformational analysis task, saving rotational and energy information on a disk file.

Data analysis options include the preparation of isoenergy contour maps (CMAPS), linear energy vs. rotation angle plots (LINPLOT) and tables of local energy minima, statistical thermodynamic probabilities, and entropy terms to enable the reporting of "free" energies in addition to "conformational" energies.

The isoenergy contour maps are plotted one "level line" at a time. The user specifies the energy level to be plotted (say 1 Kcal/mole above the global energy minimum), and all regions with energy less than or equal to that energy are indicated by an isoenergy contour line. Provision is made for labeling of

the contour lines using the joystick.

In order to determine the statistical thermodynamic probabilities and entropies for the conformational energy surface, a set of "dots" is plotted indicating the angular values of the set of conformers which define the surface. The joystick cursor control is used to select the set of conformers which occupy a given low energy region. The chosen "dots" are replaced by "asterisks" (to avoid duplication) and the probability and entropy terms are tabulated. Tables of probabilities and entropies may also be produced.

The linear energy vs. rotation angle profiles are produced directly from the conformational energy data file. Figure 7a shows a representative sample of a conformational isoenergy contour map. Figure 7b illustrates the procedure for calculating probabilities and entropies. A typical linear plot of energy vs. rotation angle is depicted in figure 7c.

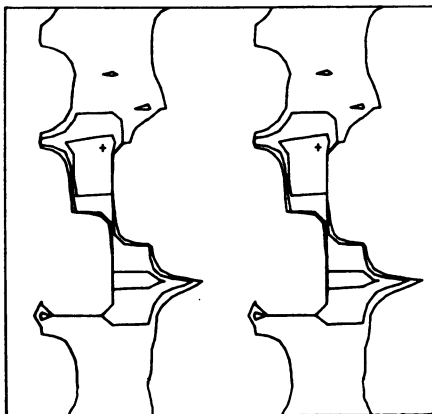
Complete conformational analyses may be performed using CAMSEQ/M. The figures in this report were prepared directly from the screen display (produced by a Tektronix hard copy device). Photographs may also be made from the screen display.

Current efforts in the development of the CAMSEQ/M system include the interfacing of a low-cost microcomputer to the 4051 to handle the actual conformational search and energy calculation. This would free the graphics terminal for other general purpose computing tasks including the set-up of additional molecules for conformational study. Conceptually, banks of these low-cost microcomputers could be linked to a single 4051 to provide a very efficient conformational analysis device.

There are many applications of conformational analysis in drug design and molecular modeling. Comparisons of the conformational energy surfaces and rotational energy profiles within congeneric series of compounds can provide valuable information for the determination of structure-activity relationships. Common low energy regions or spatial characteristics among active or inactive members of a series of molecules can lead to an understanding of some of the properties which are important factors in the observed activity or inactivity of those molecules.

There are several conformation-dependent drug design methods in use. The use of rigid analogs to probe receptors is one of these. Armed with a series of conformationally restricted analogs, one assumes that only those congeners that "fit" the receptor will be active. When several such compounds are investigated, the general spatial and electronic properties of the receptor can often be discerned. We have been applying this technique to a series of phenethylamine derivatives related to the hallucinogen 2,5-dimethoxy-4-methylphenylisopropylamine (DOM, STP). A comparison of the allowed conformational states and steric properties of these molecules with similar data

A ENTER LEVEL LINE VALUE 1 2 3



B ENTER LEVEL LINE VALUE 1 2 3 5 8

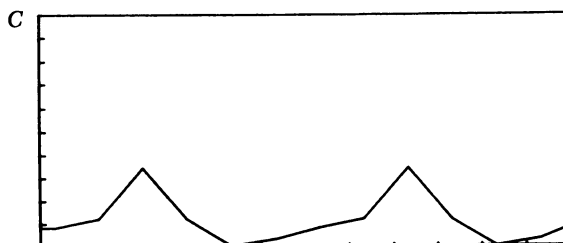
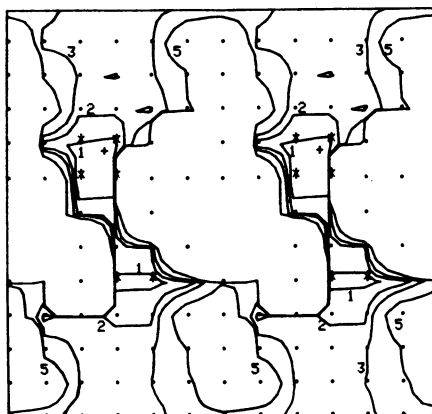


Figure 7. (A) CAMSEQ/M produced conformational isoenergy contour map; (B) procedure for calculating probabilities and entropies; (C) plot of energy vs. rotation angle

for DOM has been very useful in defining probable receptor complementarity (9,10). These studies have led to ideas for new compounds to study which are now being synthesized. CAMSEQ analysis of these new analogs, and correlation with biological activity will serve as "input" to refine further proposed receptor topography.

An excellent discussion of the application of conformational methods may be found in references 19 and 20. Receptor mapping and the rigid analogs discussed above are but two of the applications which CAMSEQ/M is designed to handle. By using CAMSEQ/M as a preprocessor for data input and using the companion batch CAMSEQ for the conformational study, semi-empirical molecular orbital methods may be employed to examine preselected regions of conformational hyperspace. As reported previously, CAMSEQ (1) has been interfaced with CNDO/2, PCILO, and other widely used MO techniques. CAMSEQ/M can therefore provide a simplified data input and data analysis package for classical-empirical potential energy calculations as well as semi-empirical molecular orbital studies.

Acknowledgements: Support from the U.S. Public Health Service (GM 25142) and the Purdue Cancer Center is acknowledged.

Literature Cited:

1. Weintraub, H.J.R.; Hopfinger, A.J. Int. J. Quantum Chem., 1975, QBS2, 203.
2. Potenzzone, R. Jr.; Cavicchi, E.; Weintraub, H.J.R.; Hopfinger, A.J. Computers and Chem., 1977, 1, 187.
3. Weintraub, H.J.R. Ph.D. Dissertation, Case Western Reserve University, Cleveland, Ohio, 1975.
4. Weintraub, H.J.R.; Hopfinger, A.J. J. Theor. Biol., 1973, 41, 53.
5. Weintraub, H.J.R.; Hopfinger, A.J., in "Molecular and Quantum Pharmacology", Bergmann, E.; Pullman, B., Eds., D. Reidel, Dordrecht, Holland, 1974; p. 131.
6. Weintraub, H.J.R.; Tsai, M.D.; Byrn, S.R.; Chang, C.-j.; Floss, H.G. Int. J. Quantum Chem., 1976, QBS3, 99.
7. Weintraub, H.J.R. Int. J. Quantum Chem., 1977, QBS4, 111.
8. Tsai, M.D.; Weintraub, H.J.R.; Byrn, S.R.; Chang, C.-j.; Floss, H.G. Biochemistry, 1978, 17, 3183.
9. Weintraub, H.J.R.; Nichols, D.E. Int. J. Quantum Chem., 1978, QBS5, 321.
10. Nichols, D.E.; Weintraub, H.J.R.; Pfister, W.R.; Yim, G.K.W., in "NIDA Research Monograph Series", Barnette, G.; Trsic, M.; Willette, R., Eds., 1978, 22, 70.
11. Hopfinger, A.J. "Conformational Properties of Macromolecules", Academic Press, New York, 1973.
12. Hesselink, F.T.; Ooi, T.; Scheraga, H.A. Macromolecules, 1973, 6 (4), 541.

13. Potenzzone, R., Jr., M.S. Thesis, Case Western Reserve University, Cleveland, Ohio, 1975.
14. Weintraub, H.J.R., presented at 1979 Sanibel Symposium, Palm Coast, Florida, March, 1979. Manuscript in preparation
15. Hopfinger, A.J.; Battershell, R.D. J. Med. Chem., 1976, 19 (5), 569.
16. Heller, S.R.; Milne, G.W.A.; Feldman, R.J. Science, 1977, 195 (4275), 253.
17. Sutton, L.E. Chem. Soc., Special Publication No. 11, 1958; Supplement No. 18, 1965.
18. Jain, S.C.; Tsai, C.-c.; Sobell, H.M. J. Mol. Biol., 1977, 114, 317.
19. Martin, Y.C., "Quantitative Drug Design, A Critical Introduction:", Dekker, New York, 1978.
20. Richards, W.G., "Quantum Pharmacology", Butterworths, London, 1977.

RECEIVED June 8, 1979.

Beyond the 2-D Chemical Structure

N. C. COHEN

Centre de Recherches ROUSSEL-UCLAF, 102 Route de Noisy,
93230 Romainville, France

The spectacular development of the QSAR theories has dominated the recent evolution of drug-design, and the conditions for the use of these theories are now well established. Being constructed on a particular chemical formula, with an implicit common invariant chemical skeleton, these methods are not always able to go beyond the chemical frame of the particular family studied.

The recent advance of molecular biology has widened the classical horizons of drug research which has been carried out until now on mainly two-dimensional lines, and based on the synthesis of chemical analogs of known compounds. This advance is due to the discovery of the three dimensional specificity of molecular interactions which is a fundamental key to the understanding of life processes. The concepts of molecular recognition and discrimination considered in the light of drug research allow one to go beyond the classical drug manipulation and optimization. Geometrical and associated electronic properties are of great importance to this new approach.

New Trends

For the conception of new molecular structures it is first necessary to be able to evaluate the geometrical properties of reference molecules. Theoretical calculations of conformations are now accurate enough so that this point can be taken as read even for molecules not yet studied experimentally. With the aid of these now classical techniques the drug-designer can go further and address himself to the three following fundamental items:

1. How to reveal, what three dimensional structural entities (pharmacophore) are responsible for the biological properties of a known active compound, and what could be the active conformation of this compound, if more than one conformer is probable.
2. How to evaluate to what extent different molecules can

0-8412-0521-3/79/47-112-371\$05.00/0

© 1979 American Chemical Society

mimic the spatial arrangement of atoms of a known active compound or its putative pharmacophore.

3. How to conceive of a method for designing original chemical compounds which conform to the desired requirements.

We have developed a computer program MIMETIC-GEMO, the purpose of which is to be of some help in answering points 1 and 2.

Mimetic-Gemo

This computer program is able to compare the molecular geometries of different molecules without any restrictions on their relative chemical structure. The only implicit assumption is that the compounds studied are supposed to act biologically by the same mechanism of action. The program is able to compare a set of at least two molecules (the actual version is programmed for a maximum of $n = 6$ molecules). The molecules are considered as free to move and to modify their geometries, provided that the global deformation of each molecule does not increase the total molecular energy by more than a pre-established amount (i.e. 10-15 kcal/mole). At the beginning of the calculations an initial position is chosen for each molecule, together with the relative orientations one to the other, the total sum of molecules forming a set of n overlapping molecular geometries.

The parameter used to characterize to what extent the "mimetism" is achieved is S_T , the total exposed molecular area of this set of overlapping molecules. The aim of the calculations is to find what are the geometries of each of the n molecules which lead to a minimal value of S_T .

In MIMETIC-GEMO the molecular energies and deformations are calculated using a previously described computer program (1), while the minimization is conducted with a standard routine (2). When the minimum is reached, it is interesting to know for each molecule:

- what is its final geometry and the corresponding molecular energy,
- what are the overlapping and the non-overlapping moieties of one compound relative to the others.

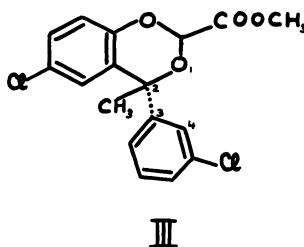
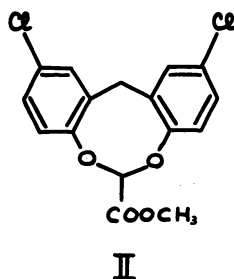
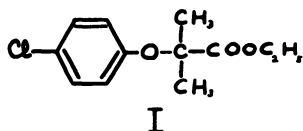
It is also useful to ascertain what are the general geometrical features common to the whole set of molecules studied.

The principle of minimizing the total exposed molecular surface area S_T parameter is a rough but practical index. Better than the manual manipulation of approximate molecular models this criterion can be a powerful tool for revealing the 3-D mimetism between several molecules. An example to illustrate the

possibilities of this approach is given below.

Example

In our research group a potent hypolipidemic agent has been found (compound III) (3) which is active as treloxinate (II) (4) and clofibrate (I) (5) in lowering the triglyceride and cholesterol levels.



Chemical Formulas: Chemical Structures of Clofibrate (I), Treloxinate (II), and RU-25961 (III).

The two compounds II and III are much more potent (in Rats) than the marketed product I. While it is easy to recognize the common p-chloro-phenoxyacetic ester in all compounds, the origin of the enhanced activities of structures II and III is questionable and MIMETIC-GEMO would appear to be useful for finding out whether these two molecules might not have more important features in common.

Results

Having separately calculated the stable conformers of the two molecules (6), MIMETIC-GEMO calculations were then performed on the acid form of the compounds (7), allowing each one a tolerance of 15 kcal/mole for its deformations. At the starting point the two p-chloro-phenoxy groups were coincident and the more stable conformer was chosen for each product. In this position the total surface area S_T was equal to 369 Å² while the isolated molecules had molecular exposed areas of 295 and 313 Å² for treloxinate and RU-25961 respectively. A first minimum was reached when S_T had decreased to 341 Å² with energies for the isolated molecules (above the minimum taken as origin) of 14.6 kcal/mole for

treloxinate and 12.2 kcal/mole for RU-25961. During the minimization the $\phi(O_1C_2C_3C_4)$ torsion angle of the free phenyl group of III has rotated from an initial value of 58.0 degrees to the final value of 26.8 degrees, indicating that a mimetism is possible, which increases toward the lowest values of ϕ .

Starting from the same point for II and a conformation of III optimized for $\phi = 0^\circ$, a second minimum was obtained with the energy values of 9.7 kcal/mole for treloxinate and 11.2 kcal/mole for RU-25961. The final values of ϕ and S_T being respectively -7.2 degrees and 345 \AA^2 .

Figure 1 shows the overlap obtained in this case. Space filling computer drawings (8) for each molecule are shown in figures 2a and 2b; they show how a remarkable mimetism is possible, which was not at all obvious from the separate 2-D chemical structures of the two compounds.

Using several other starting points it was not possible to go below the values of $341 - 345 \text{ \AA}^2$ of the S_T index already obtained. It appeared possible to improve these results by increasing the energy tolerance permitted to the molecules during the optimization. For instance with tolerance values of 25 and 40 kcal/mole the corresponding minimum values of S_T were respectively 327 \AA^2 ($\phi=9^\circ$) and 318 \AA^2 ($\phi = -10^\circ$).

Electronic Aspect

We have found it useful to have information on both molecular geometry and electronic characteristics in the same frame. For this purpose we have conceived a "four-dimensional symbolism" (9) which consists of a visualization of the envelope of the molecules studied, on the surface of which energy contours of the electrostatic molecular potential are drawn. This kind of representation is shown in figures 3a and 3b. The molecular envelope is defined by considering that each atom has a radius R_a such as

$$R_a = R_{VDW} + R_0$$

where R_{VDW} is the Van der Waals radius of the atom, and R_0 a constant value taken here to 1.5 \AA . The electrostatic molecular potential used in these calculations is the classical Coulombic expression of the interaction of a + 1e charge at the point concerned. The electronic densities of the atoms being obtained by extended Hückel molecular orbital calculations (10). From these drawings it can be seen that the similarities between the two molecules are not only geometrical but also electronic. The shape and the values of the energy contours allow an appreciation of the extent to which the two compounds can be considered as mimetics.

Conclusion

These data show that similarities between chemical compounds can exist from the point of view of the molecular geometry, which are not obvious from the 2-D structures.

What is needed for the future of drug-design is to develop new original molecular architectures conceived on the basis of spatial mimetism between geometrical and electronic features.

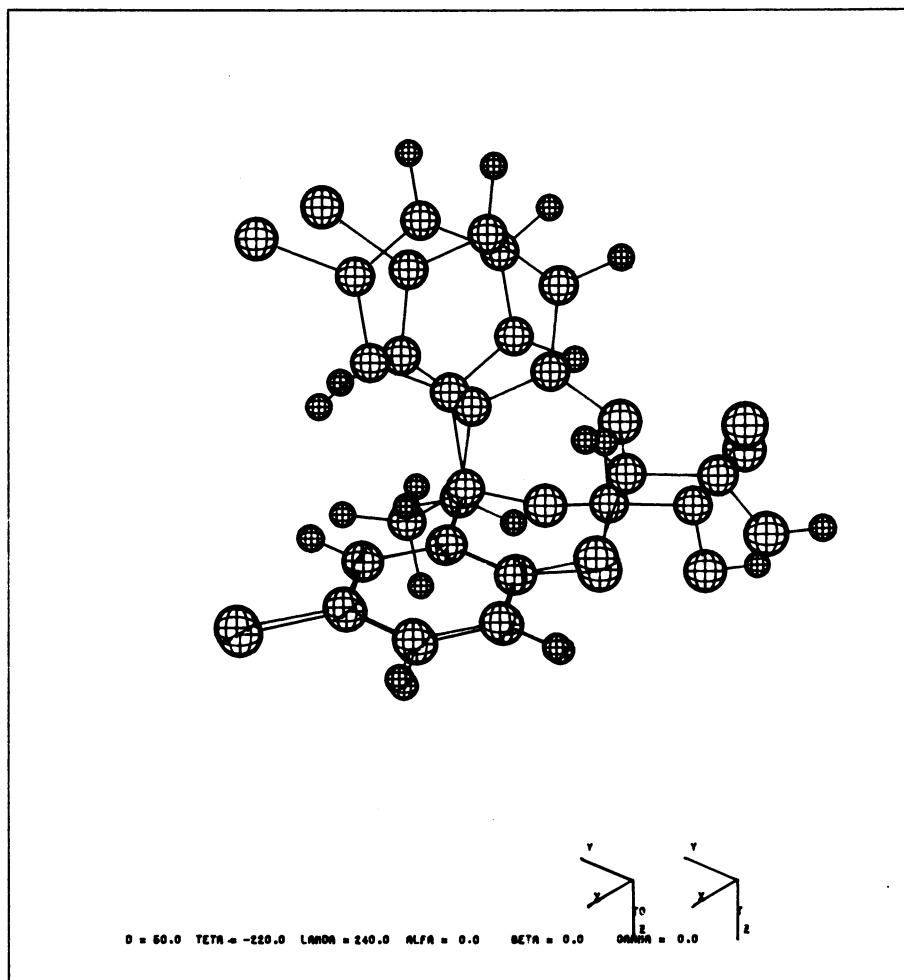


Figure 1. *Overlap reached after minimization of S_T for the acid forms of treloxinate (II) and RU-25961 (III) ($S_T = 345 \text{ \AA}^2$)*

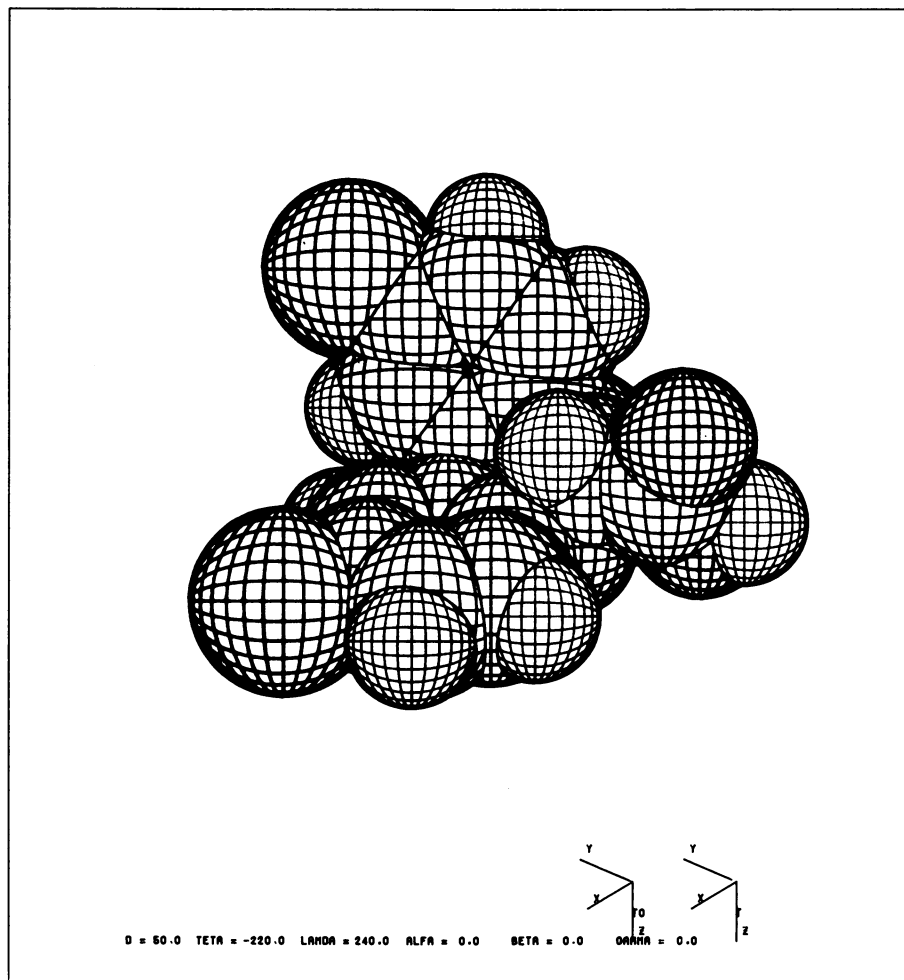


Figure 2a. Three dimensional mimetism between the acid forms of treloxinate

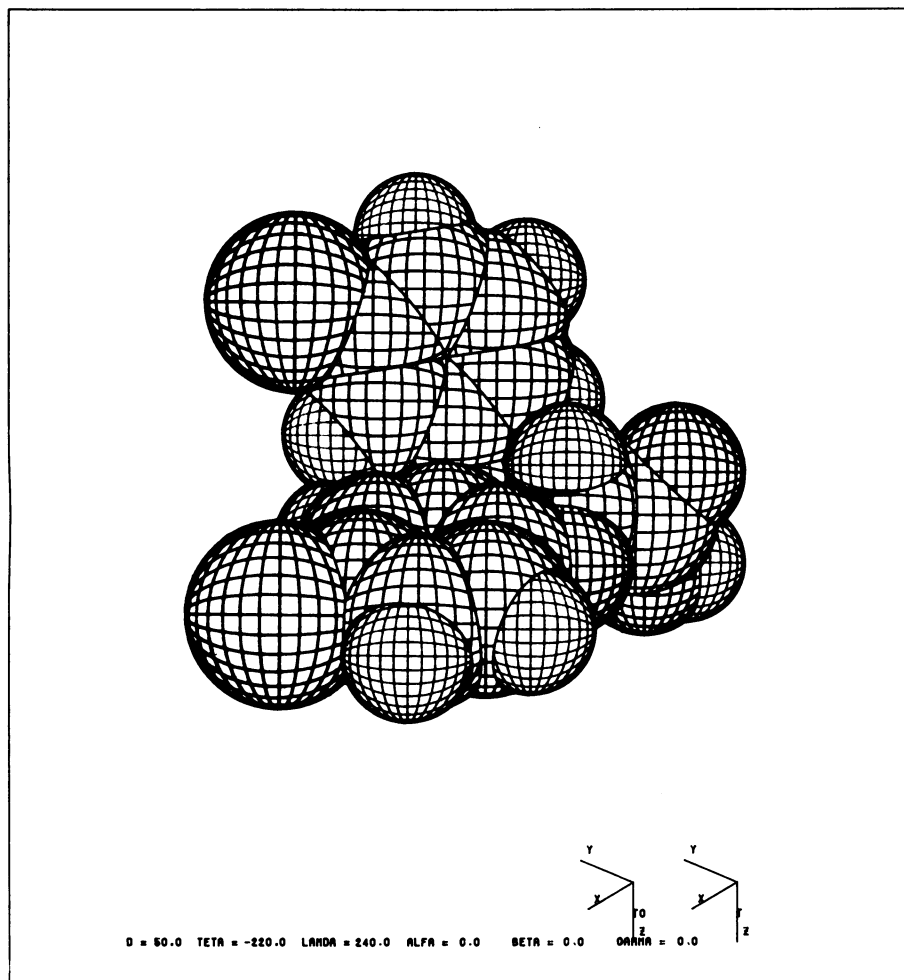


Figure 2b. Three dimensional mimetism between the acid forms of RU-25961

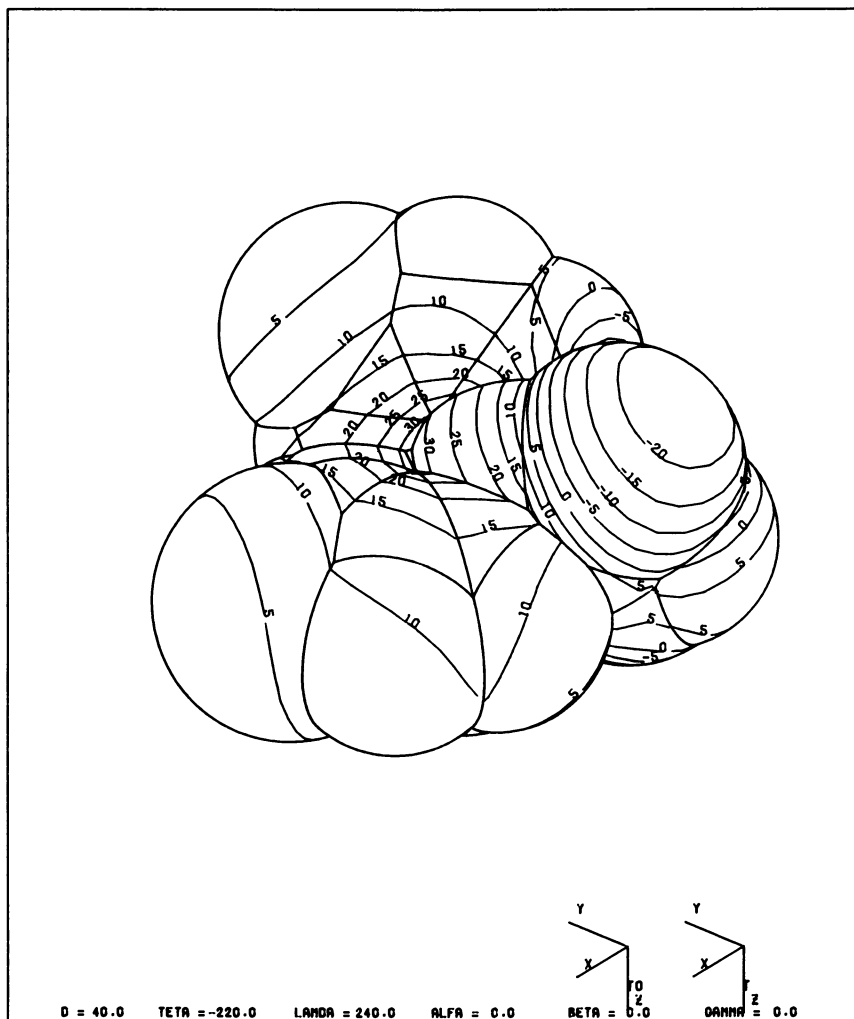


Figure 3a. Envelope of the acid forms of the molecules with the energy contours (Kcal) of the electrostatic molecular potential for treloxinate

Abstract

QSAR theories are convenient for the design of new analogs of a known active compound but these theories are not always able to go beyond the chemical frame of the particular family studied. Considering structure-activity studies in the light of the three dimensional specificity of molecular interactions between drugs and receptors, conformational properties appear to be essential. To allow the design of new molecular architectures which could be able to mimic the essential features of known active compounds, it is necessary to evaluate to what extent different molecules can mimic the spatial arrangement of atoms of a known active compound.

The criterion of minimization of the total exposed molecular area of sets of overlapping molecules appears to be a powerful tool for revealing the 3-D mimetism between several molecules. During the minimization the molecules are not rigid but free to move and modify their geometries, provided that the molecular energies do not increase by more than a pre-established amount. Such calculations can reveal the biologically relevant molecular geometries and the possible associated pharmacophores. It is possible for the 3-D mimetism to be, not only geometrical but also electronic. Both properties can be visualized in a "four dimensional symbolism" which consists of a view of the envelope of the molecules studied, on the surface of which, energy contours of the electrostatic molecular potential are drawn. An example will be presented to illustrate the possibilities of this approach.

Literature Cited

- 1 COHEN N. C. Tetrahedron (1971) 27, 789 "GEMO, a computer program for the calculation of the preferred conformations of organic molecules."
- 2 Subroutine SCOOP supplied by the IBM Company.
- 3a CLEMENCE F., HUMBERT D. and DAGNAUX M. (1977) Belgian Patent BE 860,500.
b CLEMENCE F., HUMBERT D., DAGNAUX M. and FOURNEX R. to be published.
- 4 GRISAR J. M., PARKER R. A., KARIYA T., BLOHM T. R., FLEMING R. W., PETROW V., WENSTRUP D. L. and JOHNSON R. G., J. Med Chem. (1972) 15, 1273 "Treloxinate and related hypolipidemic derivatives."
- 5 WITIAK D. T., NEWMAN H. A. I. and FELLER D. R. "Clofibrate and related Analogs. A Comprehensive Review." Marcel Dekker, New York, N.Y. 1977.

- 6 Calculated with the GEMO program (ref. 1).
- 7 It is admitted that the active principle is the free carboxylic acid form, and we refer in all the calculations to these forms.
- 8 COHEN N. C. "Computer program VIF" supplied on request.
- 9 COHEN N. C. "Computer programs VISOPOT" unpublished work.
- 10 HOFFMANN R. J. Chem. Phys. (1963) 39, 1397 "An extended Hückel theory."

RECEIVED June 8, 1979.

Conformational Analysis: A Module in a Program for the Design of Biologically Active Compounds

A. J. STUPER, T. M. DYOTT, and G. S. ZANDER

Research Laboratory, Rohm and Haas Company, Spring House, PA 19477

The last decade has seen considerable progress made in understanding the relationship between chemical structure and biological activity. Despite this progress our ability to design active molecules is still quite unsophisticated. While we are certainly unable to a priori, design molecules possessing specific biological properties, we are able to identify several factors which govern this action. These are:

- 1) steric properties
- 2) transport properties
- 3) reactivity

Certainly each factor does not operate independently of the others. Reactivity is influenced by the steric properties of a molecule, as is transport. Also these factors do not directly address the possible metabolic fate of a compound. However, such a division does allow us to concentrate on specific aspects of the problem and to direct our synthetic efforts accordingly.

How then does this relate to the practical problem of developing new, effective, and safe drugs or other biologically active materials? There are certainly two ways to approach the problem. We can assume that by its nature the design problem is intractable to any theoretical approach and therefore synthesize as many variants of a lead compound or test as many new compounds as we can economically afford. For this course of action, our success will be a function of luck aided by intuition. On the other hand, we can take the approach that control of those factors effecting activity will increase the probability of reaching our goal. While the factors influencing activity are not all under our control, an attempt to control as many as possible will reduce the arbitrary nature inherent in the first approach.

Taking the latter approach there are two ways to guide our synthetic efforts. To optimize the activity of a compound we can attempt to find mathematical models which relate changes in the above factors to changes in the observed activity. We can then

0-8412-0521-3/79/47-112-383\$08.00/0

© 1979 American Chemical Society

use these models to assist us in determining which analogues hold out the most promising chance for increased activity. This process is well documented, having been extensively developed by Hansch and co-workers (1,2).

On the other hand, we may desire a compound which has the activity of a particular lead, but is of apparently different structure. That is a new lead. In this case we would try to design compounds which possess different chemical structures but maintain identical steric, transport, and reactivity properties. It was for this purpose that we have developed a computerized tool, called MOLY, which can assist in the molecular design problem. In this paper we will briefly review what this system is and detail our efforts to parameterize the conformational analysis section of MOLY.

Overview of the MOLY Program

We present this section to provide an understanding of the context within which the conformation analysis program operates. MOLY is an interactive system which employs computer graphics as the means of communicating its results to the scientist. The system has the following characteristics:

1. large [\sim 1000K bytes]
2. modular
3. command driven
4. considerable user prompting
5. highly graphics oriented

Figure 1 shows the system's architecture. The main program, referred to as the driver, deciphers the user's directions and calls the appropriate module. This module then interacts with the user to perform a specific task. There are seven major modules whose primary functions are:

1. INPUT - Input molecules by drawing them on the screen of the graphics terminal.
2. MODEL - Build a reasonable three dimensional model of a molecule.
3. CONFOR - Perform a conformation analysis.
4. ANALYZE - Prepare contour maps of energy and population as a function of rotation about one or two bonds.
5. LOGP - Estimate the value of the octanol/water partition coefficient.

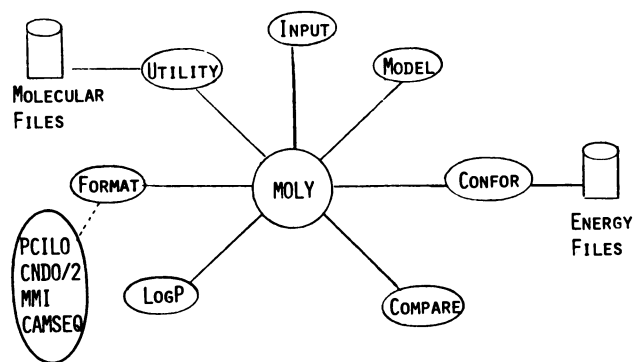


Figure 1. Architecture of MOLY

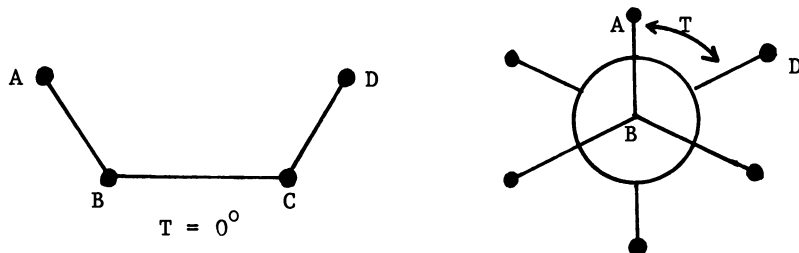
6. **FORMAT** - Provide an interface between the interactive program and batch versions of quantum mechanical programs.
7. **COMPARE** - Provide visual comparisons between different molecules by matching selected portions of their structures.

In addition MOLY contains a large number of utility programs which allow for efficient molecular storage, retrieval, orientation and display. The system is written entirely in FORTRAN and executes on an IBM 370/158 under the MVS operating system and the TSO time-sharing system. User interaction is via a Tektronix 4006 or 4010 graphics terminal. Hard copy capabilities are provided via Tektronix photocopiers as well as a Calcomp digital plotter.

The Conformation Analysis Module

The essential problem in conformational analysis is to find the proper spatial orientation of a molecule's atoms. If we consider the bond lengths and bond angles to be fixed then shape is entirely determined by the rotation about single bonds. The approximation of the fixed nature of bond lengths and angles is well accepted. These relations are not significantly altered by crystal or solvation forces. Torsional interactions which are of relatively low energy are often drastically altered by such forces. By calculating the molecular energy at all possible positions of all rotatable bonds, we are able to determine the rotation angle at which the energy is a minimum. The minimum energy conformations are those in which the molecule is most likely to be found in the gas phase. If these minima are sufficiently deep, then it is likely that the same conformations will be observed in solution or the crystalline state (3).

The torsional angles in the compounds which we shall discuss will be defined more precisely later. Here we shall recall only that the torsional angle T about the Bond B-C in the sequence of atoms A-B-C-D is the angle through which the far bond C-D is rotated relative to the near bond A-B. The cis planer position of bonds A-B and C-D represent a $T = 0^\circ$.



Torsional angles are considered positive for right-handed rotation; when looking along the bond B-C, the far bond C-D rotates clockwise relative to the near bond A-B. Alternatively, the positive angles are defined as 0° to 180° , measured for a clockwise rotation, and negative angles as 0° to -180° , measured for a counterclockwise rotation. Since the valence of an atom is generally greater than two the sequence of atoms A-B-C-D may be non-unique. However, it is sufficient to designate any four atoms which will be used to define the torsional angles used in any particular problem. As explained above, a T of 0° is always the cis planer arrangement of the atoms designated as defining the torsional angle.

Conformational analysis presents a combinatorial problem. An analysis involving 360 degree rotations in D degree increments about N different bonds requires the examination of $(360/D)^N$ conformations. The number of conformations increases exponentially with the number of bonds being rotated. For example, if D is 30 degrees and N is 2 then only 144 conformations must be examined, but if N is increased to just 5 the number of conformations jumps to 248,832. The sheer magnitude of such a problem rules out even the fastest quantum mechanical technique. Empirical methods are the only practical means of examining such a large number of conformations.

In our system, conformational analysis is performed by the module CONFOR. An analysis is defined by specifying the bonds to rotate, the increment of rotation, and the total number of degrees to rotate. On the basis of this information the program calculates the number of conformations which would result. If the problem is too large to be done interactively (>5 CPU minutes), the problem can be simplified or CONFOR can be used to initiate a background batch job which will perform the analysis. In either event once an analysis is complete the scientist can ask for the low energy conformations, look at energy contour maps, place the molecule in various conformations for visual inspection, etc.

In CONFOR, conformation energy is broken into four terms.

$$E_{\text{conf}} = E_{\text{vdw}} + E_{\text{coul}} + E_{\text{hbond}} + E_{\text{conj}}$$

where: E_{conf} is the relative energy of the conformation
 E_{vdw} is the Van der Waals term, accounting for steric interactions
 E_{coul} is the coulombic interaction term, accounting for partial charge interactions
 E_{hbond} is the hydrogen bond term
 E_{conj} is the conjugated bond term accounting for torsional preferences due to conjugation

The exact mathematical form of each energy term is given in Table I. E_{vdw} is the standard 6-12 Lennard-Jones potential function. E_{coul} is the classical coulombic potential function. The dielectric constant of the solvent, ϵ , can be specified by the scientist, otherwise it defaults to a value of 3.5. E_{hbond} is the hydrogen bond potential function developed by Scheraga et al (4). This function has no angular dependence unlike the function developed by Hopfinger et al (5) which in our experience is overly restrictive.

The E_{conj} term was developed by us in the course of this work. It is an attempt to take into account the torsional barriers caused by interaction of adjacent π systems and/or lone pairs. Such systems, e.g., amides, benzoic acids, and dienes, show a preference for planar conformations which maximize conjugation. The basic form of the function is the standard cosine relationship frequently used for torsional energy:

$$E_{conj} = B[1 - \cos(N\theta)]$$

where: B is half the barrier height

N is a symmetry constant, usually 2

θ is the torsional angle

E_{conj} has a value of 0.0 for the preferred values of θ and increases in a smooth sinusoidal manner to a maximum of +2B as θ deviates from the preferred values. If, however, either or both of the systems in conjugation are further conjugated, as is for example a benzamide, where the carbonyl is conjugated with both the amide nitrogen and the phenyl ring, the effective barrier is decreased via two mechanisms:

1. The energy gained from conjugation initially is less, i.e., in the case of the benzamide, since the carbonyl is conjugated with the phenyl it cannot conjugate as strongly with the nitrogen.
2. Energy lost by the reduction of conjugation due to rotation about one bond can be partially recovered by increased conjugation with the other conjugated groups.

This effect, however, is dynamic. For example, as the phenyl ring rotates out of conjugation with the amide the torsional barrier about the amide bond returns to full strength. Thus the effective barrier to rotation about a conjugating bond is dependent on:

1. the "normal" barrier for such a bond
2. any additional conjugated groups
3. their conjugative strength
4. their relative conformation

Table 1. Mathematical Form of Potential Functions

Nonbonded	$E_{\text{vdw}} = [B_{ij}/d^6 - A_{ij}]/d^6$
where	B_{ij} = repulsive constant dependent on atoms i, j A_{ij} = attractive constant dependent on atoms i, j d = Euclidian distance between atoms i and j
Coulombic	$E_{\text{coul}} = \frac{332.0}{\epsilon} \frac{q_i \cdot q_j}{d}$
where	ϵ = dielectric constant q_i = partial atomic charge on atom i d = Euclidian distance between atoms i and j
H-bond	$E_{\text{hbond}} = E[r_o/r_{\text{HX}}]^{12} - 2.0E[r_o/r_{\text{HX}}]^{10}$
where	E = energy constant dependent on type of atoms participating in the hydrogen bond r_o = internuclear distance parameter dependent on type of atoms r_{HX} = Euclidian distance between donor and acceptor atoms
Rotational Barriers	$E_{\text{tor}} = B'_k [1 - \cos(N\theta_k)]$ $B'_k = \frac{B_k}{2} \left[1 - \sum_{i \neq k} \frac{n_k B_i (1 + \cos(N\theta_i))}{2 \cdot \sum_{i=1}^{n_k} B_i} \right]$
where	B'_k = the effective barrier to rotation about the bond of interest B_k = the non-conjugated barrier to rotation about the bond of interest B_i = the non-conjugated barrier to rotation about the i th conjugated bond θ_k = the torsional angle for the bond being rotated θ_i = the torsional angle for the conjugate bonds n_k = number of bonds conjugated with bond k N = a symmetry factor

CONFOR automatically takes all of these factors into account. It initially finds all conjugating bonds and assigns their "normal" barrier values and symmetry constants. These values are shown to the scientist who can specify alternative values. As the molecule is being stepped through various conformations, the effective torsional barriers of any conjugating bonds are recalculated and then used to calculate E_{conj} via the formulas given in Table I.

Parameterization of the Conformational Analysis Module

Before we can use the results of conformational analysis to predict the steric constraints for a drug we must be reasonably certain that the method is properly parameterized. Our major concern is that the minima indicated are located properly and the contour energy maps produced have a reasonable shape. This requirement is rather broad. Our experience has been that the accuracy of empirical functions suffice for systems which show little or no stabilization due to resonance or other types of electronic interchange.

There are several ways to approach the parameterization problem. Previous groups have used comparisons to crystal packing as a gauge for the non-bonded interactions (6). They then fine tune the program using comparisons to all valence MO calculations. We have chosen much the same route. Our fine tuning was done by comparing our results to those obtained using the PCIO technique of Pullman (7-9). We chose this technique because of its accuracy and the availability of a large number of calculations for various small molecules.

In order to define the factors involved in parameterizing the non-bonded potential function, it is necessary to describe the functional approximations used. As have others, we used the Lennord-Jones 6-12 potential as the basis for this interaction. The form of this is:

$$E_{\text{vw}} = B/d^{12} - A/d^6 \quad (1)$$

E_{vw} is the energy of the interaction in Kcal, d is the distance, in angstroms, between the interacting atoms, A and B are constants which depend on the type of atoms involved in the interaction. A typical potential is shown in Figure 2. At large distances there is essentially no contribution to the conformational energy, but as the distance decreases this energy becomes increasingly attractive. With a further decrease in distance the atoms get too close and quite rapidly start to repel each other.

There are two facets to the Lennord-Jones potential function which interest us, the equilibrium distance d_0 and the well depth E_0 . These parameters are related to A and B by:

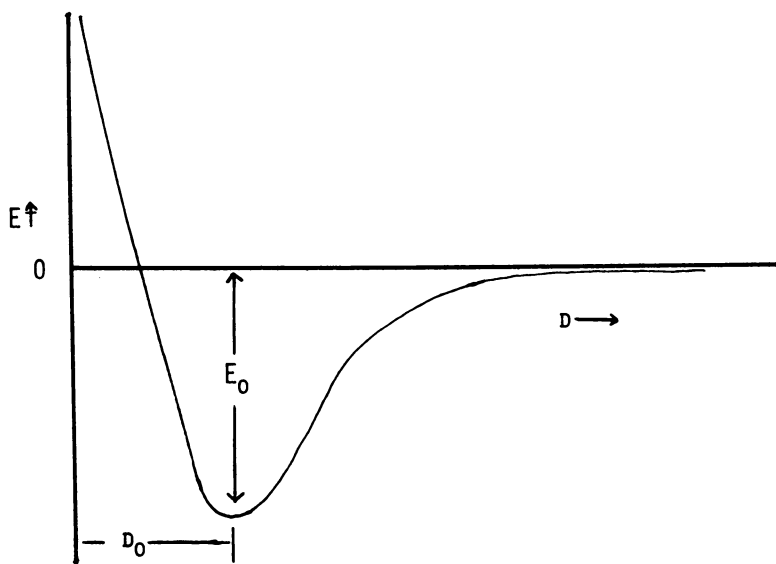
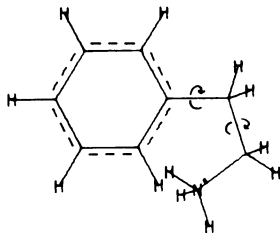
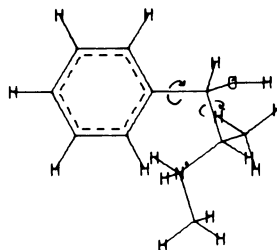


Figure 2. *Typical Lennard-Jones 6-12 potential*

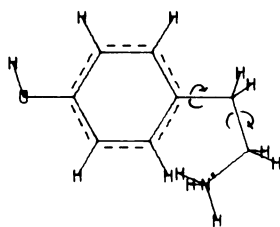
A. PHENETHYLAMINE



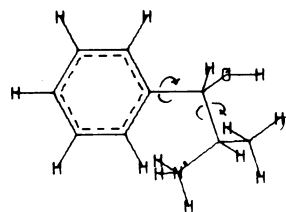
B. EPHEDRINE



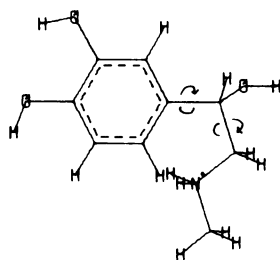
C. TYRAMINE



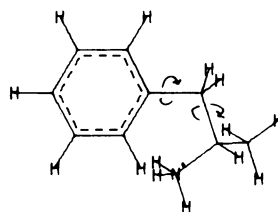
D. NOREPHEDRINE



E. EPINEPHRINE



F. AMPHETAMINE



G. SYMPATH-1

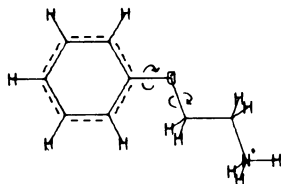
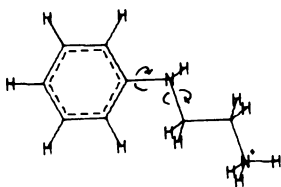
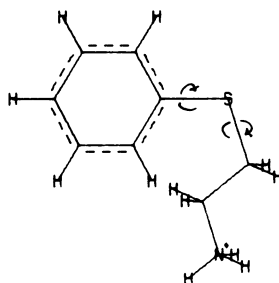


Figure 3. Thirteen molecules used for comparison to PCILO

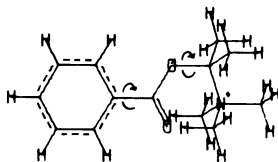
H. SYMPATH-2



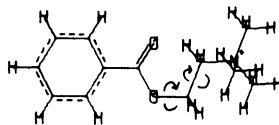
I. SYMPATH-3



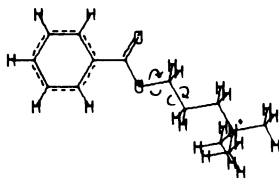
J. PSEUDOProcaine



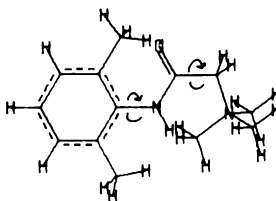
K. PSEUDOButacaine



L. PSEUDOAmyleine



M. PSEUDOlidocaine



$$A = -2E_0 d_0^6 \quad (2)$$

$$B = -E_0 d_0^{12} \quad (3)$$

Our initial settings for these constants were taken from the literature (10,11). Unlike the literature references we do not differentiate the atoms according to hybridization. The literature values thus served as an initial approximation. In order to develop a consistent set of constants we compared our calculations on the thirteen molecules in Figure 3 to those obtained by PCILO. We then adjusted our E_0 and d_0 values to obtain results which were consistent with the PCILO calculations. Table II lists the values of A, B, d_0 , and E which resulted from this comparison. Note that it was necessary to differentiate between aromatic and non-aromatic hydrogens. This parameterization is only appropriate for H, C, O, N, and S. The values for the other elements listed represent the initial approximations. Our d_0 values are all somewhat higher than those used by Hopfinger (10) and smaller than those used by Scheraga (11). The difference between our values and Hopfinger's is probably due to our not dividing the atoms according to valence. The Scheraga values are higher because the minimum energy distances are based upon contact distances somewhat larger than the Van der Waals contact distance.

We have reproduced several contour maps to demonstrate the degree of comparison between PCILO and CONFOR. These plots are shown in Figures 4 thru 8. To generate these maps the molecules were placed in the geometries specified by Pullman (7-9) and the bonds indicated in Figure 3 were rotated. The reference atoms for the torsional angles those used by Pullman. Partial atomic charges for use in the Coulombic potential function were obtained from CNDO calculations. The PCILO contour maps represent calculations having a resolution of 30 degrees about each bond. Our maps were generated using rotational increments of 10 degrees.

In general the agreement between the two methods was quite close. Phenethylamine (Figure 3a) and Pseudoprocaine (Figure 3k) are typical examples of calculations which compare quite closely. Both the location of the minima and the overall shape of the map are very close to those obtained by PCILO. The map for Ephedrine (Figure 3b) has the same general shape; however, the stabilization at $T_2 = 60$ degrees is more pronounced than that shown by PCILO. Also, the minima for $T_1 = \pm 60$, $T_2 = 0$ are about 1 Kcal higher than that found by PCILO. It is not clear why we see a larger stabilization at $T_1 = \pm 60$, $T_2 = 60$. This could be a result of improper non-bonded terms. It may also be a function of the difference in the resolution of the two maps.

A similar difference is seen in the map for Norephedrine (Figure 3d). Our calculations show that at a $T_1 = \pm 60$, $T_2 = 60$, there is a region of stability. Figure 9 shows three views of

Table II. Values for the Nonbonded Potential Function

<u>Hydrogen</u>						
	H	C	N	P	O	S
A	70.353	175.387	129.6811	264.8311	123.7239	289.9912
B	6281.149	33004.1	24024.57	78276.24	17088.45	95562.36
d _O	2.373	2.687	2.68	2.897	2.55	2.95
E _O	-0.197	-0.233	-0.175	-0.224	-0.225	-0.220

	F	Cl	Br	I	AH
A	79.67815	272.4885	299.1173	437.8515	32.80541
B	9336.187	80706.48	88410.22	166418.3	1217.415
d _O	2.783	2.898	2.897	3.021	2.05
E _O	-0.170	-0.230	-0.253	-0.288	-0.221

<u>Carbon</u>						
	H	C	N	P	O	S
A	175.387	522.0883	502.2265	875.9980	446.1175	998.1056
B	33004.1	240791.7	190507.2	616858.8	132680.5	680474.6
d _O	2.687	3.12	3.02	3.348	2.90	3.33
E _O	-0.233	-0.283	-0.331	-0.311	-0.375	-0.366

	F	Cl	Br	I	AH
A	264.9297	906.5424	1043.719	1475.163	175.3848
B	99698.52	636883.0	767148.0	1346600.	33004.10
d _O	3.016	3.346	3.372	3.496	2.687
E _O	-0.176	-0.323	-0.355	-0.404	-0.233

Table II. Values for the Nonbonded Potential Function (cont'd.)

<u>Nitrogen</u>						
	H	C	N	P	O	S
A	129.6811	502.2265	337.3978	741.0352	383.3762	934.8268
B	24024.57	190507.2	130547.3	435819.9	128927.4	483352.4
d _o	2.68	3.02	3.03	3.250	2.96	3.18
E _o	-0.175	-0.331	-0.218	-0.315	-0.285	-0.452

	F	Cl	Br	I	AH
A	224.7778	752.4295	854.4915	1216.361	129.6811
B	56642.43	450756.5	529098.3	941178.8	24024.57
d _o	2.82	3.26	3.277	3.401	2.68
E _o	-0.223	-0.314	-0.345	-0.393	-0.175

<u>Phosphorous</u>						
	H	C	N	P	O	S
A	264.8311	875.9980	41.0352	1474.461	725.8096	1562.526
B	78276.24	616858.8	435819.9	1709147.	343864.0	1436368.
d _o	2.897	3.348	3.250	3.638	3.13	3.50
E _o	-0.225	-0.311	-0.315	-0.318	-0.383	-0.425

	F	Cl	Br	I	AH
A	477.570	1032.84	1420.615	1994.231	264.8311
B	331501.9	716969.2	1233586.	2138148.	78276.24
d _o	3.34	3.56	3.467	3.591	2.897
E _o	-0.172	-0.372	-0.409	-0.465	-0.224

Table II. Values for the Nonbonded Potential Function (cont'd.)

<u>Oxygen</u>						
	H	C	N	P	O	S
A	123.7239	446.1175	383.3762	725.8096	347.0114	776.8652
B	17088.45	132680.5	128927.4	343864.0	87258.63	331604.2
d _O	2.55	2.90	2.96	3.13	2.82	3.08
E _O	-0.225	-0.311	-0.285	-0.383	-0.345	-0.455

	F	Cl	Br	I	AH
A	222.7845	743.9333	873.6900	1244.813	132.7183
B	49239.03	368957.9	462066.7	825990.9	19571.27
d _O	2.76	3.16	3.192	3.315	2.58
E _O	-0.252	-0.375	-0.413	-0.469	-0.225

<u>Sulfur</u>						
	H	C	N	P	O	S
A	289.9912	998.1056	934.8268	1562.526	776.8652	1683.851
B	95562.36	680474.6	483352.4	1436368.	331604.2	157683.0
d _O	2.95	3.33	3.18	3.50	3.08	3.5
E _O	-0.220	-0.366	-0.452	-0.452	-0.455	-0.458

	F	Cl	Br	I	AH
A	474.7878	1609.260	1561.998	2189.843	289.9912
B	231917.1	1610520.	1379998.	2383406.	95562.36
d _O	3.15	3.55	3.477	3.600	2.95
E _O	-0.243	-0.402	-0.442	-0.503	-0.220

Table II. Values for the Nonbonded Potential Function (cont'd.)

<u>Fluorine</u>						
	H	C	N	P	O	S
A	79.67815	264.9297	224.7778	477.570	222.7845	474.7878
B	9336.187	99698.52	56642.43	331501.9	49239.03	231917.1
d _o	2.783	3.016	2.82	3.34	2.76	3.15
E _o	-0.170	-0.176	-0.223	-0.172	-0.252	-0.243

	F	Cl	Br	I	AH
A	135.7213	456.2523	381.9460	544.8931	79.10228
B	37137.62	295690.5	187993.1	337396.0	9201.722
d _o	2.86	3.30	3.154	3.277	2.48
E _o	-0.124	-0.176	-0.194	-0.220	-0.170

<u>Chlorine</u>						
	H	C	N	P	O	S
A	272.4885	906.5424	752.4295	1032.84	743.9333	1609.260
B	80706.48	636883.	450756.5	716969.2	368957.9	1610520.
d _o	2.898	3.346	3.26	3.56	3.16	3.55
E _o	-0.230	-0.323	-0.314	-0.372	-0.375	-0.402

	F	Cl	Br	I	AH
A	456.2523	788.0722	1068.890	1503.2822	272.4885
B	295690.5	510738.3	855184.3	1486748.	80706.48
d _o	3.30	3.70	3.420	3.543	2.898
E _o	-0.176	-0.304	-0.334	-0.380	-0.230

Table II. Values for the Nonbonded Potential Function (cont'd.)

<u>Bromine</u>						
	H	C	N	P	O	S
A	299.1173	1043.719	854.4915	1420.615	873.6900	1561.998
B	88410.22	767148.0	529098.3	1233586.	462066.7	1379998.
d ₀	2.897	3.372	3.277	3.467	3.192	3.477
E ₀	-0.253	-0.355	-0.345	-0.409	-0.413	-0.442

	F	Cl	Br	I	AH
A	381.9460	1068.890	1388.129	1938.128	299.1173
B	187993.1	855184.3	1309037.	2246615.	88410.22
d ₀	3.154	3.420	3.515	3.638	2.897
E ₀	-0.194	-0.334	-0.368	-0.418	-0.253

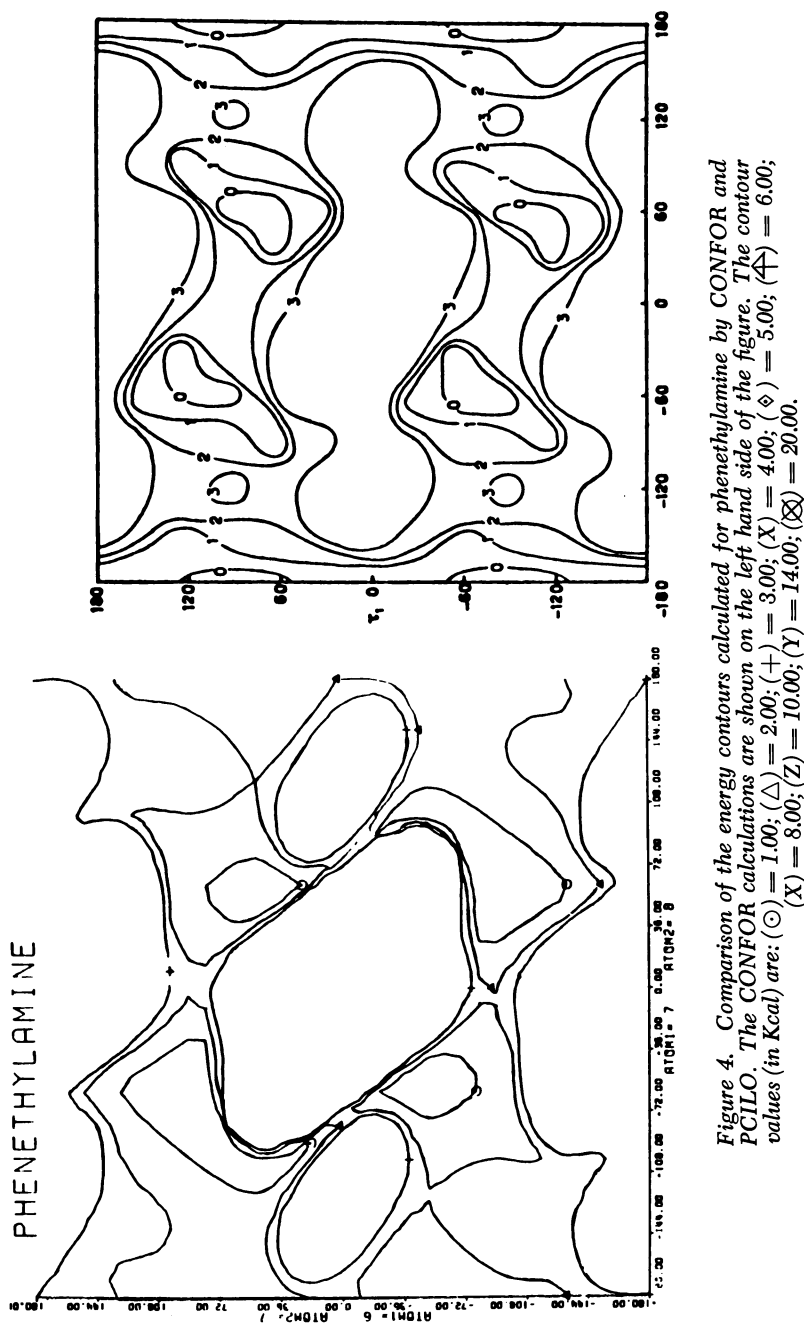
<u>Iodine</u>						
	H	C	N	P	O	S
A	437.8515	1475.163	1216.361	1994.231	1244.813	2189.843
B	166418.3	1346600.	941178.8	2138148.	825990.9	2383406.
d ₀	3.021	3.496	3.401	3.591	3.315	3.600
E ₀	-0.288	-0.404	-0.393	-0.465	-0.469	-0.503

	F	Cl	Br	I	AH
A	544.8931	1503.2822	1938.128	2693.000	437.8515
B	337396.0	1486748.	2246615.	3816973.	166418.3
d ₀	3.277	3.543	3.638	3.762	3.021
E ₀	-0.220	-0.380	-0.418	-0.475	-0.288

Table II. Values for the Nonbonded Potential Function (cont'd.)

	<u>AH</u>					
	H	C	N	P	O	S
A	32.80541	175.3848	129.6811	264.8311	132.7183	289.9912
B	1217.415	33004.10	24024.57	78276.24	19571.27	95562.36
d _O	2.05	2.687	2.68	2.897	2.58	2.95
E _O	-0.221	-0.233	-0.175	-0.224	-0.225	-0.220

	F	Cl	Br	I	AH
A	79.10228	272.4885	299.1173	437.8515	22.22401
B	9201.722	80706.48	88410.22	166418.3	629.9831
d _O	2.48	2.898	2.897	3.021	1.96
E _O	-0.170	-0.230	-0.253	-0.288	-0.196



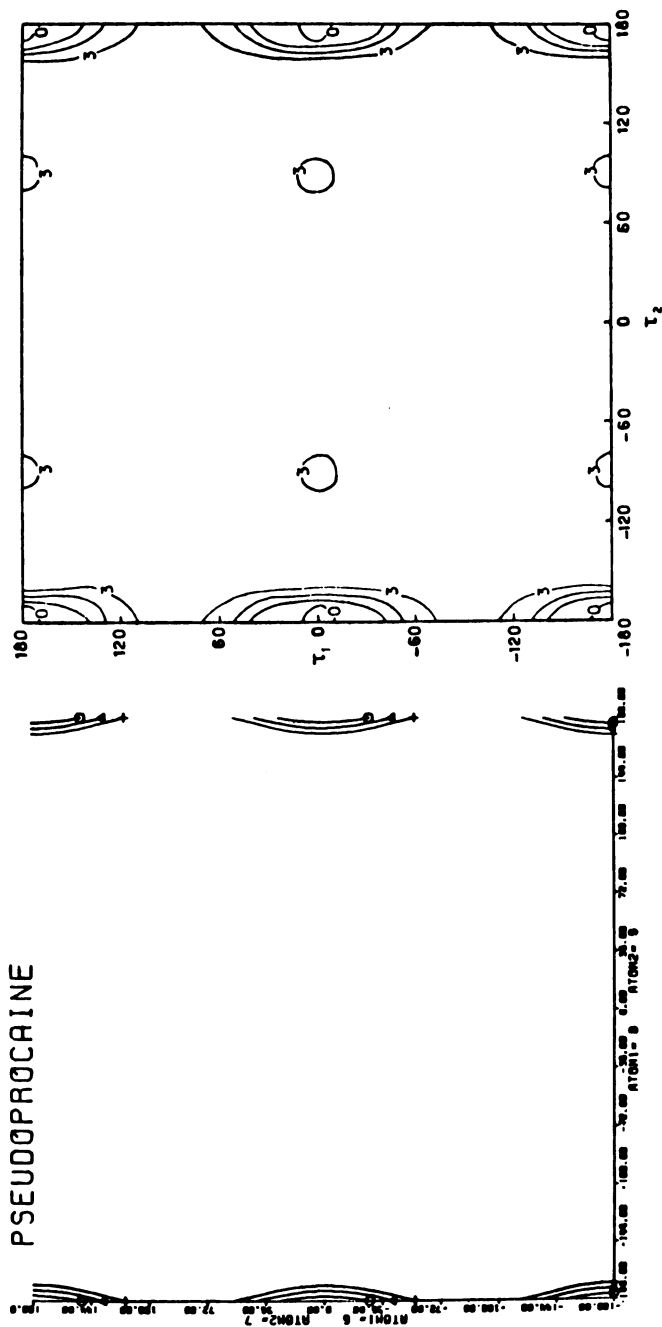


Figure 5. Comparison of the energy contours calculated for pseudoProcaine by CONFOR and PCILO (see Figure 4 for the meaning of the contour symbols)

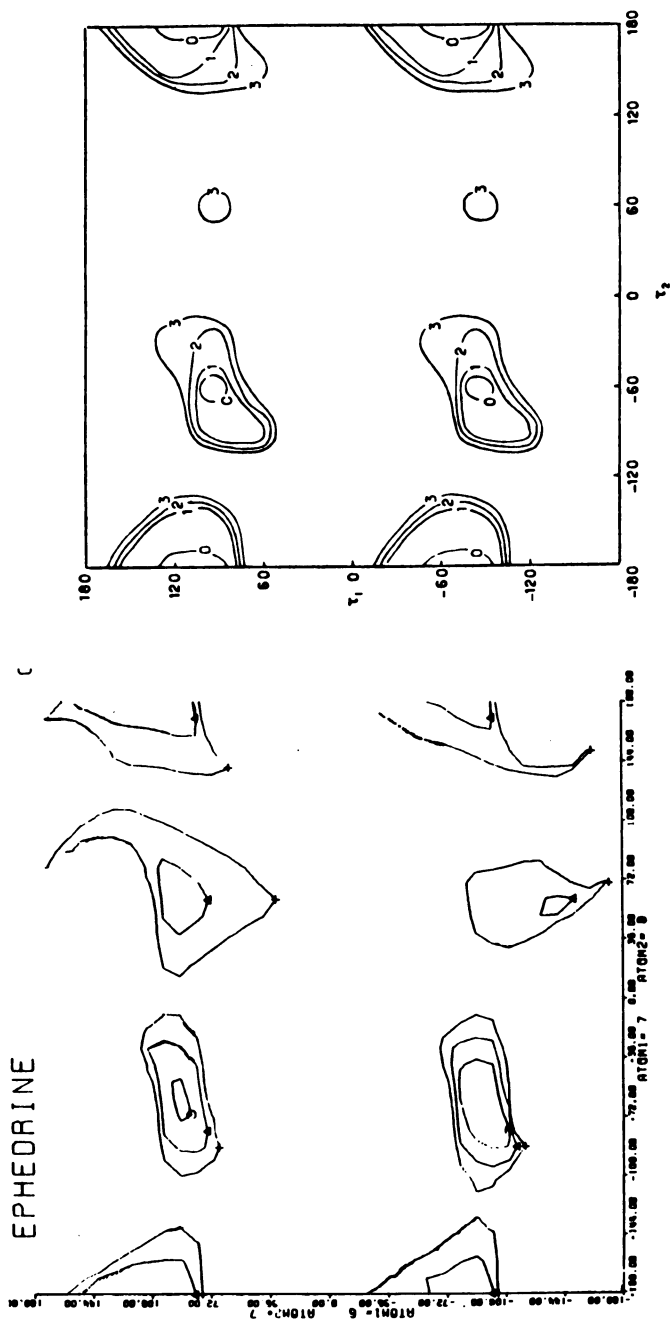


Figure 6. Comparison of the energy contours calculated for ephedrine by CONFOR and PCILO (see Figure 4 for the meaning of the contour symbols)

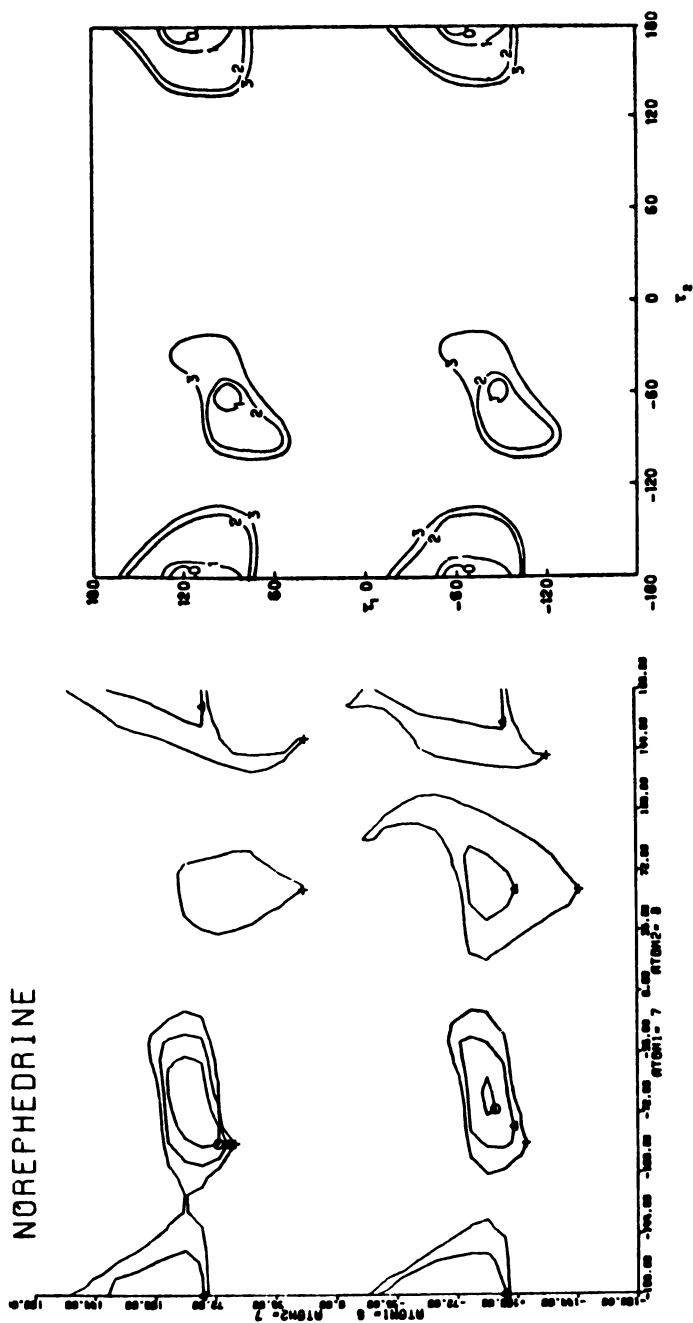


Figure 7. Comparison of the energy contours calculated for norephedrine by CONFOR and PCILO (see Figure 4 for the meaning of the contour symbols)

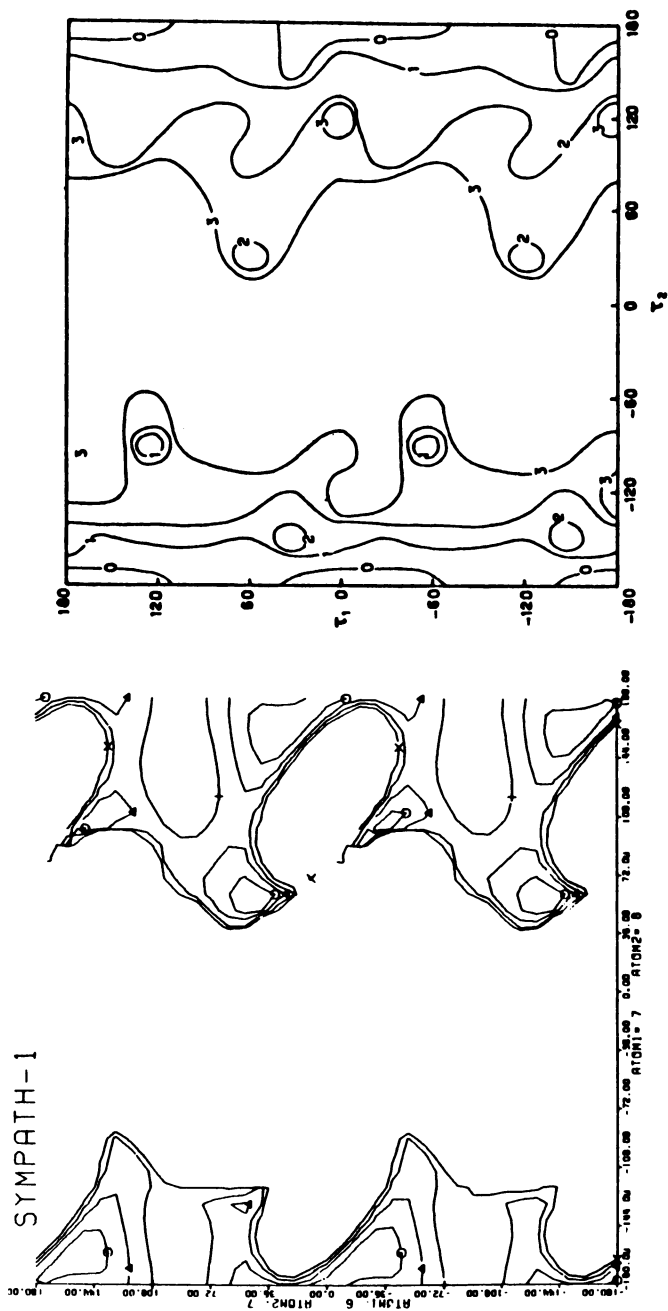


Figure 8. Comparison of the energy contours calculated for sympath-1 by CONFOR and PCILO (see Figure 4 for the meaning of the contour symbols)

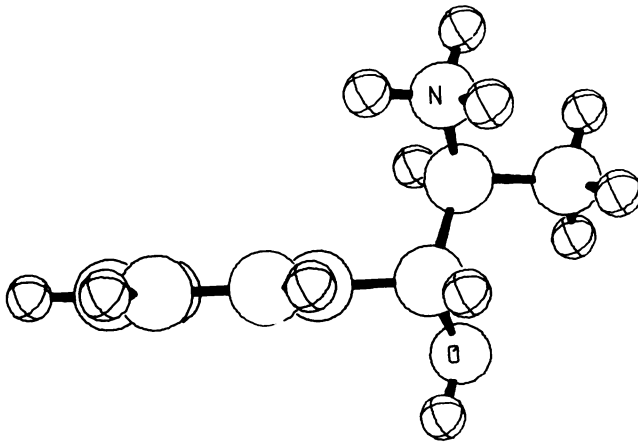


Figure 9. View of norephedrine showing the lack of any repulsive interactions

this configuration. You will note that there are no obviously repulsive interactions. It appears that PCILO simply finds the alternative states to be more attractive than does our method.

The map for Sympath-1 (Figure 3g) appears to locate the minima properly, however, the shape is somewhat different from that calculated by PCILO. We show a larger repulsion about the center of the map. This could be a function of the resolution used in the PCILO calculations, but more likely involves basic differences between the two methods. For our purposes this map is sufficiently accurate.

Comparisons for Other Compounds

For the purpose of testing the applicability of the CONFOR routine to structural types other than those used in the parameterization, we have performed many calculations. We present three such comparisons below.

A particularly interesting set of calculations are those we performed on the glycyl and valyl amino acid residues. Pullman has reported work on both molecules (12). In order to compare our results we reproduced these calculations using the CONFOR module. Figure 10 shows a comparison of the maps for the glycyl residue. The dots on the maps represent known conformations of glycyl residues in globular proteins.

In the case of the valyl residue PCILO predicts a global minimum where our method predicts a maximum. As seen in Figure 11 the conformation of the valyl residue in globular proteins occupies the minimum regions located by CONFOR and what would be a local minima in the PCILO method. The barrier between $\Psi = 270$, $\phi = 60$ and $\psi = 210$, $\phi = 300$ shown by CONFOR is due to the close approach of a hydrogen from the isopropyl and the oxygen of the carbonyl. The distance being only 1.81Å. This is well into the repulsive state of the Van der Waals interaction. Figure 12 shows one view of this interaction. Clearly for this to be a minimum some sort of bonding must be occurring. Apparently PCILO indicates a weak hydrogen bond between this hydrogen and the carbonyl oxygen. While there is some evidence for the existence of C-H---O hydrogen bonds, such bonds are experimentally observed only when the carbon also bears at least one strong electron withdrawing group (13). Initially the isopropyl group was fixed in the same conformation as that used by PCILO. If we allow the isopropyl group to relax in order to remove the carbonyl-hydrogen interaction, we see a corresponding lowering of the barrier; however, we are not able to obtain a local minimum in this region. This leads us to question whether PCILO has indicated a realistic minima for this residue. The contour map for the relaxed isopropyl group is shown in Figure 13. It is interesting to note that the calculation for the relaxed structure required the computation of 186,624 separate conformations. Such a computation is certainly impracticable using the more time consuming semi-

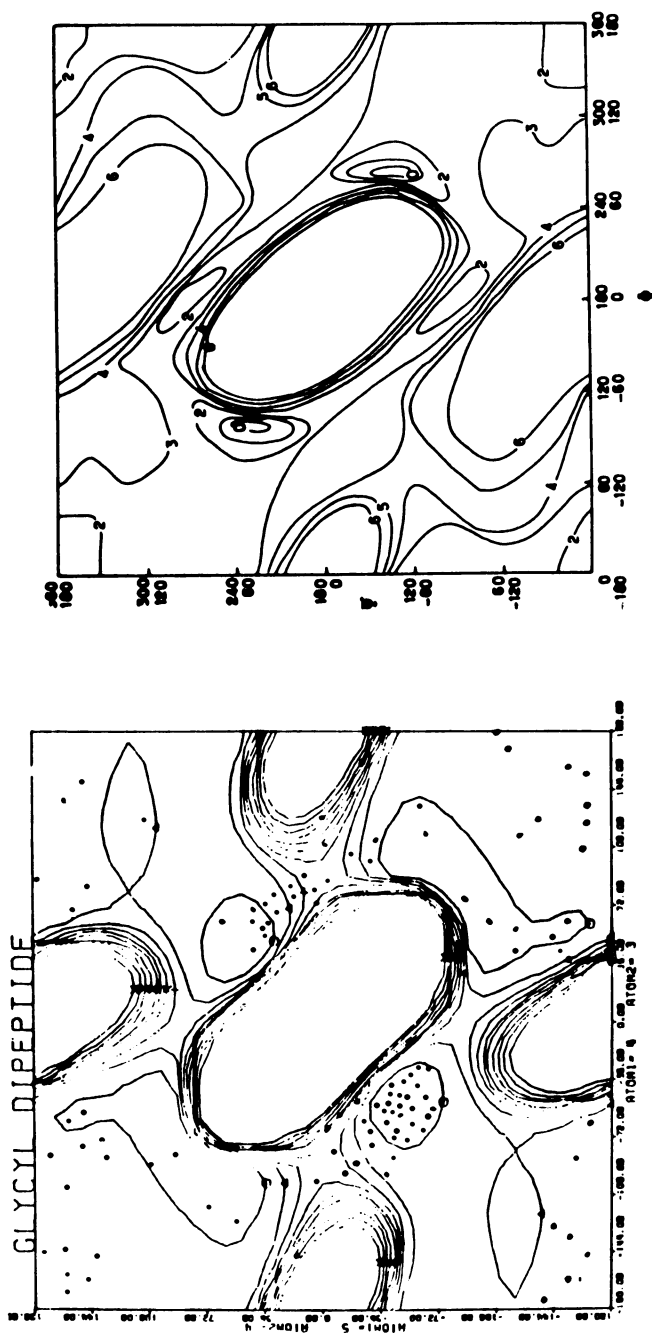


Figure 10. Comparison of the energy contours calculated for the Glycyl residue (see Figure 4 for the meaning of the contour symbols)

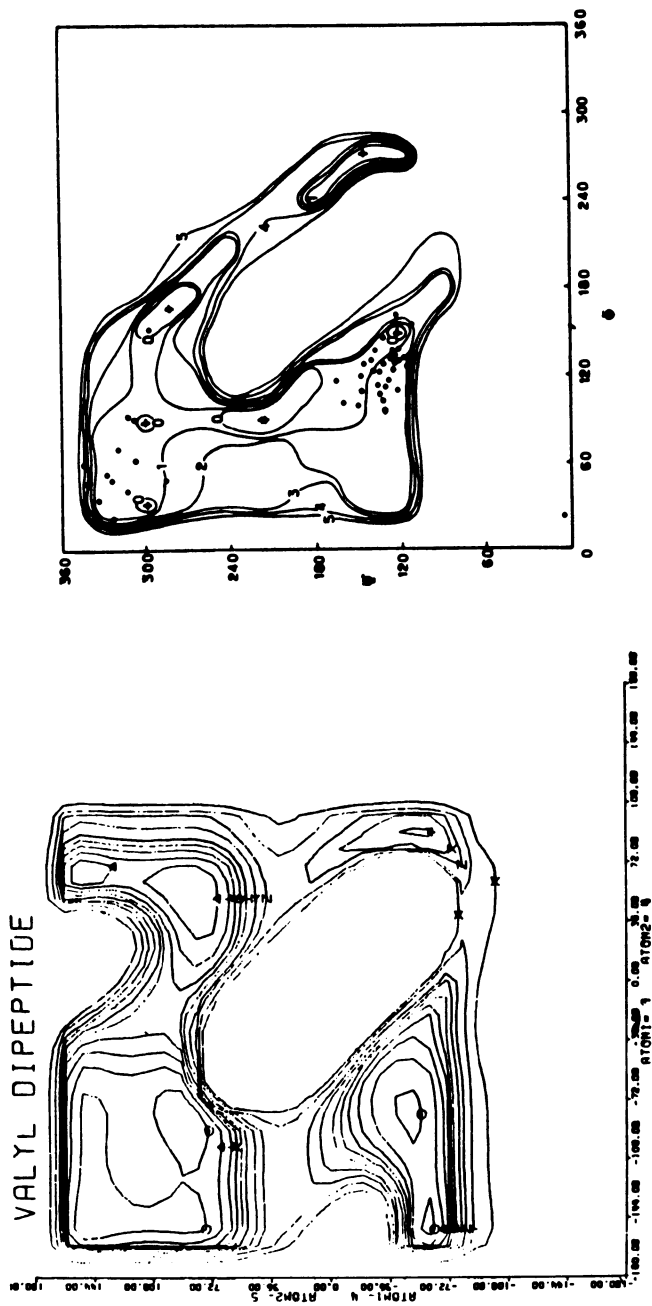


Figure 11. Comparison of the energy contours calculated for the valyl residue (see Figure 4 for the meaning of the contour symbols)

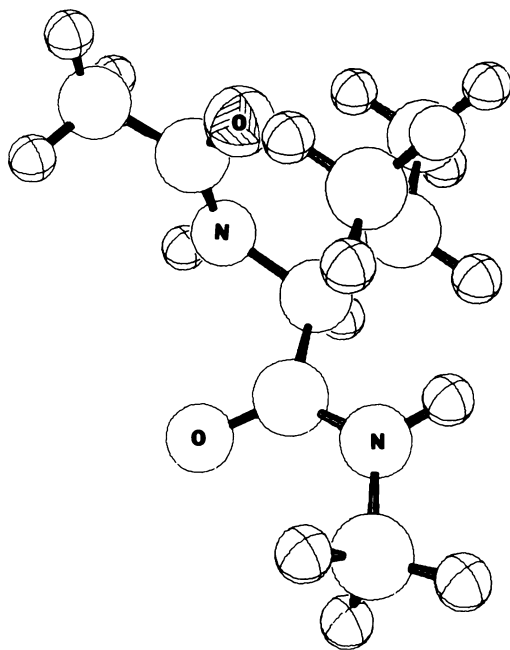
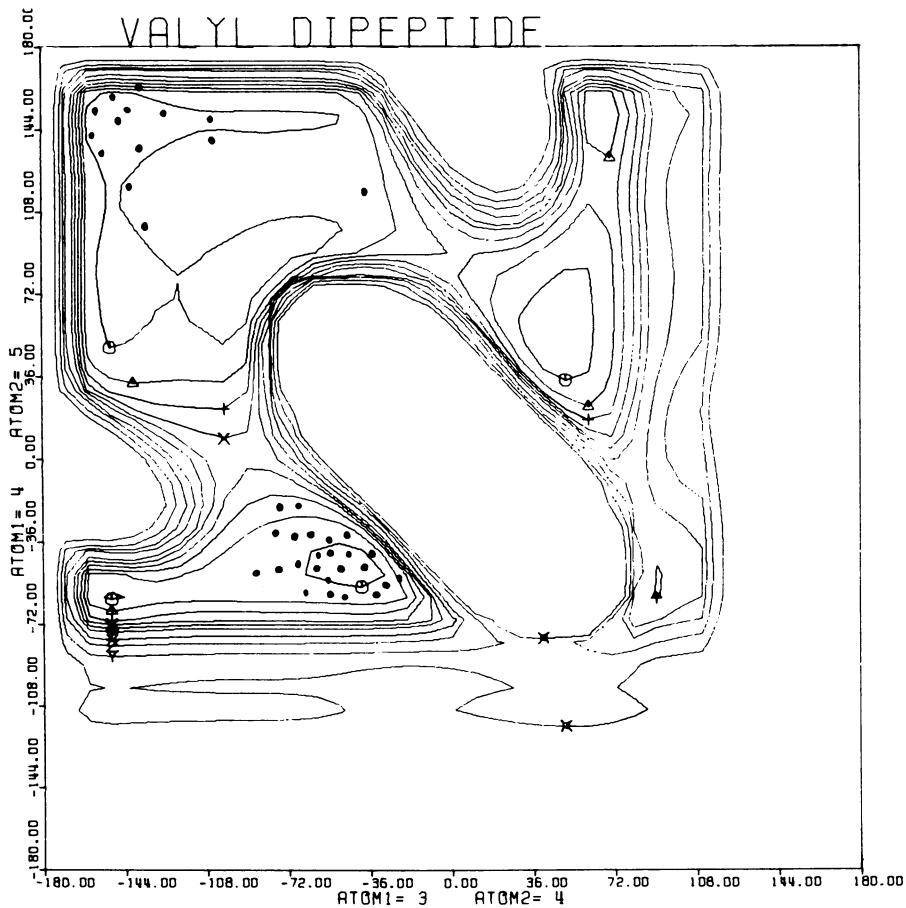


Figure 12. View of the valyl residue showing the interaction between the carbonyl and one of the hydrogens on the isopropyl. This conformation is the minimum energy conformation as calculated by PCILO (Figure 11).



CONTOUR VALUES

○	1.00
△	2.00
+	3.00
×	4.00
◇	5.00
⊕	6.00
×	8.00
Z	10.00
Y	14.00
×	20.00

Figure 13. Contour map for the valyl residue calculated using CONFOR. This map was generated by allowing the isopropyl group to relax and seek a minimum energy conformation.

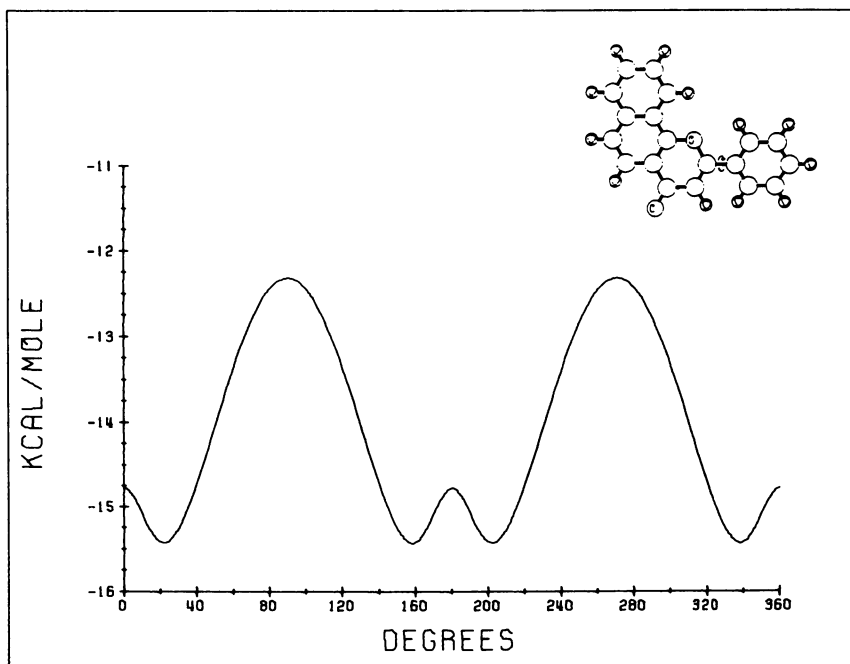


Figure 14. Energy vs. rotational angle plot for 7,8-benzflavone. The global minimum is located at 22°, the measured angle was 23°.

empirical methods.

As a final example of the utility of the program, we report a calculation on 7,8 benzflavone. The crystal structure of this molecule has been reported by Glusker et al (14). Figure 14 shows the map of energy versus rotation angle for the phenyl ring. The global minimum was predicted to be at 22 degrees. The actual value from the crystal structure was 23 degrees.

Conclusions

The parameterization of our conformational analysis program appears to give results which are largely in agreement with the semi-empirical quantum mechanical PCILO technique. The method is capable of producing data which is reasonably accurate. This tool promises to be of great aid in studying the steric requirements of a drug.

Literature Cited

1. C. Hansch, "A Quantitative Approach to Biochemical Structure-Activity Relationships," *Acc. Chem. Res.*, 2, 232 (1969).
2. W. V. Valkenburg (Ed.), "Biological Correlations - The Hansch Approach," *Advances in Chemistry Series*, No. 114, American Chemical Society, Washington, D.C., 1972.
3. B. Pullman and P. Courri re, "Molecular Orbital Studies on the Conformation of Pharmacological and Medicinal Compounds," the Jerusalem Symposium on Quantum Chemistry and Biochemistry V, 547 (1973).
4. R. F. McGuire, F. A. Momany and H. A. Scheraga, "Energy Parameters in Polypeptides V an Empirical Hydrogen Bond Potential Function Based on Molecular Orbital Calculations," *Jour. Phys. Chem.*, 76(3), 375 (1972).
5. H. J. R. Weintraub and A. J. Hopfinger, "Conformational Analysis of Some Phenethylamine Molecules," *J. Theor. Biol.*, 41, 53 (1973).
6. F. A. Momany, L. M. Carruthers, R. F. McGuire, and H. A. Scheraga, "Energy Parameters in Polypeptides VII. Geometric Parameters, Partial Atomic Charges, Nonbonded Interactions, Hydrogen Bond Interactions, and Intrinsic Torsional Potentials for the Naturally Occurring Amino Acids," *Jour. Phys. Chem.*, 79(22), 2361 (1975).
7. J. L. Coubeils and B. Pullman, "Quantum-Mechanical Study of the Conformational Properties of Drugs with Local Anesthetic Action," *Mol. Pharm.* 8, 278 (1972).

8. J. L. Coubeils, P. Courri re and B. Pullman, "Quantum-Mechanical Study of the Conformational Properties of Sympatholytic Compounds," *J. Med. Chem.* 15(5), 453 (1972).
9. B. Pullman, J. L. Coubeils, P. Courri re, and J. P. Gervois, "Quantum-Mechanical Study of the Conformational Properties of Phenethylamines of Biochemical and Medicinal Interest," *Jour. Med. Chem.* 15(1), 17 (1972).
10. A. J. Hopfinger, "Conformational Properties of Macromolecules," Academic Press, New York, 1973.
11. F. A. Momany, L. M. Carruthers, R. F. McGuire and H. A. Scherage, "Intermolecular Potentials from Crystal Data III. Determination of Empirical Potentials and Application to the Packing Configurations and Lattice Energies in Crystals of Hydrocarbons, Carboxylic Acids, Amines, and Amides," *Jour. Phys. Chem.*, 78(16), 1595 (1974).
12. C. B. Anfinsen and John T. Edsall (Ed.), "Advances in Protein Chemistry," Academic Press, New York, pp. 348-562 (1974).
13. D. June Sutor, "Evidence for the Existence of C-H---O Hydrogen Bonds in Crystals." *Jour. Chem. Soc.*, 1105 (1963).
14. Personal communication, Dr. J. P. Glusker.

RECEIVED June 8, 1979.

New Optimal Strategies for Ab-Initio Quantum Chemical Calculations on Large Drugs, Carcinogens, Teratogens, and Biomolecules

JOYCE J. KAUFMAN, HERBERT E. POPKIE, and P. C. HARIHARAN

Department of Anesthesiology, The Johns Hopkins University School of Medicine, Baltimore, MD 21205, and Department of Chemistry, The Johns Hopkins University, Baltimore, MD 21218

Our group has a long standing interest in quantum chemical calculations on large drug and biomolecules. The first all valence three-dimensional calculations reported on any such molecules were extended Hückel calculations on pyridine aldoxime antagonists to organophosphorus intoxicants we presented at the IIIrd International Pharmacological Congress, Sao Paulo, Brazil, July 1966 (1, 2). These calculations were carried out with a program we wrote ourselves in 1963 from the original papers of Wolfsberg and Helmholtz, 1952 (3), and Eberhardt, Crawford and Lipscomb, 1956 (4), the same papers from which the extended Hückel method of Hoffman (5) was derived and which he named "extended Hückel". We even wrote in 1965-1966 an iterative scheme for charge consistency in the extended Hückel method. In spite of our early computational capability our study on the pyridine aldoxime antagonists was the one and only time we ever used the extended Hückel method for calculations on drugs, biomolecules or any biological, medical or pharmacological problem. The reason that we never used the extended Hückel method again for drug and biological systems was that we were aware of the deficiencies of the method. It did not (and still to this day does not) always give a calculated energy minimum even for stretching a bond and since the method does not include electron-electron repulsion it cannot distinguish between singlet and triplet states (or doublet and quartet states, etc.). It is incredible that in this day and age extended Hückel calculations (iterative or not) are still being carried out on drugs, biomolecules or carcinogens. It is a method whose time is long since past.

Instead, in 1963 we were already beginning to carry out ab-initio LCAO-MO-SCF calculations on sizable rocket fuel molecules, such as B_2H_6 with 56 Gaussian basis functions (6), using an early version of the POLYATOM program supplied to us by the late Professor John C. Slater, which we converted for use on an IBM 7094. After using that POLYATOM program for only a few sizable molecules it became obvious for larger systems it would be advantageous to write a newer more efficient Gaussian integral program. We wrote

0-8412-0521-3/79/47-112-415\$05.25/0

© 1979 American Chemical Society

the MOSES program (Molecular Orbitals) (7) partly so named because Moses was a prophet and such a program could predict molecular properties. The MOSES program was efficient and went through several versions up through contracted d orbitals by 1967. It implemented perhaps among the most efficient algorithms for calculating integrals over contracted Gaussian basis functions and symmetry adapted contracted basis functions at that time and probably even today. When Clementi was writing his first IBMOL program he asked us to check the accuracy of his integrals program against those we calculated using our MOSES program. Subsequently the integrals section of the MOSES program was given to H. Basch when he was a postdoctoral at Bell Labs and he incorporated it into a later version of the POLYATOM program. Meanwhile, in 1964 we began using a CDC 6400 computer and in 1965 a CDC 6600 computer with an old CHIPPEWA compiler--before a Scope compiler was even available. The MOSES program was converted immediately to run on the CDC 6400 and CDC 6600 as soon as we got access to them, and from that day to this we remain devoted fans and great exponents of the CDC 6600 series and now of the CDC 7000 series and the CYBER's. We have always endeavored to make optimal use of their 60 bit word length. The MOSES program could handle 175 basis functions in a CDC 6600 computer. However, CDC 6000 computer time was a very limited commodity to us in those years and we were interested still in carrying out quantum chemical calculations on large molecules.

Thus in 1964 we derived semi-rigorous LCAO-MO-SCF methods for three-dimensional molecular calculations including electron-electron repulsion (8). This paper was presented at the January 1965 Sanibel International Symposium on Atomic, Molecular and Solid State Theory where Pople presented his CNDO method and our paper appears back to back in the same Symposium issue of the Journal of Chemical Physics in which Pople's first papers on CNDO and NDDO appeared (9, 10). There were four degrees of neglect of integrals in our paper. The first and simplest resembled CNDO, the next INDO, the third NDDO and there was yet another less approximate scheme. We programmed up the method and tried it on a few compounds. We were still convinced that ab-initio computations were the only really appropriate method for quantum chemical calculations on drugs and biological molecules. Having always only a very small group we decided to devote our major efforts to ab-initio quantum chemical calculations on large drug and biomolecules and especially to deriving and implementing new and better strategies for ab-initio calculations on such molecules.

In the intervening years for computational tractability we carried out some calculations with less rigorous quantum chemical techniques [CNDO (11), INDO (11), PCILO (12)] as well as topological* and topographical† analyses on a variety of large drugs to understand and predict their conformational profiles, electronic structures and properties and the mechanism of action of their biological effects. [*Topological similarity--indicates that in

different molecules the same kind of atoms are connected in exactly the same way. Oftentimes a common topological pattern of atoms is buried within a series of much larger molecules. †Topographical similarity--indicates that in different molecules two portions of the molecule bear the same spatial relationship to one another--although not necessarily topologically related.] These included LCAO-MO-SCF calculations on neuroleptics (the major tranquilizers) (13-20), narcotics and narcotic antagonists (13, 16, 18-23). We were able to identify the topographical and electronic requisites for neuroleptic action and we carried out quantum chemical calculations on a hypothetical series of new effective neuroleptics (14, 15, 16, 17, 18). Two years after our original presentation, six such compounds were reported as synthesized and animal tested and they were all effective neuroleptics (24). We carried out the first all-valence electron calculations on narcotics, CNDO/2 calculations on morphine, which were presented at the Sanibel International Symposium on Atomic, Molecular and Solid State Theory January 1972 and appeared in that Symposium issue of the Int. J. Quantum Chem. (13) and we followed that shortly with CNDO/2, INDO and PCILO calculations on a variety of narcotics and narcotic antagonists presented at the Sanibel International Symposium January 1974 and published in that Symposium issue of the Int. J. Quantum Chem. (21).

In connection with these earlier quantum chemical studies we had also extended the derivations of some of the semi-rigorous techniques [such as deriving the necessary expressions for including d orbitals in the INDO method (25)], suggested procedures to improve other methods [how to improve the description of lone pairs in the PCILO method (26)] and performed calculations by various techniques comparing the results of less rigorous calculations with our large basis set ab-initio calculations (27,28). We thus have an excellent grasp of the limits of validity of less than ab-initio quantum chemical calculations compared to ab-initio calculations and compared to experiment.

We also carried out a few completely ab-initio calculations on the neuroleptic chlorpromazine and on promazine (29) and on the narcotic morphine and the narcotic agonist-antagonist nalorphine (30) using IBMOL 6 (31). However, even though IBMOL 6 had been written especially to carry out ab-initio calculations on large molecules, and we had improved its use by such techniques as repacking the integral tapes so that for the first six iterations we read only integrals 10^{-4} a.u. or larger, for the next six iterations all integrals 5×10^{-6} a.u. or larger, and for the final few iterations all integrals 10^{-7} a.u. or larger, the calculations mentioned above that we did with IBMOL 6 convinced us that we were going to have to develop something still better to make computationally tractable routine ab-initio calculations on large drug and biomolecules.

Ab-Initio MODPOT/VRDDO/MERGE Technique

It was in 1974 that we began development of our new series of Gaussian integral programs especially designed for large molecules: MOLASYS (Molecular Orbital Calculations on Large Systems) (32, 33) and GIPSY (Gaussian Integral Program System) (34, 35).

MOLASYS and GIPSY calculate s,p,d and f integrals in blocks, taking full advantage of the information common to members of the same block. We also use the full 60 bit CDC word to store small two electron integrals indexing the integral also in the same word using the least significant bits for indexing and converting the integral to floating point when it is used. We have compared the timings directly with those of other general s,p, and d integral packages in current usage (and which we ourselves had used previously). Our new integral program, GIPSY, for s,p and d integrals is of the order of four times faster than Raffanetti's BIGGMOLI (36, 37) program and twice as fast as BIGGMOLI for s and p integrals only. Our GIPSY program is of the order of eight times faster for s,p and d integrals than POLYATOM (38) [improved CDC 6600 version (39)] and four times faster than POLYATOM (39) for s and p integrals only. Indirect timing comparisons (40) indicate that our GIPSY program is twice as fast as GAUSSIAN 70 (41) for general s and p integrals with different exponents and of the same speed as GAUSSIAN 70 when the same exponents are used in GAUSSIAN 70 for s and p functions. The same algorithm is used in GAUSSIAN 76 (42). Recently there has been introduced a new method especially efficient for evaluation of Gaussian integrals for basis functions of higher angular momentum (43) and the authors incorporated it into a new program called HONDO (43). Indirect timing comparisons for s,p and d integrals reported (43) for HONDO compared to PHANTOM (44) (a QCPE CDC 6600 version of POLYATOM) and to BIGGMOLI (36, 37) indicate that our GIPSY s,p and d integral program is probably a factor of 1.3 faster than HONDO. An algorithm for d orbitals similar to that of HONDO is used in the newer version of GAUSSIAN 76 (42). We plan to check the timing comparison for f orbitals between GIPSY and HONDO as soon as feasible.

But even more importantly, our programs incorporate several very desirable features as time-saving options to our ab-initio computer programs: a charge conserving integral prescreening approximation (VRDDO) especially effective for spatially extended molecules; ab-initio effective core model potentials (MODPOT) which enable one to treat only the valence electrons explicitly yet accurately; and a MERGE technique which enables us to retain all the integrals for a skeletal fragment which remains unchanged and just recalculate the new integrals for added atoms or their geometry variations both for intra- and intermolecular calculations.

Our VRDDO approximation (variable retention of diatomic differential overlap) was inspired in part by a suggestion from solid state physics by Wilhite and Euwema (45). The approximation consists of neglecting all one-electron integrals (both energy and overlap) and two-electron integrals that involve basis function pairs $\phi_i(1)\phi_j(1)$ whose pseudo-overlap:

$$S_{ij}^* = \int \phi_i^*(1) \phi_j^*(1) dv_1$$

is less than some threshold τ_1 . The pseudo-function $\phi_i^*(1)$ is represented by a single normalized spherically-averaged Gaussian function whose exponent is equal to that of the most diffuse primitive in the contracted Gaussian function $\phi_i(1)$. Our method for calculating the pseudo-overlap S_{ij}^* for the basis function pair $\phi_i(1)\phi_j(1)$ differs somewhat from that used by Wilhite and Euwema (45); however our VRDDO procedure is still charge conserving.

The VRDDO approximation has been introduced (46) into our MOLASYS (32, 33) and GIPSY (34, 35) computer programs. A second threshold τ_2 that controls the accuracy of the two-electron repulsion integrals (integrals are accurate to n decimal places where $\tau_2 = 10^{-n}$) can be specified in the input data. All two-electron integrals whose absolute magnitude is less than τ_2 are neglected. The results (46-57) show this second approximation to be numerically accurate. The computer CPU time required to generate the two-electron integrals for a large molecule is substantially reduced as τ_2 is increased.

We have investigated the effect of varying the thresholds τ_1 and τ_2 in order to arrive at values that yield a good compromise between accuracy and computer running time. We tested the accuracy of the VRDDO method by computing the completely ab-initio reference calculations and numerical results. The introduction of the VRDDO approximation with $\tau_1 = 10^{-2}$ (the pseudo-overlap integral prescreening threshold) and $\tau_2 = 10^{-4}$ hardly affects the accuracy* of the computed properties at all. [*Where accuracy implies accuracy with respect to the completely ab-initio calculations with the same atomic basis set retaining all integrals 10^{-6} a.u. or greater.] The valence orbital energies and gross atomic populations are essentially accurate* to 3 decimal places (in a.u.). The error in the total energy differences between isomers or along a potential energy curve agree to about the fourth decimal place (in a.u.) compared to the same relative energy differences from completely ab-initio calculations. The maximum error found in the total overlap population is 0.008 and in the dipole moment is 0.009 a.u.

The second desirable option in our large molecule computer programs is the introduction of ab-initio effective core model potentials for the inner shell electrons. This permits one to treat only the valence electrons explicitly yet accurately.

Recently Bonifacic and Huzinaga have given a new formulation of the model potential (MODPOT) method for carrying out LCAO-MO-SCF calculations considering only the valence electrons (59-66). The effect of the core electrons is taken into account through the use of a modified model potential (MODPOT) Hamiltonian.

As an example for molecules composed of first and second row atoms and hydrogen it has the following form:

$$\mathcal{H} = \sum_i^n h(i) + \sum_{i>j}^n 1/r_{ij}; \quad (1)$$

$$\begin{aligned}
 h(i) = & -\frac{1}{2}A_i - \sum_k^{n_h} 1/r_{ik} \quad (2) \\
 & - \sum_k^{n_c} (Z_k - N_c^k)(1 + A_{1k}e^{-\alpha_{1k}r_{ik}^2} \\
 & \quad + A_{2k}e^{-\alpha_{2k}r_{ik}^2})/r_{ik} \\
 & + \sum_k^{n_c} \{B_{1s}^k | 1s_k \rangle \langle 1s_k | + B_{2s}^k | 2s_k \rangle \langle 2s_k | \\
 & \quad + B_{2p}^k (| 2p_{xk} \rangle \langle 2p_{xk} | + | 2p_{yk} \rangle \langle 2p_{yk} | \\
 & \quad + | 2p_{zk} \rangle \langle 2p_{zk} |)\}.
 \end{aligned}$$

In equations (1) and (2), n_v is the number of valence electrons, n_h is the number of hydrogen atoms, and n_c is the number of atoms with core electrons. N_c^k is the number of core electrons in the k -th core. $|1s_k\rangle$, $|2s_k\rangle$ and $|2p_k\rangle$ are inner shell atomic orbitals for the k -th core.

Bonifacic and Huzinaga (58, 59) have introduced Gaussian terms into the one-electron part of the Hamiltonian to account for the screening of the nuclei by the core electrons and pseudopotential terms to prevent the buildup of the valence molecular orbitals in the core regions. Bonifacic and Huzinaga derived model potential expressions for inner s , p and d orbitals. We extended the derivation also to inner f orbitals. This MODPOT procedure of Bonifacic and Huzinaga has been introduced as an option into our MOLASYS and GIPSY computer programs. With the GIPSY computer program we can calculate the usual one-electron overlap, kinetic energy, nuclear attraction and two-electron repulsion integrals between s -, p -, d - and f -contracted Gaussian functions. In addition, the new type of one-electron integrals resulting from the use of the Bonifacic-Huzinaga Hamiltonian with s -, p -, d - and f -type core atomic orbitals can be calculated.

We verified the accuracy of the MODPOT method for a variety of molecules by carrying out the reference calculations with two-electron integrals accurate to 6 decimal places. Then the MODPOT calculations were carried out with the valence shell two-electron integrals accurate to 6 decimal places and the MODPOT modified one-electron integrals. The same reference atomic basis set was used for both calculations. Comparison of the MODPOT results to the completely *ab-initio* ones (including inner shell electrons) showed that the valence orbital energies and gross atomic populations are accurate* [see footnote page 419] to more than 2 decimal places (the average error is 0.005 a.u.). The maximum error in

the orbital energies for the largest molecule studied to date is only 0.11 a.u. and in the gross atomic populations is only 0.010 a.u. It should be emphasized that the accuracy obtained with the MODPOT approximation is more than acceptable if one wants to reproduce ab-initio results for large molecules. Moreover, for molecules composed even of two second row atoms, such as Cl_2 , the savings in time between the MODPOT method compared to the completely ab-initio is a factor of ten. The savings in time also increase dramatically as the size of the molecule goes up.

If both the MODPOT and VRDDO approximations are introduced, the accuracy* [*see footnote page 419] with respect to the reference calculations is about the same as that obtained using only the MODPOT approximation. The maximum error in the orbital energies is 0.012 a.u., in the gross atomic populations is 0.009, in the total overlap populations is 0.027 and in the dipole moment is 0.009 a.u.

Since one usually is interested in energy differences rather than the total energy of a single system, it is useful to compare isomer energy differences computed with the various procedures. In the 6-3G reference calculations (48-49) imidazole is more stable than pyrazole by 0.0150 a.u., pyrimidine is more stable than pyridazine by 0.0247 a.u., and pyrazine is more stable than pyridazine by 0.0222 a.u. With the VRDDO method the corresponding energy differences are 0.0153, 0.0258 and 0.0219 a.u. respectively. Use of the MODPOT method yields values of 0.0136, 0.0219 and 0.0189 respectively. Finally, the MODPOT/VRDDO method gives values of 0.0139, 0.0231 and 0.0188 a.u. respectively. Thus, even for the MODPOT/VRDDO method, the agreement with ab-initio results is excellent; 0.0011 a.u. for imidazole-pyrazole, 0.0016 a.u. for pyrimidine-pyridazine and 0.0034 a.u. for pyrazine-pyridazine.

The relative accuracy of the MODPOT/VRDDO method along a potential energy surface (such as for molecular conformational changes or interactions between two species, 0.0001 - 0.0002 a.u.) is even greater than its absolute accuracy.

For one aromatic ring with one substituent (such as NO_2) the VRDDO calculation is 1.5 times faster than the reference calculation (using our own fast ab-initio programs [32, 33, 34, 35], the MODPOT calculation is three times faster and the MODPOT/VRDDO calculation is five times faster). Moreover, the relative increase in speed goes up even more dramatically as the molecules get larger.

We verified the accuracy of the MODPOT method for a variety of molecules (47, 48, 49) by carrying out the reference calculations with two-electron integrals accurate to 6 decimal places. Then the MODPOT calculations were carried out with the valence shell two-electron integrals accurate to 6 decimal places and the MODPOT modified one-electron integrals. The same reference atomic basis set was used for both calculations. Comparison of MODPOT results to the completely ab-initio ones (including inner shell electrons) showed that the valence orbital energies and gross

atomic populations are accurate to more than 2 decimal places (the average error is 0.005 a.u.). Moreover, the accuracy of the MODPOT results compared to ab-initio is much greater (0.005 eV [50]) for calculating excitation energies, electron affinities, total energy differences between isomers, potential energy curves, etc., since these involve only energy differences. Moreover, for molecules composed even of two second row atoms, such as Cl₂, the savings in time between the MODPOT method compared to the completely ab-initio is a factor of ten. The savings in time also increases dramatically as the size of the molecule goes up.

As a test of the efficiency of the method, we ran ab-initio MODPOT/VRDDO calculations on the DNA bases (54):

<u>Molecule</u>	<u>No. Atoms</u>		<u>No. Basis Fns.</u>	<u>Time (minutes)</u>	
	<u>C,N,O</u>	<u>H</u>		<u>Integrals*</u>	<u>SCF†</u>
<u>DNA Constituents</u>					
Cytosine	8	5	37	5.1	2.0
Guanine	11	5	49	11.2	4.3
Thymine	9	6	38	7.4	3.1
Adenine	10	5	45	9.2	3.5

*CDC 6600 computer, FTN, OPT=0. (The integrals run even ~75% faster under OPT=2 on a CDC 6600 computer which supports a full OPT=2 compiler.)

†CDC 6600 computer, FTN, OPT=2.

We have obtained the MODPOT parameters for the formulation we use by matching MODPOT calculations to the completely ab-initio calculations for atoms or small molecules using the same atomic reference basis set. We are currently working on matching these parameters for higher elements. There are a number of methods for treating the effective core potentials, relativistic effective core potentials (and some comparisons and critiques). Space limitation does not permit us to list all these references. For a complete up-to-date bibliography on various model potential methods see our recent NATO Advanced Study Institute lecture (60). However, it appears that the core projection model potentials of the Bonifacic and Huzinaga type we use which yield valence orbitals with the usual oscillatory behavior in the core region may have advantages for calculation of correlation energies over the pseudopotentials giving valence pseudoorbitals which are smooth in the core region.

We have just finished writing a new MERGE technique which allows us to retain the integrals for a skeletal fragment which remains unchanged. This is especially useful in studies of large closely related molecules and their complexes since only an atom or a few atoms are changed and except for the ring which is being

attacked, the structure of the rest of the molecule remains invariant. It was not straightforward to incorporate a MERGE technique into our program since we calculate integrals very efficiently by blocks rather than one by one. We are now using this MERGE option for all our calculations since the systems we are investigating at present are quite large. The MERGE option is useful for large molecules (even one and two ring systems), and exceptionally valuable for larger molecules and where the part to be varied is small compared to a large invariant structural fragment. As an example, while the skeletal integrals for benzopyrene-7,8-dihydrodiol (a 5-ring aromatic proximate carcinogen) took 3005 seconds (CDC 6600, OPT=1) to form the ultimate carcinogen benzo-pyrene-7,8-dihydrodiol-9,10-epoxide, the extra MERGE integrals for the oxygen with the rest of the skeleton took only 560 seconds (CDC 6600, OPT=1) (57). Thus it was computationally extremely tractable for us to vary the stereochemistry of the epoxide ring and its opening to form an attacking agent.

We have recently run ab-initio MODPOT/VRDDO/MERGE calculations on the attack of CH_3^+ , the simplest ultimate carcinogen, on guanine for attack on the O^6 and $\text{N}7$ positions (58) using several different types of CDC computers (CDC 6600 and CYBER 175) with a variety of different operating systems (which did or did not support full OPT=1 and OPT=2 compilers at various times).

Some idea of comparative timings can be seen below.

Times (minutes) necessary to carry out ab-initio MODPOT/VRDDO and ab-initio MODPOT/VRDDO/MERGE calculations

Species	No. Atoms		No. Basis Fns.	Time (minutes)		
	C,N,O	H		Integrals or Skeletal Integrals	MERGE Integrals	SCF
Guanine	11	5	49	11.2*		4.3 [†]
Guanine + CH_3^+	12	8	56	6.8 ^Δ	2.3*	5.0 [†]
				2.9 [∇]	0.9 [∇]	1.9 [#]

* CDC 6600 computer, FTN, OPT=0 (The integrals run even ~75% faster under OPT=1 or OPT=2 on a CDC 6600 computer that supports a full OPT=1 or OPT=2 compiler.)

† CDC 6600 computer, FTN, OPT=2

Δ CDC 6600 computer, FTN, OPT=1

∇ CYBER 175, FTN, OPT=1

CYBER 175, FTN, OPT=2

An unusual behavior occurred in the system CH_3^+ + guanine (G). As the internuclear O or N to Me distance approached 5 bohrs, there was a mixing of electronic configurations. When we saw the electronic configuration mixing, our experience with the theory and previous large scale ab-initio configuration inter-

action calculations on ion-molecule reactions indicated the physical basis had to be due to the fact that while the reactants were $\text{CH}_3^+ + \text{G}$, at the dissociation asymptote there was a lower energy set of fragments $\text{CH}_3 + \text{G}^+$. When we initiated our calculations the experimental IP of guanine had never been reported. However, we knew the photoelectron spectra of DNA bases was being investigated by Le Breton. We called him and indeed the ionization potential of guanine, 8.3 eV (68), is considerably lower than that of CH_3 , 9.8 eV (69).

Using our efficient MERGE technique the time consuming part of computing a series of higher substituted metabolites of a common skeleton or potential energy surfaces as a function of nuclear geometry now becomes the SCF rather than the integrals. Since the Fock matrix elements corresponding to the invariant skeletal integrals remain virtually unchanged (except for a slight dependence on orthogonality) either for a family of substituted molecules or for a large molecule being attacked by a smaller species, we have derived a new strategy based on passing through for the SCF, either intra- or intermolecular, the Fock matrix contributions from the invariant skeletal integrals making up only the additional contributions to the Fock matrix elements from the new integrals, allowing only the contributions from the new integrals to the Fock matrix elements to change until the SCF calculation is about converged, then relaxing and making up new Fock matrix elements from all of the integrals for each cycle until convergence. Considerable numerical testing must now be done with this procedure. It is expected that this technique will work most effectively for systems where no convergence problems would otherwise be expected. If the scheme works well, we plan to extend a similar scheme to do transformations of integrals for CI calculations for series of congeners or for closely spaced points along a potential energy surface.

In order to obtain convergence and compute at least the diabatic continuation past 5 bohrs of the curve describing the close approach of CH_3^+ to guanine, we wrote new more efficient extrapolation and damping routines (70).

Both closed and open-shell systems can be calculated. The programs can handle several hundred basis functions in a CDC 6600 computer and more in a CDC 7600. In addition, we plan to explore further new and novel methods for subsequent calculation on large molecules the size of block regions of DNA and for interaction of reagents with helical DNA by molecular partitioning.

Configuration Interaction Calculations

The MODPOT/VRDDO LCAO-MO-SCF programs have been meshed in with the configuration interaction programs we use. In this CI (71) program, each configuration is a spin- and symmetry-adapted linear combination of Slater determinants in the terms of the spin-bonded functions of Boys and Reeves (72, 73, 74) as formu-

lated originally by Shavitt, Pipano and co-workers (75, 76, 77) and subsequently modified and extended in our laboratory (78, 79, 80).

The determinants can be obtained from single determinant or MC-SCF wave functions.

We have also added a method of calculating improved virtual orbitals. Our use of this procedure for N-electron excited state virtual orbitals (81) in the framework of the SCF calculation of the N-1 electron problem closely resembles those proposed by Huzinaga (82). We have also investigated Huzinaga's recent method for improved virtual orbitals in the extended basis function space (83). This is also a useful procedure where there are convergence problems for the Hartree-Fock calculations for the N-electron occupied space of the excited states. This should also be helpful in optimizing virtual orbitals to use them in perturbation theory expressions.

The most time and space consuming part and the limiting factor of a CI calculation is the transformation of integrals from integrals over atomic basis functions to integrals over molecular orbitals. We have implemented a technique which incorporates into an effective CI Hamiltonian operator the effect of all molecular orbitals from and to which no excitations are made (84).

Perturbation Procedures for Molecular Interactions

It would be desirable to be able to calculate the majority of the molecular interaction surface between large molecules by a perturbation technique rather than by the supermolecule technique.

Previously we had shown by comparison with accurate ab-initio calculations the non-applicability of the customary approximate monopole-transition moment long range force expressions (86, 87) as well as the non-applicability of CNDO and INDO supermolecule potential energy surfaces themselves for molecular interactions even when the interacting supermolecule system is correctly expressible by a single determinant wave function. The same applies to NDDO, MINDO (88), MNDO (89, 90) and all other ZDO methods. Even worse, in recent work presented on carcinogenesis by others the MINDO/3 method (which was derived to describe static thermochemistry) was used to describe the reaction pathway of insertion of singlet oxygen into an R-H bond. Firstly, no ZDO single determinant method is appropriate even when the lowest energy surface of the system separates correctly to closed shell ground state species. Secondly, the use of any single determinant wave function is not correct to describe potential energy surfaces for reactions of this O insertion type, since there is no way the lowest energy curve can dissociate correctly in a single determinant. Thirdly, there are five singlet potential energy surfaces arising from $O\ ^1D_g + R-H$ which lie lower in energy at the dissociation asymptote than the singlet potential energy surface arising from $O\ ^1S_g + R-H$ and all five will mix at some points in the potential

energy surfaces for the reaction with the singlet surface arising from $O\ ^1S_g + R-H$. There is no way to describe such processes correctly except by ab-initio CI (or ab-initio MODPOT/VRDDO/CI or ab-initio MODPOT/VRDDO/MERGE/CI calculations). A great problem (perhaps the greatest) in establishing and keeping the credibility of quantum chemical calculations in biology and medicine is the poor quality of much of the quantum chemical calculations carried out in biological, pharmacological and medical areas-- and what is even worse, incorrect use of quantum chemistry in these areas.

We are examining the question of various rigorous perturbation expressions (91) for intermolecular interactions and are currently doing the analysis and writing a program to calculate these (92). We shall test the validity of this latter procedure by comparison with accurate ab-initio supermolecule calculations for smaller test systems.

The electrostatic, exchange, charge-induced multipole, multipole-induced multipole, and charge-transfer contributions can also be calculated by partitioning of the SCF calculation. The dispersion energy must be calculated or estimated separately. We shall also give consideration to an alternative formulation for molecular interactions for both single determinant and multideterminant wave functions and test these against ab-initio calculations.

In addition there are correlation energy terms (ΔE_{corr}) both intra- and inter-molecular. The largest change in interaction of two non-polar molecules is in the inter-molecular dispersion energy term resulting from the instantaneous polarization of A and B.

Over the years various approximate formulas for interactions between large molecules have been derived from perturbation theory (91). The better of such perturbation theory expansions customarily include a short-range first order "exchange" term and long range terms (electrostatic, polarization and dispersion). Various approximations (such as the multi-centered multipole expansion, representation of transition densities by bond dipole and the decomposition of molecular polarizability into bond polarizabilities, the use of atomic polarizabilities, bond-bond interaction terms, etc.) have been introduced for the calculation of certain of the terms.

A careful analysis of the SCF energy decomposition for an inter-molecular interaction of A and B of various perturbation energy expressions indicate that for certain perturbation formulations there is a one-to-one correspondence between certain SCF energy decomposition terms and certain terms in the perturbation expressions. Thus one can calculate the values for the terms from energy decomposition of ab-initio or ab-initio MODPOT/VRDDO SCF wave functions and compare these to the values for the same type term resulting from the perturbation theory expressions. Care must be taken to correct for possible basis set incompleteness.

Once both the form of the various perturbation theory terms and the approximations that may be involved in estimating these terms have been validated by comparison with ab-initio MODPOT/VRDDO calculations, then the most appropriate perturbation theory expressions can be used in preliminary calculations of intermolecular interactions for the various reactants.

This same perturbation theory approach is capable of describing both inter- and intra-molecular interactions.

Electrostatic Molecular Potential Contour Maps

For molecular interactions involving molecules having net charges or permanent dipoles, useful information may be drawn by examination of the electrostatic potential, arising from one of the partners, and by simple electrostatic calculations, involving a potential and a simplified description of the charge distribution of the other molecule involved in the interaction (93, 94).

The electrostatic potential $V(r)$ gives more detailed and less ambiguous information than a population analysis, being a function computed in the overall molecular surrounding space. From the potential definition, the first order interaction energy between the molecular charge distribution and a point charge distribution is

$$W(\{r_i\}) = \sum_i q(r_i) V(r_i)$$

Such an expression has previously been used for comparative purposes, for the study of interaction between two molecular species, by computing the electrostatic potential of the first partner and by assuming some point charge model as representative of the charge distribution of the second partner. We also plan to extend this concept in a more subtle way by using an electron density contour map to describe the charge distribution of the second partner as a function of the space surrounding this second partner.

We were able to show that ab-initio large basis set and ab-initio optimized small basis set wave functions generate electrostatic molecular potential contour maps almost identical in shape and magnitude (95). We also confirmed the earlier findings by other researchers that CNDO/2 and INDO wave functions, either orthogonal or deorthogonalized, did not generate maps which compared correctly with those generated from ab-initio wave functions (95, 96, 97). Those CNDO/2 and INDO maps are especially inaccurate for aromatic compounds and even for isolated double bonds (96, 97).

Such maps are reflective of dynamic reaction indices and prove more accurate indicators of the positions favored for electrophilic attack than do such static indices as charges on the atoms.

Abstract

The optimal strategy for ab-initio quantum chemical calculations on large molecules (drugs, carcinogens, teratogens and biomolecules) is somewhat different than for smaller molecules. It is possible to take advantage of certain computational properties of large molecules and of series of closely related congeners of large molecules. We have implemented several desirable options into our fast ab-initio Gaussian integral programs (SCF, MC-SCF and CI): the use of ab-initio effective core model potentials (MODPOT) which enable one to calculate only the valence electrons explicitly yet accurately, a charge conserving integral prescreening evaluation (VRDDO) especially effective for spatially extended molecules and a new MERGE technique which enables reuse of integrals from a common skeletal fragment and only recalculation of integrals involving new positions of certain atoms or new atoms. We have also recently derived a new SCF strategy which passes through Fock matrix elements as a starting Fock matrix, reutilizes the Fock matrix elements from the common skeletal integrals, takes advantage of the fact that the common Fock matrix elements change little during the preliminary iterations. We have also derived new MC-SCF and CI integral transformation strategies which take advantage of behavior similar to that utilized above. Examples are given from our recent research. Supported in part by NCI Contract N01-CP-75929 and by NINCDS Contract N01-NS-5-2320.

Acknowledgment

This research was supported in part by NCI under Contract N01-CP-75929. We thank our NCI contract monitor, Dr. Kenneth Chu, for his perceptiveness in appreciating the contributions quantum chemical studies could make to the field of chemical carcinogenesis.

This research was supported in part by NINCDS under Contract N01-NS-5-2320. We thank our NINCDS contract monitor, Dr. Ernst Freese, for his appreciation of the fundamental insight quantum chemical calculations, both intra- and inter-molecular, could make to the processes leading to teratogenesis and to the general problems of membrane transport and biomolecular interactions.

We also thank CDC for a grant of CYBER 175 computer time. The timings for ab-initio MODPOT/VRDDO/MERGE calculations on CH_3^+ + guanine using the CYBER 175 are very encouraging for the future of ab-initio calculations on large drugs and biomolecules as well as on carcinogens and teratogens.

Literature Cited

1. Giordano, W., Hamann, J. R., Harkins, J. J., and Kaufman, Joyce J., "Quantum Mechanically Derived Electronic Distributions in the Conformers of 2-PAM," in "Physico-Chemical Aspects of Drug Action," Ed. Ariens on the Proceedings of the Third International Pharmacological Congress, Sao Paulo, Brazil, July 1966, Volume 7, 327-354, Pergamon Press, 1968.
2. Giordano, W., Hamann, J. R., Harkins, J. J., and Kaufman, Joyce J., "Quantum Mechanical Calculations of Stability in 2-Formyl N-Methyl Pyridinium (Cation) Oxime (2-PAM⁺) Conformers," Mol. Pharmacol. (1967), 3, 307-317.
3. Wolfsberg, M., and Helmholtz, L., J. Chem. Phys. (1952), 20, 837.
4. Eberhardt, W. H., Crawford, B., Jr., and Lipscomb, W. N., J. Chem. Phys. (1954), 22, 989.
5. Hoffman, R., J. Chem. Phys. (1963), 39, 1397.
6. Burnelle, L., and Kaufman, Joyce J., "Molecular Orbitals of Diborane in Terms of a Gaussian Basis," J. Chem. Phys. (1965), 43, 3540-3545.
7. Sachs, L. M., and Geller, M., "MOSES, A Fortran IV System for Polyatomic Molecules," Int. J. Quantum Chem. (1967), 15, 445.
8. Kaufman, Joyce J., "Semi-rigorous LCAO-MO-SCF Methods for Three-Dimensional Molecular Calculations Including Electron-Electron Repulsion," J. Chem. Phys. (1965), 43, S152-S156.
9. Pople, J. A., Santry, D. P., and Segal, G. A., J. Chem. Phys. (1965), 43, S129.
10. Pople, J. A., and Segal, G. A., J. Chem. Phys. (1965), 43, S136.
11. Pople, J. A., and Beveridge, D. A., Approximate Molecular Orbital Theory, McGraw-Hill, New York, 1970.
12. Diner, S., Malrieu, J. P., and Gilbert, M., Theor. Chim. Acta (1968), 15, 100 and references therein.
13. Kaufman, Joyce J., and Kerman, Ellen, "Quantum Chemical Calculations on Antipsychotic Drugs and Narcotic Agents," Int. J. Quantum Chem. (1972), 6S, 319-335.
14. Kaufman, Joyce J., "Quantum Chemical and Theoretical Techniques for the Understanding of Action of Psychoactive Drugs," Proceedings of the Symposia at the VIII Congress of the Collegium Internationale Neuro-Psychopharmacologicum on "Psychopharmacology, Sexual Disorders and Drug Abuse," 31-42, North-Holland Publishing Co., Amsterdam-London, Avicenum, Czechoslovak Medical Press, Prague, 1973.
15. Kaufman, Joyce J., and Kerman, Ellen, "Quantum Chemical and Other Theoretical Techniques for the Understanding of the Psychoactive Action of the Phenothiazines," International Conference on Phenothiazines and Related Drugs, Rockville, Maryland, June 25-28, 1973, in "Advances in Biochemical Pharmacology, Vol. 9: Phenothiazines and Structurally Related Drugs," 55-75, Eds. Irene S. Forrest, C. J. Carr, and E. Usdin, Raven Press, New York, 1974.

16. Kaufman, Joyce J., and Kerman, Ellen, "Quantum Chemical and Theoretical Techniques for the Understanding of Action of Drugs Which Affect the Central Nervous System: Antipsychotics, Narcotics and Narcotic Antagonists," International Symposium on Chemical and Biochemical Reactivity, Jerusalem, Israel, April 9-13, 1973, in "The Jerusalem Symposia on Quantum Chemistry and Biochemistry, VI," 523-547, Jerusalem, Israel, 1974.
17. Kaufman, Joyce, and Kerman, Ellen, "The Structure of Psychotropic Drugs (Including Theoretical Prediction of a New Class of Effective Neuroleptics)," Int. J. Quantum Chem.: Quantum Biology Symp. (1974), 1, 259-289.
18. Kaufman, Joyce J., and Koski, W. S., "Physicochemical, Quantum Chemical and Other Theoretical Techniques for the Understanding of the Mechanism of Action of CNS Agents: Psychoactive Drugs, Narcotics and Narcotic Antagonists and Anesthetics," in "Drug Design," Vol. V., 251-340, Ed. E. J. Ariens, Academic Press, New York, 1975.
19. Kaufman, Joyce J., "Quantum Chemical and Theoretical Techniques for the Understanding of Action of Psychoactive Drugs and Narcotic Agents," International Conference on Computers in Chemical Research and Education, Ljubljana, Yugoslavia, July 12-17, 1973, in Comput. Chem. Res. Educ. Proc. (1973), 3, 5/1-5/20.
20. Kaufman, Joyce J., and Koski, W. S., "Physicochemical, Quantum Chemical and Other Theoretical Studies of the Mechanism of Action of CNS Agents: Anesthetics, Narcotics and Narcotic Antagonists, and Psychotropic Drugs," Int. J. Quantum Chem.: Quantum Biology Symp. (1975), 2, 35-57.
21. Kaufman, Joyce J., Kerman, Ellen, and Koski, W. S., "Quantum Chemical, Other Theoretical and Physicochemical Studies on Narcotics and Narcotic Antagonists to Understand Their Mechanism of Action," Int. J. Quantum Chem.: Quantum Biology Symp. (1974), 1, 289-313.
22. Saethre, L. J., Carlson, T. A., Kaufman, Joyce J., and Koski, W. S., "Nitrogen Electron Densities in Narcotics and Narcotic Antagonists by X-Ray Photoelectron Spectroscopy and Comparison with Quantum Chemical Calculations," Mol. Pharm. (1975), 11, 492-500.
23. Kaufman, Joyce J., and Kerman, Ellen, "Conformational Profile of Nalorphine by PCILO Calculations," Int. J. Quantum Chem. (1977), 11, 181-184.
24. Janssen, P. A. J., CINP Meeting, Congress Internationale de Neuropsychopharmacologie, Paris, France, July, 1974.
25. Kaufman, Joyce J., and Predney, R., "Extensions of INDO Formalism to d Orbitals and Parameters for Second Row Atoms," Int. J. Quantum Chem. (1972), 6S, 231-242.
26. Kaufman, Joyce J., "A Suggested Procedure to Improve the Description of Lone Pairs in the PCILO or More General Ab-Initio Perturbative Configuration Interaction Schemes Based

- on Localized Orbitals," Workshop Letter, Int. J. Quantum Chem.: Quantum Biology Symp. (1974), 1, 197.
27. Preston, H. J. T., and Kaufman, Joyce J., "Ab-Initio SCF Calculations on Pyrrole and Pyrazole," Int. J. Quantum Chem. (1973), 7S, 207-215.
 28. Preston, H. J. T., Kerman, Ellen, Kaufman, Joyce J., and Cusachs, L. C., "Comparison for Pyrrole and Pyrazole of Orbital Energies and Population Analyses from Ab-Initio SCF, CNDO/2, INDO, Extended Hückel and ARCANA Calculations," Int. J. Quantum Chem. (1973), 7S, 249-260.
 29. Popkie, H. E., and Kaufman, Joyce J., "Ab-Initio LCAO-MO-SCF Calculations of Chlorpromazine and Promazine," Int. J. Quantum Chem. (1976), 10, 569-580.
 30. Popkie, H. E., Koski, W. S., and Kaufman, Joyce J., "Ab-Initio LCAO-MO-SCF Calculations on Morphine and Nalorphine and Comparison with Their Measured Photoelectron Spectra," J. Am. Chem. Soc. (1976), 98, 1342-1345.
 31. Clementi, E., and Popkie, H. E., IBMOL, 6.
 32. Popkie, H. E., "MOLASYS: A Computer Program for Molecular Orbital Calculations on Large Systems," The Johns Hopkins University, 1974.
 33. Popkie, H. E., "MOLASYS-MERGE," The Johns Hopkins University, 1978.
 34. Popkie, H. E., "GIPSY (Gaussian Integral Program System): A System of Computer Programs for the Evaluation of One- and Two-Electron Integrals Involving s-, p- and d-type Contracted Cartesian Gaussian Functions," The Johns Hopkins University, 1975.
 35. Popkie, H. E., "GIPSY (Gaussian Integral Program System): A System of Computer Programs for the Evaluation of One- and Two-Electron Integrals Involving s-, p- and d- and f-type Contracted Cartesian Gaussian Functions," The Johns Hopkins University, 1976.
 36. Raffenetti, R. C., J. Chem. Phys. (1973), 58, 4452.
 37. Raffenetti, R. C., Chem. Phys. Letts. (1973), 19, 335.
 38. Csizmadia, I. G., Harrison, M. C., Moskowitz, J. W., and Sutcliffe, B. T., Theoret. Chim. Acta (1966), 6, 191.
 39. Hornbeck, C., New York University, private communication.
 40. Petrongolo, C., private communication, 1975.
 41. Hehre, W. J., Lathan, W. A., Ditchfield, R., Newton, M. D., and Pople, J. A., "Gaussian 70, Program 236, QCPE," Indiana University, Bloomington, Indiana.
 42. Pople, J. A., and coworkers, "Gauss 76," National Resource for Computational Chemistry, Berkeley, California.
 43. Dupuis, M., Rys, J., and King, H. F., J. Chem. Phys. (1976), 65, 111.
 44. Goutier, D., Macauley, R., and Duke, A. J., "PHANTOM: Program 241, QCPE," Indiana University, Bloomington, Indiana.
 45. Wilhite, D. L., and Euwema, R. N., Chem. Phys. Letts. (1973), 20, 610.

46. Popkie, H. E., and Kaufman, Joyce J., "Test of Charge Conserving Integral Approximations for a Variable Retention of Diatomic Differential Overlap (VRDDO) Procedure for Semi Ab-Initio Molecular Orbital Calculations on Large Molecules," Int. J. Quantum Chem.: Quantum Biology Symp. (1975), 2, 279-288.
47. Kaufman, Joyce J., "Quantum Chemical Calculations on Small, Medium and Large Molecules," Presented at the Summer Conf. in Theoretical Chemistry, Boulder, Colorado, June 1975.
48. Popkie, H. E., and Kaufman, Joyce J., "Molecular Calculations with the VRDDO, MODPOT and MODPOT/VRDDO Procedures. I. HF, F₂, HCl, Cl₂, Formamide, Pyrrole, Pyridine and Nitrobenzene," Int. J. Quantum Chem. Symp. Issue (1976), 10, 47-57.
49. Popkie, H. E., and Kaufman, Joyce J., "Molecular Calculations with the VRDDO, MODPOT and MODPOT/VRDDO Procedures. II. Cyclopentadiene, Benzene, Diazoles, Diazines and Benzonitrile," J. Chem. Phys. (1977), 66, 4827-4831.
50. Popkie, H. E., and Kaufman, Joyce J., "Molecular Calculations with the MODPOT, VRDDO and MODPOT/VRDDO Procedures. III. MODPOT/SCF + CI Calculations to Determine Electron Affinities of Alkali Metal Atoms," Chem. Phys. Letts. (1977), 47, 55-58.
51. Kaufman, Joyce J., Popkie, H. E., and Preston, H. J. T., "Ab-Initio and Approximately Rigorous Calculations on Small, Medium and Large Molecules," Int. J. Quantum Chem. (1977), 11, 1005-1015.
52. Popkie, H. E., and Kaufman, Joyce J., "Molecular Calculations with the MODPOT, VRDDO and MODPOT/VRDDO Procedures. IV. Boron Hydrides and Carboranes," Int. J. Quantum Chem. (1977), 12, 937-961.
53. Popkie, H. E., and Kaufman, Joyce J., "Molecular Calculations with the MODPOT, VRDDO and MODPOT/VRDDO Procedures. V. Ab-Initio and MODPOT LCAO-MO-SCF Calculations on the Chloro-fluoromethanes," Int. J. Quantum Chem. (1977), S11, 433-443.
54. Kaufman, Joyce J., and Popkie, H. E., "Molecular Calculations with Non-Empirical Ab-Initio MODPOT, VRDDO and MODPOT/VRDDO Procedures. VI. Nucleic Acid Constituents: A, G, C, T," Presented at the Sixth Canadian Conference on Theoretical Chemistry, Fredericton, N. B., Canada, June 1977. Manuscript in preparation for publication.
55. Kaufman, Joyce J., Popkie, H. E., and Preston, H. J. T., "Molecular Calculations with the Non-Empirical Ab-Initio MODPOT, VRDDO and MODPOT/VRDDO Procedures. VII. Prototype Normal Neurotransmitters and Their Metabolites," Presented at the Sixth Canadian Conference on Theoretical Chemistry, Fredericton, N. B., Canada, June 1977.
56. Kaufman, Joyce J., Popkie, H. E., and Preston, H. J. T., "Molecular Calculations with the Ab-Initio Non-Empirical MODPOT, VRDDO and MODPOT/VRDDO Procedures. VIII. Charge Delocalization in the Anions of Aromatic Carboxylic Acids and Phenolic Compounds," Int. J. Quantum Chem. (1978), S12, 283-291.

57. Kaufman, Joyce J., Popkie, H. E., Palalikit, S., and Hariharan, P. C., "Molecular Calculations with the Ab-Initio Non-Empirical MODPOT, VRDDO and MODPOT/VRDDO Procedures. IX. Carcinogenic Benzo(a)pyrene and Its Metabolites Using a MERGE Technique," Int. J. Quantum Chem. (1978), 14, 793-800.
58. Hariharan, P. C., Popkie, H. E., and Kaufman, Joyce J., "Molecular Calculations with the Ab-Initio Non-Empirical MODPOT, VRDDO, and MODPOT/VRDDO Procedures. X. The Attack of the Simplest Ultimate Carcinogen, CH_3^+ , on Guanine by a MERGE Technique and the Fundamental Difference Between Methylating versus Ethylating Carcinogens," Presented at the Sanibel International Symposium on Quantum Biology and Quantum Chemistry, Palm Coast, Florida, March 1979. In press Int. J. Quantum Chem.: Quantum Biology Symp. Issue.
59. Bonifacic, V., and Huzinaga, S., J. Chem. Phys. (1974), 60, 2779.
60. Ibid., J. Chem. Phys. (1975), 62, 1507.
61. Ibid., J. Chem. Phys. (1975), 62, 1509.
62. McWilliams, D., and Huzinaga, S., J. Chem. Phys. (1975), 63, 4678.
63. Huzinaga, S., personal discussion, August 1975.
64. Huzinaga, S., "Atomic and Molecular Calculations with Model Potential Method," Presented at the Quantum Chemistry Symposium, first North American Chemical Congress, Mexico City, December 1975.
65. Bonifacic, V., and Huzinaga, S., J. Chem. Phys. (1976), 64, 956.
66. Huzinaga, S., private communication, July 1977.
67. Kaufman, Joyce J., "Potential Energy Surfaces for Ion-Molecule Reactions," Presented at the NATO Advanced Study Institute on Ion-Molecule Reactions, La Baule, France, September 1978. In press in the Proceedings of the Meeting.
68. Le Breton, P. R., private communication, November 1978.
69. Rosenstock, H. M., Draxl, K., Steiner, B. W., and Herron, J. T., J. Phys. Chem. Ref. Data (1977), 6, Supplement 1, 1-87.
70. Hariharan, P. C., The Johns Hopkins University, 1979.
71. Shavitt, I., "The Method of Configuration Interaction," in "Modern Theoretical Chemistry, Vol. II, Electronic Structure Ab-Initio Methods," Ed. H. F. Schaefer, Plenum Press, New York, 1976.
72. Reeves, C. M., Ph.D. Thesis, Cambridge University, 1957.
73. Reeves, C. M., Commun. ACM (Assoc. Comput. Mach.) (1966), 9, 276.
74. Cooper, I. L., and McWeeny, R., J. Chem. Phys. (1966), 45, 226.
75. The basic CI programs described below are those written originally by I. Shavitt and A. Pipano and continued by A. Pipano while he was a postdoctoral in our group at The Johns Hopkins University 1970-1971. In addition we have subsequently written a number of additional features to these CI programs 1971-present.

76. Gershgorin, Z., and Shavitt, I., Int. J. Quantum Chem. (1967), 1S, 403.
77. Pipano, A., and Shavitt, I., Int. J. Quantum Chem. (1968), 2, 741.
78. Pipano, A., and Kaufman, Joyce J., The Johns Hopkins University, unpublished, 1971.
79. Preston, H. J. T., and Kaufman, Joyce J., The Johns Hopkins University, 1972-present.
80. Raffanetti, R. C., Preston, H. J. T., and Kaufman, Joyce J., The Johns Hopkins University, 1973.
81. Hunt, W. J., and Goddard, W. A., III, Chem. Phys. Letts. (1969), 3, 413.
- 82a. Huzinaga, S., and Arnau, C., Phys. Rev. (1970), A1, 1285.
 - b. Huzinaga, S., and Arnau, C., J. Chem. Phys. (1971), 54, 1948.
 - c. Hirao, K., and Huzinaga, S., Chem. Phys. Letts. (1977), 45, 55.
83. Huzinaga, S., and Hirao, K., J. Chem. Phys. (1977), 66, 2157.
- 84a. Raffanetti, R. C., The Johns Hopkins University, 1973.
 - b. Raffanetti, R. C., ICASE, 1974-1976.
85. Kaufman, Joyce J., and Predney, R., "Non-Applicability for the $\text{Li}^+ - \text{H}_2$ Ion Molecule System of an INDO Potential Surface or of the Approximate Monopole-Transition Moment Long Range Force Expressions," Int. J. Quantum Chem. 5, 235.
- 86a. Rein, R., and Pollak, M., J. Chem. Phys. (1967), 47, 2039.
 - b. Pollak, M., and Rein, R., J. Chem. Phys. (1967), 47, 2045.
 - c. Rein, R., Claverie, P., and Pollak, M., Int. J. Quantum Chem. (1968), 2, 129.
 - d. Claverie, P., and Rein, R., Int. J. Quantum Chem. (1969), 3, 537.
87. Pullman, B., Claverie, P., and Caillet, J., Proc. Natl. Acad. Sci. USA (1966), 55, 911.
88. Dewar, M. J. S., and Dougherty, R. C., "The PMO Theory of Organic Chemistry," Plenum Publishing Corp., New York, 1975.
89. Dewar, M. J. S., and Thiel, W., J. Am. Chem. Soc. (1977), 99, 4899.
90. Ibid., (1977), 99, 4970.
91. Claverie, P., "Elaboration of Approximate Formulas for the Interactions Between Large Molecules: Applications in Organic Chemistry," in "Intermolecular Interactions: From Diatomic to Biopolymers," 69-305, Ed. B. Pullman, John Wiley and Sons, New York, 1978. (This review is extensive and contains several hundred references covering both rigorous and approximate treatments.)
92. Sokalski, W. A., Hariharan, P. C., and Kaufman, Joyce J., The Johns Hopkins University, research in progress, 1979.
93. Scrocco, E., and Tomasi, J., Top. Curr. Chem. (1973), 21, 97, and references therein.
94. Petrongolo, C., Gazz. Chim. Ital. (1978), 108, 445, and references therein.
95. Petrongolo, C., Preston, H. J. T., and Kaufman, Joyce J.,

- "Ab-Initio LCAO-MO-SCF Calculations of the Electrostatic Molecular Potential of Chlorpromazine and Promazine," Int. J. Quantum Chem. (1978), 13, 457-468.
96. Kaufman, Joyce J., "Recent Physicochemical and Quantum Chemical Studies on Drugs of Abuse and Relevant Biomolecules," in "Quantitative Structure Activity Relationships of Analgesics, Narcotic Antagonists and Hallucinogens," 250-277, Eds. G. Barnett, M. Trsic and R. E. Willette, NIDA Research Monograph 22, DHEW, NIDA, 1978.
97. Petrongolo, C., Preston, H. J. T., Popkie, H. E., and Kaufman, Joyce J., The Johns Hopkins University, 1975-1978.

RECEIVED June 8, 1979.

Application of CNDO/2 Calculations and X-Ray Crystallographic Analysis to the Design of Conformationally Defined Analogs of Methamphetamine

GARY L. GRUNEWALD, MARY WEIR CREESE, and D. ERIC WALTERS

Department of Medicinal Chemistry, University of Kansas, Lawrence, KS 66045

A major objective of our research program is to elucidate the molecular events involved in the interaction of sympathomimetic amines with the adrenergic neuroeffector complex. A major portion of this work is aimed at the delineation of the conformational requirements for optimal interaction of sympathomimetic amines with their various physiologically significant interactive sites within the synaptic area (pre- and post-synaptic receptors, presynaptic uptake and storage sites and metabolic enzymes). Such knowledge would be prerequisite for rational design of specific chemical agents to selectively modify neurotransmitter-receptor interactions and to eventually replace drugs of less specificity.

Any discussion of structure-activity relationships among sympathomimetic amines must be undertaken in terms of the multiplicity of physiologically important sites within the noradrenergic synapse with which such amines can potentially interact. A generalized synapse of this type is depicted in Figure 1, and consists of the following generally accepted points of molecular physiology.

Consider first the presynaptic events. Depolarization of the nerve ending results in elevated intraneuronal Ca^{++} levels and concomitant exocytotic release of norepinephrine from the storage granule (1,2,3) where norepinephrine is stored in high concentrations as an osmotically inert complex with ATP (4,5). Although most neuronal norepinephrine exists in this vesicular pool, there is also a small extravesicular pool which has been interpreted in terms of norepinephrine ionically bound to soluble cytoplasmic constituents (6). The cytoplasmic pool of norepinephrine in fact is maintained at its low level by vesicular storage of norepinephrine and by oxidative deamination

0-8412-0521-3/79/47-112-439\$12.25/0

© 1979 American Chemical Society

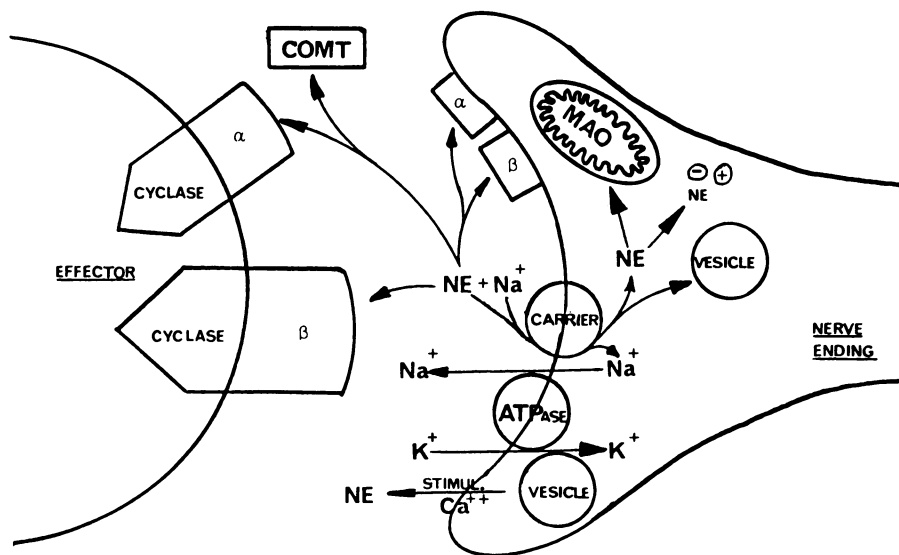


Figure 1. The noradrenergic nerve ending

of norepinephrine via mitochondrial monoamine oxidase (MAO). Thus, inhibition of MAO (pargyline) provides a means of swelling the extravesicular norepinephrine compartment (7). The vesicular pool can be obliterated altogether by treatment with reserpine which irreversibly inhibits the vesicular uptake of norepinephrine (8,9). While depolarization releases only vesicular norepinephrine in a Ca^{++} -dependent manner (10), indirectly acting sympathomimetic agents will displace norepinephrine from both storage forms in a Ca^{++} -independent manner (11,12).

Norepinephrine released into the synaptic area is rendered inactive either by O-methylation (primarily meta but some para) via catechol O-methyltransferase (13) or by uptake by the neuronal amine uptake system. Utilizing the inwardly-directed Na^+ concentration gradient maintained by the neuronal membrane ($\text{Na}^+ + \text{K}^+$)-ATPase, norepinephrine is co-transported with Na^+ in a facilitated diffusion process which appears to be carrier-mediated (14,15).

Inhibitors of ($\text{Na}^+ + \text{K}^+$)-ATPase will antagonize norepinephrine accumulation as will sympathomimetic amines which competitively bind to the uptake carrier (15). Cocaine and desipramine (DMI) both display a high affinity for the carrier system, thus inhibiting the uptake of norepinephrine.

Postsynaptically norepinephrine may interact with α - or β -receptors, which couple with an adenylate (or guanylate) cyclase to initiate postsynaptic events (16,17). Presynaptic α - and β -receptors have also been implicated in the modulation of transmitter release. While an adenylate cyclase has been associated with the presynaptic β -receptor, the presynaptic α -receptor appears to be involved only with modulation of Ca^{++} fluxes (18,19).

Structure-activity relationships (SAR) in sympathomimetic amines must then deal with the effect of molecular structure on:

- 1) The inhibition of neuronal transmitter uptake.
- 2) The uptake and retention of the agents themselves.
- 3) The release of transmitter from its storage sites.
- 4) Direct receptor activity of the amines.
- 5) Competitive binding to the active sites of metabolic enzymes.

Unless these effects are carefully sorted out, efforts to design selective agents are thwarted from the outset. A review of SAR within each division of action would be impossible within the confines of the present

discussion. The reader is referred elsewhere for the general conclusions of SAR on the competitive inhibition of the neuronal uptake of norepinephrine (20,21,22,23), transport of norepinephrine by the carrier (24), release of previously accumulated norepinephrine (25), interaction of directly-acting sympathomimetic agonists on the α - and β -receptors (26) and the substrate specificity for catechol O-methyltransferase (27,28). These conclusions of SAR illustrate that selectivity of pharmacological action is not always possible through gross structural manipulation alone. Several interactive sites can display similar structural preference. Despite the detail of these SAR studies, the exact nature of the neurotransmitter (or drug)-receptor complex remains unknown.

The possibility that different sites of drug action require different conformational arrangements of the agonist, thereby uncovering a new dimension of drug selectivity not possible through gross structural manipulation alone, will be discussed below.

Techniques for Assessing Conformation-Activity Relationships

Conformation-activity relationships are considerably more difficult to delineate than are structure-activity relationships. In the latter case, one needs only to modify the structure of an active substance and to determine the pharmacological effects induced by the structural change. Most active adrenergic agents, however, are conformationally mobile; rotation about single bonds is usually facile and rapid. For this reason, most adrenergic agents provide very little direct information regarding conformational effects.

A number of techniques have been applied to the study of conformation-activity relationships: x-ray crystal structure determination, molecular orbital calculations of low-energy conformations, nuclear magnetic resonance determinations of solution conformations, and the preparation of conformationally-restricted analogs. The first three methods are directed toward determining which conformation (or which conformations) is (are) most favorable; the assumption is made that an energetically preferred conformation will be the biologically active conformation. This need not be the case; Portoghese's suggestion (29) regarding a "high-energy" conformation would not be consistent with such an assumption. Nevertheless, the possibility that a favorable

conformation might correspond to an active one has served as a starting point for a large amount of research. The use of conformationally-restricted analogs precludes the necessity of such an assumption, but this method also has its limitations. We wish to combine a number of these techniques in a single laboratory with the hope that the advantages of each approach will be magnified and the disadvantages minimized.

X-Ray Crystallography. This approach is based on the assumption that a low-energy conformation should be the conformation recognized by the receptor. (Receptor is here used loosely to refer to any site of interaction with the agonist or antagonist molecule). The crystalline form of a substance should certainly be a relatively low-energy and stable conformation. In addition, structural and conformational features are delineated with a high degree of accuracy by x-ray diffraction. Several negative factors should be pointed out, however. First, the molecule in the solid state has a much different environment from that in dilute solution. Also, if a substance has two or more relatively stable conformations, only one of these is likely to exist in the crystal, while the other(s) would go undetected. Finally, crystal packing forces may cause distortions of the molecule whenever the increased strain energy is compensated for by enhanced intermolecular interactions.

At least eighteen phenylethylamines have been examined by x-ray crystallography. Among these are phenylethylamine hydrochloride (30), ephedrine hydrochloride (31), ephedrine monohydrogen phosphate monohydrate (32), ephedrine dihydrogen phosphate (33), dopamine hydrochloride (34), 5-hydroxydopamine hydrochloride (35), 6-hydroxydopamine hydrochloride (36), epinephrine hydrogen tartrate (37), norepinephrine hydrochloride (38), isoproterenol sulfate (39), 2,4,5-trimethoxyamphetamine hydrochloride (40), 4-ethyl-2,5-dimethoxyamphetamine (41), mescaline hydrobromide (42), and mescaline hydrochloride (43).

Using the conventions shown in Figure 2, several observations may be made regarding the solid state conformations of these phenylethylamines. First, fifteen of the eighteen compounds have the nitrogen atom and the aromatic ring in an extended conformation ($\tau_2 \cong 180^\circ$). It is interesting to note that mescaline hydrochloride exists in an extended conformation, while the hydrobromide has a gauche conformation. Since the ionic radii of the anions (44) are very

Figure 2. 2-Phenylethylamine (unless otherwise stated, the numbering system shown is used throughout most of this paper in describing phenylethylamines)

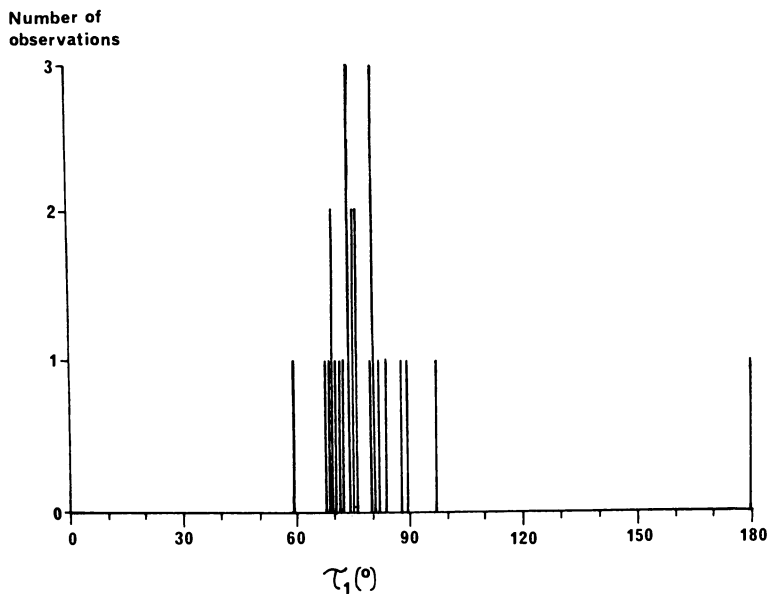
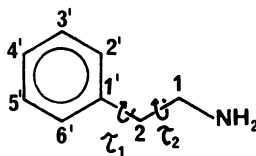


Figure 3. Distribution of values of τ_1 (see text)

nearly equal ($\text{Cl}^- = 1.80 \text{ \AA}$; $\text{Br}^- = 1.95 \text{ \AA}$), it would appear that both conformations are reasonably stable and might exist in dilute solution. Among the three compounds existing in a gauche conformation ($\tau_2 \cong 60^\circ$), mescaline hydrobromide and 2,4,5-trimethoxyamphetamine hydrochloride are potent hallucinogens, while 4-ethyl-2,5-dimethoxyamphetamine is reported to cause euphoria and a feeling of enhanced self-awareness (45).

In Figure 3, the τ_1 values are represented graphically. It may be seen that most of these values fall in the range of 67° to 97° , indicating that the side chain is usually directed away from the plane of the aromatic ring. The two exceptions are 5-hydroxydopamine (59°) and epinephrine hydrogen tartrate (179°). In the latter case, the authors (37) suggest that the extremely dense packing of the structure and the numerous hydrogen bonds may cause the molecule to assume this unusual conformation.

Molecular Orbital Calculations. Theoretical calculations of molecular properties have become reasonably accurate and relatively easy to carry out in recent years. Atomic coordinates from x-ray structure determinations or standard bond lengths and angles can serve as the input, and a wide range of properties may be determined. Of particular interest to us is the determination of the relative energies of various conformations of a molecule, or the relative populations of a series of conformers. Rotational barriers, which can affect interconversion among conformers, can also be estimated. The assumption is again made that a conformation which predominates in solution should be biologically active. Other parameters (such as charge distribution) which reflect the capacity for electrostatic interactions at various points on the molecule can be calculated.

Five computational methods have been applied to conformational studies of phenylethylamines. These are Extended Hückel Theory (EHT) (46), Complete Neglect of Differential Overlap (CNDO) (47), Intermediate Neglect of Differential Overlap (INDO) (48), Perturbative Configuration Interactions using Localized Orbitals (PCIO) (49), and some ab initio (AI) (50) procedures. EHT is relatively easy and inexpensive to use, but it tends to distort electron distribution in systems containing heteroatoms. In EHT, electron-repulsion terms are neglected, causing a tendency to predict unrealistically low energies for gauche and eclipsed conformations. CNDO and INDO take into account some

important electron-repulsion terms, but gauche and eclipsed conformations are again predicted to be too stable relative to staggered conformations. CNDO and INDO are referred to as "semi-empirical" because they can be "parameterized"; certain variables in the computation are assigned values which will give good results for a given class of compounds or for a given chemical property. All of the above methods except PCILO employ the Hartree-Fock approximation in their calculations; this approximation is valid only at or near low-energy states of the molecule. For this reason, rotational barriers are often overestimated. PCILO, on the other hand, predicts slightly lower-than-normal rotational barriers. AI methods use much more sophisticated computations, but the basis sets from which molecules are constructed are often either too small to give good results or too expensive to use for molecules of more than a few atoms. A more mathematical comparison of methods has been presented by Hoyland (51). Table I lists the compounds which have been studied by molecular orbital methods. This list includes directly and indirectly acting agonists as well as antagonists. Although the energies determined by these methods are not free energies since entropic effects are not taken into consideration, energy differences among closely related compounds appear to be quite accurately determined.

Conformational results are often sensitive to the input geometry used. For instance, Pullman *et al.* (68) carried out two PCILO calculations for mescaline, using the x-ray crystal geometries obtained for the hydrobromide and hydrochloride salts; these salts exist in gauche and extended conformations, respectively. In both cases the calculations showed relative minima for both conformations, and these minima differed by less than one kcal/mol. In each case, however, the global (overall) minimum corresponded to the crystal geometry.

Martin *et al.* (70) have published a detailed study of phenylethylamine, utilizing EHT, CNDO, INDO, PCILO and AI. These authors used semi-empirical methods to generate a series of conformational energy surfaces and then used an AI procedure with a large basis set for a few selected points. EHT, CNDO and INDO all indicated a minimum for τ_1 at 90° ; EHT predicts a rotational barrier of 7 kcal/mol, while CNDO and INDO show a barrier of about 2 kcal/mol. PCILO gives a rotational barrier of about 2 kcal/mol, but it shows a very broad minimum for τ_1 , ranging from 45° to 135° . With respect to rotation about τ_2 , these authors found the extended

form to be about 1 kcal/mol more stable than the gauche using EHT and AI, while the gauche form appeared to be about 1 kcal/mol more stable using CNDO, INDO and PCILO.

The most extensive studies of phenylethylamines have been done by Pullman and coworkers (59,65,67,68). Their PCILO calculations indicated that gauche and trans conformers have nearly identical energies in the majority of phenylethylamines and phenylethanolamines studied. AI studies were carried out using three different basis sets (67); these all indicated a slight but definite preference for gauche conformers. These authors noted the preponderance of extended conformers in crystal structures and in solution; they attributed this to environmental forces. To test this hypothesis, computations were repeated with the inclusion of water molecules, and a tendency toward the extended conformation was indeed observed.

Thus, most of the evidence available from molecular orbital calculations suggests that both the extended and gauche conformations are reasonably stable. It would appear unwise to deduce that one of these conformations can be recognized in vivo and the other not, solely on the basis of such computations.

Nuclear Magnetic Resonance. Through a careful analysis of coupling constants, it has sometimes been possible to derive conformational information from nuclear magnetic resonance (NMR) spectra. This offers the advantage of directly determining solution conformations, which should be more relevant to the biological situation than would be solid state or in vacuo conformations. Once again, the underlying assumption is that a highly populated conformer is the active species. Table II summarizes the conformational studies of phenylethylamines which have been carried out using NMR. In describing conformations, we will use the following conventions (see Figure 4):

- 1) I will designate that conformation in which the nitrogen atom is antiperiplanar with respect to the aromatic ring.
- 2) II will designate that conformation in which the aromatic ring and the nitrogen atom are in a gauche relationship; in compounds bearing a hydroxyl group on C-2, II will also have the nitrogen and oxygen atoms gauche to each other, and in amphetamines, II will have the C-1 methyl group antiperiplanar with respect to the aromatic ring.
- 3) III will designate a gauche relationship

**American Chemical
Society Library
1155 16th St. N. W.**

Table I. Summary of molecular orbital studies of phenylethylamines.

<u>Compound(s)</u>	<u>Method</u>	<u>Ref.</u>
1. Ephedrine Pseudoephedrine	EHT	52
2. Norepinephrine	EHT	53
3. Dopamine	EHT	54
4. Dopamine	EHT	55
5. Norepinephrine Epinephrine N-Ethylnorepinephrine Isoproterenol	EHT, CNDO	56
6. Norepinephrine Norepinephrine·Li·H ₂ O	INDO	57
7. Dopamine	EHT	58
8. Phenylethylamine Norepinephrine Ephedrine Dopamine Tyramine Norephedrine Epinephrine Amphetamine	PCILO	59
9. Phenoxyethylamine	PCILO	60
10. Practolol 1-(4'-methylphenyl)-2-isopropyl- aminoethanol 1-(4'-methylphenyl)-2-isopropyl- aminopropanol	CNDO	61
11. Dopamine	EHT	62
12. Norepinephrine Dopamine (including <u>meta</u> - and <u>para</u> -anions)	CNDO	63

Table I (cont'd.)

13.	2,5-Dihydroxyphenylethylamine 2,3,4-Trihydroxyphenylethylamine 3,4,5-Trihydroxyphenylethylamine 2,3,6-Trihydroxyphenylethylamine 2,4,6-Trihydroxyphenylethylamine	CNDO	64
14.	2,4,5-Trimethoxyamphetamine 2,4,6-Trimethoxyamphetamine	PCILO	65
15.	Norepinephrine	CNDO	66
16.	Norepinephrine Amphetamine Epinephrine Isoproterenol Ephedrine	AI, PCILO	67
17.	Mescaline 2,3,4-Trimethoxyamphetamine 2,4,5-Trimethoxyamphetamine 2,4,6-Trimethoxyamphetamine 2,4,5-Trihydroxyphenylethylamine 3,4,5-Trihydroxyphenylethylamine	PCILO	68
18.	Isoproterenol INPEA	CNDO	69
19.	Phenylethylamine	AI EHT CNDO INDO PCILO	70
20.	Phenylethylamine Amphetamine	AI	71
21.	Dopamine	INDO	72

Table II. NMR conformational studies of adrenergic agents.

<u>Compound(s) Studied</u>	<u>Solvent</u>	<u>Ref.</u>
1. Ephedrine Pseudoephedrine (free bases)	CDCl_3	73
2. Ephedrine Pseudoephedrine (free bases and HCl salts)	$\text{CCl}_4, \text{C}_6\text{H}_6$ $\text{CDCl}_3, \text{DMSO}$ $\text{CF}_3\text{COOH}, \text{D}_2\text{O}$	74
3. Isoproterenol Epinephrine (TMS ethers)	CCl_4	75
4. Dopamine HCl	D_2O	55
5. Amphetamine (free base and HCl) Methamphetamine HCl o-Methoxy-methamphetamine HCl Benzphetamine HCl	D_2O	76
6. Five dimethoxyamphetamines (free bases)	CDCl_3	77
7. Eighteen aryl-substituted amphetamines (free bases) Nine aryl-substituted amphetamines (HCl salts)	CDCl_3	78
8. N-Isopropyl-p-nitrophenylethyl- amine		79
9. Amphetamine (free base)	CDCl_3	80
10. Eleven adrenergic agents (HCl salts)	D_2O	81

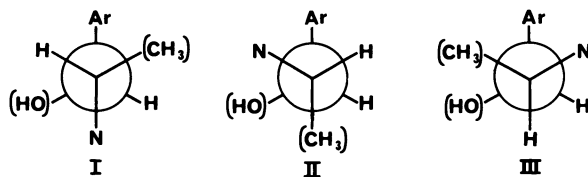


Figure 4. Newman projections of the major conformations of phenylethylamines

between the aromatic ring and the nitrogen atom; further, in compounds bearing a hydroxyl group on C-2, III will have the oxygen atom antiperiplanar with respect to the nitrogen atom, while for amphetamines, III will indicate that the C-1 methyl group is also gauche with respect to the aromatic ring.

Hyne (73) investigated the conformations of ephedrine and pseudoephedrine as the free bases in deuteriochloroform (CDCl_3). For ephedrine, the ratio of conformations I, II and III was found to be 40:40:20, while for pseudoephedrine the ratio was 62:30:8. Portoghesi (74) subsequently reinvestigated the ephedrines. For ephedrine hydrochloride in D_2O , he found 90% of the material to be in conformations which permitted hydrogen bonding between the hydroxyl and amino groups (*i.e.*, 90% in conformations I and II); pseudoephedrine hydrochloride was reported to have 84% of the extended conformation I and 16% gauche conformers (II and III). These results are accounted for by two factors. First, hydrogen bonding interactions favor conformations I and II for both ephedrine and pseudoephedrine. Secondly, the methyl substituent α to the nitrogen atom causes conformers I and II to be sterically less favorable in ephedrine, while it makes II and III less favorable in pseudoephedrine.

A large number of substituted amphetamines have been studied by NMR. These results are summarized in Table III. In all cases, the extended conformation (I) predominated, making up 47-76% of the conformer population. Conformer II, in which the methyl group on C-1 is directed away from the phenyl ring, is also present to a significant extent in all cases, making up 21-45% of the sample. Conformation III, in which the aromatic ring has gauche interactions with both the methyl group and the nitrogen atom, makes up the remaining 0-12%.

Results for a variety of other adrenergic agents have been compiled in Table IV. As was the case for the amphetamines, the extended conformation is by far the predominant one, with conformer II making a substantial contribution in compounds bearing a hydroxyl group on C-2. The notable exception is dopamine hydrochloride, in which 43% of the extended conformation and 57% of the gauche conformation were observed.

In general, NMR experiments have been in agreement with molecular orbital calculations. Both extended and gauche conformers are usually present to a significant extent, with the extended conformer predominating. Again, the evidence available is not sufficient to

Table III. Amphetamine conformer distributions as determined by NMR. (a) fb = free base. (b) See Figure 4 and the text for conventions used in specifying conformations.

Aromatic Substitution	N-Alkyl	Salt ^a	Solvent	Ratio of conformers ^b			Reference
				I : II	III	III	
1. ---	---	fb	D ₂ O	50	39	11	76
2. ---	---	HCl	D ₂ O	50	45	5	76
3. ---	CH ₃	HCl	D ₂ O	55	39	6	76
4. 2-OCH ₃	CH ₃	HCl	D ₂ O	47	36	17	76
5. ---	CH ₃ , CH ₂ C ₆ H ₅	HCl	D ₂ O	64	35	1	76
6. 2,3-(OCH ₃) ₂	---	fb	CDCl ₃	67	21	12	77
7. 2,4-(OCH ₃) ₂	---	fb	CDCl ₃	62	26	12	77
8. 2,5-(OCH ₃) ₂	---	fb	CDCl ₃	64	24	12	77
9. 3,4-(OCH ₃) ₂	---	fb	CDCl ₃	64	25	11	77
10. 3,5-(OCH ₃) ₂	---	fb	CDCl ₃	65	23	12	77
11. 2,3-(OCH ₃) ₂	---	HCl	CDCl ₃	68	32	0	78
12. 2,4-(OCH ₃) ₂	---	HCl	CDCl ₃	66	34	0	78
13. 2,5-(OCH ₃) ₂	---	HCl	CDCl ₃	65	35	0	78
14. 3,4-(OCH ₃) ₂	---	HCl	CDCl ₃	76	23	1	78
15. 3,5-(OCH ₃) ₂	---	HCl	CDCl ₃	75	25	0	78
16. 2,3,4-(OCH ₃) ₃	---	HCl	CDCl ₃	65	35	0	78
17. 3,4,5-(OCH ₃) ₃	---	HCl	CDCl ₃	72	28	0	78
18. ---	---	fb	CDCl ₃	72	28	0	80

Table IV. NMR conformational analysis of some phenylethylamines. (a) fb = free base. (b) See Figure 4 and the text for conventions used in specifying conformation. (c) Determined as the TMS ether. (d) Sum of conformers II and III. (e) II and III are equivalent; sum of II and III. (f) Sum of conformers I and II.

Compound	Salt ^a	Solvent	Ratio of conformers ^b		Ref.
			I : II	: III	
1. Isoproterenol ^c	fb	CCl ₄	50	----50-----d	75
2. Epinephrine ^c	fb	CCl ₄	70	----30-----d	75
3. Dopamine	HCl	D ₂ O	43	----57-----d	55
4. N-Isopropyl-p-nitrophenylethylamine					79
5. Norepinephrine	HCl	D ₂ O	76	14 10	81
6. Epinephrine	HCl	D ₂ O	77	17 6	81
7. Isoproterenol	HCl	D ₂ O	83	11 6	81
8. Phenylethanolamine	HCl	D ₂ O	84	10 6	81
9. Phenylephrine	HCl	D ₂ O	81	16 3	81
10. Synephrine	HCl	D ₂ O	81	10 9	81
11. N,N-Dimethyl-o-bromophenylethanol-amine	HCl	D ₂ O	96	0 4	81
12. Norephedrine	HCl	D ₂ O	---	79---f 21	81
13. Metaraminol	HCl	D ₂ O	---	79---f 21	81
14. Butanephrine	HCl	D ₂ O	---	72---f 28	81
15. Phenylethylamine	HCl	D ₂ O	56	---44---e	81

assess which conformation is biologically active.

Conformationally-Restricted Analogs. In preparing conformationally-restricted analogs of a biologically active substance, the essential structural features of the agonist are built into a molecular framework which has limited conformational mobility. With this approach it is no longer necessary to assume that a predominating conformation is active, since the conformation of the message molecule is restricted. When a conformationally-restricted analog exhibits an activity, the conformational possibilities responsible for that activity are well defined. Unfortunately, the converse is not true -- when such analogs are inactive, there may be a number of reasons other than incorrect conformation. The extra atoms necessary to make the molecule less flexible may sterically interfere with the drug-receptor recognition process. In addition, the extra atoms may alter absorption, tissue distribution, metabolism and other cellular processes. The ideal conformationally-restricted analog should provide well-defined and undistorted geometry with as few extra atoms as possible. Although a plethora of rigid and semi-rigid analogs of adrenergic agents has appeared in recent years, few have approached this ideal; most have shown little or no biological activity except at high concentrations, and few show pronounced conformational selectivity. Comparisons are often difficult because the pharmacological testing performed on these compounds ranges from simplistic to highly sophisticated. Some compounds would be expected to show direct activity (receptor activation), some indirect activity (release of neurotransmitter), and some might have both kinds of activity. Table V is a compilation of these ring systems and their applications. A detailed review is beyond the scope of this discussion, but some comments on the analogs of amphetamine are pertinent.

Smismann and coworkers prepared a large number of trans-decalin compounds (82-90). The trans-decalin system is reasonably rigid and may be expected to provide nearly ideal dihedral angles; however, the very large number of extra atoms makes these compounds highly lipophilic and creates a great deal of steric bulk. The decalin amphetamine analogs (skeleton 1 in Table V) were tested for their effect on motor activity (86), since amphetamine itself is known to increase motor activity. Paradoxically, all four compounds decreased motor activity. In addition, the extended conformer caused lachrymation. Later studies (87,88,89) on

Table V. Ring systems which have been utilized in preparing conformationally-restricted analogs of adrenergic agents.


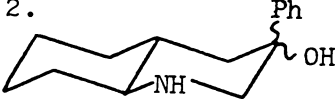
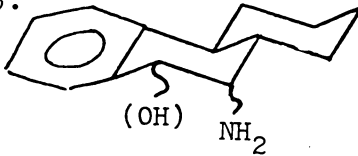

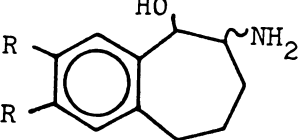
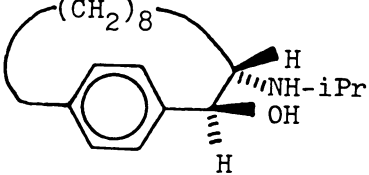
Structural Skeleton	Analogues of:	Number of extra atoms	Ref.
1.  <u>trans-decalin</u>	Norephedrine Amphetamine Dopamine Norepinephrine	7-8	82- 90
2.  <u>trans-decahydroquinoline</u>	Phenylethanol-amine	7	91
3.  octahydrophenanthrene	Norephedrine Amphetamine	5	92, 93
4.  cyclohexane	Amphetamine	3	94
5.  benzocycloheptene	Norephedrine Norepinephrine	2-3	95, 96, 97
6.  [10]-paracyclophane	N-Isopropyl-norephedrine	7	98

Table V (cont'd.)

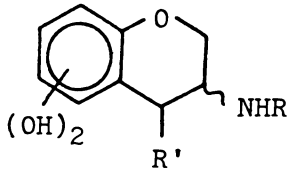
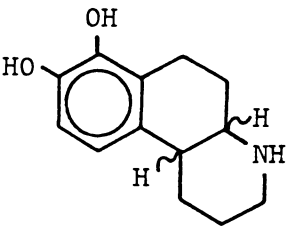
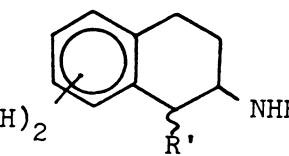
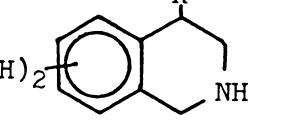
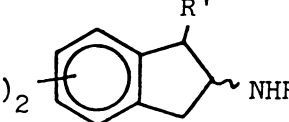

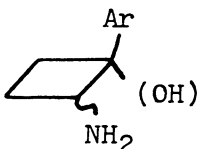
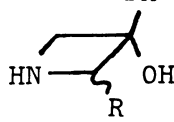

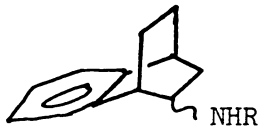
7.		Ephedrine Isoproterenol	1-2	99
	chromane			
8.		N-Methyl- dopamine	5	100, 101
	octahydrobenzo[f]- quinoline			
9.		Norephedrine Dopamine Norepinephrine Epinephrine	1-2	102- 119
	2-aminotetralin			
10.		Phenylethylamine Dopamine Norepinephrine Epinephrine	1	106 120- 127
	tetrahydroiso- quinoline			
11.		Norephedrine Amphetamine Dopamine	0-1	103, 105, 123, 128- 130
	2-aminoindane			
12.		Amphetamine	0	129, 131- 133
	cyclopropane			

Table V (cont'd.)

13.	 <p>cyclobutane</p>	Norephedrine Amphetamine Dopamine Metaraminol	1-2	130, 132, 134, 135
14.	 <p>azetidine</p>	Ephedrine Phenylethanol- amine	1	136
15.	 <p>benzonorbornene</p>	Amphetamine Methamphet- amine	2	137- 140
16.	 <p>benzobicyclo[2.2.2]- octene</p>	Amphetamine Methamphet- amine	3	130, 141- 144

inhibition of dopamine uptake into striatal tissue showed that a gauche conformer (a-amino, e-phenyl) was considerably more potent than the other three isomers. Blockade of the uptake of biogenic amines into platelets by the four amphetamine conformers showed little selectivity (the least potent isomer being 60% as effective as the most potent).

Amphetamine analogs in skeleton 3 (92,93) showed no amphetamine-like behavioral effects and no amphetamine-like hyperthermia in animals. This skeleton is rigid, but shows some distortions from normal bond angles for antiperiplanar and gauche conformations. Skeletons 4, 5, 7-11, 13 and 14 have varying degrees of conformational flexibility which make any conformation-activity relationships difficult, if not impossible, to obtain. Skeleton 6 does not mimic any of the low-energy conformations of the phenylethylamine skeleton. Although skeleton 12 has the fewest possible number of extra atoms (one extra for a phenylethylamine and none for an amphetamine) the bond angles are so distorted that an extended conformation cannot be achieved by the trans-substituted cyclopropyl system and a gauche conformer cannot be approximated (the cis-isomer is actually an eclipsed conformer).

Choice of a Conformationally-Defined Analog

In our search for the most suitable system for the study of sympathomimetic amines, we chose initially to look at a system with the minimum of synthetic problems in obtaining suitable quantities for initial pharmacological evaluation. To this end we chose to look at amphetamine analogs, with the ultimate aim of adding the β -hydroxyl and catechol hydroxyls to produce the catecholamines. We felt that it was essential that the system chosen have a minimum number of "extra" atoms not present in amphetamine (or its N-methyl derivative, methamphetamine). It was also necessary that the system not be highly strained and that the bond angles and bond lengths found in the low-energy conformations of amphetamine be very closely approximated in the conformationally-defined model. It was desired that any system chosen would show similar physical properties to those of amphetamine -- lipophilicity (partition coefficient), pKa and charge distribution in particular.

Examination of these points suggested skeletons 15 and 16 of Table V. Since the benzobicyclo[2.2.1]-heptene analogs (skeleton 15) are considerably more strained than the benzobicyclo[2.2.2]octene (skeleton

16), the latter was chosen for our initial study. However, derivatives of skeleton 15 and its oxo-bridged analog, were also planned (see Figure 7). Figures 5 and 6 show the relationship of the methamphetamine or amphetamine analogs in skeleton 16 to amphetamine and methamphetamine.

The benzobicyclo[2.2.2]octene system has only three "extra" atoms and thus appeared to have a number of advantages over the other systems in Table V and Figure 7. First, the two low-energy conformations of the phenylethylamines can be clearly fixed (the systems are almost completely rigid, in contrast to the more extensively-studied systems like skeletons 9 and 13 of Table V). Second, molecular models show there is minimal distortion in bond angles or bond lengths from those in the mobile system, amphetamine. This is not so for systems III-VIII of Figure 7. Third, a relatively simple synthetic procedure would afford both the *endo*- and *exo*-isomers of amphetamine (NH-N and NH-X) or methamphetamine (NM-N and NM-X) so the system could be evaluated by x-ray crystallography to determine the precise bond angles and bond lengths in comparison to those of the mobile (amphetamine) system. Using the atomic coordinates of the x-ray study, molecular orbital calculations could be carried out to compare the rigid system with the mobile system (*i.e.*, charge distributions). And, most importantly, it could be evaluated pharmacologically in comparison with the mobile drug to see if conformational differences in activity were present. Any differences could then be interpreted in terms of precise conformational arguments. To date the benzobicyclo[2.2.2]-octene system has lived up to its expectations. In the remaining part of this discussion we will address the above points.

Evaluation of Some Bicyclic Systems as Conformationally-Defined Phenylethylamines. X-Ray Crystallography. Perhaps the ultimate test of the suitability of a chemical structure for preparing conformationally-defined analogs would be their ability to produce biological effects identical to those of the parent compound. Since this ideal is rarely achieved, and since a "wrong" conformer should be inactive, it is useful to have other criteria with which to evaluate these systems. The bicyclic molecules to be examined in the present study are compounds I-VIII (Figure 7), benzobicyclo[2.2.2]octenes, benzobicyclo[2.2.1]heptenes, and 1,2,3,4-tetrahydro-1,4-epoxynaphthalenes. A number of structural parameters

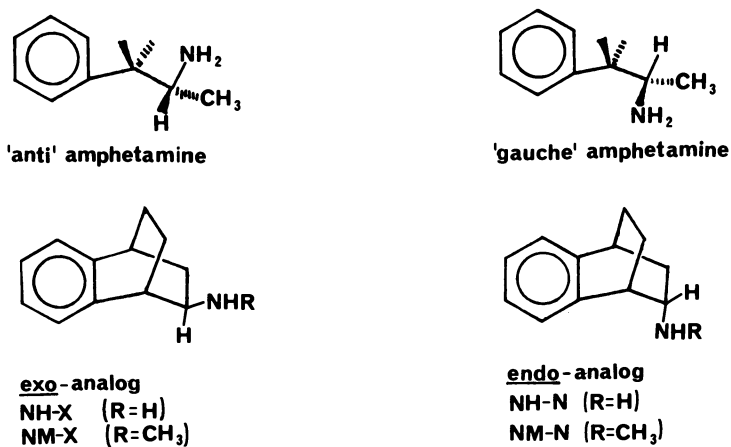


Figure 5. Relationship of the *exo*- and *endo*-2-amino- and 2-methylaminobenzo-bicyclo[2.2.2]octenes to the two low energy conformations of amphetamine and methamphetamine. The abbreviations shown for the conformationally defined analogs are used throughout the text.

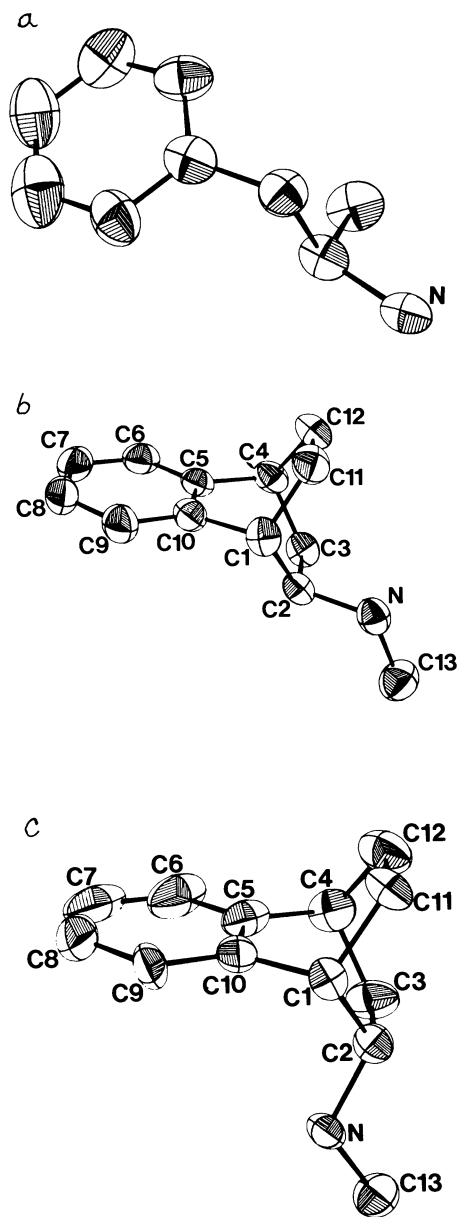


Figure 6. ORTEP drawings of (a) amphetamine (146) drawn in the same perspective as that for (b) 2-methylaminobenzobicyclo[2.2.2]octene (*exo*-isomer NM-X) and (c) *endo*-isomer NM-N (145).

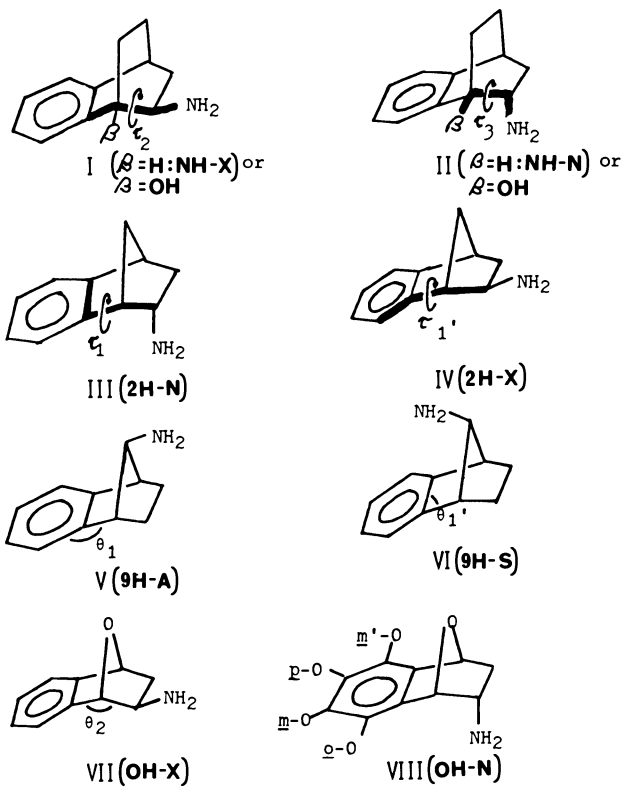


Figure 7. Conformationally defined analogs of phenylethylamines (see text)

(dihedral angles, bond angles and interatomic distances) can be determined for these bicyclic phenylethylamine analogs since the carbon skeletons of these compounds have almost no flexibility. These structural parameters can be compared with the corresponding values derived from x-ray structure determinations and molecular orbital calculations of phenylethylamines. The dihedral angles τ_1 , τ_1' , τ_2 and τ_3 (illustrated on structures I-IV of Figure 7) correspond to those in Figure 2. The significant bond angles (θ_1 , θ_1' and θ_2) are defined in structures V-VII of Figure 7. Interatomic distances between functional groups provide a measure of how accurately the rigid structures approximate the spatial placement of these groups as observed in conformationally-mobile phenylethylamines. The functional groups of interest (shown on structure VIII) include the nitrogen atom, the oxygen atom on C-2, the aromatic oxygen substituents and the aromatic ring. Determination of these distances and angles was based on x-ray structure determinations of compounds having the benzobicyclo[2.2.2]octene (145) or benzobicyclo[2.2.1]heptene (147) ring systems. Where necessary, extra atoms were added, using standard bond lengths and angles (44). These calculations were carried out with the computer programs QCLM/ICOORD and QCLM/COORD (made available to us by the Quantum Chemistry Group at the University of Kansas) which calculate bond lengths, bond angles, dihedral angles and interatomic distances from atomic coordinates. The position of the aromatic ring was represented by the center of the ring (CR), which was determined by averaging the coordinates of the six-carbon atoms which make up the ring.

The x-ray structure determinations of exo- and endo-2-methylamino-1,2,3,4-tetrahydro-1,4-ethanonaphthalene hydrochlorides (145) served as the basis for calculations of the benzobicyclo[2.2.2]octene compounds (NM-X and NM-N).

The benzobicyclo[2.2.1]heptene and 1,2,3,4-tetrahydro-1,4-epoxynaphthalene systems were constructed from the coordinate data for benzobicyclo[2.2.1]heptene-syn- and anti-bromobenzenesulfonates (147). Positions for the exo and endo amino groups were selected by using a standard C-N bond length with the bond angles and dihedral angles calculated from a crystallographic study of 2-exo-aminonorbornane-2-carboxylic acid (148).

The dihedral angle τ_1 describes the rotation of the aromatic ring with respect to the ethylamine side chain (see Figures 2 and 7). Compared with values found in x-ray studies and molecular orbital calcula-

tions, the benzobicyclo[2.2.1] compounds provide a better approximation of this angle than do the benzo-bicyclo[2.2.2] compounds. This is represented graphically in Figure 8. Values of τ_1 , observed in x-ray studies (see above) are depicted along with the result predicted by molecular orbital (see Table VII) calculations (90°). Compounds III-VIII form an angle of 69° , while the bicyclo[2.2.2] compounds form a 60° dihedral angle.

The comparison of τ_2 values may be divided into two groups: those for extended conformers ($\tau_2 \cong 180^\circ$) and those for gauche conformers ($\tau_2 \cong 60^\circ$).² Both groups are depicted in Figure 9. The analogs having an extended conformation all give τ_2 values well within the range of observed values² for extended phenylethylamines: the exo-benzobicyclo[2.2.2]octene isomer gives a nearly ideal τ_2 dihedral angle (179°); the exo-[2.2.1] compounds are nearly as good (176°), while the anti-[2.2.1] compound forms an angle of 168° . The gauche conformers are more difficult to evaluate, since observed τ_2 values are available for only three gauche phenylethylamines. All four conformationally-defined structures are within 10° of the observed and calculated values, as illustrated in Figure 9. The endo-[2.2.1] structure ($\tau_2 = 59^\circ$) most nearly approximates the predicted value of 60° . The endo-[2.2.2] and syn-[2.2.1] compounds form angles of 69° and 70° , respectively.

Building the C-2 oxygen atom into the bicyclo-[2.2.1] ring system seriously distorts τ_3 , the O-C-C-N dihedral angle. As shown in Figure 10, the benzo-bicyclo[2.2.2]octenes provide excellent values for τ_3 (endo, 59° ; exo, 56°), but VII forms an 83° angle. The endo isomer VIII, which approximates an antiperiplanar arrangement of the oxygen and nitrogen atoms, forms an angle of 160° , compared to a predicted angle of 180° .

The bond angles θ_1 , θ_1' and θ_2 provide an indication of the amount of ring strain in the bicyclic ring structures. Ideally, θ_1 and θ_1' should be 120° ; in the [2.2.2] ring system, this is¹ distorted by about 7° , and in the [2.2.1] structures it is nearly twice as distorted (by about 12°). The Cl'-C2-Cl angle (θ_2) is within three degrees of the ideal tetrahedral angle (109°) for compounds I, II, III, IV, VII and VIII. The syn- and anti-[2.2.1] structures, however, have θ_2 values of 100° and 96° , respectively.

The spatial arrangement of functional groups would appear to be particularly important in order for conformationally-restricted analogs to exhibit biological

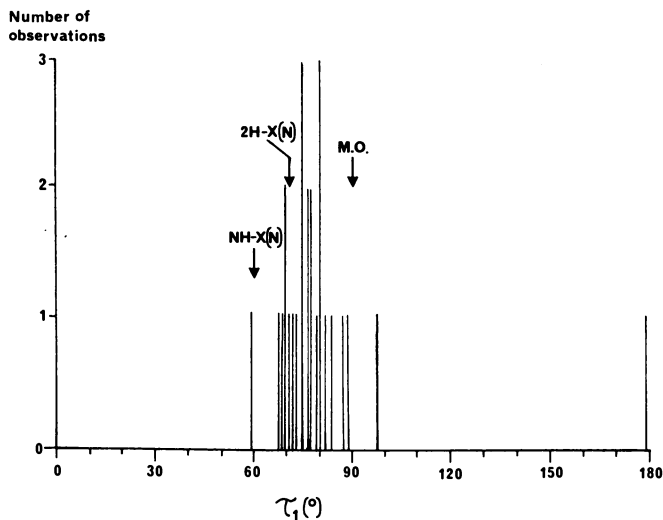


Figure 8. Observed values of τ_1 from x-ray crystallography of phenylethylamines showing the agreement with the value of this angle from molecular orbital studies and some conformationally defined analogs (see text)

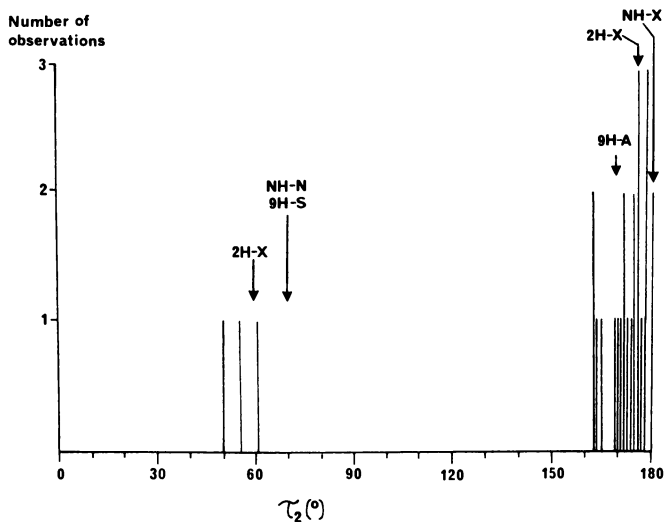


Figure 9. Observed values of τ_2 from x-ray crystallography of phenylethylamines showing the agreement with the value of this angle from molecular orbital studies and some conformationally defined analogs (see text)

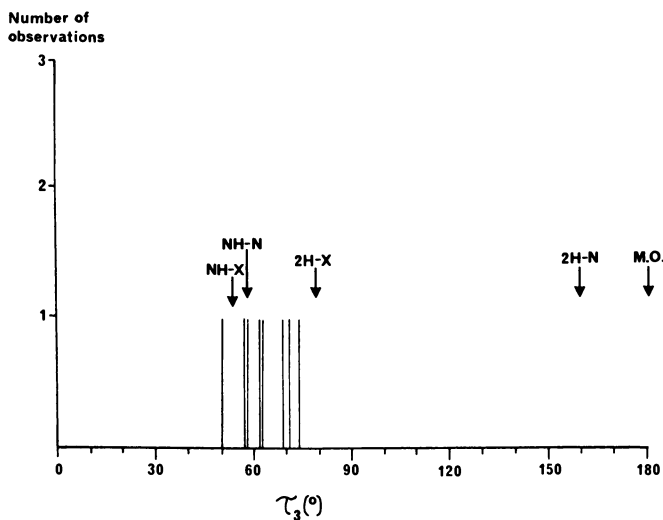


Figure 10. Observed values of τ_3 from x-ray crystallography of phenylethylamines showing the agreement with the value of this angle from molecular orbital studies and some conformationally defined analogs (see text)

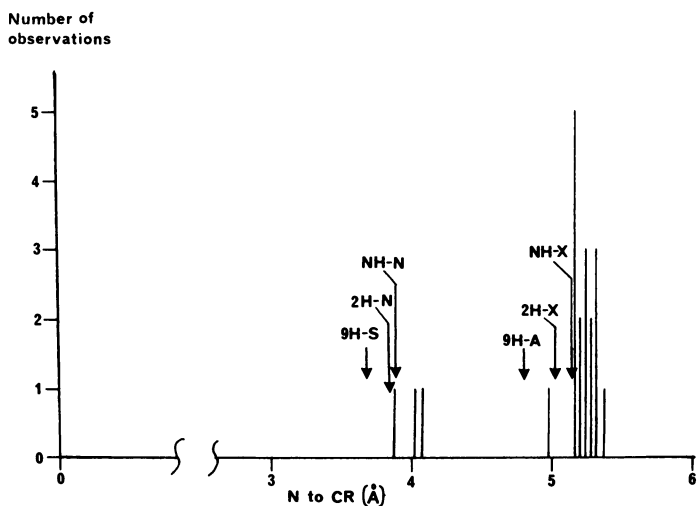


Figure 11. Comparison of the distances from the nitrogen to the center of the aromatic ring (CR) in some conformationally defined analogs of phenylethylamines with those found from x-ray crystallographic studies reported for phenylethylamines (see text)

activity. This spatial arrangement is reflected in the interatomic distances among these functional groups. In the case of catecholamines, the functional groups of particular interest include the nitrogen atom, the C-2 oxygen atom, the aromatic ring and the aromatic oxygen substituents (see Figures 2 and 7). For compounds I-VIII, the interatomic distances are tabulated in Table VI. These may readily be compared to the corresponding values from x-ray studies of three representative conformationally mobile phenylethylamines, norepinephrine hydrochloride (IX), amphetamine sulfate (X) and mescaline hydrobromide (XI).

X-ray studies of phenylethylamines in the extended conformation exhibit nitrogen-to-aromatic ring (N to CR) distances ranging from 4.98 to 5.21 Å, with a mean value of 5.13 Å (see Figure 11). In comparison, the exo-[2.2.2] compound has an N to CR distance of 5.07 Å, the exo-[2.2.1] compound a distance of 5.03 Å, and the anti-[2.2.1] compound a distance of 4.74 Å. Among phenylethylamines in the gauche conformation, N to CR distances observed were 3.89, 4.01 and 4.07 Å (mean = 3.99 Å). The endo-[2.2.2] structure most nearly approximates this distance (N to CR = 3.89 Å). The endo and syn-[2.2.1] compounds show N to CR distances of 3.85 and 3.69 Å, respectively.

The nitrogen-to-para-oxygen distances (N to p-O) follow a similar pattern (see Figure 12). For extended phenylethylamines, observed distances range from 7.78 to 7.90 Å (mean = 7.83 Å). The exo-[2.2.2] and exo-[2.2.1] compounds approximate this reasonably well (7.68 and 7.66 Å, respectively), while the anti-[2.2.1] compound shows a significantly shorter distance (7.25 Å). Only two of the gauche phenylethylamines studied crystallographically have para-oxygen atoms; these have N to p-O distances of 6.26 and 6.27 Å, which are matched by the endo-[2.2.2] structure (6.27 Å). The endo and syn-[2.2.1] systems have N to p-O distances of 6.17 and 6.03 Å.

Table VI also lists the N to C2-O distances observed in x-ray studies of phenylethylamines. In each case, the nitrogen atom is gauche with respect to the C2 oxygen atom. The values range from 2.64 to 2.96 Å (mean = 2.84 Å). The endo-[2.2.2] compound gives a N to C2-O distance of 2.87 Å, well within the observed range. The exo-[2.2.2] and exo-[2.2.1] systems are fairly close to the observed values, having distances of 3.01 and 3.09 Å, respectively. The endo-[2.2.1] structure VIII, which corresponds to an extended arrangement of the nitrogen and oxygen atoms, has a distance of 3.82 Å; there are no known phenyl-

Table VI. Comparison of bicyclic ring systems as frameworks for phenylethylamines.
 (a) A = N to CR; B = N to P-O; C = N to m-O, N to m'-O; D = N to o-O; E = N to C2-O;
 F = CR to C2-O; G = C2-O to p-O; H = C2-O to m-O, C2-O to m'-O; J = C2-O to o-O.

Compound	Dihedral Angles (°)				Bond Angles (°)			
	I ₁ -	I ₁ '-	I ₂ -	I ₃ -	θ ₁ -	θ ₁ '-	θ ₂ -	θ ₂ '-
I	60	123	179	56	127	113	107	107
II	60	123	69	59	124	113	109	109
III	69	108	59	---	131	107	106	106
IV	69	108	176	---	131	107	106	106
V	69	108	168	---	131	107	96	96
VI	69	108	70	---	131	107	100	100
VII	69	108	176	83	131	107	106	106
VIII	69	108	59	160	131	107	106	106
IXa (norepin- ephine HCl)	97	82	176	64				
x ^b (amphetamine sulfate)	i	71	176	---				
	ii	77	173	---				
	iii	69	172	---				
	iv	80	166	---				
XI (mescaline HBr)	80	98	56	---				

TABLE VI (cont'd.)

Compound	<u>A</u>	<u>B</u>	<u>C</u>	<u>D</u>	<u>E</u>	<u>F</u>	<u>G</u>	<u>H</u>	<u>J</u>
I	5.07	7.68	6.37 7.38	5.33	3.01	3.78	6.28	5.08 6.02	2.39
II	3.89	6.27	5.76 5.54	3.95	2.87	3.78	6.28	5.08 6.02	2.39
III	3.85	6.17	5.49 5.51	3.84	-----	-----	-----	-----	-----
IV	5.03	7.66	7.15 6.34	5.09	-----	-----	-----	-----	-----
V	4.74	7.25	7.23 5.47	5.52	-----	-----	-----	-----	-----
VI	3.69	6.03	5.94 4.65	4.53	-----	-----	-----	-----	-----
VII	5.03	7.66	7.15 6.34	5.09	3.09	3.31	5.83	5.80 4.32	4.33
VIII	3.85	6.17	5.49 5.51	3.84	3.82	3.31	5.83	5.80 4.32	4.33
IX ^a	5.10	7.79	6.85	-----	2.85	3.72	-----	-----	-----
X i	5.18	-----	-----	-----	-----	-----	-----	-----	-----
ii	5.17	-----	-----	-----	-----	-----	-----	-----	-----
iii	5.21	-----	-----	-----	-----	-----	-----	-----	-----
iv	5.12	-----	-----	-----	-----	-----	-----	-----	-----
XI	3.89	6.27	5.25 5.93	-----	-----	-----	-----	-----	-----

(a) Reference 38; (b) reference 146. The authors of this reference report τ_1 values of 720, 770, 700 and 830 for i-iv, respectively; (c) reference 42. The authors report a τ_2 value of 56°.

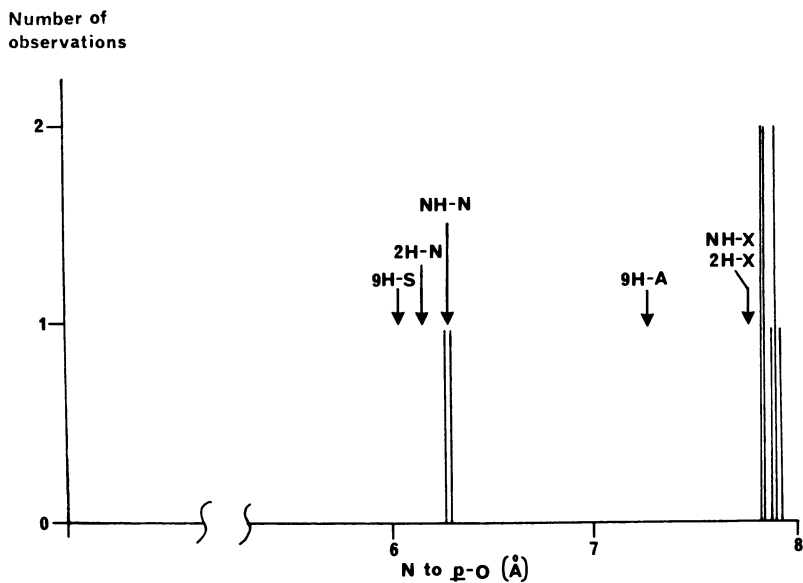


Figure 12. Comparison of the distances from the nitrogen to the para-oxygen in some conformationally defined analogs of phenylethylamines with those found from x-ray crystallographic studies reported for phenylethylamines (see text)

ethylamines which have such a conformation in the crystal structure, so comparisons are not possible in this instance.

The observed interatomic distance from the center of the aromatic ring to the C2-oxygen atom (CR to C2-O) ranges from 3.53 to 3.76 Å, with a mean value of 3.68 Å. The benzobicyclo[2.2.2]octene system gives a reasonably good CR to C2-O distance (3.78 Å). As might be expected, compounds VII and VIII, in which the oxygen atom is built into a somewhat strained structure, provide a much less accurate CR to C2-O distance, 3.31 Å.

In general, the benzobicyclo[2.2.2]octene ring system affords the best model of gauche and extended phenylethylamine conformers. Its main drawback would appear to be the slightly low value of τ_1 . The exo and endo-benzobicyclo[2.2.1]heptenes (structures III and IV) and the 1,2,3,4-tetrahydro-1,4-epoxynaphthalenes (VII and VIII) provide moderately good approximations of phenylethylamine conformations, although the placement of the C2-oxygen atom is somewhat distorted in structures VII and VIII. The syn- and anti-benzobicyclo[2.2.1]heptenes (V and VI) seem to be the least suitable of the structures studied; dihedral angles, bond angles and interatomic distances differ substantially from predicted and observed values. All things considered, the best systems appear to be I and II (NM-X and NM-N for methamphetamine analogs).

Evaluation of Some Bicyclic Systems as Conformationally-Defined Phenylethylamines. CNDO/2 Calculations on Amphetamines. Quantum mechanical studies on phenylethylamine and amphetamine have been carried out previously by Pullman and coworkers (59,67) using both the semi-empirical PCILO and an ab initio method, and by Hall and coworkers (71) using an ab initio method. The results of these two groups were quite similar.

Our aim being to investigate both the energy differences and the charge distributions of a variety of phenylethylamines, we chose the CNDO/2 method of Pople *et al.* (48) as the most appropriate. The PCILO method would have provided a more efficient route to the energy maps, but it appears to be less satisfactory from the point of view of investigating charge distribution in systems of this type, where it has been shown (70) that the CNDO method gives results in best agreement with those obtained by ab initio methods. We considered that, for a series of molecules of this size, ab initio methods were too expensive; and since it is

relatively inexpensive and since its strengths and limitations are well established (48,149,150), we have used the CNDO/2 method. Preliminary investigations on amphetamine, the parent molecule of our series, allowed us to establish that in fact the CNDO/2 method produced energy data very similar to those obtained in the two previous studies.

Data for the basic molecular geometry were taken from the x-ray crystallographic studies cited above (44,145,146,147,148). Hydrogen atoms and substituents were added using standard values for bond lengths and bond angles (44). The results for amphetamine are summarized in Table VII.

The conformation of the aminopropyl side chain may be described in terms of four torsion angles (*atom numbers and torsion angles in this and the following sections are shown in Figure 13, not Figure 2*): τ_1 ($C^1-C^7-C^8$), τ_2 ($C^6-C^7-C^8-N^{10}$), τ_3 ($C^7-C^8-N^{10}-R_2$) and τ_4 ($C^8-N^{10}-C^{11}-C^{12}$). (The usual convention for torsion angles was followed: thus, τ_1 is the angle between the planes through $C^1C^7C^8$ and $C^6C^7C^8$. The angle is positive for clockwise rotations about C^6-C^7 looking from C^6 to C^7 .) When N is unsubstituted, the energy is nearly independent of τ_3 (71), and so it is ignored. When $R_2 = -C_2H_5$ (Figure 13), τ_3 must be chosen with care, but reasonable values for τ_4 are readily found from an examination of molecular models, and were verified by a minimal number of trial calculations. Energy surfaces for amphetamine and N-ethylamphetamine have been plotted in the conventional way as functions of τ_1 and τ_2 (Figures 14 and 15). In addition, the variation of the energy surface of N-ethylamphetamine as a function of τ_2 and τ_3 is presented (Figure 16).

Because of the ring symmetry, the maps of amphetamine and N-ethylamphetamine (Figures 14 and 15) are periodic in τ_1 , with period 180° .

The energy map of amphetamine is presented in Figure 14. Rotations were carried out about τ_1 and τ_2 in approximately 30° increments, points being added to the basic grid where it was thought necessary, particularly in the area $80^\circ \leq \tau_1 < 100^\circ$. Seventy-one points were calculated. The main features are two fairly wide energy wells around $\tau_1 = \pm 90^\circ$ and $-60^\circ < \tau_2 \leq 60^\circ$. These contain local hollows, and two of these hollows, at $\tau_1 = \pm 90^\circ$, $\tau_2 = 30^\circ$, are the global minima. They represent the folded form of the molecule with the nitrogen over the ring. Local minima also occur at $\tau_1 = \pm 90^\circ$, $\tau_2 = 180^\circ$, which correspond to the fully extended form in which the aminopropyl side chain is

Table VII

Compound	Energies		Approximate Locations					
	Barriers	ΔE_b	Global Minimum		Local Minimum ^c		Barrier	
			τ_1	τ_2	τ_1	τ_2	τ_1	τ_2
Amphetamine (CNDO)	~ 6	2	60° -120°	30°	+90°	180°	100° -80°	120°
Amphetamine (PCILO) (59,67)	3-4	0	+90°	+60° 180°	+90°	+60° 180°	+90°	+120°
Amphetamine (ab initio) (67)	~ 5	2	+90°	-60°	+90°	60° 180°	+90°	-120°
Amphetamine (ab initio) (71)	~ 8	5	+90°	60°	120° -60°	180°	+90°	120°
N-Ethylamphetamine (CNDO)	9-10	6	100° -80°	15°	100° -80°	180°	+90°	120°

^aBarrier height between folded and extended forms (kcal/mol).

^bEnergy difference between folded and extended forms (kcal/mol).

^cThis refers to the extended form except for the PCILO case, where the minima are equivalent.

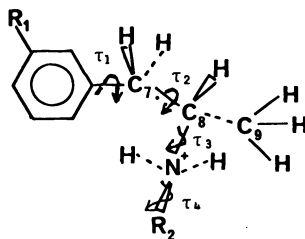


Figure 13. Numbering scheme used for the CNDO calculations on amphetamine and N-ethylamphetamine

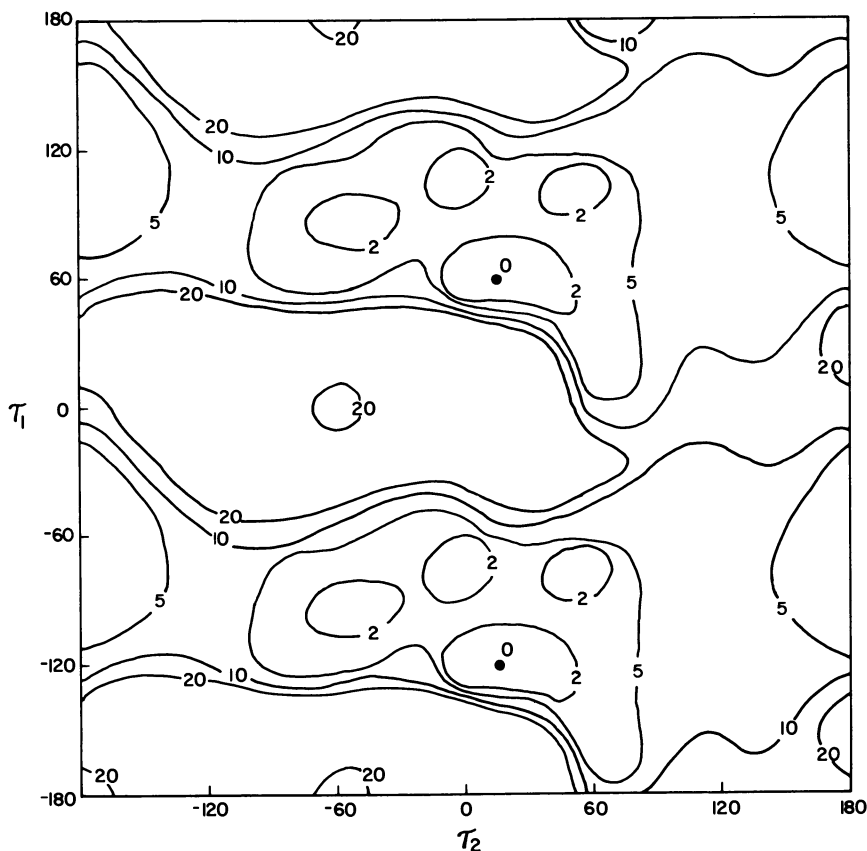


Figure 14. Conformational energy map of amphetamine as calculated using the CNDO/2 method (Kcal/mol with respect to the global minimum). See Figure 13 for the numbering scheme.

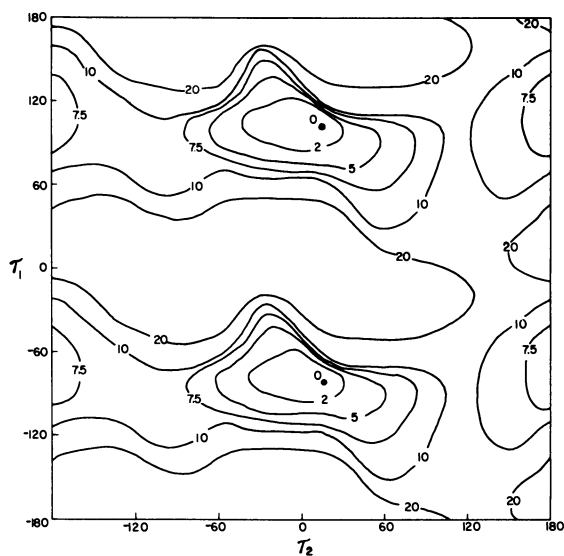


Figure 15. Conformational energy map of N-ethylamphetamine as calculated using the CNDO/2 method (Kcal/mol with respect to the global minimum). See Figure 13.

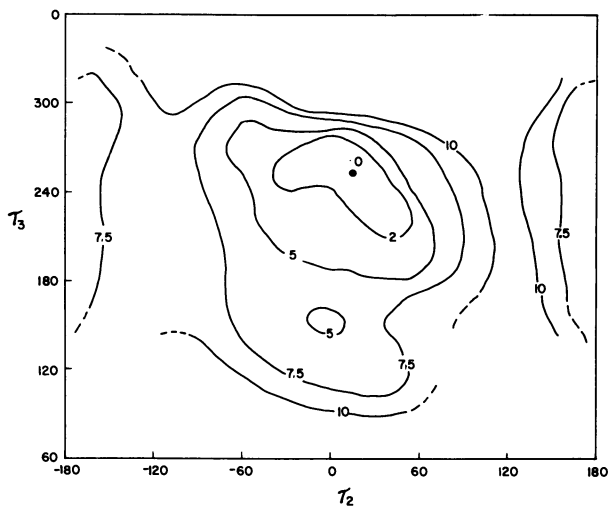


Figure 16. Conformational energy map of N-ethylamphetamine for τ_2/τ_3 as calculated using the CNDO/2 method (Kcal/mol with respect to the global minimum). See Figure 13 for the numbering scheme.

in the more stable staggered conformation. There is an energy difference of about 2 kcal/mol between the folded and extended forms. The barrier between them is about 5.5 kcal/mol. These values are quite similar to those reported previously, though the positions of the minima vary somewhat (see Table VII). These calculations suggest that use of CNDO/2 to explore the conformationally-defined systems would be appropriate.

CNDO/2 Calculations on Bicyclic Systems. Calculation of the energy differences between the exo- and endo-isomers in either the benzobicyclo[2.2.2]octene or benzobicyclo[2.2.1]heptene system (e.g., I, II, III and IV in Figure 7) is very dependent on the molecular geometries. If x-ray crystallographic coordinates are used, the normal crystal packing distortions will make such a comparison of little value. However, if the x-ray coordinates of the heavy atoms of one member of each pair (e.g., NM-X in Figure 5) and standard hydrogen positions are used as the basis of the calculations, and if the same coordinates are used for the other isomer (NM-N) except that the exo-methylamino group is switched to the endo-position, then an energy difference giving very good agreement with those in Table VII is found. Using this approach in the benzobicyclo[2.2.2]octene system (NM-X and NM-N), a ΔE value of 0.1 kcal/mol is found. Repeating this process in reverse (using the heavy atom coordinates for NM-N and standard hydrogen positions) and then switching the position of the methylamino group into the exo-position, gives a ΔE value of 6.9 kcal/mol. For the benzobicyclo[2.2.1]heptene system, using literature values for the heavy atom coordinates of a related system (147,148) and proceeding as above, a ΔE value of 0.01 kcal/mol is found between the 2-exo-methylamino (2M-X) and 2-endo-methylamino (2M-N) systems. This is totally consistent with the nearly equal energy of extended and gauche conformers of amphetamine noted above.

Of more interest is a calculation of the net atomic charge on each atom of the rigid systems and a comparison to that of an N-substituted amphetamine (we used the values for N-ethylamphetamine). The results of this analysis are presented graphically in Figure 17 using the atom numbering scheme of Figure 13.

The similarity in charge distribution on the rigid systems compared to that of N-ethylamphetamine (in a conformation corresponding to that of the rigid analogs) is encouraging for it allows one to predict that when

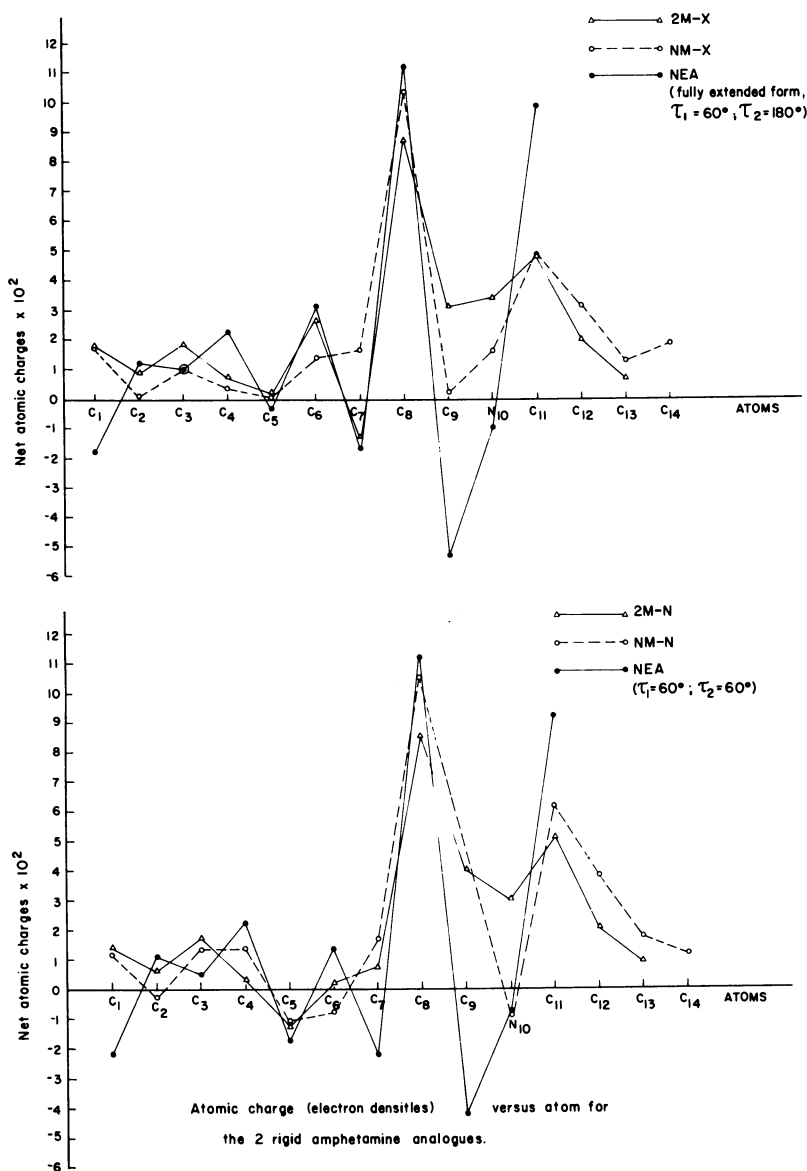


Figure 17. Comparison of the atomic charges as calculated using the CNDO/2 method on the protonated amines.

NEA is N-ethylamphetamine, NM-X is the *exo*- and NM-N the *endo*-isomer of 2-methylaminobenzobicyclo[2.2.2]octene, and 2M-X is the *exo*- and 2M-N the *endo*-isomer of 2-methylaminobenzobicyclo[2.2.1]heptene. The numbering for C₁ through C₉ and N₁₀ is as shown in Figure 13. C₁₁ is the carbon attached to the nitrogen (the N-methyl substituent in the bicyclic compounds and the methylene group of the ethyl substituent in NEA), C₁₂ is the bridgehead atom of the bicyclic system that is not a part of the amphetamine skeleton, and C₁₃ and C₁₄ are the bridge atoms (not part of the amphetamine skeleton) bonded to atoms C₇ and C₁₂.

one of the rigid systems approaches a pharmacological site of action, it will do so by presenting a similar charge distribution as would the parent drug (in this case, amphetamine or methamphetamine).

Synthesis. The synthesis of the amphetamine and methamphetamine derivatives in the benzobicyclo[2.2.2]-octene system has been described elsewhere (141 and references cited therein). To be able to study certain of the pharmacological aspects of these compounds, ³H-derivatives were prepared of NM-X and NM-N by reducing the N-formyl derivatives of the amphetamine analogs NH-X and NH-N with lithium aluminum tritide (142).

Tests of the Model

Physical Properties. For the rigid methamphetamine analogs NM-X and NM-N to be good models for pharmacological studies and of any value in the determination of conformation-activity relationships, it is necessary that they show similar physical properties to the parent. Two properties of importance are the basicity and the lipophilicity, for these properties will greatly influence transport to the various sites of action.

The pK_a values were determined (142) and show a close agreement to those expected, namely that amphetamine has a value intermediate between those of NM-X and NM-N (NM-X, 10.01 ± 0.03; amphetamine, 9.97 ± 0.03; NM-N, 9.76 ± 0.04).

A similar situation is found when the *n*-octanol/buffer (pH 7.4) partition coefficients were determined (142). The value for amphetamine is intermediate between that for either NM-N or NM-X.

Thus, comparison of the charge distribution calculated by CNDO/2, the pK_a values and the partition coefficients shows that both NM-X and NM-N are excellent models for a study of conformation-activity relationships of methamphetamine.

Pharmacology. In order to assess the importance of conformational preferences in the action of amphetamine at the adrenergic nerve endings, the effect of the amphetamine and methamphetamine analogs (NH-X, ³H-NH-N, NM-X and NM-N) on the uptake and release of ³H-norepinephrine (NE) and ³H-dopamine (DA) in chopped tissues from various regions of rat brain was examined. NH-X and NM-X were nearly as potent as amphetamine in the inhibition of uptake of ³H-NE into chopped cortex

and ^3H -DA into striatal tissue. The endo isomers NH-N and NM-N were much less active. NH-X and NM-X were more effective than the endo analogs in the release of previously accumulated ^3H -NE and ^3H -DA from the same tissue preparation although both were considerably less potent than amphetamine. ^3H -NM-X and ^3H -NM-N were accumulated to the same extent into synaptosomes from rat hypothalamus, and the accumulation was sensitive to desipramine. Behavioral studies indicated that if deamination of released amine is inhibited by pargyline, NM-X is as potent as methamphetamine (MA) in increasing locomotor activity in mice, whereas NM-N is much less active. In reserpine and pargyline pretreated rat atrial slices, NM-X was as potent as NE and MA in the release of previously accumulated ^3H -NE. The release was sensitive to desipramine. NM-N was inactive in this system. Similarly, MA and NM-X were equipotent in the inhibition of uptake of ^3H -NE into pargyline and reserpine treated rat atrial tissue while NM-N was a thousand-fold less active. In untreated rat atria (primarily vesicular NE storage) MA produced a pronounced increase in atrial rate, while NM-X and NM-N were inactive. Following exposure of the atria to pargyline, NM-X was nearly as potent as MA in producing rate acceleration. When reserpine and pargyline treated rat atria were exposed to cold NE followed by extensive washout to remove extraneuronal NE (primarily extravesicular neuronal NE) NM-X and MA were equipotent in the release of NE, producing the same rate increases at concentrations which were ineffective under conditions of primarily vesicular neuronal NE storage. As the releaseable pool of NE was changed from primarily vesicular to primarily extravesicular, the ability to release NE decreased ten-fold for MA and disappeared for NM-X. The results suggest that MA and NM-X do not differ in their relative abilities to interact with the neuronal amine uptake site, to be accumulated in a desipramine sensitive fashion, or to release extravesicular NE. However, it appears that either conformational mobility or a conformation of MA not approximated by either rigid analog is required for the release of vesicular NE (141,142,143).

Conclusion. Thus, conformation-activity relationships have clearly been found using the endo- and exo-2-amino (or 2-methylamino) isomers of benzocyclo[2.2.2]octene as models for the gauche and extended conformations of amphetamine and methamphetamine. Further work to extend these studies into substituted amphetamines and catecholamines is in

progress using all of the ring systems (I-VIII) in Figure 7.

List of Abbreviations

<u>a</u>	axial
<u>AI</u>	<u>ab initio</u>
CNDO (CNDO/2)	Complete Neglect of Differential Overlap
COMT	catechol O-methyltransferase
CR	center of the aromatic ring
DA	dopamine
DMI	desipramine
<u>e</u>	equatorial
<u>EHT</u>	Extended Hückel Theory
INDO	Intermediate Neglect of Differential Overlap
MA	methamphetamine
MAO	monoamine oxidase
M.O.	molecular orbital
NE	norepinephrine
NEA	N-ethylamphetamine
NH-N	<u>endo</u> -2-aminobenzobicyclo[2.2.2]- octene
NH-X	<u>exo</u> -2-aminobenzobicyclo[2.2.2]octene
NM-N	<u>endo</u> -2-methylaminobenzobicyclo- [2.2.2]octene
NMR	nuclear magnetic resonance
NM-X	<u>exo</u> -2-methylaminobenzobicyclo[2.2.2]- octene
OH-N	<u>endo</u> -2-amino-1,2,3,4-tetrahydro- 1,4-epoxynaphthalene
OH-X	<u>exo</u> -2-amino-1,2,3,4-tetrahydro-1,4- epoxynaphthalene
PCILO	Perturbative Configuration Inter- actions with Localized Orbitals
SAR	structure-activity relationships
2HN	<u>endo</u> -2-aminobenzobicyclo[2.2.1]- heptene
2HX	<u>exo</u> -2-aminobenzobicyclo[2.2.1]- heptene
9H-A	<u>anti</u> -9-aminobenzobicyclo[2.2.1]- heptene
9H-S	<u>syn</u> -9-aminobenzobicyclo[2.2.1]- heptene

Acknowledgements

We wish to acknowledge the assistance of Mr. Michael A. Pleiss, Dr. Thomas J. Reitz and Ms. Carolyn

J. Kelly in the preparation of this manuscript. The fruitful pharmacological collaboration with Professors P. N. Patil, C. O. Rutledge and J. A. Ruth is also acknowledged. Financial support of this work was provided by the U.S. Public Health Service (Training Grant GM 1341, Research Grants GM 22988 for the chemical, x-ray and pharmacological studies and DA 01990 for the CNDO/2 studies, Biomedical Research Support Grant RR 05606 for pilot studies), the University of Kansas for computation time, and the Kansas Affiliate of the American Heart Association for support of some of the pilot studies leading up to the work described above.

Literature Cited

1. Geffen, L.B. and B.G. Livett, Physiol. Rev., (1971), 51, 98.
2. Kirshner, N. and O.H. Viveros, Pharmacol. Rev., (1972), 24, 385.
3. Holtzman, E., Neurosci., (1977), 2, 327.
4. Carlsson, A., N.-A. Hillarp and B. Waldeck, Acta Physiol. Scand. Suppl., (1963), 59, 1.
5. Philippu, A. in "The Mechanism of Neuronal and Extraneuronal Transport of Catecholamines", Ed. by Paton, D.M., Raven Press, New York, (1976), p. 215.
6. Giachetti, A. and R. A. Hollenbeck, Br. J. Pharmacol., (1976), 58, 497.
7. Giachetti, A. and P.R. Shore, Biochem. Pharmacol., (1966), 15, 607.
8. Weiner, N. and C.O. Rutledge in "Mechanism of Release of Biogenic Amines," Pergamon Press, Oxford, (1966), p. 307.
9. Stitzel, R, Pharmacol. Rev., (1977), 28, 179.
10. Farnebo, L.-O., Biochem. Pharmacol., (1971), 20, 2715.
11. Farnebo, L.-O., Acta Physiol. Scand. Suppl., (1971), 371, 45.
12. Kalisker, A., J.C. Waymire and C.O. Rutledge, J. Pharmacol. Exp. Ther., (1975), 193, 64.
13. Axelrod, J., Pharmacol. Rev., (1966), 18, 95.
14. Bogdanski, D.F. and B.B. Brodie, J. Pharmacol. Exp. Ther., (1969), 165, 181.
15. Paton, D.M., Br. J. Pharmacol., (1973), 49, 614.
16. De Robertis, E., Science, (1971), 171, 963.
17. Nathanson, J.A., Physiol. Rev., (1977), 57, 157.
18. Langer, S.Z., Br. J. Pharmacol., (1977), 60, 481.
19. Starke, K., Rev. Physiol. Biochem. Pharmacol., (1977), 77, 1.
20. Maxwell, R.A., R.M. Ferris and J.E. Buresu in

- "The Mechanism of Neuronal and Extraneuronal Transport of Catecholamines," Ed. by Paton, D.M., Raven Press, New York, (1976), p. 95.
21. Burgen, A.S.V. and L.L. Iversen, Br. J. Pharmacol., (1965), 25, 34.
 22. Grunewald, G.L., J.M. Grindel, P.N. Patil and K.N. Salman, J. Med. Chem., (1976), 19, 10.
 23. Grunewald, G.L., T.J. Reitz, J.A. Ruth, S. Vollmer, L.E. Eiden and C.O. Rutledge, Biochem. Pharmacol., (1979), 28, 417.
 24. Ross, S.B. in "The Mechanism of Neuronal and Extraneuronal Transport of Catecholamines," Ed. by Paton, D.M., Raven Press, New York, (1976), p. 67.
 25. Paton, D.M. in "The Mechanism of Neuronal and Extraneuronal Transport of Catecholamines," Ed. by Paton, D.M., Raven Press, New York, (1976), p. 155.
 26. Triggle, D.J.: "Neurotransmitter-Receptor Interactions," Academic Press, New York, (1971), p. 209.
 27. Borchardt, R.T., J. Med. Chem., (1973), 16, 387.
 28. Smissman, E.E. and R.T. Borchardt, J. Med. Chem., (1971), 14, 377.
 29. Portoghese, P.S., J. Pharm. Sci., (1971), 60, 806.
 30. Tsoucaris, G., Acta Crystallogr., (1961), 14, 909.
 31. Bergin, R., Acta Crystallogr. Sect. B., (1971), 27, 381.
 32. Hearn, R.A., G.R. Freeman and C.E. Bugg, J. Amer. Chem. Soc., (1973), 95, 7150.
 33. Hearn, R.A. and C.E. Bugg, Acta Crystallogr. Sect. B., (1972), 28, 3662.
 34. Bergin, R. and D. Carlstrom, Acta Crystallogr. Sect. B. (1968), 24, 1506.
 35. Andersen, A.M., A. Mostad and C. Romming, Acta Chem. Scand., (1972), 26, 2670.
 36. Kolderup, M., A. Mostad and C. Romming, Acta Chem. Scand., (1972), 26, 483.
 37. Carlstrom, D., Acta Crystallogr. Sect. B., (1973), 29, 161.
 38. Carlstrom, D. and R. Bergin, Acta Crystallogr., (1967), 23, 313.
 39. Mathew, M. and G.J. Palenik, J. Amer. Chem. Soc., (1971), 93, 497.
 40. Baker, R.W., C. Chothia, P. Pauling and H.P. Weber, Mol. Pharmacol., (1973), 9, 23.
 41. Kennard, O., C. Giacobozzo, A.S. Horn, R. Mongiorgi and L. Riva Di Sanseverino, J. Chem. Soc. Perkin Trans., Sect. B., (1974), 2, 1160.
 42. Ernst, S.R. and F.W. Cagle, Jr., Acta Crystallogr. (1973), B29, 1543.
 43. Tsoucaris, D., C. De Rango, G. Tsoucaris, C. Zelwer R. Parthasarathy and F.E. Cole, Cryst. Struct.

- Commun., (1973), 2, 193.
44. Pauling, L.: "The Nature of the Chemical Bond," Cornell University Press, Ithaca, New York, (1960), p. 257.
 45. Snyder, S.H., H. Weingartner and L.A. Faillace, Arch. Gen. Psychiatry, (1971), 24, 50.
 46. Lohr, L.L. and W.N. Lipscomb, J. Chem. Phys., (1963), 38, 1607.
 47. Pople, J.A. and G.A. Segal, J. Chem. Phys., (1966), 44, 3289.
 48. Pople, J.A., D.L. Beveridge and P.A. Dobosh, J. Chem. Phys., (1967), 47, 2026.
 49. Diner, S., J.P. Malrieu, F. Jordan and M. Gilbert, Theor. Chim. Acta., (1969), 15, 100.
 50. Christoffersen, R.E. and G.M. Maggiora, Chem. Phys. Lett., (1969), 3, 419.
 51. Hoyland, J.R. in "Molecular Orbital Studies in Chemical Pharmacology," Ed. by Kier, L.B., Springer-Verlag, New York, (197), p. 31.
 52. Kier, L.B., J. Pharmacol. Exp. Ther., (1968), 164, 75.
 53. Kier, L.B., J. Pharm. Pharmacol., (1969), 21, 93.
 54. Kier, L.B. and E.B. Truitt, Jr., J. Pharmacol. Exp. Ther., (1970), 174, 94.
 55. Bustard, T.M. and R.S. Egan, Tetrahedron, (1971), 27, 4457.
 56. George, J.M., L.B. Kier and J.R. Hoyland, Mol. Pharmacol., (1971), 7, 328.
 57. Pedersen, L., R.E. Hoskins and H. Cable, J. Pharm. Pharmacol., (1971), 23, 216.
 58. Rekker, R.F., D.J.C. Engel and G.G. Nys, J. Pharm. Pharmacol., (1972), 24, 589.
 59. Pullman, B., J.-L. Coubeils, P. Courriere and J.-P. Gervois, J. Med. Chem., (1972), 15, 17.
 60. Coubeils, J.-L., P. Courriere and B. Pullman, J. Med. Chem., (1972), 15, 453.
 61. Germer, H.A., Jr., Tex. Rep. Biol. Med., (1973), 31, 189.
 62. Kier, L.B., J. Theor. Biol., (1973), 40, 211.
 63. Katz, R., S.R. Heller and A.E. Jacobson, Mol. Pharmacol., (1973), 9, 486.
 64. Katz, R. and A.E. Jacobson, Mol. Pharmacol., (1973), 9, 495.
 65. Pullman, B., and P. Courriere, Compt. Rend., (1973), 276D, 1907.
 66. Germer, H.A., Jr., J. Pharm. Pharmacol., (1974), 26, 467.
 67. Pullman, B., H. Berthod and P. Courriere, Int. J. Quantum Chem., QBS, (1974), 1, 93.
 68. Pullman, B., H. Berthod and A. Pullman, An. Quim.,

- (1974), 70, 1204.
69. Petrongolo, C., J. Tomasi, B. Macchia and F. Macchia, J. Med. Chem., (1974), 17, 501.
70. Martin, M., R. Carbo, C. Petrongolo and J. Tomasi, J. Amer. Chem. Soc., (1975), 97, 1338.
71. Hall, G.G., C.J. Miller and G.W. Schnuelle, J. Theor. Biol., (1975), 53, 475.
72. Grof, C.J. and H. Rollema, J. Pharm. Pharmacol., (1977), 29, 153.
73. Hyne, J.B., Can J. Chem., (1961), 39, 2536.
74. Portoghese, P.S., J. Med. Chem., (1967), 10, 1057.
75. Forrest, J.E., R.A. Heacock and T.P. Forrest, J. Pharm. Pharmacol., (1970), 22, 512.
76. Neville, G.A., R. Deslauriers, B.J. Blackburn and I.C.P. Smith, J. Med. Chem., (1971), 14, 717.
77. Bailey, K., J. Pharm. Sci., (1971), 60, 1232.
78. Bailey, K., A.W. By, K.C. Graham and D. Verner, Can. J. Chem., (1971), 49, 3143.
79. Ceccarelli, G., A. Balsamo, P. Crotti, B. Macchia and F. Macchia, 11th Congresso Nazionale della Societa Chimica Italiano, Perugia, (October, 1972).
80. Wright, G.E., Tetrahedron Lett., (1973), 1097.
81. Ison, R.R., P. Partington and G.C.K. Roberts, Mol. Pharmacol., (1973), 9, 756.
82. Smissman, E.E. and R.T. Borchardt, J. Med. Chem., (1971), 14, 377.
83. Smissman, E.E. and W.H. Gastrock, J. Med. Chem., (1968), 11, 860.
84. Smissman, E.E. and S. El-Antably, J. Med. Chem., (1971), 14, 30.
85. Smissman, E.E. and R.T. Borchardt, J. Med. Chem., (1971), 14, 383.
86. Smissman, E.E. and T.L. Pazdernik, J. Med. Chem., (1973), 16, 18.
87. Tuomisto, J., L. Tuomisto and E.E. Smissman, Ann. Med. Exp. Biol. Fenn., (1973), 51, 51.
88. Tuomisto, L., J. Tuomisto and E. E. Smissman, Eur. J. Pharmacol., (1974), 25, 351.
89. Tuomisto, J., E.E. Smissman, T.L. Pazdernik and E. J. Walaszek, J. Pharm. Sci., (1974), 63, 1708.
90. Hava, M., J. Bernstein, E.E. Smissman and S. El-Antably, J. Med. Chem., (1976), 19, 52.
91. Smissman, E.E. and G.S. Chappell, J. Med. Chem., (1969), 12, 429.
92. Nelson, W.L. and D.D. Miller, J. Med. Chem., (1970), 13, 807.
93. Nelson, W.L. and B.E. Sherwood, J. Pharm. Sci., (1974), 63, 1467.
94. Smissman, E.E. and T.L. Pazdernik, J. Med. Chem., (1973), 16, 14.

95. Yupraphat, T., H.-J. Rimek and F. Zymalkowski, *Justus Liebig's Ann. Chem.*, (1970), 738, 79.
96. Lal, B., J.M. Khanna and N. Anand, *J. Med. Chem.*, (1972), 15, 23.
97. Singh, G.B., R.C. Srimal and B.N. Dhawan, *Jpn. J. Pharmacol.*, (1974), 24, 5.
98. Hagishita, S. and K. Kuriyama, *Chem. Pharm. Bull.*, (1976), 24, 1724.
99. Sugihara, H. and Y. Sanno, *Chem. Pharm. Bull.*, (1977), 25, 859.
100. Cannon, J.G., G.J. Hatheway, J.P. Long and F.M. Sharabi, *J. Med. Chem.*, (1976), 19, 987.
101. Sharabi, F.M., J.P. Long, J.G. Cannon and G.J. Hatheway, *J. Pharmacol. Exp. Ther.*, (1976), 199, 630.
102. Zymalkowski, F. and H.-J. Rimek, *Arch. Pharm.* (Weinheim, Ger.), (1961), 294, 581.
103. Thrift, R.I., *J. Chem. Soc., Sect. C.*, (1967), 288.
104. Pinder, R.M., D.A. Buxton and G.N. Woodruff, *J. Pharm. Pharmacol.*, (1972), 24, 903.
105. Cannon, J.G., J.C. Kim, M.A. Aleem and J.P. Long, *J. Med. Chem.*, (1972), 15, 348.
106. Horn, A.S., *J. Pharm. Pharmacol.*, (1974), 26, 735.
107. Nishikawa, M., M. Kanno, H. Kuriki, H. Sugihara, M. Motohashi, K. Itoh, O. Miyashita, Y. Oka and Y. Sanno, *Life Sci.*, (1975), 16, 305.
108. McDermed, J.D., G.M. McKenzie and A.P. Phillips, *J. Med. Chem.*, (1975), 18, 362.
109. Elkhawad, A.O. and G.N. Woodruff, *Br. J. Pharmacol.*, (1975), 54, 107.
110. McDermed, J.D., G.M. McKenzie and H.S. Freeman, *J. Med. Chem.*, (1976), 19, 547.
111. Ilhan, M., J.P. Long and J.G. Cannon, *Arch. Int. Pharmacodyn. Ther.*, (1976), 223, 215.
112. Ilhan, M., J.P. Long and J.G. Cannon, *Arch. Int. Pharmacodyn. Ther.*, (1976), 219, 193.
113. Ilhan, M., J.P. Long and J.G. Cannon, *Arch. Int. Pharmacodyn. Ther.*, (1976), 222, 70.
114. Kohli, J.D., P.H. Volkman, L.I. Goldberg, T. Lee and J.G. Cannon, *Pharmacologist*, (1976), 18, 228.
115. Oka, Y., M. Motohashi, H. Sugihara, O. Miyashita, K. Itoh, M. Nishikawa and S. Yurugi, *Chem. Pharm. Bull.*, (1977), 25, 632.
116. Woodruff, G.N., K.J. Watling, C.D. Andrews, J.A. Poat and J.D. McDermed, *J. Pharm. Pharmacol.*, (1977), 29, 422.
117. Costall, B., R.J. Naylor, J.G. Cannon and T. Lee, *Eur. J. Pharmacol.*, (1977), 41, 307.
118. Cannon, J.G., T. Lee, H.D. Goldman, B. Costall and R.J. Naylor, *J. Med. Chem.*, (1977), 20, 1111.

119. Ilhan, M., J.M. Kitzen, J.G. Cannon and J.P. Long, Eur. J. Pharmacol., (1977), 41, 301.
120. Toth, E., G. Fassina and E. SantiSocin, Arch. Int. Pharmacodyn. Ther., (1967), 169, 375.
121. Heikkilä, R., G. Cohen and D. Dembiec, J. Pharmacol. Exp. Ther., (1971), 179, 250.
122. Cohen, G., C. Mytilineau and R.E. Barrett, Science, (1972), 175, 1269.
123. Gray, A.P., E. Reit, J.A. Ackerly and M. Hava, J. Med. Chem., (1973), 16, 1023.
124. Greenberg, R.S. and G. Cohen, J. Pharmacol. Exp. Ther., (1973), 184, 119.
125. Locke, S., G. Cohen and D. Dembiec, J. Pharmacol. Exp. Ther., (1973), 187, 56.
126. Simpson, L.L., J. Pharmacol. Exp. Ther., (1975), 192, 365.
127. Smissman, E.E., J.R. Reid, D.A. Walsh and R.T. Borchardt, J. Med. Chem., (1976), 19, 127.
128. Rimek, H.-J., T. Yuraphat and F. Zymalkowski, Justus Liebig's Ann. Chem., (1969), 725, 116.
129. Horn, A.S. and S.H. Snyder, J. Pharmacol. Exp. Ther., (1972), 180, 523.
130. Komiskey, H.L., F.L. Hsu, F.J. Bossart, J.W. Fowble, D.D. Miller and P.N. Patil, Eur. J. Pharmacol., (1978), 52, 37.
131. Kaiser, C., B.M. Lester, C.L. Zirkle, A. Burger, C.S. Davis, T.J. Delia and L. Zirngibl, J. Med. Chem., (1962), 5, 1243.
132. Zirkle, C.L., C. Kaiser, D.H. Tedeschi, R.E. Tedeschi and A. Burger, J. Med. Chem., (1962), 5, 1265.
133. Hendley, E.D. and S.H. Snyder, Nature, (1968), 220, 1330.
134. Miller, D.D., F.-L. Hsu, K.N. Salman and P.N. Patil, J. Med. Chem., (1976), 19, 180.
135. Komiskey, H.L., F.J. Bossart, D.D. Miller and P. N. Patil, Pharmacologist, (1977), 19, 225.
136. Miller, D.D., J. Fowble and P.N. Patil, J. Med. Chem., (1973), 16, 177.
137. Burn, P., P.A. Crooks and J.M.H. Rees, J. Pharm. Pharmacol., (1976), 27 (Supplement), 80P.
138. Wood, L.E., Res. Commun. Chem. Path. Pharmacol., (1978), 21, 169.
139. Thompson, E. and L. Wood, Res. Commun. Chem. Path. Pharmacol., (1978), 21, 447.
140. Grunewald, G.L., T.J. Reitz, C.O. Rutledge and J.A. Ruth, unpublished results.
141. Grunewald, G.L., J.A. Ruth, T.R. Kroboth, B.V. Kamdar, P.N. Patil and K.N. Salman, J. Pharm. Sci., (1976), 65, 920.

142. Bartholow, R.M., L.E. Eiden, J.A. Ruth, G.L. Grunewald, J. Siebert and C.O. Rutledge, J. Pharmacol. Exp. Ther., (1977), 202, 532.
143. Ruth, J.A., G.L. Grunewald and C.O. Rutledge, J. Pharmacol. Exp. Ther., (1978), 204, 615.
144. Faraj, B.A., Z.H. Israili, N.E. Kight, E.E. Smissman and T.J. Pazdernik, J. Med. Chem., (1976), 19, 20.
145. Grunewald, G.L., D.E. Walters, D.L. Flynn, S.D. Atwood, M.W. Creese, B.A. Frenz and J.M. Troup, Acta Crystallogr. Sect. B., (1978), 34, 3462.
146. Bergin, R. and D. Carlstrom, Acta Crystallogr. Sect. B., (1971), 27, 2146.
147. Koyama, H. and K. Okada, J. Chem. Soc., Sect. B., (1969), 940.
148. Apgar, P.A. and M.L. Ludwig, J. Amer. Chem. Soc., (1972), 94, 964.
149. Davidson, R.B. and C.R. Williams, J. Amer. Chem. Soc., (1978), 100, 73.
150. Halgren, T.A., D.L. Kleier, J.H. Hall, Jr., L.D. Brown and W.N. Lipscomb, J. Amer. Chem. Soc., (1978), 100, 6595.

RECEIVED July 10, 1979.

Configurational Analysis, Inversion, and Reduction of Some Pyridine Carbaldoximes

RODNEY PEARLMAN and NICHOLAS BODOR¹

Department of Pharmaceutical Chemistry, University of Kansas,
Lawrence, KS 66045

The complex process of reduction of pyridine and protonation of the various dihydro species has successfully been treated recently by MINDO/3 SCF MO calculations (1). The obtained results encouraged us to extend a similar treatment to some of the biologically important pyridine derivatives, particularly some pyridinenealdoximes, one of which the N-methyl-2-pyridiniumaldoxime (2-PAM) is the drug of choice in the treatment of organophosphate poisoning. The dihydropyridine \rightleftharpoons pyridine redox system was recently applied for the delivery of this drug through the blood-brain barrier (2, 3, 4, 5). The present work reports detailed studies on a) the controversial issue of the structure of the active form of 2-PAM chloride (syn or anti), on b) the mechanism of syn-anti interconversion, and c) the process and results of the reduction of 2-PAM chloride.

Experimental evidence indicates (6) that the simple pyridine-2-aldoxime is predominantly in the syn configuration. On the other hand, the authors concluded that both syn and anti isomers can be obtained for the corresponding quaternary oxime (i.e. 2-PAM) derivatives. (It is necessary to specify that in order to avoid confusion with the literature, the old nomenclature was maintained throughout the present work, according to which a syn isomer refers to cis configuration between the pyridine ring and the N-lone pair.) It was also concluded that in the N-methyl case, the more stable isomer is the anti one. The less stable isomer could be isolated, but it readily reverts to the stable form. Wilson's complementarity theory (7) supports this observation that the more stable and the active form of 2-PAM chloride is the anti isomer. Some aspects of the general complementarity theory of Wilson were criticized by Poziomek, et al. (8), who determined that in the case of N-methyl-4-pyridiniumaldoxime the syn configuration is the active form. Due to the obvious structural

¹Current address: College of Pharmacy (Box J-4), J. Hillis Miller Health Center, U. of Florida, Gainesville, Fla. 32610

differences (the distance between the oxime and the quaternary head), however, the 4- and the 2-aldoximes cannot really be compared for their activity. This conclusion is nicely supported in the case of the corresponding phenylketoximes (9). Thus, in the series of phenyl-N-methylpyridinium ketoximes, the anti-2 derivative is twenty times more active than the syn isomer, whereas the anti-4 derivative is six times less active than the corresponding syn isomer. These findings for the anti-syn potency relationship were recently reconfirmed (10) in the case of the phenyl-N-methyl-2-pyridinium ketoximes. All this evidence would imply that the active form of 2-PAM chloride is in the anti configuration. The only contradictory example, however, is given by crystallographic studies of N-methylpyridine-2-aldoxime iodide. It was found (11) that the crystalline form of 2-PAM·I contains the syn isomer and that the oxime is coplanar with the pyridinium ring.

Theoretical Procedure

The calculations reported here were carried out with the latest version of the MINDO/3 SCF MO program (12, 13) which is an improved version in the series of MINDO (14)-MINDO/2 (15, 16)-MINDO/2' (17) methods. The powerful and fast DFP (12) optimization procedure allowed us to calculate the various structures without any geometrical restrictions. The same method was very successful in the case of the reduction and protonation processes of pyridine and related compounds (1).

Results and Discussions

First, the simple pyridine-2-aldoximes were calculated in the syn and anti forms. As shown in Table I, it was found that the syn configuration (II) is slightly more stable than the anti (I) one, in agreement with the experimental results (6). The calculated geometries shown in Figure 1 indicate that the oxime portion is rotated off the plane of pyridine to a position of about 40°, both in the anti and the syn forms. The rotational barrier, however, is relatively small as indicated in Figure 2. The rotation around the C-C exocyclic bond requires maximum 3.0 or minimum .5 kcal/mol as compared to the most stable form. These values are certainly more in the expected range, than the high rotational barrier calculated in the case of 2-PAM cation, using EHMO procedure (18). The smallest rotational barrier calculated of about 30 kcal/mol indicates the inadequacy of the EHMO theoretical procedure used for this kind of quantitative study.

The corresponding anions of the anti and syn forms were also calculated and found to be essentially coplanar, indicating a more extensive interaction between the two unsaturated systems (pyridine ring and oximate). Accordingly, the bond length in the anion between the N-O is significantly shorter (1.25 Å from 1.36 Å), as well as the CNO angle is significantly larger (131°) than

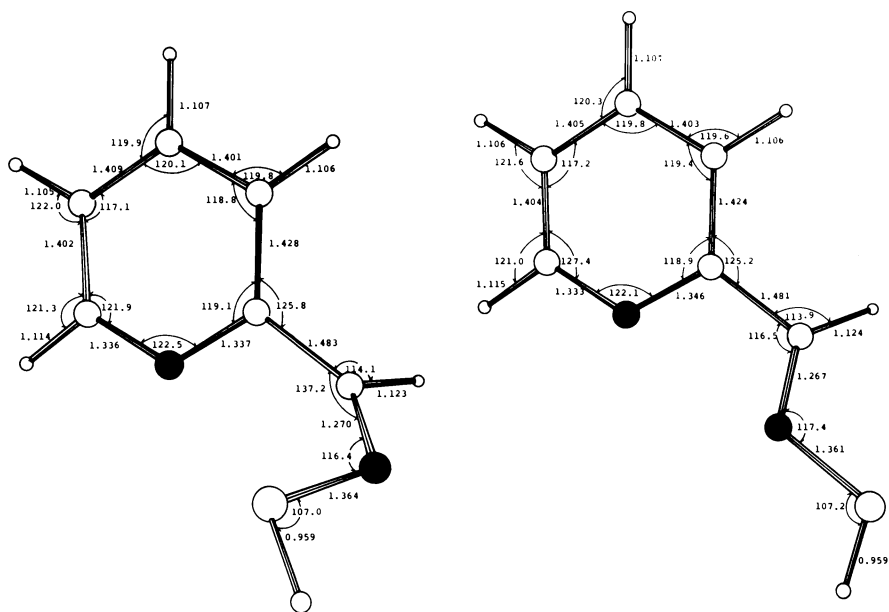


Figure 1. Calculated geometries of the most stable (optimized) forms of anti-(I) and syn (II)-2-pyridinecarbaldoxime. The pyridine ring is essentially planar in both cases, while the oxime function is rotated 35.4° out of plane in I and 39.5° out of plane in II, and the O—H is coplanar and trans with respect to the double bond in both structures.

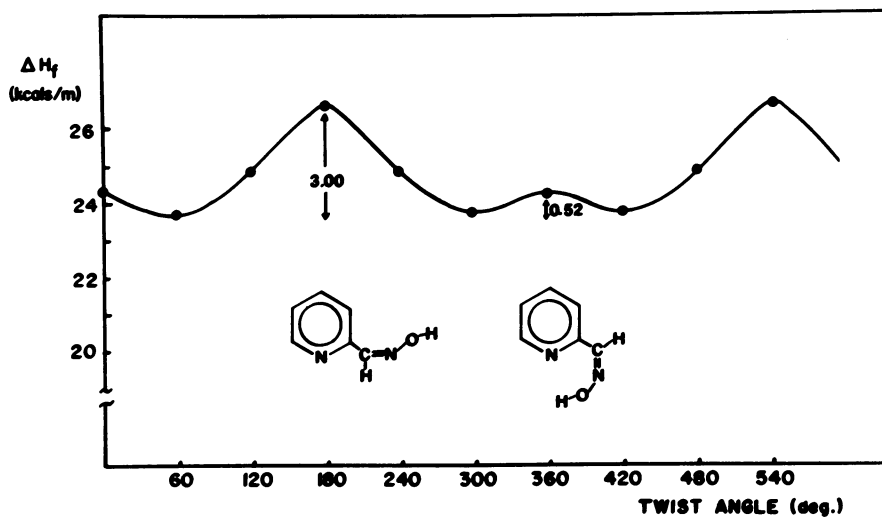


Figure 2. Calculated energies in Kcal/mol vs. degree of rotation of the oxime function about the pyridine ring in anti-2-pyridinecarbaldoxime (1)

the original of 117° in the neutral form. Contrary to the undissociated forms, the anti anion (III) is slightly more stable than the syn (IV) isomer.

TABLE I

Calculated Heats of Formation (ΔH_f) and Ionization Potentials of Some Tertiary and Quaternary Pyridinium Aldoximes

Nr.	Compound Name	Isomer	ΔH_f kcal/mol (25°)	Vertical Ionization Potentials eV
I	2-pyridine carbaldoxime	anti	23.7	8.19 (σ) 8.72 (π) 9.07
II	2-pyridine carbaldoxime	syn	22.4	8.55 (σ) 8.77 (π) 9.14
III	2-pyridine carbaldoxime anion	anti	17.3	----
IV	2-pyridine carbaldoxime anion	syn	17.8	----
V	1-methyl-2-pyridinium carbaldoxime cation	anti	172.9	----
VI	1-methyl-2-pyridinium carbaldoxime cation	syn	177.5	----

Next, the syn and anti forms of 2-PAM, the drug of the main interest, were analyzed. The calculated heats of formation are given in Table I while the calculated geometries can be seen in Figure 3 for the anti form (V) and Figure 4 for the syn isomer (VI), respectively. It is clear that contrary to the tertiary oximes (I) and (II), in the N-methylated case the anti configuration is about 4.5 kcal/mol more stable than the syn isomer. This is in agreement with most of the experimental evidence and would suggest that the anti form indeed is the active species of this drug. Since, however, both the syn and the anti forms are apparently available, and one of them converts to the more stable conformer relatively easily (6), the energetics and structural

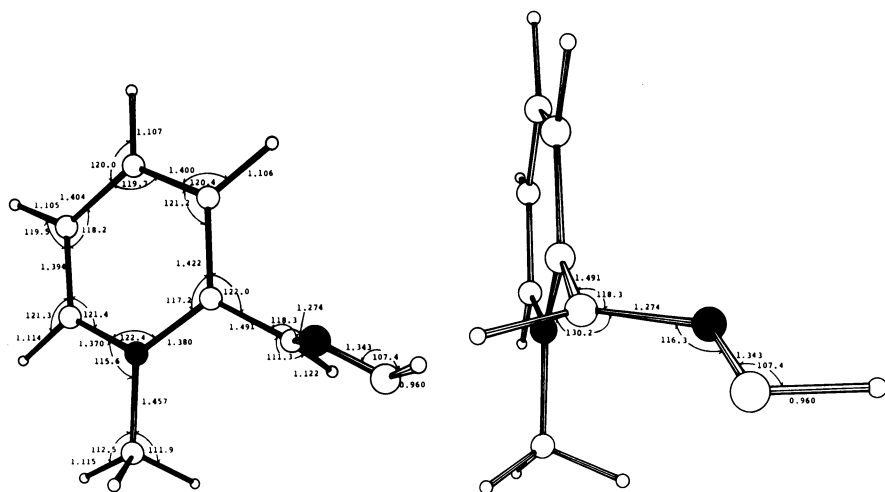


Figure 4. Two structural representations of the calculated geometry of *syn*-1-methyl-2-pyridiniumcarbaldoxime ion (VI) in its most stable form. The second structure is a perpendicular view of the first one. The oxime group is 86° out of the plane to the ring. The oxime hydrogen is *trans* and essentially planar with respect to the C—N—O plane. The methyl group hydrogens are rotated 10° , 130° , and 250° respectively from the ring, thus occupying a position of least steric hindrance to the oxime.

features of syn-anti interconversion was also studied. As a simple model case, first we studied acetaldoxime.

The two possible routes for the syn-anti interconversion of oximes are shown in Figure 5. In the first case, a rotation around the C-N double bond takes place through a 90° rotated transition state, from which the other configuration can be obtained. The intermediate is either a biradical or a structure in which the carbon carries the positive while the nitrogen, the negative charge. The other possible route is the so-called inversion in which case the oxygen will migrate in the same plane going through a C-N-O linear structure. The calculated energies and geometries for the various structures are given in Table 2. As in the case of the pyridine-2-aldoximes, the syn acetaldoxime (VIII) is slightly more stable than the corresponding anti isomer (VII). (The calculated geometries are in excellent agreement with the experimental values (19) obtained for the oxime of acetone: CC = 1.49 Å; C-N = 1.29 Å; N-O = 1.36 Å; CNO = 131°). As observed before in the corresponding anions, the stability order is reversed as indicated by the values for structures (IX) and (X). The two routes for interconversion of syn and anti forms have then been investigated. The two intermediates, the linear anion (XI) and the 90° rotated (XII), were calculated. As expected, the linear route corresponding to the inversion is significantly more favorable than the double bond breaking rotation route. The low inversion energy obtained of about 8-9 kcal/mol is in agreement with the facile interconversion, as well as with the values of 10 to 11 kcal/mol obtained for the inversion in guanidines (20). The rotation requires a much higher energy of about 37 kcal/mol, which is again in good agreement with the energies of 22 to 26 kcal/mol for the rotation in guanidinium salts (20).

Similarly, the interconversion of the syn-anti forms of pyridine-2-aldoximes in the anion forms III and IV, respectively, through the linear intermediate (XIII) requires only about 11 kcal/mol. It is expected that if there is an additional positive charge on the pyridine ring, it would further facilitate the formation of the linear intermediate by abstracting electrons from the oxime portion. (Attempts to calculate the zwitterions corresponding to the oximes V and VI and the corresponding intermediates have failed; no convergence in the SCF calculations were obtained.)

It can be concluded, thus, that the syn-anti conversion in the case of 2-PAM is indeed a facile process and the more stable form, and supposedly the active form of the drug, is the anti isomer.

The process of reduction of 2-PAM was then studied. The preferred (3) 1,6- and 1,4-dihydro derivatives were calculated both in the case of the anti and syn forms. The results are given in Table III.

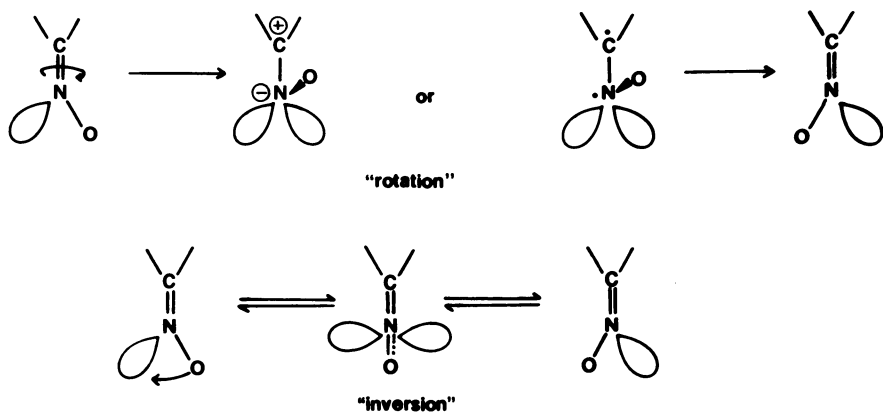


Figure 5. Possible pathways for the syn-anti interconversion in oximes

TABLE II

Calculated Heats of Formation (ΔH_f) and Geometries of Some Oximes, Their Anions and Transition States

Number	Compound	Isomer	ΔH_f kcal/mol	Calcd Bond Lengths (ab, Å) Bond Angles (abc) and Dihedral Angles (abcd)
VII	$\overset{1}{\text{H}}_3\overset{2}{\text{C}}\text{CHNOH}$	anti	-17.8	HC_1 , 1.1; CC , 1.47; HC_2 , 1.12; CN , 1.27; NO , 1.32; OH , 0.96; HC_1C_2 , 112.3°; $\text{C}_1\text{C}_2\text{H}$, 114.1°; HC_2N , 107.1°; CNO , 130.0°; NOH , 119.6°; $\text{HC}_2\text{C}_1\text{H}$, 41.6°; ONCH , 180.0°; HONC , 2.2°.
VIII	$\overset{1}{\text{H}}_3\overset{2}{\text{C}}\text{CHNOH}$	syn	-18.4	HC_1 , 1.11; CC , 1.47; HC_2 , 1.13; CN , 1.27; NO , 1.32; OH , 0.96; HC_1C_2 , 112.2°; $\text{C}_1\text{C}_2\text{H}$, 111.7°; HC_2N , 126.8°; CNO , 127.6°; NOH , 119.6°; $\text{HC}_2\text{C}_1\text{H}$, 55.8°; ONCH , 0.0°; HONC , 1.7°.
IX	$\overset{1}{\text{H}}_3\overset{2}{\text{C}}\text{CHNO}^\ominus$	anti	- 9.3	HC_1 , 1.13; CC , 1.48; HC_2 , 1.13; CN , 1.29; NO , 1.26; HC_1C_2 , 115.0°; $\text{C}_1\text{C}_2\text{N}$, 108.2°; HC_2N , 112.1°; CNO , 132.0°; $\text{HC}_2\text{C}_1\text{H}$, 51.6°; ONCH , 180.0°.
X	$\overset{1}{\text{H}}_3\overset{2}{\text{C}}\text{CHNO}^\ominus$	syn	-10.6	HC_1 , 1.13; CC , 1.46; HC_2 , 1.15; CN , 1.28; NO , 1.26; HC_1C_2 , 115.3°; $\text{C}_1\text{C}_2\text{H}$, 107.1°; HC_2N , 131.1°; CNO , 132.7°; $\text{HC}_2\text{C}_1\text{H}$, 59.16°; ONCH , 0.0°.

TABLE II (Continued)

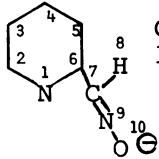
Number	Compound	Isomer	H _f kcal/mol	Calcd Bond Lengths (ab, Å) Bond Angles (abc) and Dihedral Angels (abcd)
XI	H_3CCHNO^-	C-N-O linear	- 1.6	HC ₁ , 1.13; CC, 1.48; HC ₂ , 1.17; CN, 1.22; NO, 1.24; HC ₁ C ₂ , 114.8°; C ₁ C ₂ H, 105.5°; HC ₂ N, 123.6°; CNO, 180.0°; HC ₂ C ₁ H, 55.2°.
XII	H_3CCHNO^-	Rotated Oxime 90°	27.1	HC ₁ , 1.14; CC, 1.42; HC ₂ , 1.11; CN, 1.38; NO, 1.23; HC ₁ C ₂ , 115.1°; C ₁ C ₂ H, 117.4°; HC ₂ N, 111.7°; CNO, 127.0°; HC ₂ C ₁ H, 55.3°; ONC ₁ H, 90.0°.
XIII		C-N-O linear	28.6	CH _(ring) , 1.1; C-C _(ring) , 1.4; C ₆ C ₇ , 1.47; C ₇ H, 1.16. CN, 1.23; NO, 1.23; C ₇ C ₆ C ₅ , 126.0°; HC ₇ C ₆ , 107.7°; NC ₇ C ₆ , 129.4°; ONC, 180.0°; H ₈ C ₇ C ₆ C ₅ , 0.41°; ONCH, -0.1°; i.e., oxime in plane of ring.

TABLE III

Calculated Heats of Formation (ΔH_f) of Various Dihydro Forms of 1-Methyl-2-pyridinium Carbaldoxime Isomers

Number	Compound Name	Isomer	ΔH_f kcal/mol
XIV	1-methyl-1,4-dihydropyridine-2-carbaldoxime	anti	32.87
XV	1-methyl-1,4-dihydropyridine-2-carbaldoxime	syn	32.35
XVI	1-methyl-1,6-dihydropyridine-2-carbaldoxime	anti	38.48
XVII	1-methyl-1,6-dihydropyridine-2-carbaldoxime	syn	38.30
XVIII	C-protonated 1-methyl-3,6-dihydropyridinium-2-carbaldoxime	anti	170.23
XIX	C-protonated 1-methyl-5,6-dihydropyridinium-2-carbaldoxime	anti	167.09
XX	N-protonated 1-methyl-1,4-dihydropyridinium-2-carbaldoxime	anti	184.76

It is interesting to see that the approximate 4.5 kcal/mol difference between the anti and syn forms disappears as either dihydro derivative is formed. Thus, the 1,4-dihydro forms XV and XVI have essentially the same heat of formation and so do the corresponding 1,6-dihydro derivatives XVI and XVII.

Although both in the syn and anti cases the 1,4-dihydro derivatives are about 6.0 kcal/mol more stable than the corresponding 1,6-dihydro derivatives, it was found that the 1,6-dihydro derivative is the structure corresponding to the synthesized (3) Pro-2-PAM. The explanation for this comes from analysis of the mechanism of the reduction. As shown in the case of the reduction of pyridine, the attack of the hydride ion occurs on the more positive carbon which in the case of 2-PAM in the anion (zwitterionic) form should be position 6 (carbon 2 is highly hindered), based on the analysis of the charge distribution in the anion III as indicated in Figure 6. Under the conditions of the synthesis,

the dihydro derivative obtained undergoes cyanide addition on carbon atom 2, while the elimination of HCN results again in the 1,6-dihydro derivative such as structure XVI. The corresponding structures for the two dihydro derivatives of the anti oxime are given in Figures 7 and 8.

The corresponding protonated forms which can be obtained by C-protonation of dihydro derivative XVI are shown in Figures 9 and 10. The N-protonated form (Figure 11) which can be obtained from the 1,4-dihydro derivative is, as in the case of pyridines (1), much less stable than the C-protonated forms. As in previous cases, the protonated 3,6-dihydro derivative which forms experimentally is somewhat less stable than the corresponding protonated 5,6-dihydro derivative (XIX). However, it is easy to see, based on the relative charge distribution (Figure 6), that the protonation of (XVI) will occur on carbon 3 rather than carbon 5, resulting in the protonated dihydro derivative XVIII instead of XIX.

In conclusion, a MINDO/3 study on the pyridine-2-aldoximes and related derivatives revealed that the syn-2-pyridinealdoxime is more stable than the corresponding anti isomer. The difference in the isomers is reverted when the corresponding anions are formed. In agreement with the experimental observations, in the case of N-methylated forms, the anti isomer is more stable than the syn form and based on this, it was concluded that the anti-2-PAM is the active form of this well-known drug. It was found that in all of the above oximes, the oxime portion is near perpendicular to the ring of the pyridine except in the anions of the pyridine-2-aldoximes in which a coplanar structure is the most stable.

It was found that the mechanism of interconversion of these oximes and probably any aldoximes (as also indicated by the case of acetaldoxime), involves "inversion" type mechanism in which the key intermediate is the one in which the carbon, nitrogen and oxygen atoms of the oxime are colinear. The relatively low inversion energy is in good agreement with the observation of the facile interconversion of the less stable form of 2-PAM to the more stable anti isomer.

Calculations on the dihydro forms corresponding to 2-PAM indicate that the structural assignment (3) is supported by the calculations according to which the 1,6-dihydro form is the base form of the 1-methyl-1,6-dihydropyridine-2-carbaldoxime (Pro-2-PAM) which corresponds to the protonated 3,6-dihydro structure XVIII. It was also found that the initial syn-anti energy difference in the case of 2-PAM disappeared when the corresponding dihydro derivatives are formed.

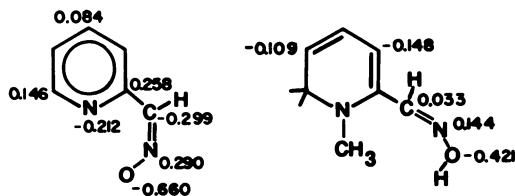


Figure 6. Calculated charge distribution in the anti-2-pyridinecarbaldoxime anion (III) and in the anti-1-methyl-1,6-dihydropyridinecarbaldoxime (XVI)

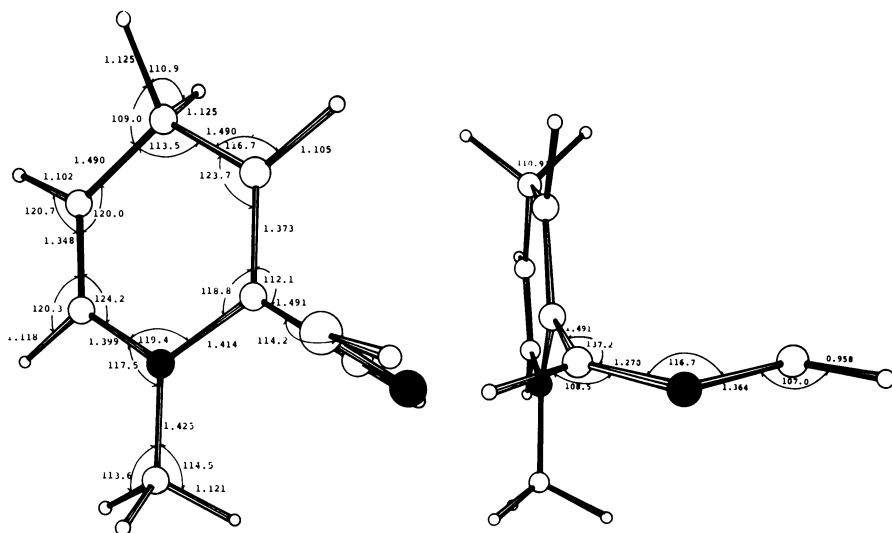


Figure 7. Two structural representations of the calculated geometry of anti-1-methyl-1,4-dihydropyridine-2-carbaldoxime (XIV) in its most stable form. The second structure is a perpendicular view to the first one. The oxime group is rotated 85° from the plane of the ring and the oxime hydrogen is trans and virtually planar with the oxime function. The hydrogens of the methyl group are rotated 16, 126, 256° with respect to the ring. The ring is essentially planar.

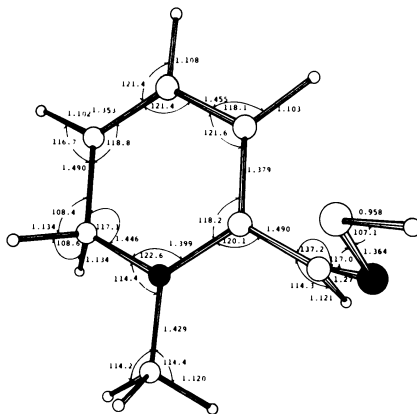


Figure 8. Calculated geometry of the most stable form of anti-1-methyl-1,6-dihydropyridine-2-carbaldoxime (XVI). The oxime group is rotated 86° from the plane of the ring and the oxime hydrogen is trans and planar with the oxime function. The methyl hydrogens are rotated 25 , 145 , and 265° with respect to the ring. The ring is essentially planar.

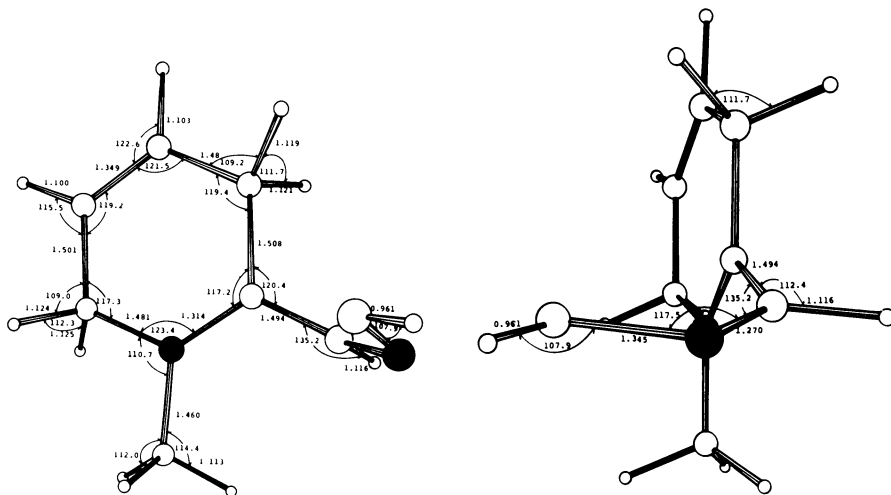


Figure 9. Two structural representations of the calculated geometry of the most stable form of C-protonated anti-1-methyl-3,6-dihydropyridine-2-carbaldoxime ion (XVIII). The second structure is a perpendicular view. The oxime group is rotated 90° from the plane of the ring with the oxime hydrogen trans and in the plane of the oxime group. The methyl hydrogens are rotated 33 , 153 , 273° with respect to the ring. The ring is essentially planar.

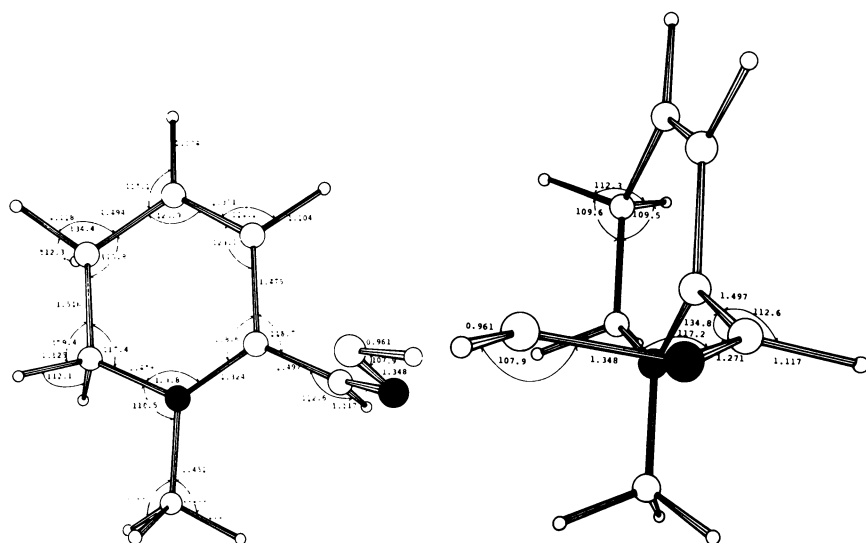


Figure 10. Two structural representations of the most stable form of C-protonated anti-1-methyl-5,6-dihydropyridine-2-carbaldoxime ion (XIX). The oxime function is rotated by 90° with respect to the ring and the oxime hydrogen is trans and in the plane of the oxime. The methyl group hydrogens are rotated 33, 153, 273° with respect to the virtually planar ring.

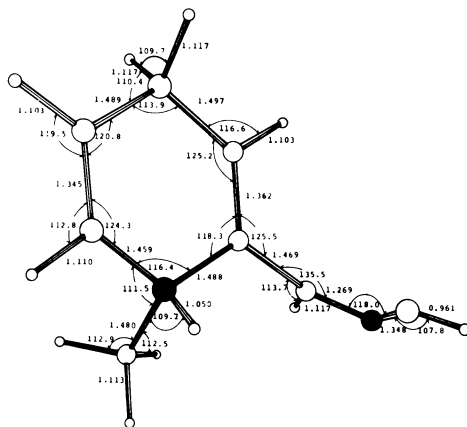


Figure 11. Calculated geometry of the most stable form of N-protonated anti-1-methyl-1,4-dihydropyridine-2-carbaldoxime ion (XX). The ring is essentially planar except the methylene carbon which is out of plane by 6° . The oxime function is rotated 93° with respect to the ring. The oxime hydrogen is trans and planar with respect to the oxime group. The methyl hydrogens are rotated 16, 136, 256° with respect to the plane of the ring.

LITERATURE CITED

1. Bodor, N. and Pearlman, R., J. Am. Chem. Soc. (1978), 100, 4946-4953.
2. Bodor, N., Shek, E., and Higuchi, T., Science (1975), 190, 155-156.
3. Bodor, N., Shek, E., and Higuchi, T., J. Med. Chem. (1976), 19, 102-107.
4. Shek, E., Higuchi, T., and Bodor, N., ibid. (1976), 19, 108-112.
5. Shek, E., Higuchi, T., and Bodor, N., ibid. (1976), 19, 113-117.
6. Ginsburg, S., and Wilson, I. B., J. Am. Chem. Soc. (1957), 79, 481-485.
7. Wilson, I. B., Federation Proc. (1959), 18, 752-758.
8. Poziomek, E. J., Kramer, D. N., Mosher, W. A., and Michel, H. O., J. Am. Chem. Soc. (1961), 83, 3916-3917.
9. Kitz, R. J., Ginsburg, S., and Wilson, I. B., Biochem. Pharmacol. (1965), 14, 1471-1477.
10. Barfknecht, C. F., Long, J. P., and Benz, F. W., J. Pharm. Sci. (1971), 60, 138-139.
11. Carlstrom, D., Acta Chem. Scand. (1966), 20, 1240-1246.
12. Bingham, R. C., Dewar, M. J. S., and Lo, D. H., J. Am. Chem. Soc. (1975), 97, 1285-1293.
13. Bingham, R. C., Dewar, M. J. S., and Lo, D. H., ibid. (1975), 97, 1302-1306.
14. Baird, N. C., and Dewar, M. J. S., J. Am. Chem. Soc. (1969), 91, 352-357.
15. Dewar, M. J. S., and Haselbach, E., ibid. (1970), 92, 590-598.
16. Bodor, N., Dewar, M. J. S., Harget, A., and Haselbach, E., ibid. (1970), 92, 3854-3859.
17. Bodor, N., Dewar, M. J. S., and Lo, D. H., ibid. (1972), 94, 5304-5310.
18. Giordano, W., Hamann, J. R., Harkins, J. J., and Kaufman, J. J., Mol. Pharmacol. (1967), 3, 307-317.
19. Bierlein, T. K., and Lingafeller, E. C., Acta Cryst. (1951), 4, 450-453.
20. Kessler, H., Angew. Chem. Internat. Edit. (1970), 9, 219-235, and ref. cited.

RECEIVED June 8, 1979.

The Analysis of Electronic Factors in Quantitative Structure-Activity Relationships Using Distribution Coefficients

ROBERT A. SCHERRER and SUSAN M. HOWARD

Riker Laboratories, St. Paul, MN 55101

In an earlier paper (1) we illustrated some of the advantages of using distribution coefficients instead of partition coefficients for the QSAR analysis of ionizable compounds. We wish to expand on one particular aspect: assessing the role electronic effects play in the manifestation of the biological activity of some compounds.

Before what I say on this subject will be meaningful to you, you have to be convinced of one basic concept. That is that regression analyses in $\log D$ [and $(\log D)^2$ if required] relate to the amount of compound at a site, or in a membrane, etc.

By distribution coefficient we mean the ratio of lipid to aqueous concentration without regard to species form in the aqueous phase. The lipid phase is assumed to contain the undissociated compound. You are already familiar with $\log D$ by another name, "apparent $\log P$," the measured $\log P$ at the pH of interest. You know that many times one gets good correlations with "apparent $\log P$ " or " $\log P$ " instead of $\log P$. (2,3,4) The problem is that "apparent $\log P$ " is too often viewed as a property someone measures with little thinking about the implications behind it.

From the relation, eq 1, one can see that the distribution coefficient, D , is equal to the partition coefficient P , times the fraction unionized. So if an acid or base is 3/4th unionized, $\log D$ will be 3/4 of $\log P$. There will be 3/4th as much compound in the octanol phase than if it were undissociated.

0-8412-0521-3/79/47-112-507\$05.00/0

© 1979 American Chemical Society

$$(1) \quad D = f_u \cdot P \quad \text{where } f_u \text{ is the fraction unionized.}$$

From eq 1, it follows (eq 2) that

$$(2) \quad \log D = \log P + \log f_u \\ = \log P + C_D.$$

The complex term f_u is equal to $\frac{1}{1 + 10^{\text{pH} - \text{p}K_a}}$ for acids and $\frac{1}{1 + 10^{\text{p}K_a - \text{pH}}}$ for bases. To simplify the use of $\log D$ and to emphasize that C_D is characteristic of a compound under a given set of experimental conditions, we have tabulated some C_D values (Table I). C_D is $\log f_u$; it is either zero or negative. When a $\text{p}K_a$ is more than about 1 unit away from the pH of a system on the side of ionization, C_D is approximated by the difference, $\text{p}K_a - \text{pH}$ (acids), $\text{pH} - \text{p}K_a$ (bases). One can write eq 3 for highly ionized compounds.

$$(3) \quad \log D = \log P + \text{p}K_a - \text{pH} \text{ (acids)}$$

$$(3) \quad \log D = \log P - \text{p}K_a + \text{pH} \text{ (bases)}$$

Absorption Studies. No Electronic Component of Mechanism.

Some of the best illustrations of regression analyses requiring $\log D$ terms alone are those related to absorption. This is a simple process, generally, with no "mechanism of action" or receptor site interactions to explain. One is just looking at how much compound passes through a tissue by passive transfer.

Colonic Absorption. In our earlier paper (1) we analyzed the absorption of acids and bases from the rat colon and compared these results with alternative analyses. We reported eq 4 for the bases (Table II), which compares favorably with eq 5 (5). (The absolute t-value is shown below each parameter.)

$$(4) \quad \log \% \text{ Abs.} = 0.362 \log D + 1.83 \\ (7.2)$$

$$(n = 10, r = 0.930, s = 0.143, F = 51)$$

$$(5) \quad \log \% \text{ Abs.} = -0.330 (\log P)^2 + 0.869 \log P \\ + 0.059 (6.8 - \text{p}K_a) + 0.817$$

$$(n = 10, r = 0.910, s = 0.187, F = 9.6)$$

Table I
Correction-for-Dissociation Values

pK _a - pH (acids)		pK _a - pH (acids)	
pH - pK _a (bases)	C _D	pH - pK _a (bases)	C _D
2.0	<u>ca.</u> 0	-0.05	-0.327
		-0.10	-0.354
1.50	-0.01	-0.15	-0.382
1.40	-0.017	-0.20	-0.412
1.30	-0.021	-0.25	-0.444
1.20	-0.027	-0.30	-0.476
1.10	-0.033	-0.35	-0.510
		-0.40	-0.546
1.00	-0.041	-0.45	-0.582
0.95	-0.046	-0.50	-0.619
0.90	-0.051	-0.55	-0.658
0.85	-0.057	-0.60	-0.697
0.80	-0.064	-0.65	-0.738
0.75	-0.071	-0.70	-0.779
0.70	-0.079	-0.75	-0.821
0.65	-0.088	-0.80	-0.864
0.60	-0.097	-0.85	-0.907
0.55	-0.108	-0.90	-0.951
0.50	-0.119	-0.95	-0.996
0.45	-0.132	-1.00	-1.041
0.40	-0.146		
0.35	-0.160	-1.10	-1.133
0.30	-0.176	-1.20	-1.227
0.25	-0.194	-1.30	-1.321
0.20	-0.212	-1.40	-1.417
0.15	-0.232	-1.50	-1.514
0.10	-0.254		
0.05	-0.277	-2.00	<u>ca.</u> -2.00
0.00	-0.301		

Table II
Colonic Absorption of Acids and Bases and Physicochemical Constants Used^a

Acids	Log P	pK _a -pH	C _D 6.8	Log D	Log % Abs. Obsd. / Calcd. b
5-Nitro salicylic acid	1.98	-4.5	-4.5	-2.52	0.30 / 0.57
m-Nitrobenzoic acid	1.83	-3.4	-3.4	-1.57	1.00 / 0.82
Salicylic acid	2.26	-3.8	-3.8	-1.54	1.08 / 0.83
Benzoic acid	1.85	-2.6	-2.6	-0.75	1.28 / 1.04
Phenylbutazone	3.22	-2.4	-2.4	0.82	1.58 / 1.45
o-Nitrophenol	1.79	0.2	-0.21	1.58	1.74 / 1.65
Thiopental	2.50	0.8	-0.06	2.44	1.70 / 1.88
p-Hydroxypropiofenone	1.85	1.0	-0.04	1.81	1.66 / 1.71
m-Nitrophenol	2.00	1.4	0	2.00	1.64 / 1.76
Phenol	1.46	3.1	0	1.46	1.55 / 1.62
		pH-pK _a			
Acetanilide	1.16	6.5	0	1.16	1.56 / 1.54
p-Nitroaniline	1.39	5.8	0	1.39	1.70 / 1.60
Antipyrine	0.23	5.4	0	0.23	1.30 / 1.30
m-Nitroaniline	1.37	4.3	0	1.37	1.68 / 1.60
Aniline	0.90	2.2	0	0.90	1.64 / 1.48
Aminopyrine	0.76	1.8	0	0.76	1.32 / 1.44
p-Toluidine	1.39	1.5	0	1.39	1.71 / 1.60
Quinine	1.83	-1.6	-1.6	0.23	1.30 / 1.30
Ephedrine	1.56	-2.8	-2.8	-1.24	0.95 / 0.91
Tolazoline	2.65	-3.5	-3.5	-0.85	0.60 / 1.01
Levorphan	3.02	-3.0	-3.0	0.02	1.11 / 1.24

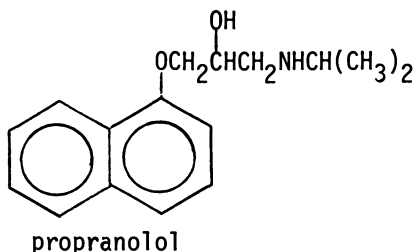
a) The groups have been analyzed separately (1). b) Equation 6.

Table III
Buccal Absorption of Organic Acids and Physicochemical Constants Used^a

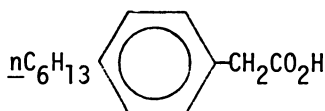
	Log P	C_D ($pK_a - 6.0$)	Log D	Log % Abs.	
				Obs.	Calcd. ^b
Phenylacetic	1.30	-1.69	-0.39	0.00	0.29
<i>p</i> -Nitro	1.54	-2.15	-0.61	0.00	0.10
<i>p</i> -Fluoro	1.44	-1.75	-0.31	0.18	0.36
<i>p</i> -Methoxy	1.31	-1.64	-0.33	0.48	0.34
<i>p</i> -Methyl	1.75	-1.63	0.12	0.85	0.69
<i>p</i> -Chloro	2.00	-1.81	0.19	0.85	0.74
<i>p</i> -Bromo	2.20	-1.81	0.39	0.90	0.88
<i>p</i> - <i>n</i> -Propyloxy	2.31	-1.64	0.67	1.00	1.05
<i>p</i> -Ethyl	2.25	-1.63	0.62	1.00	1.02
<i>p</i> -Iodo	2.53	-1.82	0.71	1.00	1.08
<i>p</i> - <i>n</i> -Propyl	2.75	-1.64	1.11	1.40	1.30
<i>p</i> - <i>t</i> -Butyl	2.98	-1.64	1.34	1.40	1.40
<i>p</i> - <i>n</i> -Butyl	3.25	-1.64	1.61	1.53	1.52
<i>p</i> - <i>t</i> -Pentyl	3.48	-1.64	1.84	1.48	1.60
<i>p</i> -Cyclopentyl	3.44	-1.64	1.80	1.48	1.59
<i>p</i> - <i>n</i> -Pentyl	3.75	-1.64	2.11	1.69	1.68
<i>p</i> -Cyclohexyl	3.81	-1.64	2.17	1.64	1.70
<i>p</i> - <i>n</i> -Hexyl	4.25	-1.64	2.61	1.79	1.79
<i>n</i> -Valeric	1.42	-1.19	0.23	0.83	0.77
<i>n</i> -Hexanoic	1.92	-1.15	0.77	1.15	1.11
<i>n</i> -Heptanoic	2.42	-1.11	1.31	1.46	1.39
<i>n</i> -Octanoic	2.92	-1.15	1.77	1.60	1.58
<i>n</i> -Nonanoic	3.42	-1.15	2.27	1.72	1.72
<i>n</i> -Decanoic	3.92	-1.15	2.77	1.76	1.81
<i>n</i> -Undecanoic	4.42	-1.15	3.27	1.83	1.84
<i>n</i> -Dodecanoic	4.92	-1.15	3.77	1.90	1.81
Benzoic	1.81	-1.79	0.02	0.90	0.62
<i>p</i> -Methyl	2.27	-1.67	0.60	1.22	1.02
<i>m</i> -Methyl	2.37	-1.76	0.61	1.08	1.02
<i>o</i> -Methyl	2.81	-2.08	0.73	0.82	1.09
2,4-Dimethyl	3.01	-1.72	1.29	1.30	1.38
2,4,6-(CH ₃) ₃	3.69	-2.44	1.25	0.60	-
2,3,5,6-(CH ₃) ₄	4.19	-2.51	1.68	0.12	-

^aSee ref. 7. ^bCalculated from eq 7.

Table IV
 Buccal Absorption of an Acid and a Base
 at Six pH Values and Physicochemical
 Constants Used for Eq 9, 10



Log P	pK _a	pH	C _D	Log D	Obsd. ^a	Log K _{app.}	
						Calcd. ^b	Calcd. ^c
3.33	9.45	5.08	-4.37	-1.04	-2.19	-2.01	-2.22
		6.02	-3.43	-0.10	-1.71	-1.69	-1.70
		7.00	-2.45	0.88	-1.22	-1.34	-1.25
		7.93	-1.52	1.81	-0.79	-1.02	-0.90
		8.94	-0.63	2.70	-0.53	-0.71	-0.65
		9.93	-0.12	3.21	-0.35	-0.54	-0.54



4.25	4.36	4.0	-0.16	4.20	-0.46	-0.19	-0.40
		5.0	-0.73	3.63	-0.54	-0.39	-0.47
		6.0	-1.64	2.72	-0.72	-0.70	-0.64
		7.0	-2.64	1.72	-1.07	-1.05	-0.93
		8.0	-3.64	0.72	-1.44	-1.40	-1.31
		9.0	-4.64	-0.28	-1.78	-1.75	-1.79

^aReference 8. ^bCalculated from eq 9. ^cCalculated from eq 10.

The remaining portion of this paper illustrates a useful approach to this kind of separation. Four activities of phenols will be examined, then uncoupling activity by a variety of agents and the bacteriostatic activity of carboxylic acids will be analyzed.

Activities of Phenols.

Uncoupling Oxidative Phosphorylation. The importance of an acidic proton for activity has long been recognized. A mechanism proposed for classical uncouplers (11) is that such agents act as ionophores (possibly synergistically with natural ionophores) transporting protons into mitochondria and cations out. In such a case it is reasonable that ionic character is important. The quantitation of the contribution of pK_a to activity vis-a-vis partition coefficients has been subjective and conflicting in many cases (1, 12).

The 23 phenols in Table V (13) cover a broad range of pK_a 's. Their uncoupling activity is described by eq 12. The presence of the pK_a term is highly significant and required for a good correlation. Some factor related to acidity is important for activity. From eq 12 one can also say that for these phenols, the more acidic the phenol (keeping log P constant) the more active it will be as an uncoupler (even though the more ionized phenol will be present in lower concentration in lipid phases). One can say this because the coefficient of the pK_a term is greater than that of the log D term. For highly dissociated phenols ($pH-pK_a > 1$) eq 12 can be rewritten as eq 12'.

$$(12) \quad \log 1/C = 0.471 \log D - 0.618 pK_a + 7.584$$

(6.88) (12.66)

$$(n = 23, r = 0.946, s = 0.351, F = 85)$$

$$(12') \quad \log 1/C = 0.471 (\log P + pK_a - pH) - 0.618 pK_a + 7.584$$

At a given pH there is a net negative pK_a term. The more acidic a phenol, the smaller this term and the greater the calculated activity. For the undissociated phenols, $\log D = \log P$, and more acidic phenols are still calculated to be more active.

It should be noted that one is not restricted to log D and pK_a combinations. By using log D and σ parameters for example, the σ -term will represent only the electronic component related to mechanism. On the subject of sigma values, care should be taken to use ortho values for o-substituents [see Norrington et al (14)] when acidity is a property of interest. Both Stockdale and Selwyn (13) and Motais et al (15) for the phenols of Table V used, for example, 0.23 instead of 0.68 for o-chloro; 0.80 instead of 1.24 for o-nitro.

Table V
 Uncoupling Activity, Inhibition of Chloride Ion Transport and Physicochemical
 Constants Used for Eq 12, 20.

Phenol Substitution	Log P	pKa	C _D 7.5a		Log D 7.5a		Uncoupling, log 1/C		Inhibition of Cl ⁻	
			C _D 7.5a	Log D 7.5a	Obsd. ^b	Calcd. ^c	Obsd. ^d	Calcd. ^e		
H	1.48	9.99	0	1.48	2.10	2.11	3.00	3.00		
2-C1	2.17	8.56	0	2.17	2.79	3.32	3.92	3.63		
4-C1	2.41	9.43	0	2.41	3.05	2.89	4.05	3.85		
3-C1	2.52	9.12	0	2.52	3.38	3.13	4.28	3.95		
2,4-C12	3.08	7.98	-0.12	2.96	4.17	4.05	4.58	4.48		
2,5-C12	3.22	7.51	-0.30	-0.30	4.17	4.32	4.62	4.63		
2,4-Br2	3.24	7.94	-0.13	3.11	4.17	4.14				
2,4,6-C13	3.64	6.22	-1.28	2.36	4.34	4.85	4.88	5.15		
2,4,6-Br3	3.98	6.44	-1.09	2.89	5.14	4.97				
2,3,4,6-C14	4.12	5.74	-1.76	2.36	5.48	5.15				
Cl5	4.12	4.95	-2.55	1.57	5.70	5.26	5.88	5.77		
2-NO2	1.81	7.23	-0.46	1.35	2.84	3.75	3.27	3.36		
4-NO2	1.98	7.16	-0.51	1.47	3.96	3.85	3.50	3.52		
3-NO2	2.02	8.36	0	2.02	3.49	3.37				
2,4-(NO2)2	1.52	4.10	-3.40	-1.88	4.72	4.16	4.07	3.51		
2,5-(NO2)2	2.02	5.22	-2.28	-0.26	4.32	4.23	3.44	3.81		
2,6-(NO2)2	1.57	3.71	-3.79	-2.22	3.94	4.24	3.40	3.61		
3-CH3	1.98	10.10	0	1.98	2.30	2.27	3.22	3.46		
4-CH3	1.97	10.28	0	1.97	2.28	2.16	3.10	3.45		
β-naphthol	2.72	9.59	0	2.72	3.17	2.94				
F5	3.25	5.33	-2.17	1.08	4.47	4.80	4.85	4.91		
2,6-Cl24-NO2	2.96	3.54	-3.96	-1.00	4.80	4.92				
2,6-Br2,4-NO2	3.07	3.39	-4.11	-1.04	5.11	5.00				

^aValues for pH 7.4 for eq 20 not shown. ^bReference 13. ^cEquation 12. ^dReference 15. ^eEquation 20.

The same acidity : activity relationship holds for uncoupling by six nitrophenols at four pH's, eq 13 (1). The more acidic the phenol of a given log P, the more active it will be as an uncoupler.

$$(13) \quad \log 1/C = 0.409 \log D - 0.604 pK_a + 6.970$$

(14.25) (13.66)

$$(n = 24, r = 0.968, s = 0.231, F = 154)$$

Bacteriostatic Activity. The bacteriostatic activity (vs. *E. coli*, 16) of some of the same phenols as in eq 12 and 13 is described by eq 14 and 15. Again, a pK_a term is present and required for high correlations. But note that for a given log P, the more acidic the phenol, the less active an antibacter-ial it will be. While some factor related to acidity, such as stability of an anion, electron density in the ring or on the oxygen, hydrogen bonding, etc. must be important for activity (or the term wouldn't be there), the contribution of this property is not great enough to overcome the negative influence of dissociation on the concentration at the site of action. This is consistent with the general observation that many commercial bacteriostatic phenols are not highly acidic. Equation 15 combines bacteriostatic activity measurements determined at four pH values. The value for trinitrophenol at pH 8.5 was omitted because it was so far out of line from the rest (F for the equation increased from 104 to 157).

pH 6.5

$$(14) \quad \log 1/C = 0.751 \log D - 0.582 pK_a + 5.850$$

(6.71) (5.90)

$$(n = 8, r = 0.956, s = 0.147, F = 26)$$

pH 5.5 - 8.5

$$(15) \quad \log 1/C = 0.594 \log D - 0.461 pK_a + 5.231$$

(17.54) (14.50)

$$(n = 31, r = 0.958, s = 0.179, F = 157)$$

It is not possible to use just log P and pK_a to get a good correlation of the data (Table VI) described by eq 14. The result is that neither term is significant at the 95% level (eq 16).

$$(16) \quad \log 1/C = 0.701 \log P + 0.196 pK_a + 1.500$$

(1.57) (0.31)

$$(n = 8, r = 0.650, s = 0.380)$$

Table VI
Bacteriostatic Activity of Nitrophenols and Physicochemical
Constants Used for Eq 14

Phenol Substitution	Log P	pK _a	C _D 6.5	Log D	Log 1/MIC	
					Obsd. ^a	Calcd. ^b
3-NO ₂	2.02	8.36	0	2.02	2.59	2.50
4-NO ₂	1.98	7.16	-0.09	1.89	3.00	3.10
2-NH ₂ ,4-NO ₂	1.07	7.00	-0.12	0.95	2.48	2.49
2,5-(NO ₂) ₂	2.02	5.22	-1.28	0.74	3.46	3.37
2-NH ₂ ,4,6-(NO ₂) ₂	1.40	4.4	-2.1	-0.70	2.52	2.76
2,4-(NO ₂) ₂	1.52	4.10	-2.40	-0.88	2.96	2.80
2,6-(NO ₂) ₂	1.57	3.71	-2.79	-1.22	2.80	2.78
2,4,6-(NO ₂) ₃	1.34	0.38	-6.12	-4.78	2.05	2.04

^aReference 16 ^bEquation 14.

given log P, the more acidic the phenol, the more active it will be. This can be said because the coefficient of the C_D term is greater than the coefficient of log D. It is possible there are two mechanisms for the inhibition of chloride ion transport. By contrast, in no case has the use of a C_D term improved the correlation for uncoupling activity.

$$(19) \quad \log 1/C = 0.547 \log D - 0.430 pK_a + 6.393$$

(4.91) (5.54)

$$(n = 16, r = 0.843, s = 0.465, F = 16)$$

$$(20) \quad \log 1/C = 0.915 \log D - 1.058 C_D + 1.645$$

(10.05) (10.44)

$$(n = 16, r = 0.947, s = 0.278, F = 56)$$

Uncoupling of Oxidative Phosphorylation by a Variety of Chemical Classes.

As an alternative test of distribution coefficients it is informative to look at a single biological activity, uncoupling oxidative phosphorylation, by a variety of classes of agents.

N-Arylanthranilic Acids (2). In equations 21 and 22 it can be seen that the N-arylanthranilic acids are similar to phenols (eq 12, 13) in the effect of pK_a on activity. The more acidic the anthranilic acid, the more potent it will be as an uncoupler for any given log P. In contrast, binding to bovine serum albumin for some of the same compounds is satisfied by an equation in log D alone (eq 23). The pK_a terms in eq 21 and 22 may not be significant enough to make anything of the difference in magnitude of the pK_a coefficients between heart and liver mitochondria, but this is a potential tool for this type of comparison.

(heart)

$$(21) \quad \log 1/C = 0.417 \log D - 0.864 pK_a + 7.462$$

(4.12) (2.24, 94.4%)

$$(n = 11, r = 0.873, s = 0.275, F = 12.8)$$

(liver)

$$(22) \quad \log 1/C = 0.488 \log D - 0.572 pK_a + 5.374$$

(7.64) (2.30, 95.6%)

$$(n = 13, r = 0.932, s = 0.189, F = 33)$$

(binding to BSA)

$$(23) \quad \log K = 0.160 \log D + 5.373$$

(8.12)

$$(n = 8, r = 0.957, s = 0.041, F = 66)$$

$$(25) \quad \log 1/C = 0.508 \log D - 0.308 pK_a + 5.348$$

(15.82) (3.26)

$$(n = 24, r = 0.967, s = 0.165, F = 154)$$

van den Berg et al. report eq 26 for the same 24 derivatives.

$$(26) \quad \log 1/C = 0.504 \pi - 0.347 \sigma + 3.813$$

$$(n = 24, r = 0.968, s = 0.164)$$

The negative sigma coefficient in eq 26 indicates that electron withdrawing groups are unfavorable. But it isn't clear from this equation that this is due to an unfavorable effect on distribution and that the σ component relating to mechanism is positive (i.e. a $+\sigma$ corresponds to a $-pK_a$ term in eq 25). ortho-Derivatives required an E_s^o term. Since ortho substitution also reduces acidity, the true influence of steric factors on uncoupling, separated from influences on acidity-distribution, can only be determined by using E_s^o with log D.

Comparison of eq 27 for the inhibition of prostaglandin synthesis with eq 25 for uncoupling allows one to state that the effect of increasing acidity (lowering pK_a) on the two biological activities is in opposite directions. For each pK unit the acidity of a 2-arylindanedione is increased (with the same log P), the difference between the two activities will increase 15 fold. (Note that it doesn't matter, for this comparison, that the values for the two equations were obtained at different pH's. The equations should hold for any pH not affecting the system itself.)

[Inhibition of prostaglandin synthesis (17) Table VII.]

$$(27) \quad \log 1/IC_{50} = 0.382 \log D - 1.354 pK_a + 9.420$$

(8.58) (10.36)

$$(n = 24, r = 0.920, s = 0.228, F = 58)$$

Bacteriostatic Activity of Carboxylic Acids

These final examples are included to emphasize one particular point: the importance of varying pK_a in a series to be able to detect a possible role for electronic effects. The bacteriostatic activity of α -hydroxy fatty acids at several pH's was reported by Eggerth (23) and analyzed by Hansch and Clayton (24). They reported the following correlations (for B. lepisepcticus):

Table VII
 Uncoupling Activity, Inhibition of Prostaglandin Synthetase
 and Physicochemical Constants for 2-Phenyl-1,3-indanediones (3)

Substitution on 2-phenyl	Log P	pK _a	Log D _{7.5}	Log D _{8.3}	Log 1/C ₅₀ PGS	
					Obsd. a	Calcd. b
H	2.90	3.98	-0.62	-1.42	3.74	3.81
3-CH ₃	3.46	4.09	0.05	-0.75	4.01	4.12
3-C ₂ H ₅	3.92	4.09	0.51	-0.29	4.42	4.35
3-tPr	4.43	4.09	1.02	0.22	4.69	4.61
3-t-Bu	4.88	4.14	1.52	0.72	4.95	4.85
3,5-(CH ₃) ₂	4.02	4.21	0.73	-0.07	4.42	4.42
3,5-(C ₂ H ₅) ₂	4.94	4.21	1.65	0.85	5.06	4.89
3,5-(i-Pr) ₂	5.96	4.21	2.67	1.87	5.41	5.41
3,5-(t-Bu) ₂	6.86	4.31	3.67	2.87	5.61	5.89
3-OCH ₃	2.88	3.84	-0.78	-1.58	3.62	3.77
3,5-(OCH ₃) ₂	2.86	3.70	-0.94	-1.74	3.38	3.73
3-CF ₃	3.78	3.18	-0.54	-1.34	3.99	4.10
3-Cl	3.61	3.35	-0.54	-1.34	4.02	4.04
3,5-Cl ₂	4.32	2.73	-0.45	-1.25	4.31	4.28
4-CH ₃	3.46	4.26	0.22	-0.58	4.10	4.15
4-C ₂ H ₅	3.92	4.23	0.65	-0.15	4.53	4.38
4-1-Pr	4.43	4.23	1.16	0.36	4.87	4.64
4-t-Bu	4.88	4.31	1.69	0.89	5.03	4.88
4-t-C ₈ H ₁₇	6.93	4.25	3.68	2.88	5.66	5.91
4-C ₆ H ₅	4.79	3.96	1.25	0.45	4.98	4.76
4-OCH ₃	2.88	4.43	-0.19	-0.99	3.75	3.89
4-CF ₃	3.78	3.06	-0.66	-1.46	4.11	4.07
4-Cl	3.61	3.59	-0.30	-1.10	4.20	4.09
4-Br	3.76	3.52	-0.22	-1.02	4.31	4.15

^aReference 21 ^bEquation 25 ^cReference 22 ^dEquation 27

$$(28) \quad \frac{\text{pH } 7.5}{\log 1/C} = -0.41 (\log P)^2 + 0.21 \log P + 3.75$$

$$\log P_0 = 0.26$$

$$(n = 7, r = 0.963, s = 0.245)$$

$$(29) \quad \frac{\text{pH } 8.5}{\log 1/C} = -0.16 (\log P)^2 + 0.58 \log P + 3.06$$

$$\log P_0 = 1.85$$

$$(n = 6, r = 0.997, s = 0.108)$$

The log P's in these equations are for the calculated partition coefficients of the sodium salts (where $\log P_{\text{salt}} = \log P - 4.90$). These can also be analyzed in terms of distribution coefficients to obtain eq 30. The log D_0 obtained should apply at any pH in contrast to log P_0 . Because the α -hydroxy acids all have the same pK_a there is no way to assess the effect of acidity (or correlated properties) on bacteriostatic activity. Any electronic factor is buried in the constant.

$$(30) \quad \frac{\text{pH } 7.5, 8.5}{\log 1/C} = -0.102 (\log D)^2 + 0.483 \log D + 2.962$$

$$(6.91) \qquad (9.55)$$

$$\log D_0 = 2.36$$

$$(n = 13, r = 0.950, s = 0.288, F = 46)$$

A few years ago Sheu and associates (25) reported the bacteriostatic activity of a variety of carboxylic acids with pK_a 's covering a range of 2 1/2 units. The monocarboxylic acid activities vs. *B. subtilis*, are listed in Table VIII. Regression analysis shows that acidity is important for activity. Equation 31, with eq 3, predicts that the greater the acidity of an acid of a given log P, the more potent it will be in inhibiting bacterial growth. Even though the more acidic compound will be present in lower concentrations in bacterial lipid biophases, some aspect related to ionization is important enough for activity to more than overcome this disadvantage. Carboxylic acids differ in this respect from phenols, at least nitrophenols (eq 14,15). Two diacids (Table VIII) are more active than predicted by eq 31.

$$(31) \quad \log 1/IC_{50} = 0.479 \log D - 0.713 pK_a + 5.539$$

$$(8.58) \qquad (8.53)$$

$$(n = 15, r = 0.955, s = 0.241, F = 63)$$

Table VIII
Inhibition of *B. subtilis* by Carboxylic Acids and
Physicochemical Constants for Eq 31.

Acid	Log P	pK _a	Log D _{6.5}	Log 1/IC ₅₀	
				Obsd. ^a	Calcd. ^b
CH ₃ CO ₂ H	-0.17	4.76	-1.91	1.19	1.23
C ₂ H ₅ CO ₂ H	0.33	4.87	-1.30	1.40	1.45
C ₃ H ₇ CO ₂ H	0.79	4.87	-0.84	1.52	1.67
C ₅ H ₁₁ CO ₂ H	1.88	4.87	0.25	1.96	2.19
C ₇ H ₁₅ CO ₂ H	2.98	4.88	1.36	2.74	2.71
C ₉ H ₁₉ CO ₂ H	4.09	4.90	2.49	3.60	3.24
CH ₃ CH=CHCO ₂ H	0.72	4.69	-1.09	1.70	1.68
CH ₃ CH=CHCH=CHCO ₂ H	1.46	4.70	-0.34	1.96	2.03
C ₆ H ₅ CO ₂ H	1.87	4.20	-0.43	2.20	2.34
4-HOC ₆ H ₄ CO ₂ H	1.58	4.58	-0.34	1.92	2.11
C ₆ H ₅ CH ₂ CO ₂ H	1.41	4.29	-0.80	2.10	2.10
C ₆ H ₅ CH=CHCO ₂ H	2.13	4.44	0.07	2.52	2.41
2-HOC ₆ H ₄ CO ₂ H	2.26	2.97	-1.27	3.00	2.82
2-AcOC ₆ H ₄ CO ₂ H	1.23	3.50	-1.77	2.70	2.20
3,5-I ₂ -2-HOC ₆ H ₂ CO ₂ H	4.56	2.33	0.39	3.70	4.07
HO ₂ C(CH ₂) ₄ CO ₂ H	0.08	4.42,5.41	-3.20	1.46	(0.86)
HO ₂ C(CH ₂) ₅ CO ₂ H	0.61	4.48,5.42	-3.13	1.82	(0.85)

^aReference 25. ^bEquation 31.

Concluding Remarks

The wide range of examples we have presented using log D tend to support the broad view that distribution coefficients for ionizable compounds are the natural counterpart of partition coefficients for neutral compounds. The use of log D is particularly advantageous in the following circumstances: (i) when analyzing series containing ionized and unionized compounds, or containing both acids and bases; (ii) when consolidating data obtained at more than one pH; and (iii) when sorting out electronic, steric or other effects on distribution from those associated with a mechanism or site of action.

Literature Cited

1. Scherrer, R.A. and Howard, S.M., J. Med. Chem. (1977), 20, 53.
2. Timmermans, P.B.M.W.M. and van Zwieten, P.A., J. Med. Chem. (1977), 20, 1636.
3. Coats, E.A., Milstein, S.R., Pleiss, M.A. and Roesener, J.A., J. Med. Chem. (1978), 21, 804.
4. Al-Gailany, K.A.S., Houston, J.B. and Bridges, J.W., Biochem. Pharmacol. (1978), 27, 783.
5. Lien, E.J. in "Drug Design," Vol. V, E.J. Ariëns, Ed., Academic Press, New York, N.Y., 1975, pp 81-132.
6. Beckett, A.H. and Moffat, A.C., J. Pharm. Pharmacol. (1968), 20, 239S; (1969), 21, 139S.
7. Lien, E.J., Koda, R.T. and Tong, G.L., Drug Intel. (1971), 5, 38.
8. Schürmann, W. and Turner, P., J. Pharm. Pharmacol. (1978), 30, 137.
9. Büchel, K.H. and Draber, W., in "Biological Correlations-The Hansch Approach", pp. 141-154. Advances in Chemistry Series 114, American Chemical Society, Washington, D.C., 1972.
10. Draber, W., Büchel, K.H. and Schäfer, G., Z. Naturforsch (1972), 27, 159-171.
11. Kessler, R.J., Tyson, C.A. and Green, D.E., Proc. Natl. Acad. Sci. USA (1976) 73, 3141.
12. Tollenaere, J.P., J. Med. Chem. (1973), 16, 791.
13. Stockdale, M. and Selwyn, M., Eur. J. Biochem. (1971), 21, 565.
14. Norrington, F.E., Hyde, R.M., Williams, S.G. and Wootton, R., J. Med. Chem. (1975) 18, 604.
15. Motais, R., Sola, F. and Cousin, J.L., Biochim. Biophys. Acta (1978), 510, 201.
16. Cowles, P.B. and Klotz, I.M., J. Bacteriol. (1948), 56, 277.
17. Fujita, T., J. Med. Chem. (1966), 9, 797.
18. Weinbach, E.C. and Garbus, J., J. Biol. Chem. (1965), 240, 1811.
19. Hansch, C., Kiehs, K. and Lawrence, G.L., J. Am. Chem. Soc. (1965), 87, 5770.
20. Terada, H., Muraoka, S. and Fujita, T., J. Med. Chem. (1974), 17, 330.
21. van den Berg, G., Bultsma, T., Rekker, R.F. and Nauta, W.T., Eur. J. Med. Chem. (1975), 10, 242.
22. van den Berg, G., Bultsma, T. and Nauta, W.T., Biochem. Pharmacol. (1975), 24, 1115.
23. Eggerth, A.H., J. Exp. Med. (1929), 50, 299.
24. Hansch, C. and Clayton, J.M., J. Pharm. Sci. (1973), 62, 1.
25. Sheu, C.W., Salomon, D., Simmons, J.L., Sreevalsan, T. and Freese, E., Antimicrob. Agts. Chemother. (1975), 7, 349.

RECEIVED June 8, 1979.

Computer-Assisted Synthetic Analysis: The Merck Experience

P. GUND, E. J. J. GRABOWSKI, and G. M. SMITH—Merck Sharp & Dohme Research Laboratories, Rahway, NJ 07065

J. D. ANDOSE and J. B. RHODES—Management Information Systems, Merck & Co., Inc., Rahway, NJ 07065

W. T. WIPKE—Board of Studies in Chemistry, University of California, Santa Cruz, CA 95060

In the decade since the first published description of a computer program capable of deriving synthetic routes to complex molecules,⁽¹⁾ the field has grown into a lively discipline.^(2,3,4) Nevertheless, it cannot yet be said that computer assisted synthesis is an accepted research tool for many practicing synthetic chemists.

It is easy to see the appeal of computer aids to synthesis. Organic synthesis is complex; in principle several thousand reactions could be applied in several steps to millions of potential starting materials in order to prepare a desired product.^(2,3) The computer has the capability of extending the chemist's perception and memory, and of systematizing and organizing his synthetic knowledge.

While Merck has for a long time recognized the computer's potential in this area,⁽⁵⁾ the decade was half over before the field was sufficiently advanced for an in-house effort to be considered. When a computer assisted synthesis project was begun, a number of major design decisions were quickly made. To assure data security and to encourage optimal use, we wanted to run the program in-house. We wished to involve the synthetic chemists directly in the analyses; this meant interactive time-shared program operation on graphics terminals remote from the computer. This also meant that synthetic analyses would not be run exclusively as a scientific information service. Finally, by using the same graphics terminals for computer assisted synthesis and for the Merck Molecular Modeling Project, we could make more efficient use of our resources. A brief survey of extant programs indicated that the Simulation and Evaluation of Chemical Synthesis (SECS) program⁽⁶⁾ was most suited to Merck's needs.

Today we are running SECS on our corporate IBM computer, with the program accessed by four graphics terminals in three research laboratories (Rahway, New Jersey, West Point, Pennsylvania and Montreal, Canada). About 60 chemists have been checked out on running SECS analyses, and the program is

0-8412-0521-3/79/47-112-527\$06.25/0

© 1979 American Chemical Society

used almost daily. About 20 volunteer chemists have contributed chemistry for the program, and two chemists are adding chemistry to the file.

At this stage in the program's development, it is appropriate to report our experiences in implementing, evaluating and enhancing the SECS program at Merck, and to review plans for future development.

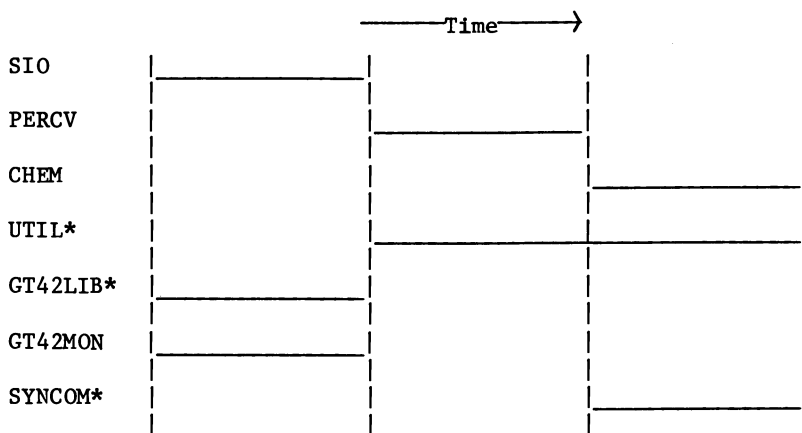
Implementation of SECS at Merck

This section gives a rather technical discussion of how SECS was converted to run in the Merck environment. We believe that our experiences may prove helpful to others faced with the problem of converting large programs to operate on alien computers. Chemists uninterested in the programming details may skip to the next section, on evaluation of program utility.

In October of 1974, we received version 1.0⁹ of the SECS program (6,8,10) with supporting documentation. The source program consisted of approximately 25,000 lines of code written for a Digital Equipment Corporation (DEC) PDP-10 computer (ca. 20,000 lines of FORTRAN written for DEC's F-40 compiler and about 5,000 lines of assembly language code written in MACRO-10) together with another 5,000 lines of code to support a DEC GT-40 series graphics display terminal (written in MACRO-11 assembly language for a PDP-11 minicomputer). System level documentation consisted of short (1-2 pages) write-ups for each of the ca. 250 subroutines comprising the SECS system and brief descriptions of the major data structures used throughout. In addition, we received a preliminary data base of approximately 250 chemical reactions written in the ALCHEM language (8) and treated as data by the program. Our immediate objective was to translate the program into code that would operate on our corporate IBM computer - at that time a model 370/155 - under the Time Sharing Option (TSO). (11)

General Implementation Approach. We divided the program into a number of smaller program modules, and then implemented each of these in an order which would permit testing of the program during the intermediate stages of assembly. Where possible, sections were tested independently prior to integrated testing.

For the purposes of our implementation, the SECS program was divided into the modules listed in Table I. (10) Implementation order dependencies (for testing purposes) for each module are shown graphically in Figure 1. Of these modules, UTIL, GT42LIB, and SYNCOM were separately and independently testable. GT42MON could be tested once GT42LIB was available, following which SIO could be tested. PERCV could not be tested until SIO and all of UTIL was available, whereas CHEM required the availability of all of the preceding plus SYNCOM.



(*Separately Testable)

Figure 1. SECS module testing dependencies

Implementation of the program, especially through the intermediate stages, was greatly facilitated by three factors. First, the program contained extensive built-in debugging facilities which could be selectively activated by setting appropriate "debug" switches as part of the input options in SIO. Second, a working version of the SECS program, available through a commercial time-sharing service,⁽¹²⁾ could be used to compare debug output from our version, allowing the detection of subtle errors introduced during the implementation process. Finally, Professor W. T. Wipke was available as a project consultant for assisting with difficult problems.

Implementation Decisions. A number of decisions were made during implementation of the SECS program which reflect differences between our computer environment and objectives, and those of the program originators. We decided to reimplement much of the assembly language code in FORTRAN in order to simplify future program maintenance and modifications. We also decided to omit packing of the connection table for the target molecule (which saves about 10 K bytes) since we had less stringent core limitations. Similarly, we chose not to implement the disk-based virtual array storage package that came with the SECS program. Instead, we expanded the memory allocation for the dynamic array allocation routines and diverted all virtual array storage package calls to the dynamic array storage package. Again for similar reasons, we chose not to overlay our implementation of the SECS program. The net effect of all these decisions was to increase the memory requirements for the

Table I

SECS Program Modules

<u>Module</u>	<u>Description</u>
SIO	Graphical and keyboard input/output routines; includes main command processor and driver routines.
PERCV	Perception routines; analyze structure for basic structural features (rings, functional groups, etc.) and assign a unique stereochemical name.
CHEM	Strategy routine, transform interpreter, precursor generator, precursor evaluation routines, synthesis tree generator.
UTIL	General purpose utility routines for dynamic array allocation, list processing, bit packing and unpacking, and logical set manipulation.
GT42LIB	General purpose, FORTRAN-callable subroutine package to support graphics on a GT-40 series graphics display terminal, based on software developed by Digital Equipment Corporation.
GT42MON	GT-40 series terminal monitor program controls communication between graphics terminal and host computer. Based on software developed by Digital Equipment Corporation.
SYNCOM	Stand-alone program for converting descriptions of chemical reactions written in ALCHEM into an object code usable by the SECS program.

SECS program from 240 K bytes (48 K words) on the PDP-10 to approximately 620 K bytes on our IBM 370/168 computer. This does not present a problem, however, since under IBM's virtual operating system,⁽¹⁾ each application program has available 16,000 K bytes of storage.

The most important implementation decision made was to make the minimal program changes necessary to get the program to run. Code optimization and program enhancements were deferred until after a working version was available. Failure to strictly adhere to this principle in the early stages significantly added to the conversion effort and resulted in the introduction of several errors which required additional time to locate and remove.

Implementation Problems. A discussion of problems encountered during implementation of the SECS program follows in three parts: problems in converting the assembly language routines, problems in interfacing the DEC GT-40 series graphics display terminal to our IBM computer, and problems in converting from DEC to IBM FORTRAN. Each of these problem areas will be discussed separately, below.

1. Assembly Language Implementation. In general, no problems were encountered in converting the DEC assembly language portions of the SECS program to run on the IBM computer. The program authors had anticipated eventual conversion of their program to run on machines with reduced word size. Care had been taken throughout the program to use no more than 32 of the 36 available bits of a computer word when packing data for storage. Had this not been the case, conversion of both the assembly language and the FORTRAN routines would have been significantly more complicated.

With the exception of the functional group definition routine (FGTAB) from the PERCV module, all other assembly language routines (from the UTIL and GT42LIB modules) are of the "software tool"⁽¹³⁾ variety - *i.e.*, independent, general purpose routines which perform a single, well-defined function. Rather than attempt a line by line translation from DEC's MACRO-10 assembly language to IBM's assembly language, we started with the functional definition of each routine and recoded each from scratch.

This gave us the flexibility to implement these routines in the most convenient language. Where we felt that execution speed would not be impacted too badly, we opted to code in FORTRAN because of the greater ease of program maintenance. Thus, the dynamic array allocation routines, the list processing routines, and most of the GT42LIB routines were rewritten in FORTRAN. The bit packing and unpacking routines and the logical set manipulation routines were rendered into IBM assembly language.

In the process of implementing the UTIL routines, whether in FORTRAN or assembly language, extensive error checking of passed arguments was incorporated to facilitate the location of errors introduced in implementing other parts of the program. The UTIL routines are extensively called throughout both the PERCV and CHEM modules. Since it was likely that errors introduced in implementation would result in "bad" arguments being passed to these routines, then by trapping such events and forcing a program termination with subroutine traceback, location of the error in PERCV or in CHEM was simplified.

2. GT-40 Interface. Our version of the SECS program uses a DEC GT-40 series graphics display terminal for all input and output (14) The software we received to support the terminal was a modified version of DEC's GT-40L program. Actually, GT-40L is two programs. The first program, which we designate as GT42LIB, consists of a set of FORTRAN-callable subroutines written in the DEC MACRO-10 assembly language. These subroutines are called by the application program (SECS) and send graphics commands to the terminal as encoded strings of ASCII characters. The second program, GT42MON, written in MACRO-11 assembly language, resides in the PDP-11 minicomputer which is an integral part of the graphics terminal. This program decodes the strings of characters, converts them into graphical images, and handles all communications between the terminal and the host computer.

Our task in interfacing the GT-40 series graphics terminal to our IBM 370 computer involved much more than a simple conversion of the MACRO-10 code into IBM equivalent code. It was necessary to modify the TeleCommunications Access Method (TCAM) portion of our IBM operating system to allow transmission of the full ASCII character set (including all control characters) to our terminal (15) It was necessary to modify all of the communication code in both GT42LIB and GT42MON to reflect differences between the IBM and DEC protocols for communication with asynchronous terminals (16) As mentioned earlier, most of GT42LIB was rewritten in FORTRAN with those portions that actually transmit character data written in IBM assembly language (17) Since there is no MACRO-11 cross compiler available on our IBM system, changes to GT42MON were made on a PDP-10 computer accessed through a commercial time-shared network (12)

3. FORTTRAN Implementation. Implementation of the FORTRAN sections of code involved removal of obvious syntax errors resulting from differences between the DEC FORTRAN-40 and the IBM FORTRAN G1 language specifications. Each section of code was compiled using the FORTRAN G1 compiler and the resulting output was scanned line-by-line. Compiler detected incompatibilities were removed by recoding.

Table II summarizes the major language differences which we encountered between the DEC and IBM FORTRAN dialects. In

general, the DEC language specification provided many added, non-standard features which were not available in the IBM version, necessitating significant recoding effort.

One particular difference in the IBM and DEC FORTRAN implementations was not flagged by the compiler but was isolated only with great difficulty during program testing. In the DEC computer, the expression "A.EQ.B" involves a bit by bit comparison of A with B for all data types. Under the IBM compiler, however, for single and double precision real (floating point) variables, if the fractional components of A and B both evaluate as zero, then A and B are considered equal no matter what the bit pattern for the exponents (the first eight bits). In ordinary usage, this would create no problems. In the SECS program, however, double precision variables represent not floating point numbers but sets of data, with each bit position representing a different element in the set. Therefore, under the IBM compiler, two sets with bit positions 9 through 63 equal to zero always compare as equal no matter what the individual bit patterns for positions 1 through 8. Once this problem was identified, it was easily rectified by coding an assembly language function EQV (SET 1, SET 2) which returns .TRUE. when SET 1 is equivalent to SET 2 and .FALSE. otherwise.

Implementation of the SECS program in our IBM computer environment spanned an 18 month period. This included time to learn IBM, PDP-10 and PDP-11 assembler languages; to learn the details of how the SECS program was operating; and to solve the IBM-GT40 communications problem. Also, time was lost when our heavily used time-sharing system gave consistently poor response. Nevertheless, far too much time was spent in converting incompatible dialects of the FORTRAN language. The authors would like to add their voice to that of Professor Beebe(8) in proposing strict adherence to ANSI standard FORTRAN. We feel, however, that the 1966 standard(9) is far too restrictive, especially for non-numerical programming applications, and we urge computer manufacturers to rapidly adopt the 1977 ANSI standard FORTRAN.(20) Most importantly, we urge all manufacturers who extend this standard in the future to build an option into their compilers to flag all use of extensions to the standard.

We believe that it would take much less time to reimplement the program today - both because of the problems which we have solved in a general way, and because the current version of SECS is closer to ANSI standard FORTRAN. It is possible that a pre-processor program could automatically remove many of the FORTRAN inconsistencies.

It is noteworthy that the data base of chemical reactions written in ALCHEM which we received, needed no modifications for use with our IBM version of SECS. The description of chemistry was thus independent of machine differences described above.

Table II

FORTRAN Differences

Feature	IBM FORTRAN G1	DEC F-40 FORTRAN
<u>Word Length:</u>		
Single Precision	32 bits	36 bits
Double Precision	64 bits (molecule of 64 atoms accommodated)	72 bits (molecule of 72 atoms accommodated)
<u>Memory Initialization:</u>	None	Initialized to zero.
<u>Data Initialization:</u>		
Variables in COMMON	Requires Separate BLOCK DATA routine.	DATA statements can appear in any routine.
<u>Boundary Alignment:</u>	Required (imposes restrictions on ordering of variables in COMMON)	No
<u>Input/Output:</u>		
Number of Characters/Word	4	5
Dynamic File Allocation	No	Yes
Free Format Input of Character Data	No	Yes
Partial Input of Variables in Free Format	No	Yes
Required Input for Each Item in Input List	Yes	No; omitted items default to zero.
Special Carriage Control Characters in FORMAT Statement	No	Yes; \$ in FORMAT statement suppresses carriage return and line feed.
ENCODE/DECODE	No	Yes

Table II (Continued)

Feature	IBM FORTRAN G1	DEC F-40 FORTRAN
<u>Data Types:</u>		
<u>Hollerith Constants</u>	Allowed in DATA and FORMAT statements only.	Also allowed in ASSIGNMENT and IF statements.
<u>Integer vs. Logical</u>	Strict separation enforced. In logical IF, operators .AND., .OR., .NOT. must be used with logical variables only while .EQ., .LT., .GT. must be used with integer variables only.	No strict separation. No need to declare logical variables. Operators .AND., .OR., .NOT., .EQ. operate bit by bit comparison for integer variables. Integer value \emptyset corresponds to .FALSE. while -1 evaluates as .TRUE.
<u>DO Loop Bounds:</u>	Integer expressions not permitted. Loop increment must be positive integer.	Integer expressions permitted. Loop increment can be positive or negative.
<u>Array Indices:</u>	Lower bound is one.	Can specify lower and upper bounds for range (positive or negative).
<u>Implementation of .EQ.:</u>	Bit by bit comparison except for floating point variables whose fractions both evaluate as zero. (Refer to text for consequences in SECS program.)	Bit by bit comparison for all data types.

Following our implementation of the SECS program and based on our experiences in using the program in a production environment (see below), we incorporated a number of enhancements into our version of SECS. These enhancements include a more streamlined (continuous drawing) format for molecule input using a light pen, the option to selectively display at the terminal only those structures which successfully pass evaluation, brightening of lines in the synthesis tree corresponding to structures which have been rated "GOOD", and lastly, a hard copy (Calcomp) summary of the results of an analysis. All of these enhancements have also been incorporated into subsequent versions⁽⁹⁾ of the parent SECS program at Santa Cruz by Professor Wipke, although several (such as hard copy) were implemented in a different manner.

Evaluation of Program Utility

Once SECS was implemented and running at Merck, it was necessary to involve the chemists in evaluating the program's utility. However, we were concerned that without being given background information, a chemist might see one or two poor SECS analyses and dismiss the methodology out of hand. In essence, we were asking the chemists to evaluate the utility of an eventual "production" system based on the current developmental system. In order to make such an evaluation validly, we felt the chemists needed some background in how the program actually operated.

Consequently, in the Fall of 1976 we organized a course for participating chemists. A total of 25 senior level chemists from six departments in Rahway (Basic Synthetic Research; Process Research; New Leads Discovery; Arthritis & Anti-inflammatory Research; Membrane Chemistry; and Radiochemistry) attended. The material (Table III), presented in 15 one-hour

Table III

Merck Course on Computer Assisted Organic Synthetic Analysis

1. Overview of the field.
2. Overview of SECS program (W. T. Wipke - 3 hours).
3. Program operation; introduction to IBM time-sharing (TSO).
4. Review of typical synthetic analysis.
5. Program modules, molecule drawing, structure representation.
6. Molecule perception.
7. What is a transform?
8. Transform types.
9. Detailed walk-through of a coded transform.
10. Precursor evaluation, strategy.
11. Model building, steric and aromatic effects.
12. Writing a transform.

lectures was detailed enough to dispel some of the mystery about how a computer can "perceive" a chemical structure, "remember" chemical reactions and "suggest" possible precursors. Despite the heavy technical content and the chemists' busy schedules, all participants but one completed the course. One chemist who was transferred to Merck's West Point, Pennsylvania laboratories commuted one day a week to Rahway in order to complete the course. A spin-off from preparing the lecture materials was a comprehensive SECS user's manual. The first lecture was also published (3)

In addition to attending the lectures, each participant attended workshops in groups of five and used the program on real synthetic problems. Each chemist performed at least 2 analyses, for a total of over 50. For perhaps two out of three compounds analyzed, nothing useful was discovered. Occasionally, however, really novel approaches were found to current synthetic objectives. Some noteworthy analyses are reviewed in the next section.

When the course was finished, we asked the participants to complete questionnaires and write evaluations of the project. We received 22 completed (anonymous) questionnaires and 15 reports. In the Fall of 1977, after a terminal was installed in West Point, the course was repeated for 21 Medicinal Chemistry Department chemists. At the conclusion of this course, 16 questionnaires were returned.

The following conclusions emerged from the chemists' evaluations.

1. The methodology offers great promise for aiding synthetic design. The system can improve and extend the chemist's own analytical abilities, and stimulate the consideration of new approaches. However, it can neither supplant the synthetic chemist nor relieve him of the task of working out a detailed synthetic plan.

2. Our system hardware is acceptable. Using the light-pen to "point" to the GT42-GT43 display screen is a convenient input method for the chemist. The IBM time-shared system (TSO) presented some problems for the computer-naive chemist (e.g., cryptic error messages), and often suffered from poor response times; at other times the system was quite satisfactory. Similarly, the SECS program design was considered excellent. There was a strong desire for hard copy summary of analyses (one member of each workshop recorded each analysis; automatic analysis summary on the plotter became available after the Rahway course was over).

3. The version of SECS (1.0) which we implemented⁽⁹⁾ was considered to need extensive further development. The chemists found that the preliminary data base needed expansion and revision. This was expected, since the supplied transforms had been written by the system developers to exercise various program features rather than in a systematic attempt to cover the field of synthetic organic chemistry. The chemists also felt that the analysis was too automated; they wanted more control over what chemistry was applied (more recent versions of SECS have high level strategies for this purpose). Many other, relatively minor complaints and suggestions were made.

4. It was widely believed that such a detailed course was not necessary for learning how to operate the SECS program.

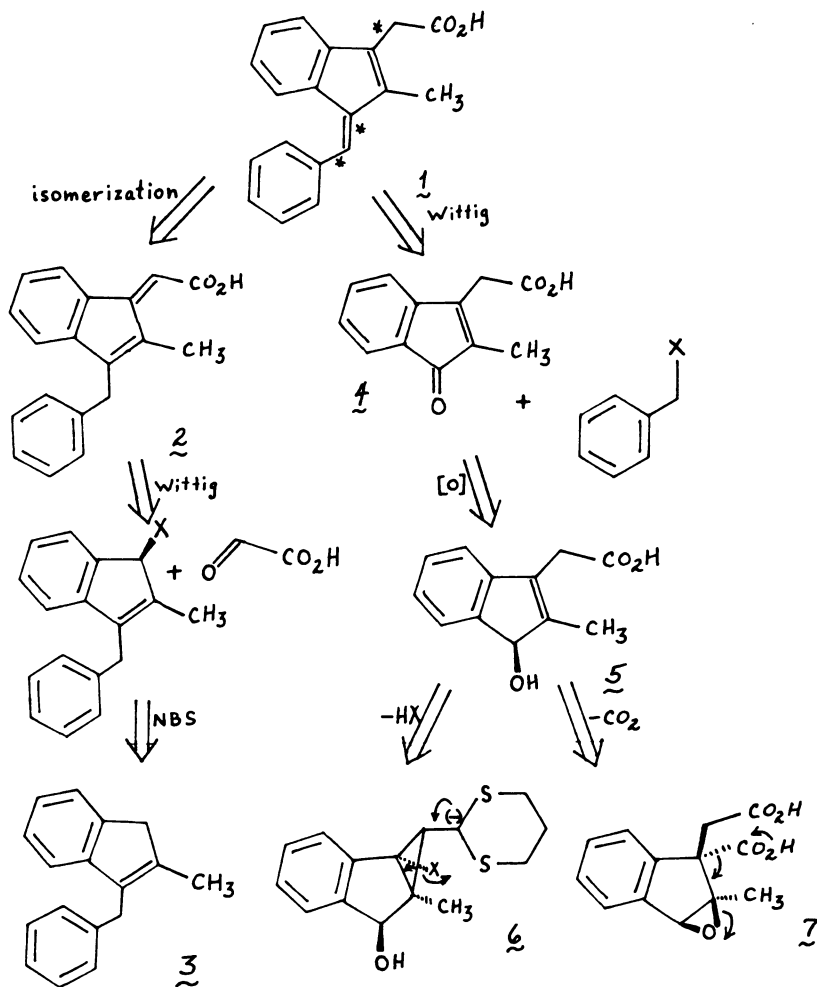
Since the completion of the courses, chemists have continued to participate in the project, by improving the data base of reactions (as discussed below) and by using the program. During 1978, monthly use of the program ranged from a low of 6 to a high of 71 runs. This year, we have offered first level SECS analyses as a service, similar to the way Dr. Beryl Dominy runs the program at Pfizer Corporation;⁽²¹⁾ this has been well received.

We plan to implement a more up-to-date version of SECS at Merck, and (as discussed below) we are working to improve and expand the program's chemical knowledge. As the software improves, we expect SECS to gain increasing acceptance here as a research tool.

Discussion of Selected Analyses

A review of the Merck experience with computer assisted synthesis would be incomplete without some discussion of actual results which have been generated by SECS. In doing so, one must remember that we are not working with a "finished product", but instead with a developing research tool which is the subject of current research efforts, and that our analyses are based on software which was "state-of-the-art" in 1974.⁽⁹⁾ Still, the best way to gain insight into the strengths and weaknesses of the synthesis program at this stage in its development is by considering the results of a few specific analyses. Structure 1 represents the nucleus for the new Merck product CLINORIL[®], and was our first candidate for a computer synthesis trial. It was an excellent test, because the SECS program was not particularly developed for analyzing structures such as 1. Furthermore an extensive body of chemistry had been developed for structures like 1 prior to the SECS analysis, so that the computer suggestions could be evaluated in light of laboratory experience. Finally, 1 remains an excellent target for new synthetic approaches. Part of the analysis developed by the chemists with the aid of the computer is summarized in Scheme I.

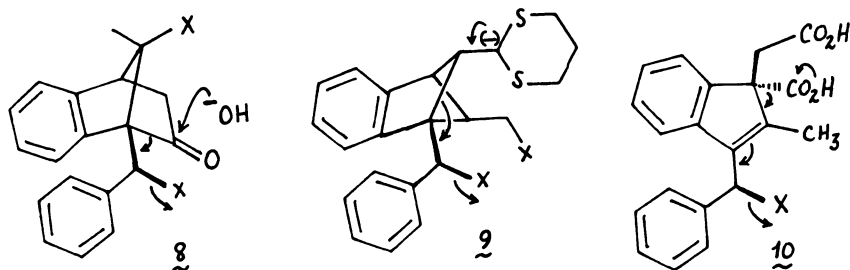
During the initial processing of 1 the starred bonds (*) were indicated to be "strategic" bonds (22). This proved most satisfying since one of the recently reported syntheses of CLINORIL® is based on the formation of these bonds (23,24). However, valence isomer 2 was not recognized by SECS 1.0 as a precursor of 1, and had to be processed separately. The combined processing of 1 and 2 did generate the basic elements of the synthesis of CLINORIL® discussed in reference 23. However,



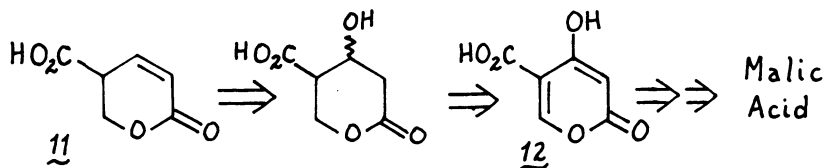
Scheme I

one key point missed by SECS is the acidity of 3 and its amenability to base catalyzed condensations. SECS did propose 3 as a precursor of 2, but by way of retro-synthetic Wittig and NBS reactions. The failure of SECS to recognize either the acidity of 3 or the viability of 2 as a precursor to 1 simply reflect deficiencies in the reaction data base at the time of the analysis, as opposed to any defects in the program's overall strategy.

Processing of intermediate 5, which derived from 1 via retro-synthetic Wittig and oxidation reactions, gave rise to precursors 6 and 7. *A priori*, the complex structures 6 and 7 (having 4 and 3 chiral centers, respectively), would appear to be poor precursors of the achiral target structure 1. However, the suggestion of 6 and 7 by the synthesis program incited chemists working on the problem to think about fragmentation processes as a route to 1. Such processes had not been considered previously because of the planar nature of 1. As an outgrowth of the SECS analysis of 1, reactions suggested by the arrows in structures 8, 9 and 10 were proposed by the chemists as possible precursors to 1. Although these proposals did not in fact lead to a significant new synthesis of 1, this example illustrates the program's ability to induce chemists to think about new approaches to familiar synthetic problems.

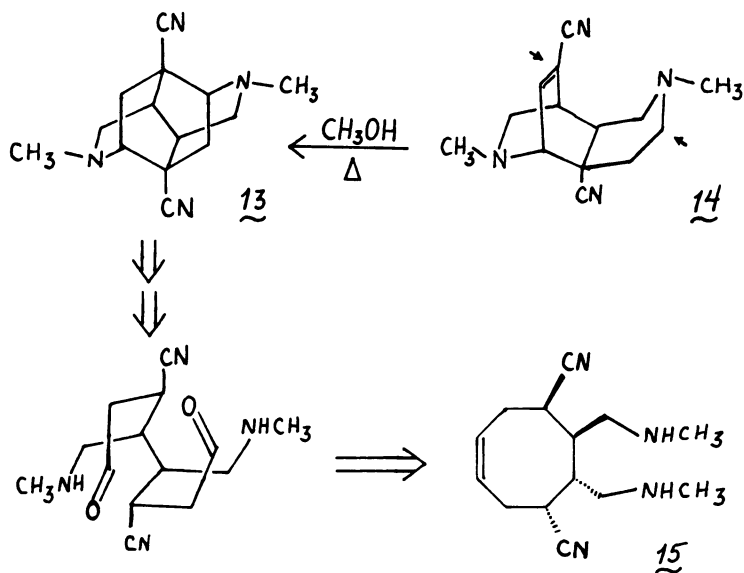


In another example, a simple route to structure 11, which has not been reported in the literature, was desired to support a possible synthetic program. A session on the terminal led the chemist to propose compound 12 as a likely precursor (Scheme II). A literature search indicated its ready preparation from malic acid, thus, providing a simple solution to the problem.



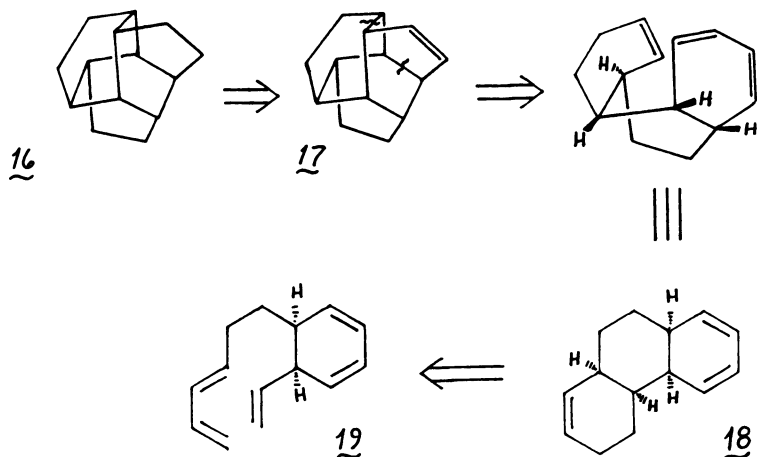
Scheme II

The potential of computer synthesis for fully and systematically analyzing routes to complex, polycyclic molecules has been noted before.⁽⁶⁾ Recently the chemistry of 5,11-diazaditwistanes has been extensively studied in these laboratories.⁽²⁵⁾ Compound 13, the key analog providing entry into this series, has been prepared from 14 as indicated in Scheme III. Not surprisingly, this approach was not suggested by SECS; indeed, the necessary hydride-transfer reaction was only discovered when thermolysis of 14 unexpectedly gave rise to 13. However, an intriguing route via a double Mannich reaction and ozonolysis of an appropriate cyclooctene derivative (15) did result (Scheme III). Functional group protection is unavailable in our current version of SECS; consideration of protecting groups makes the analysis of Scheme III somewhat longer.



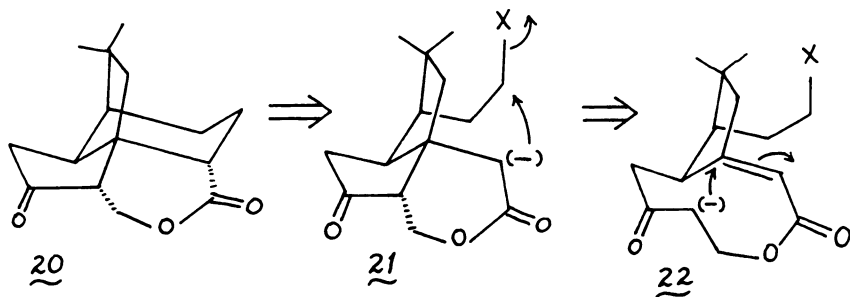
Scheme III

Work with the 5,11-diazaditwistanes brought to mind the unknown, all-carbon analog in the homologous series—tritwistane (16). Computer processing of 16 produced the appealing Diels-Alder route shown in Scheme IV. At present this route and obvious related approaches have not been reduced to practice.



Scheme IV

Recently our SECS program was demonstrated to Professor P. Magnus while he was visiting these laboratories.⁽²⁶⁾ He chose as a target the natural product Quadrone (20), which is of current interest in his research program. SECS projected preparation of 20 from 21 via anionic alkylation (Scheme V), and preparation of 21 from 22 via an intramolecular Michael reaction. The latter step solves the difficult problem of creating the tetrasubstituted center and could be combined with the alkylation to provide 20 in a single step. Using the basic approach of this SECS analysis, and taking into account protection and possible side reactions, Professor Magnus quickly proposed a different route to 20 which had not been considered before and which is a reasonable alternate to the routes which he had already projected.



Scheme V

The few examples shown here, among many that have been done, are representative of our current program's strengths and weaknesses. Without exception these examples show the importance of the computer-chemist interaction, with the computer suggesting approaches and the chemist thinking them through and refining them.

Further Development of SECS at Merck

Our experiences in using the SECS program indicated that further development was necessary for SECS to become routinely useful to the synthetic chemist. While development of new program features and strategies was deemed essential, these tasks have so far been left mainly in the hands of Professor Wipke's research group. We at Merck have concentrated on a different task: expanding and correcting the base of chemical reaction information available to the program, and establishing a system for efficiently writing reaction transforms.

Reaction Descriptions. Before discussing the work on data base expansion, a short description of the representation used for data storage is in order. The basic unit of information in the data base is the transform,^(6,7) which describes, in the retro-synthetic direction, the structural changes caused by a chemical reaction. Each transform consists of three sections;

1. Header
2. Qualifiers
3. Manipulations

The header contains the substructure (groups or pattern) which must be present in the molecule being processed (target molecule) in order for a transform to be applicable.

The qualifiers are used to determine if the structure being processed has features which will affect the course of the reaction. They can change the final priority (arbitrary rating score) of a precursor generated, or may prevent a transform from being applied at all. The qualifiers are optional; by defining the scope and limitations of a reaction, they give a transform its "chemical intelligence".

The manipulations symbolically modify the structure under consideration to give a proposed starting material.

The transforms are described in a special language, ALCHEM.⁽⁸⁾ This was designed to allow simple chemical statements to access all of the perceived structural information in the SECS program. An example of a statement is:

If aldehyde is anywhere then add 100

The organization of the reaction information as data, separate from the SECS program, and having its own language, allows transforms to be written, tested and modified easily. This has enabled us to concentrate on adding and correcting transforms without having to also make changes to the SECS program itself.

Supplied Reactions. The preliminary data base which was received with the SECS 1.0 program, containing approximately 250 transforms, was used unchanged in the initial evaluation phase described above. While many of these transforms gave excellent descriptions of chemical reactions, other transforms were not fully developed at the time we obtained them. The results of the workshop analyses highlighted some specific deficiencies. The problem given highest priority by the participating chemists was the presence of transforms which gave unacceptable results. The next highest priority item was missing reactions. After some discussion a list of 15 transforms was drawn up which represented the worst offenders. The most common features of these transforms were the chemically incorrect application of a reaction, excessive priority for an unimportant reaction, and/or application of a reaction despite an interfering functional group.

A prime example of this was the PINACOL reaction. The "bare-boned" transform which represented this reaction contained few qualifying statements, and required only the presence of a carbonyl group in the structure being analyzed. Since no restrictions were placed on the environment of that group, it could be a carbonyl in an acid, ketone, aldehyde, amide, or ester. The result was that many unrealistic starting materials were proposed as being converted to a target structure by the PINACOL reaction. Once the problem was understood, it was easy to add qualifying statements to properly restrict the application of this transform. The current version of SECS (2.7) uses higher level strategies to restrict the application of simplistic transforms. Since our version has a primitive strategy section, we have been forced to put strategic constraints into the transforms.

Transform Review Workshops. Workshops were organized to deal with the list of worst transforms. Attendance at the workshops, which met for an hour per week, was voluntary and ranged from a low of 3 to a high of 15. The format of the workshops was informal but followed the following outline:

The transform to be reworked was announced in advance, with literature references to the reaction being covered. A person familiar with ALCHEM (a "coder") gave a step by step walk-through of a transform. The chemical and structural meaning of each statement was explained, triggering a wider discussion of the scope and limitations of the reaction. As new aspects of

the reaction were covered, the coder made notes on qualifiers to be added or changed. One key objective of the coder was to prevent detailed discussion of experimental conditions which cannot be represented in a transform. From the notes on the meeting, a new transform was written and tested by the coder. Finally, this transform replaced the original. After the 15 transforms had been modified, a noticeable reduction in the number of bad structures was observed in analyses produced by SECS. From this experience two points became clear: the chemists did not want to learn ALCHEM, even though they were willing to volunteer their time and knowledge to improve the chemistry data base; and some other format was required to efficiently collect chemical information.

Reaction Review Seminars. At this point it was decided to change the emphasis from strictly transform correcting to writing new transforms. Reaction review seminars were instituted, in which a volunteer chemist would select any reaction of which he had special knowledge and prepare a seminar on its synthetic utility. The coder attended the seminar, and took notes; the chemist also supplied his notes. The coder then wrote a transform or transforms and did preliminary testing. Next the chemist tested the performance of the transforms on target structures of his choosing. Based on the chemist's suggestions, the transform was corrected as often as necessary, and finally added to the data base.

The advantages of this approach were that the speaker selected the reactions and so had more interest in them, and that the seminar served to increase the current awareness of other chemists. The major disadvantage was our difficulty in inducing chemists to volunteer, since preparation for the seminar required a great deal of time. While this format lasted for only a short while, it resulted in significant upgrading of the data base, as summarized below:

<u>Speaker</u>	<u>Length of Seminar (Hours)</u>	<u>Resulting Transform(s)</u>
R. A. Firestone	4	1-3 Dipolar Additions (30)
R. W. Ratcliffe	2	-Lactam Preparations (12)
E. D. Thorsett	1	Ni-Coupling Reactions (3)
E. J. J. Grabowski	1	Enamine Rearrangement (1)

Current Transform Writing Procedures. At this point we held a project review meeting with the chemists to assess progress and point future directions. The chemists felt they would like to continue to improve the data base, and Dr. Tom Beattie suggested that we send a memo to each Merck chemist asking for volunteers to summarize reactions for transform writing. Response was gratifying; about 60 chemists volunteered to help.

To facilitate this approach, we created a "reaction summary form" which was designed to assist the chemist and the coder in focusing on those aspects of a reaction which are important for transform writing. European companies running SECS have independently developed a similar form.⁽²⁷⁾ To further clarify the type of information required, a seminar was given on how to fill out the form - using an example of a reaction which had been written up as a transform. At the seminar a request was made for volunteers and 19 chemists agreed to participate.

The 15-page summary form consists of several sections, each focusing on a different type of structural feature. The first section requests reaction name, literature references, and a general equation. The next section is for a detailed description of the product substructure. Here the effect of possible substituents on the basic substructure and on reaction yield should be considered. Next the starting materials are similarly described. The following section is concerned with how rings affect the reaction, and whether the reaction is intermolecular, intramolecular or either. The next section requests information on interfering functional groups, and how unsaturation affects the reaction. The last section concerns experimental detail, which is described as comments in the transform. There is also a request for several test structures and how well the reaction works for their synthesis.

It takes chemists 1-5 days to summarize one reaction; approximately one week is required for a coder to incorporate the information on the form into a working transform. Additional time is required for the chemist to test the transform and the coder to make corrections. The procedure at this point seems to be accepted by the chemists. One point which is emphasized by the chemists is the need for current references to each reaction.

This process of collecting chemical information from experienced chemists and incorporating it into transforms has dramatically improved the quality of the analyses produced by SECS. However there are still gaps in the program's knowledge of most reactions and this is reflected in incorrect applications of some transforms. The result is that the process of transform writing must be interactive and continuing. If a major point has been overlooked or a new fact is learned about a reaction and the deficiency comes to light in an analysis, the transform must be modified and retested. As new facts are incorporated and corrections made based on experience, the program's chemical knowledge can be expected to become continually more sophisticated.

Writing a transform is similar to writing a computer program. When there is a large amount known about a reaction, the result can be a very long and logically complex structure. In order to streamline this procedure, it has been divided into several

steps, allowing the structure of the transform to grow in a way which facilitates testing at each stage.

The first step is to write the framework which consists of preceding comments (references, conditions, etc.), name, sub-structure, initial priority and manipulations. This is then tested with several test structures to make sure that the basic transform is correct. Next, qualifiers are added to exclude obvious problems, such as valence violations or simple steric requirements. Finally more sophisticated structural effects are noted and coded into qualifiers. The transform is tested at each step with structures which will exercise each new feature.

At some point in the coding of a transform the question arises: how much detail is required to give new ideas without producing nonsense results? To even consider this question, a second question must be answered; what are the goals of the chemist who is using the program? If he is searching for new ideas in the synthesis of a compound, then the transforms should be written without too much detail. The resulting analysis will be a larger and more general (but not strictly practical) presentation of ideas which would have to be refined further by the chemist. However if a chemist is looking for a reaction to perform a specific task, he wants to know the known limits of the reaction. Here the results will be a simpler analysis showing the application of a few specific transforms. Since a single data base is preferred and since the chemists who use the system have a wide range of requirements, the transforms written at Merck tend to be somewhere between the extremes described. One problem with attempting to include too much detail is that there are limits to what can easily be expressed in the ALCHEM language, as discussed below.

Several points have become clear during the two years which this work covered. First and foremost, experienced chemists must be involved in the development of the transform data base - not only to supply the information on reactions but also to recommend the types of reactions to be included. It is clear from the current analyses produced by SECS that many more transforms - perhaps 2-3 times the present number - must be added before a reasonably thorough analysis of a structure can be produced. We have also learned that chemists are not generally willing to learn ALCHEM, even enough to understand why a transform is misbehaving; but they will tell us which transforms need further work, and why.

One of the key items which a transform should contain if it is to be truly useful is a list of up-to-date references. This raises the major question of how best to maintain and retrieve the references in a large and growing data base.

American Chemical
Society Library
1155 16th St., N.W.

Our efforts to write new transforms have increased substantially in the past year with the addition of to our group of a synthetic chemist, Dr. Dale Hoff, who can both summarize reactions and code transforms. The current Merck SECS data base contains over 500 transforms.

The ALCHEM language⁽⁸⁾ has proved to be general enough and flexible enough to describe every reaction which we have considered, in English-like sentences which are reasonably intelligible to chemists. Nevertheless the vocabulary of organic chemists is very rich, and some concepts are presently difficult to express in ALCHEM. For example, the chemist fairly easily perceives the effect of heterocyclic functionality remote from a reaction center, but ALCHEM statements to accomplish such perception are quite complex. Besides being less easily comprehended, complex ALCHEM statements generally require extensive testing to verify that they are interpreted by the program as intended. We expect that further development of the ALCHEM vocabulary will make the language even more powerful and flexible.

Conclusion

If a convenient, routinely useful synthetic analysis program is developed, it will be accepted. Although many chemists resisted IR, NMR and mass spectral techniques when they were introduced, no modern chemist would dream of performing structure determination exclusively by degradation experiments. Complex organic synthesis is difficult, and the chemist can use help. If computer aided planning can reduce the number of unsuccessful experiments run, research managers will embrace the methodology.

Merck is participating in development of the SECS software in order to gain experience in this area; in order to help assure that the program develops into a useful tool for pharmaceutical research; and for what it can contribute to Merck's synthetic effort, even at this early stage.

Development of a synthesis may more or less be divided into the "synthetic planning" stage, where a rough synthetic approach is blocked out, and a "reaction retrieval" stage, where analogous sequences and specific reaction conditions are sought. An optimal "synthetic planning" tool is probably not best for the "reaction retrieval" function.⁽²³⁾ SECS is primarily a "synthetic planning" tool. There is a real need for "reaction retrieval" aids, but we suspect that this will have to be met primarily by other systems. Perhaps the new Derwent Reaction Retrieval Service will fill this need.

SECS program improvements must occur in several areas. We hope to implement an up-to-date version of SECS at Merck, to overcome many present limitations. In addition, high-level strategies must be developed to guide analyses, especially as the reaction data base expands, in order to remove uninteresting precursors and keep to a reasonable size the total number of structures generated. While our experience may help define successful strategies, this is not the area of Merck's initial interest.

Merck's main emphasis at present is in building an improved reaction data base. Already it is clear to us that this must be done slowly and carefully, by expert chemists, with continual refinement of transforms based on experience. The total number of synthetic reactions is rather large, and it is not at all clear that one company can commit enough resources to cover much of synthetic chemistry in a reasonable time period. Some sort of exchange of transforms by participating companies, such as is currently underway in Europe,⁽²⁾ would considerably speed system development.

Acknowledgement

We are grateful to M. J. Pensak for his contributions to an improved molecule input routine and the hard copy summary program; to Dr. Dale R. Hoff for his careful analysis of system capabilities and for his continuing efforts to improve the reaction data base; to the many Merck chemists who participated in evaluating and improving the program; and to several far-sighted administrators who initiated this project and sustained it in its early stages. SECS program development at Santa Cruz has been supported by the National Institutes of Health, Research Resources Grant No. RR-01059, and through computer support from the Stanford University SUMEX-AIM Project Grant No. RR-00785.

Literature Cited

1. Corey, E. J. and Wipke, W. T., Science (1969) 166, 178.
2. Review: Bersohn, M. and Esacks A., Chem. Revs. (1976) 76, 269.
3. Review: Gund, P., Ann. Repts. Med. Chem. (1977) 12, 288.
4. Symposium: Computer-Assisted Organic Synthesis, W. T. Wipke and W. J. Howe, Edits., ACS Symposium Series, No. 61, 1977.
5. Sarett, L. H., unpublished speech before Synthetic Organic Chemical Manufacturers Association, June 1964; quoted in Ref. 6.
6. Wipke, W. T., "Computer Representation and Manipulation of Chemical Information", W. T. Wipke, S. R. Heller, R. J. Feldmann and E. Hyde, Edits., J. Wiley, New York, p. 147, 1974.
7. Gund, P., Andose, J. D. and Rhodes, J. B. in Contributed Papers to III Internat. Conf. on Computers in Chemical Res., Educ. & Technol., E. V. Ludeña and F. Brito, Edits., Centro de Estudios Avanzados del Instituto Venezolano de Investigaciones Cientificas (IVIC), Caracas, p. 226, 1977.
8. Review: Wipke, W. T., Braun, H., Smith, G., Choplin, F. and Sieber, W. in Ref. 4, p. 97.
9. Different releases of the SECS program are identified by different version numbers, starting with 1.0. The current release of the program corresponds to version 2.7.
10. "Utilization of Stereochemistry and Other Aspects of Computer-Assisted Synthetic Design", T. M. Dyoth, Ph.D. Thesis, Princeton University, 1973 (University Microfilms, No. 74-9677).
11. Our current computer environment consists of an IBM 370/168 computer running under the OS/VS2 MVS-3.7A operating system. We have also run SECS on TSO under the OS-MFT and OS-MVT operating systems.
12. Initially available through First Data Corporation, Waltham, Massachusetts, the program is now available on ADP Network Services, Inc., Ann Arbor, Michigan.
13. Kerneghan, B. W. and Plauger, P. J., "Software Tools", Addison-Wesley, Reading, Massachusetts, 1976.
14. At Merck, we are using DEC GT-42 and GT-43 graphics display terminals with 16 K words of core memory. Communication with our host computer is at 1200 baud using VADIC 3400 series modems.
15. Detailed instructions for this modification to the operating system were already available to us through purchase of the PLOT-10 Terminal Control System software from Tektronix, Inc., for support of Tektronix 4010 series terminals on IBM computers.

16. DEC uses a full-duplex communications protocol with XON/XOFF control over transmission in both directions. "Full duplex" permits simultaneous transmission and reception of data. "XON/XOFF" control means that either host computer or terminal can turn off transmission from the other device by sending an "XOFF" control character, and can request resumption of transmission by sending an "XON" control character. By contrast, IBM uses a half-duplex communications protocol with the host computer the master and the terminal the slave device. When the host computer is transmitting data, it cannot receive, and when receiving, it cannot send (without loss of data). To prevent line conflict, the host computer invites the terminal to send data by transmitting a special character sequence. The terminal must wait until receiving this "invitation" before sending data to the host.
17. Reprogramming of GT42LIB was performed on a contract basis by William S. Price Associates, Inc., New York, New York.
18. Beebe, N. H. F. in Quantum Chemistry Program Exchange Newsletter No. 63, Indiana University (November, 1978), p. 5.
19. American National Standards Institute. American National Standard Programming Language FORTRAN, X3.9-1966.
20. American National Standard Programming Language FORTRAN, X3.9-1978; see also, W. Brainerd, Ed., *Comm. ACM*, 21, 806 (1978) and "FORTRAN 77", H. Katzan, Jr., Van Nostrand Reinhold, New York; New York, 1978.
21. Dominy, B., "SECS and the Information Scientist", presented at the Science Information Subsection of the Pharmaceutical Manufacturer's Association Meeting, Washington, D.C., March 6, 1977.
22. Corey, E. J., Howe, W. J., Orf, H. W., Pensak, D. A. and Petersson, G., *J. Amer. Chem. Soc.* (1975) 97, 6116 (1975).
23. Shuman, R. F., Pines, S. H., Shearin, W. E., Czaja, R. F., Abramson, N. L. and Tull, R., *J. Org. Chem.* (1977) 42, 1914.
24. Pines, S. H., Czaja, R. F. and Abramson, N. L., *J. Org. Chem.* (1975) 40, 1920.
25. Ten Broeke, J., Douglas, A. W. and Grabowski, E. J. J., *J. Org. Chem.* (1976) 41, 3159.
26. We are indebted to Professor Magnus for his permission to use this example.
27. Bruns, H., private communication.
28. Gund, P., Andose, J. D. and Rhodes, J. B., in Ref. 4, p. 179.
29. Bruns, H., Textes Conf. Cadre Congr. Int. "Contrib. Calc. Electron Dev. Genie Chim. Ind.", A, Soc. Chim Ind., Paris, 1978, p. 83.

RECEIVED June 8, 1979.

A Hierarchal QSAR Molecular Structure Calculator Applied to a Carcinogenic Nitrosamine Data Base

B. PETIT, R. POTENZONE, JR., and A. J. HOPFINGER—Department of
Macromolecular Science, Case Western Reserve University, Cleveland, OH 44106

G. KLOPMAN—Department of Chemistry, Case Western Reserve University,
Cleveland, OH 44106

M. SHAPIRO—Division of Computer Research and Technology,
National Institutes of Health, Bethesda, MD 20014

Lijinsky and coworkers (1) have reported a data base containing two quantitative measures of carcinogenic potency for a set of nitrosamines. The first carcinogenic measure, TD50, is the average time, in weeks, for 50% of a set of Sprague-Dawley rats to die relative to a control set of equal population. An upper limit of 100 weeks has been placed on each experiment. The second measure is an overall assessment of relative carcinogenicity, RC, based upon observations made over the course of the experiments. RC is recorded on a discrete scale of 0 to 4 with zero being "noncarcinogenic" and four being "extremely carcinogenic".

Exposure to the nitrosamine was achieved by placing a fixed dosage in the animal's drinking water. The sites of tumor formation were recorded for each compound tested. For specific testing details see Lijinsky and Taylor reports 2-4. It is to be emphasized that this data base was not constructed from biological experiments designed for quantitative structure activity relationship, QSAR, studies. The experiments were developed only to ascertain whether or not a compound is "carcinogenic". Consequently, relatively high doses have been used which lead to a narrow range of responses.

Nevertheless, Singer et al. (5) and Wishnok et al. (6) have used biological activities of the type and quality as TD50 to establish QSAR's for some sets of nitrosamines. Singer et al. (5) found a linear correlation between experimentally measured water/octanol partition coefficient, P, and the percentage of Sprague-Dawley rats bearing olfactory carcinomas induced by substituted N-nitrosopiperidines. In addition, a parabolic correlation was established between P and the relative mean lifetime of the animals bearing hepatocellular carcinomas induced by a small, but moderately diverse, set of nitrosamines.

Wishnok et al. (6) have found the water/hexane partition coefficient P' and "electronic factors" as measured by Taft σ^* to correlate with carcinogenic activity for a series of nitroso-

0-8412-0521-3/79/47-112-553\$07.25/0

© 1979 American Chemical Society

compounds. However, the applicability of σ^* to nitroso group substituents is questionable. Nevertheless, both groups of researchers concluded that transport of the carcinogen to the active site, as is assumed modeled through partition coefficient, plays a significant role in specifying carcinogenic potency.

We have carried out a series of investigations in pursuit of a QSAR for a set of cyclic nitrosamines from the Lijinsky data base (1). TD₅₀ has been used as the biological measure. All the cyclic nitrosamines considered produced tumors of the esophagus. In several instances tumors were found at additional sites in the animals. It is to be emphasized that the occurrence of tumors, and the reduced average life of the animals, due to exposure to nitrosamines, must be assumed indicative of the carcinogenic potency of a compound.

Our nitrosamine QSAR investigations can be classified in terms of four increasing levels of structural sophistication;

1. determination of group-additive (7) molecular descriptors,
2. calculation of descriptors using molecular mechanics methods (8),
3. calculation of quantum mechanical (molecular orbital) (9) descriptors, and
4. mechanism of action computations which employ all of the above mentioned descriptors.

In general QSAR models can be constructed from any combination of the four levels of structure calculation. We are currently developing a software package to compute the different classes of molecular descriptors and to construct action models. This software package is being called "A Hierarchal QSAR Molecular Structure Calculator".

Method

1. Group Additive Molecular Descriptors: Log P values were calculated using Rekker fragment constants (10) applying non-additive corrections according to Leo (11). Fragment constants are not available for the nitroso group. To estimate this fragment constant, $f(\text{NNO})$, the cyclic nitrosamines were grouped according to the types of atoms in the ring. Those compounds containing only aliphatic units, besides a single nitroso group, were grouped together as were the nitrosomorpholines, the nitropiperazines, and di-N-nitroso-derivatives. In constructing these clusters we assume that the electronic properties of the nitroso group are altered by in-ring substituents. In turn, changes in the electronic properties of the nitroso group alters its solute-solvent behavior and, consequently, log P. A linear least-square fit of the experimental log P versus the calculated log P (not containing a contribution from the nitroso group) has been carried out for each of these four groups. The negative of the log P_{cal} intercept of the least-square fit line can be identified as $f(\text{NNO})$. The legend of Table 1 contains the $f(\text{NNO})$ of each

of the four compound groups.

The experimental log P are reported by Singer et al. (5) and were measured by optical density spectroscopy. We have carried out a linear regression analysis between log P_{exp} and log P_{cal} for the complete data base in Table 1. The correlation equation is

$$\log P_{\text{exp}} = 1.0049 \log P_{\text{cal}} + .0108 \quad (1)$$

$$N = 28 \quad R = .97 \quad S = \pm .203$$

where N is the number of compounds, R is the correlation coefficient, and S the standard deviation.

Log P, at a fixed temperature, is a measure of the free energy difference between a solute in water and 1-octanol. At room temperature,

$$\log P = .735 (F_{\text{H}_2\text{O}} - F_{\text{Oct}}). \quad (2)$$

Eq. (2) indicates that $F_{\text{H}_2\text{O}}$ and F_{Oct} can both vary, while log P remains constant. Consequently, if the chemical activity (solution free energy) of a compound in the aqueous and/or lipid medium is a significant feature(s) in controlling biological activity, log P may not be an adequate descriptor. Fortunately, a large number of aqueous activity coefficients of organic compounds have been measured (12). Using a group additive formalism and the thermodynamic relationship,

$$a_{\text{H}_2\text{O}} = \exp(-F_{\text{H}_2\text{O}}/RT) \quad (3)$$

it has been possible to construct a set of aqueous free energy fragment constants analogous to Rekker f-constants (10) and Hansch π -constants (13). $a_{\text{H}_2\text{O}}$ is the measured aqueous activity coefficient. The corresponding aqueous free energy fragment constants, w(X), are listed in Table 2. No w(NNO) constants were available from the analysis of the aqueous activity coefficient. Consequently, we carried out computer simulation calculations to model the interaction of water molecules with an NNO group. These simulation studies are identical to those employed to estimate the aqueous and 1-octanol free energies of some simple organic compounds (14). Interestingly, these calculations suggest that the octanol contribution to log P for the NNO group is essentially zero. The electrostatic interactions between the water molecules and NNO dominate in the solvation energetics.

The set of molecular aqueous free energies of the cyclic nitrosamines are in Table 1. Eq. (2) can be solved for F_{Oct} and these descriptor values for each of the cyclic nitrosamines are also reported in Table 1.

2. Molecular Mechanics Descriptors: A consistent force field (CFF) method, adapted from the MMI program of Allinger, (15) of

Table 1 (1 of 5)

Table 1
List the Nitroamines, Carcinogenic Indices, and Physicochemical
Features Considered

Compound	TD50	RC	Log P _{exp}	Log P _{cal}	F _{H₂O} kcal/mol	F _{oct} kcal/mol	E[boat] ^H					
							Eq		Ax			
							Cis (Eq-Eq)	trans (Eq-Ax-Cis)	Cis	trans (Ax-Ax)		
Nitrosopiperidines^A												
Nitroso-piperidine	38	3	.63	.72	-.94	-1.92	11.85	11.85	11.85	11.85	11.85	11.85
2-Methyl-	80	2	.71	.80	-.58	-1.67	53.81	22.46	27.25	25.40	25.40	25.40
3-Methyl-	55	3	.99	-2.05	34.27	37.17	35.98	33.80	33.80	33.80
4-Methyl-	40	3	1.04	-2.57	23.74	23.74	28.78	28.78	28.78	28.78
2,6-Dimethyl-	>100	0	1.36	-1.85	(55.94)	(54.46)	(54.46)	(39.92)	(39.92)	(39.92)
3,5-Dimethyl-	100	1	1.53	-2.23	(39.61)	(38.42)	(38.91)	(40.80)	(40.80)	(40.80)
2,2,6,6,- Tetraamethyl	>100	0	2.42	-2.89	86.15	83.92	84.55	70.30	70.30	70.30
4-Phenyl-	>60	2	2.59	-6.27	34.68	34.68	64.19	64.19	64.19	64.19
4-t-Butyl-	>100	0	2.54	-3.07	46.33	46.33	72.91	72.91	72.91	72.91
3-Hydroxy-	43	3	-.47	-5.60	22.35	34.80	33.86	32.85	32.85	32.85
2-Carboxy-	>100	0	-	-7.37	HI	30.55	31.90	33.62	33.62	33.62
4-Carboxy-	>100	0	-	-7.01	26.58	26.58	18.93	18.93	18.93	18.93
4-Chloro-	41	4	-	-2.85	20.63	20.63	25.21	25.21	25.21	25.21
3,4-Dichloro-	20	4	1.04	-3.08	(15.39)	(29.23)	(29.97)	(13.15)	(13.15)	(13.15)
3,4-Dibromo-	36	4	1.23	-5.54	-	-	-	-	-	-
Nitrosopyrrolidines^A												
2,5-Dimethyl-	>100	0	.86	.67	-.42	-1.33	(43.51)	(61.36)	(35.51)	(26.61)	(26.61)	(26.61)
3,4-Dichloro-	44	4	.83	.27	-1.92	-2.29	(40.95)	(28.89)	(27.10)	(39.55)	(39.55)	(39.55)
2-Carboxy-	>100	0	-	-.88	-5.93	-4.73	46.51	30.50	31.92	31.36	31.36	31.36
2-Carboxy-C	>100	0	-	-2.08	-12.09	-9.25	HI	24.49	32.36	32.95	32.95	32.95
4-Hydroxy-	>100	0	-	-	-	-	-	-	-	-	-	-

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

Table 1 (2 of 5)

Compound	TD50	RC	Log P _{exp}	Log P _{cal}	F _{H₂O} kcal/mole	F _{oct} kcal/mole	E[boat] ^M			
							Eq		Ax	
							Cis	trans	Cis	trans
							(Eq-Eq) Eq-Eq	(Eq-Ax-Cis) Eq-Ax	(Eq-Ax-trans) Eq-Ax	(Ax-Ax) Ax-Ax
<u>Nitrosomorpholine^D</u> 2,6-Dimethyl- -Thiomorpholine -Phenothiazine ^{E, F} N-Nitroso-heterocycles	30 90 90	4 2 0	.32 .40 2.09	.45 .70 1.65	-2.38 -.82 -4.38	-2.99 -1.77 -6.62	(58.11) 42.41 [74.29]	(52.60) 42.41 [48.93]	(48.83) 42.41 [56.21]	(37.45) 42.41 [70.92]
Nitrohexam- thyleneimine	28	4	.92	1.25	-.74	-2.45	N ₁ and (C ₄ C ₅) above C ₂ C ₃ C ₆ C ₇ plane			
Nitrosopentam- thyleneimine	25	4	1.47	1.68	-.54	-2.83	C ₄ C ₅ C ₆ vertex same direction as C ₂ H ₃ C ₈			
Nitrosocotam- thyleneimine	43	3	2.04	2.21	-.34	-3.35	N ₁ and (C ₄ C ₅ C ₆ C ₇) both above (C ₂ C ₃ C ₆ C ₉) plane			
<u>Nitrosopiperazines^C</u> Nitroso- piperazine 4-Methyl- Di-N-Nitroso- _{H, I} Derivatives	>100 >100	0 0	.18 .20	.03 .37	-5.53 -1.70	-5.58 -2.20	26.93 31.37	26.93 31.37	26.93 31.37	26.93 31.37
Dinitroso- piperazine 2-Methyl- 2,5-Dimethyl- ^J 2,6-Dimethyl-	51 33 46	3 4 3	-.85 -.28 .15	-.79 -.28 .02	-3.16 -2.80 -2.44	-2.09 -2.42 -2.47	7.69 59.42 (43.85)	7.69 31.93 (37.21)	7.69 24.19 (37.21)	7.69 24.38 (30.53)
2,3,5,6- Tetraethyl- ^K -Homopiperazine ^L	>100 36	0 4	.92 -.51	.83 -.26	-1.72 -2.96	-2.85 -2.61	[Eq-Ax-Eq-Ax] trans			
							46.60	46.60	46.60	46.60

Table 1 (3 of 5)

Part 2-A

Compound	E[chair]M						ΔE^M		Boat		Chair	
	Eq		trans		Cis		AX		Qu	Q _B	Qu	Q _B
	Cis	trans	(Eq-Ax-Cis)	(Eq-Ax-trans)	(Eq-Ax-trans)	trans	(boat _{min})	Qu	Q _B	C _{2v}	C _{2v}	
	(Eq-Eq)	(Eq-Ax-Cis)	(Eq-Ax-trans)	(Eq-Ax-trans)	(Ax-Ax)	(chair _{min})	(eu)	(eu)	(eu)	(eu)	(eu)	(eu)
Nitrosopiperidines^A												
Nitroso-piperidine	9.56	9.56			9.56	9.56	2.29	.111	-	.137	-	
2-Methyl-	81.53	16.39			17.83	14.25	8.21	.123	.145	.138	.145	
3-Methyl-	36.32	35.51			33.47	32.28	1.52	.106	-	.111	-	
4-Methyl	21.73	21.73			27.45	27.45	2.01	.112	-	.113	-	
2,6-Dimethyl-	(69.13)	(58.83)			(45.58)	(30.31)	9.61	-	.141	-	.142	
3,5-Dimethyl-	(42.65)	(49.69)			(49.31)	(52.86)	-4.23	.107	-	.111	-	
2,2,6,6-Tetraethyl-	80.53	77.62			78.50	60.58	9.72	-	.149	-	.138	
4-Phenyl-	30.89	30.89			42.25	42.25	3.79	.133	-	.128	-	
4-t-Butyl-	59.46	59.46			57.13	57.13	-10.80	.108	-	.108	-	
3-Hydroxy-	25.00	22.06			19.61	34.20	2.74	.082	-	.080	-	
2-Carboxy-	HI	31.45			32.25	33.95	.90	.136	.166	.130	.152	
4-Carboxy-	24.31	24.31			21.92	21.92	-2.99	.137	-	.138	-	
4-Chloro-	17.67	17.67			19.68	19.68	2.96	.113	-	.130	-	
3,4-Dichloro-	(23.91)	(11.46)			(11.02)	(10.73)	2.03	.114	-	.132	-	
3,4-Dibromo-	-	-			-	-	-	-	-	-	-	
Nitrosopyrrolidines^A												
2,5-Dimethyl-	-	-			-	-	10.75 ^B	-	.150	-	.153	
3,4-Dichloro-	-	-			-	-	1.79 ^B	.158	-	.158	-	
2-Carboxy-	-	-			-	-	.86 ^B	.140	.169	.141	.170	
4-Hydroxy-	HI	27.39			31.39	32.58	-2.90	.129	.173	.131	.175	

Table 1 (4 of 5)

Part 2-B

Compound	E[chair] ^M				ΔE^M (boat _{min} ^M - chair _{min} ^M) (eu)	Boat			Chair			
	Eq		Ax			Q _u	Q _s	C _a	Q _u	Q _s	C _a	
	Cis (Eq-Eq)	Trans (Eq-Ax-Cis)	Cis (Eq-Ax-trans)	Trans (Ax-Ax)		Q _u	Q _s	C _a	Q _u	Q _s	C _a	
	(eu)	(eu)	(eu)	(eu)		(eu)	(eu)	(eu)	(eu)	(eu)	(eu)	
<u>Nitrosomorpholines^D</u>												
2,6-Dimethyl- -thiomorpholine	(63.39)	(50.95)	(46.32)	(33.93)	3.53	-	.153	-	.150	-	.137	-
-Phenmetrazine ^E	[53.95]	[40.55]	[51.22]	[36.39]	8.28	.109	1.44	.116	.147			
<u>N-Nitroso-heterocycles^{A,F}</u>												
Nitrosobenzene- thyleneimine	N1 below and (C ₄ ,C ₅) above C ₂ C ₃ C ₆ C ₇ plane	29.14			2.78	.111	-	.129	-			
<u>Nitrosopiperazines^G</u>												
Nitrosopiperazine- thyleneimine	C ₄ C ₅ C ₆ vertex opposite direction as C ₂ N ₁ C ₈	25.86			2.60	.109	-	.127	-			
Nitrosocotame thyleneimine	N1 below and (C ₄ C ₅ C ₆ C ₇) above (C ₂ C ₃ C ₈) plane	37.32			2.33	.107	-	.130	-			
<u>Nitrosopiperazines^G</u>												
Nitroso- piperazine	18.29	18.29	18.29	18.29	8.64	.142	-	.144	-			
4-Methyl- DI-N-Nitroso- Derivatives ^{H,I}	21.81	21.81	21.81	21.81	9.56	.133	-	.131	-			
Dinitroso- piperazine	6.28	6.28	6.28	6.28	1.41	.148	-	.146	-			
2-Methyl- 2,5-Dimethyl- 2,6-Dimethyl- 2,3,5,6- tetramethyl- K	63.92 (37.35) (58.21)	34.95 (34.83) (57.83)	22.03 (34.83) (34.71)	21.88 (37.19) (37.71)	4.30	.149	.155	.151	.154			
Nitrosobenzene- thyleneimine ^L	[Eq-Ax-Eq-Ax] trans	52.38			-3.01	-	.169	-	.172			
Nitrosobenzene- thyleneimine ^L	42.38	42.38	42.38	42.38	3.22	.138	-	.138	-			

Table 1 (5 of 5)

- A) Least square fit for $f(\text{NNO})$ constant is -1.94 .
- B) There are no boat or chair conformers. The energy difference between the two lowest ring puckered conformers arising from axial, equatorial, and cis/trans NNO configurations replaces $\Delta E[\text{boat-chair}]$.
- C) Only axial 4-OH substitutions were considered.
- D) Least square fit for $f(\text{NNO})$ constant is -0.90 .
- E) The 3-CH₃ is the first unit referred to in the configurational definitions. The phenyl ring was only considered as trans to the nitroso group.
- F) Trial ring structures were selected which could be identified as analogous to boat, chair conformers for 6-member rings. These trial structures were then energy minimized using CFF calculations. It is quite probable that additional conformers exist.
- G) Least square fit for $f(\text{NNO})$ constant is -0.27 .
- H) Least square fit for $f(\text{NNO})$ constant is -1.46 .
- I) The N-O of the two NNO were held trans to one another across the ring.
- J) Only trans configuration was considered for NNO relative to 2-CH₃.
- K) The only stable configuration for both the boat and chair is $2(\text{Eq}) - 3(\text{Ax}) - 5(\text{Eq}) - 6(\text{Ax})$
- L) The nitrosohexamethyleneimine trial structures were used.
- M) The CFF model has been refined on an arbitrary energy scale which is close, but not precisely, kcal/mole units.

Table 2
Current set of aqueous free energy
fragment constants $w(s)$.

Solute Group Number	Solute Group Type, s	$w(s)$ kcal/mole
1	CH ₃ (Aliphatic Methyl)	.80
2	OH (Aliphatic Hydroxyl)	-5.82
3	CH ₂ (Aliphatic Methylene)	.20
4	CH (Aliphatic Methine)	-.24
5	C (Aliphatic Tert-Butyl)	-.72
6	C=C (Aliphatic Vinyl)	-.66
7	F (Aliphatic Fluoro)	.80
8	CL (Aliphatic Chloro Single CL on Atom)	-.94
9	CL (Aliphatic Multiple Cl's on an Atom)	-.05
10	BR (Aliphatic Bromo)	-1.14
11	-O- (Aliphatic Ester)	-3.97
12	O=C (Aliphatic Carbonyl)	-.65
13	NH (Aliphatic Amide)	-5.96
14	H (Vinyl)	.29
15	-O- (Aliphatic Ether)	-3.97
16	CH (Aromatic)	-.17
17	C (Aromatic)	-.90
18	OH (Aromatic Hydroxyl)	-5.45
19	-O- (Aromatic ether)	-3.75
20	C (Aromatic-Fused)	-.65
21	NH (Aromatic Amide)	-5.83
22	CH ₃ (Aromatic Methyl)	.82
23	CH ₂ (Aromatic Methylene)	.26
24	CH (Aromatic Methine)	-.30
25	C (Aromatic Tert-Butyl)	-1.03
26	CL (Aromatic Chloro)	.75
27	C=N (Aliphatic Cyano)	-2.25
28	-S- (Aliphatic Sulfure)	-.40

our molecular structure processor system, CAMSEQ-II, (16) was used to determine the cyclic nitrosamine ring conformers. The valence geometry and force-field constants were taken from earlier calculations (17,18) and experiments (19). Lijinsky and coworkers (1) did not resolve the configurational isomers used in the biological experiments. Consequently, we found it necessary to structurally analyze possible configurational isomers.

The majority of compounds in the cyclic nitrosamine data base are six-membered rings. We quickly realized that these six-membered rings could adopt two general conformer states corresponding to the boat and chair conformations of cyclohexane. Hence, we constrained the CFF calculations to refining trial boat and chair conformations as a function molecular configuration. As a typical example, "stick" models of the preferred boat and chair conformers of nitrosopiperidine are shown in Fig. 1. Ring substituents can be substituted either axial or equatorial and, for ortho- and meta substitutions, either cis or trans to the (NO) of the nitroso group. All unique configurations (maximum of eight) were considered for the hexacyclic nitrosamine data base. The CFF relative energy for the energy minimized structures are given in Table 1. It should be pointed out that the CFF energies do not have significance on an absolute energy basis, but are meaningful measures of relative conformer stability.

The relative stabilization energies of ring puckering, as a function of ring substituent configuration, have been used as entries in Table 1 for the nitrosopyrrolidines. For the seven, and larger size rings, the specific conformer states cited in Table 1 were used as initial trial structures. Each of these classes of conformers can be identified as analogous to a boat, or a chair, structure in that the pseudo para ring groups are, respectively, cis or trans to the nitroso group across the ring. All energy minimized structures are very similar to their corresponding trial conformers. The alternating (eq-ax-eq-ax) tetramethyl configuration was used to determine boat/chair conformers of 2,3,5,6 tetramethyl di-N-nitrosopiperazine. We cannot rigorously ascertain the energy-minimized ring conformer states to be the lowest possible energy states. More generally, for any compound we cannot be sure that a resulting structure is the actual lowest energy conformer for a given configuration. The resultant energy minimized structure depends both upon the initial trial structure and minimizer algorithm. Ultimately, the internal consistency and reasonableness of the resulting QSAR is the only indirect evidence for the reliability of the calculations.

3. Quantum Mechanical Descriptors: One set of quantum mechanical descriptors used in this study is the atomic charge densities. Charge density is a reasonable indicator of the potential chemical reactivity of a given site of a molecule as a result of electrostatic interaction (9). In particular, it has been



Figure 1. Molecular stick models of the chair (left) and boat (right) conformer states of nitrosopiperidine

suggested that the C_{α} ring atoms chemically react as part of the carcinogenic mechanism (3,20). Therefore, determination of the excess positive charge density for the unsubstituted/substituted C_{α} atoms in the set of cyclic nitrosamines appears to be a potentially useful QSAR descriptor.

Two different molecular charge distributions per configurational conformer were determined using the CNDO/2 method (21). An initial charge distribution was calculated for a stereochemically allowable initial structure. This set of charges was used in the subsequent CFF energy minimization, CNDO/2 was then reapplied to determine the charge densities for the energy minimized structure. It was found that individual atomic charge densities differed between the trial and final calculations, for all compounds in the data base, by less than .07 eu. The residual charge densities for the C_{α} atoms are given in Table 1.

The other set of quantum mechanical descriptors computed for some of the cyclic nitrosamines is related to a proposed model of action. This model is described in the next section and is based upon the reaction pathway shown in Fig. 2 for 2-methyl nitrosopiperidine. The hydroxylation is postulated to occur from a non-specific interaction with cytochrome P-450. This ultimately leads to the formation of carbonium ions. It is further postulated that the carcinogenicity arises from an alkylation process, most likely involving DNA. Consequently, the carbonium ion metabolites of the nitrosamines are identified as potentially "active" species. The relative stabilization energies of the different possible carbonium ion species, corresponding to the cyclic nitrosamine, constitutes the remaining set of quantum mechanical descriptors. These quantities were computed using the MINDO/3 (22) method with a fixed valence geometry constraint. Torsional bond rotations were allowed in which the chain conformer state was used as the initial trial structure. The relative stabilization energies of the model compounds are reported in Table 3. $\Delta E(\text{II-I})$ is the formation energy of a primary carbonium ion while $\Delta E(\text{II-III})$ is the formation of a secondary (or tertiary) carbonium ion.

4. Mechanism of Action: We began with the classic action model used by Hansch (23) to develop his linear free energy relationships and make some generalizations:

$$\frac{d(\text{response})}{dt} = ACFK, \text{ where} \quad (4)$$

A represents the probability of a molecule to reach the critical site in a given time interval

C is the applied concentration, and

$\frac{d(\text{response})}{dt}$ is the biological effect.

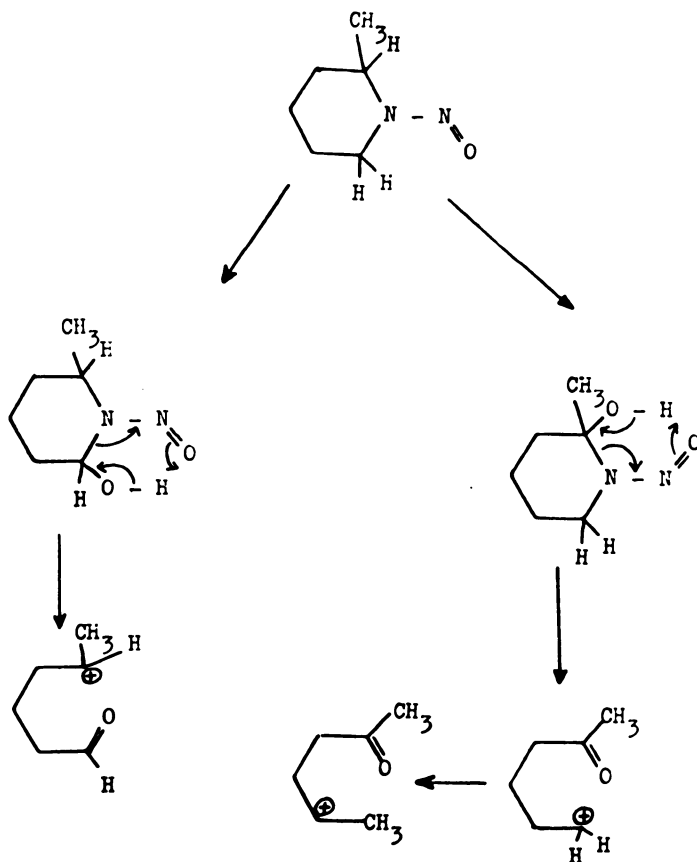


Figure 2. Possible mechanism of metabolic activation of 2-methyl nitrosopiperidine to carbonium ion metabolites

Table 3

Relative Stability of alcohol, primary, secondary and tertiary models for possible cyclic nitrosamine metabolites as proposed in figure 2. All energies are in kcal/mole.

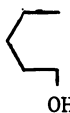





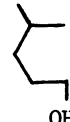


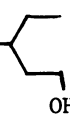
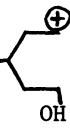
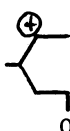
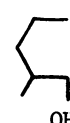



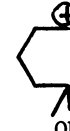
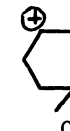
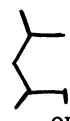
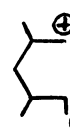







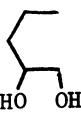
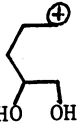
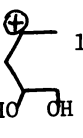
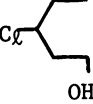
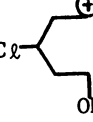
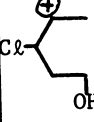
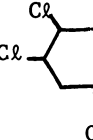
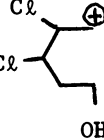
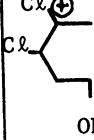
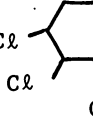
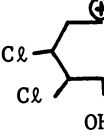
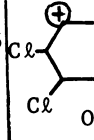
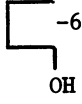
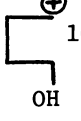
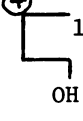


I	E	II	E	III	E	$\Delta E(\text{II-I})$	$\Delta E(\text{II-III})$
	-64.7		162.8		144.4	(227.5)	(18.4)
	-68.0		139.0		137.3	(207.0)	(1.7)
	-56.7		168.2		129.5	(224.9)	(38.7)
	-51.7		173.0		155.8	(224.7)	(17.2)
	-53.5		173.2		158.3	(226.7)	(14.9)
	-64.1		162.7		144.0	(226.8)	(18.7)
	-44.3		179.8		143.8	(224.1)	(36.0)
	-100.4		124.0		110.6	(224.4)	(13.4)

Table 3

I	E	II	E	III	E	$\Delta\epsilon$ (II-I)	$\Delta\epsilon$ (II-III)
	-101.8		128.9		81.7	(230.7)	(47.2)
	98.4		131.3		111.5	(229.7)	(19.8)
	-69.9		161.5		144.8	(231.4)	(16.7)
	-76.2		160.9		130.0	(237.1)	(30.9)
	-75.2		158.5		142.6	(233.7)	(15.9)
	-61.6		168.6		150.2	(230.2)	(18.4)
$\text{H}_3\text{C}-\text{CH}_2\text{C}\ell$	-25.6		$\text{H}_2\text{C}^+-\text{CH}_2\text{C}\ell$ 214.8		$\text{H}_3\text{C}-\text{CHC}\ell$ 200.0	(189.2)	(14.8)

The generalization introduced into eq. (4) is \underline{F} which is defined as:

\underline{F} is the fraction of molecules reaching the critical site which have been transformed into an active metabolite.

Generally, \underline{F} has been set equal to one, that is explicit metabolic considerations have been neglected. In turn, \underline{K} , which accounts for drug-target interactions is defined in our formalism as:

\underline{K} is the activity of the metabolite, i.e. its reactivity toward the target molecule.

We further note that several metabolites are allowed in the postulated metabolic pathway of Fig. 2. Therefore, eq. (4) becomes:

$$\frac{d(\text{response})}{dt} = AC \sum_i \underline{F}_i \underline{K}_i \quad (5)$$

where \underline{i} refers to the i^{th} metabolite and the constraint $\sum \underline{F}_i = 1$ is present. Eq. (5) can be solved under three different¹biological boundary conditions:

1. Concentration and time are constant,
2. Concentration and response are constant, and
3. Response and time are constant.

The TD_{50} and RC measures fall into the second classification. Consequently, the final form of the drug action equation becomes

$$\frac{1}{t} = A \sum_i \underline{F}_i \underline{K}_i \quad (6)$$

An expression for \underline{K}_i was formulated in terms of a product of structural switches:

$$\underline{K}_i = \prod_u \underline{S}_{ui} \quad , \text{ where} \quad (7)$$

$$\underline{S}_{ui} = (1 - \underline{V}_{ui} X_{ui}) \quad (8)$$

in which

$X_{ui} = 1$ or 0 depending upon the presence or absence of descriptor u in metabolite i .

\underline{V}_{ui} measures the importance of descriptor u in metabolite i in the production of the observed response.

The descriptors employed include those given in Tables 1 and 3. \underline{A} in eq. (6) was separately considered as represented by a gaussian function in $\log P$ multiplied by the population of the boat conformer. The parent compound was used to determine the values of $\log P$ in \underline{A} . Our program MULFIT which is a form of non-

linear regression analysis was used to determine the coefficients for the A term, the V_{ui} and the F_i . The values of the F_i , in turn, provide the information to determine which metabolic species are most important in the action model. The V_{ui} indicate which molecular descriptors are most important in the action of a particular metabolite.

Results

Two distinct types of structure activity relationships were carried out; one independent of the action model, the other based upon the proposed mechanism of action.

1. QSAR-Independent of an Action Model:

The TD50 were initially correlated against the complete sets of group additive molecular features, CFF absolute conformer energies, and equivalent atom residual charge densities using linear regression analysis. Both linear and quadratic values were considered for each descriptor. The systematic and repetitive deletion of a single descriptor term from the general correlation equation indicated that all CFF conformer energies and atomic charge densities, except for those of the C_α , did not contribute meaningfully to the correlation equation.

Additional molecular mechanics descriptors were constructed by taking all combinations of energy differences between the different conformer states considered. Several degeneracies arose because of the energy equivalence of some conformers for certain compounds. Nevertheless some of the sets of energy differences enhanced the correlation when both linear and quadratic values were included in the regression analysis. The descriptor sets of energy differences are highly colinear with respect to correlation with TD50. Consequently, we sought the one set of energy differences which maximized the significance of the correlation. The energy difference between the most stable boat and most stable chair form of each compound yielded the highest correlation. These energy differences, ΔE , are reported in Table 1. It should be noted that these ΔE may, in part at least, correlate best with TD50 because this data set contains the least degeneracy (largest number of unique entries).

When the ΔE are included in the correlation analysis, the atomic charge densities of the C_α no longer are significant. This suggests that the contribution of the C_α charge densities to the correlation are included within the ΔE . This is consistent with the fact that the charge densities are used to determine the CFF energies. If this is the case, then the C_α atomic charge densities do not contribute to the TD50 correlation through chemical reactivity (an intermolecular process), but through conformer stability (an intramolecular process).

All combinations of linear and quadratic values of the remaining descriptors were used in linear regression fits to the TD₅₀. The most significant correlation equation found is,

$$\text{TD}_{50} = 97.7 + 29.7 (F_{\text{Oct}}) + 3.8 (F_{\text{Oct}})^2 - 6.61 (\Delta E) + 1.23 (\Delta E)^2 \quad (9)$$

$$N = 22 \quad R = .92 \quad S = \pm 8.4$$

Eq. (9) has been judged significant because the constant term (97.7) is at the upper end of the TD₅₀ activity range [0,100]. Contributions of F_{Oct} and ΔE from compounds in the data base to eq. (9) must be to lower TD₅₀. This cannot occur randomly as compared to the case where the constant term is 50, and "random fluctuation" contributions from F_{Oct} and ΔE to reproduce the range of TD₅₀, [0,100]. The number of descriptor terms in eq. (9), four, two pairs of interfunctional descriptors, is small with respect to the number of observations, 22. There is a "rule-of-thumb" that the ratio of observations to descriptor terms must be four, or more, to yield a statistically significant relationship. Thus this is additional evidence in support of the statistical validity of eq. (9).

Each of the two descriptors F_{Oct} and ΔE appear in the correlation equation (eq. 9) with parabolic dependence. This means that each descriptor possesses an optimum value which yields a maximum, or minimum, contribution to determining TD₅₀. These optimum values can be determined by simple partial differentiation of eq. (9),

$$\frac{\partial \text{TD}_{50}}{\partial F_{\text{Oct}}} = 29.7 + 7.6 (F_{\text{Oct}}) = 0$$

$$(F_{\text{Oct}})_{\text{opt}} = -3.9 \text{ kcal/mole} \quad (10)$$

Many cyclic nitrosamines possess F_{Oct} values both less than, and greater than, -3.9 kcal/mole (see Table 1). Thus the compound data base used to construct eq. (9) samples a significant range of F_{Oct}-space with respect to TD₅₀.

Eq. (3) yields a (ΔE)_{opt} of

$$\frac{\Delta \text{TD}_{50}}{\partial (\Delta E)} = -6.6 + 2.46 (\Delta E) = 0$$

$$(\Delta E)_{\text{opt}} = 2.69 \text{ kcal/mole} \quad (11)$$

The (ΔE)_{opt} = 2.69 kcal/mole is also located near the middle of (ΔE)-space for the range of (ΔE) values in the compound data set. Thus the parabolic TD₅₀ dependence on (ΔE) is also judged significant.

We did not set aside some compounds to use as "tests" of the QSAR since the size of the data base is small. However, there are

eleven compounds in Table 1 which have $TD_{50} > 100$. These compounds could not be used in constructing eq. (9). However, we can use these compounds to qualitatively test (at least in terms of extending the range of) eq. (9) by predicting TD_{50} . The predicted TD_{50} of the eleven compounds have observed TD_{50} 's > 100 are listed in Table 4. Only two compounds, 2-carboxy-nitrosopyrrolidine and 2,3,5,6 tetramethyldinitrosopiperazine, are predicted to have TD_{50} 's considerably less than 100. The 2-carboxy compound, as well as all the carboxyl-containing compounds, will be predicted, from eq. (9), to have higher TD_{50} 's if the charged form of the carboxyl group is assumed.

The poor prediction of TD_{50} for 2,3,5,6 tetramethyldinitrosopiperazine might be due to the wrong selection of the molecular configuration. If, for example, the all-equatorial configuration were selected, ΔE is estimated to be near 10 which would lead to a predicted TD_{50} of about 110.

We reiterate that we do not have knowledge of the experimental configurational states, or the extent of carboxyl ionization, of the cyclic nitrosoamines. The self-consistency of the QSAR is the only indirect evidence for predicting configurational, conformer, and ionization states of the molecules.

Linear regression analysis has pitfalls. There is always the possibility of chance correlations. Hence, we opted to analyze the data using an alternate statistical method, namely cluster analysis. The data were scaled so that each of the descriptors ranged in value between 0 and 1. Minimal tree spanning methods was employed in the determination of clusters (24).

We did make use of some guidelines established in the linear regression study: Only the four most significant descriptors (F_{Oct} , F_{H_2O} , ΔE , and $\log P$) were used in the cluster analysis. The results are shown in Fig. 3. This is a non-linear map of the descriptors in their four-dimensional space projected into two-dimensions. Nine different clusters can be identified. However, one cluster plus one member of a nearby cluster (the enclosed area within the map) contains all active compounds except for the 3-hydroxy nitroso-piperidine. That is, highly carcinogenic compounds cluster together while the "non-carcinogenic" cyclic nitrosamines are scattered about in the four-dimensional descriptor space.

The values of the four descriptors were also pairwise-correlated against one another. In addition, the single variable (no testing for the effect of combinations) predicting ability was measured by putting the data into categories, based upon their TD_{50} value being less than, or greater than, 75. The Fisher ranking method was then applied to the data base. The results are given in Table 5. Table 5a indicates that F_{H_2O} and $\log P$ are highly correlated (.76) as well as F_{Oct} and F_{H_2O} (.88). Interestingly, however, F_{Oct} and $\log P$ are much less correlated (.35). Table 5b indicates that F_{Oct} is the most

Table 4

Predicted TD_{50} for compounds possessing observed $TD_{50} > 100$ using eq. (9).

Compound	Predicted TD_{50}
<u>Nitrosopiperidines</u>	
2,6 Dimethyl -	105.9
2,2,6,6, Tetramethyl -	99.6
4-t-Butyl -	257.2
2-Carboxyl -	85.7
4-Carboxyl -	98.9
<u>Nitrosopyrrolidines</u>	
2,5 Dimethyl -	145.7
2-Carboxy -	37.3 (120.2) ^a
2-Carboxy - 4-hydroxy	139.2
<u>Nitrosopiperazines</u>	
Nitrosopiperazine	85.0
4-Methyl -	100.0
<u>2,3,5,6 Tetramethyl-</u> <u>dinitrosopiperazine</u>	35.2 ^b

a) the carboxyl group is charged.

b) the low TD_{50} is probably indicative of selecting the wrong configuration(s) and/or conformer states.

Table 5

a) Correlation Table of Log P, F_{H_2O} , F_{oct} , and ΔE .				
	Log P	F_{H_2O}	F_{oct}	ΔE
Log P	1			
F_{H_2O}	.76	1		
F_{oct}	.35	.88	1	
ΔE	.12	.24	.26	1
b) Single Variable Predictive Capacity of Log P, F_{H_2O} , F_{oct} , and ΔE .				
Variable	Unnormalized Predictive Weight			
F_{H_2O}	2.24			
F_{oct}	1.19			
ΔE	.07			
Log P	.009			

important predictive descriptor followed by F_{H_2O} . ΔE and $\log P$, individually, are of very little predictive value. This finding, in conjunction with eq. (9), suggests that F_{oct} in eq. (9) is the dominant predictive descriptor and ΔE only becomes important in those few cases where F_{oct} 's predictive capacity fails.

2. QSAR-Using an Action Model

The set of equations (6-8) employing combinations of molecular descriptors from Table 1 and the descriptors from Table 3 were used to describe the action model. The optimum action model achieved to date for the nitrosamines listed in Table 6 is,

$$\frac{100}{TD_{50}} = A(F_1P_1 + F_2P_2) \text{ where,} \quad (12)$$

100 is an arbitrary scaling constant chosen equal to the maximum time duration of the experiments, and

$$A = (8.34 \exp[0.05(\log P - 1.32)^2])(1 + \exp(-.26 \Delta E))^{-1}$$

The fraction of each metabolite reaching the target site is chosen to be uniform, e.g. $F_1 = (\text{set}) = F_2 = (\text{set}) = .5$. That is, we have assumed a non-selective oxidizing agent (P-450?) in setting equal weights to the two metabolic pathways (see Fig. 2). P_1 and P_2 are the switch variable products defined as;

$$P_1 = \prod_{j=1}^6 (1 - V_{j1}X_{j1}) \text{ and,}$$

$$P_2 = \prod_{j=1}^6 (1 - V_{j1}X_{j2}), \text{ where, lastly,}$$

$$X_{11} \text{ or } X_{12} = 1 \text{ if } C-C^{\oplus} \text{ is formed, else } = 0.$$

$$X_{21} \text{ or } X_{22} = 1 \text{ if } (C)_n-C-C^{\oplus}, \text{ for } n \geq 1, \text{ is formed, else } = 0.$$

$$X_{31} \text{ or } X_{32} = 1 \text{ if } C \begin{array}{l} \diagup \\ \diagdown \end{array} C-C^{\oplus} \text{ is formed, else } = 0.$$

$$X_{41} \text{ or } X_{42} = 1 \text{ if } C-C-C^{\oplus}-C \text{ is formed, else } = 0.$$

$$X_{51} \text{ or } X_{52} = 1 \text{ if } X-C-C^{\oplus} \text{ (X=halogen) is formed, else } = 0.$$

$$X_{61} \text{ or } X_{62} = 1 \text{ if } C=C-C^{\oplus} \text{ is formed, else } = 0.$$

The assignment of the X_{ji} were made from an analysis of the ΔE values given in Table 3 for the model metabolites.

The resulting nonlinear regression fit of the V_{j1} are

$$V_{11} = 0.50, V_{21} = .438, V_{31} = .484, V_{41} = 1.000, V_{51} = .021, V_{61} = -0.007.$$

Figure 4 is a plot of predicted TD_{50} , using eq. (12), versus the observed TD_{50} . The major outlier is 4-keto-nitrosopiperidine. It is not possible to compute a correlation coefficient for this fitting procedure since nonlinear regression methods have been employed.

Table 6

Nitrosamines Used in the Mechanism of
Action Model QSAR.

MOLECULE	TD50 (WEEKS)	EXP.
NITROSO-PIPERIDINE	38.0	
2-METHYL -	80.0	
3-METHYL -	55.0	
4-METHYL -	40.0	
2,6-DIMETHYL	125.0	
3,5-DIMETHYL	100.0	
2,2,6,6-TETRAMETHYL -	125.0	
4-T. BUTYL -	125.0	
3-HYDROXY -	43.0	
4-HYDROXY -	44.0	
4-KETO	45.0	
2-CARBOXY -	125.0	
4-CARBOXY -	125.0	
4-CHLORO -	41.0	
3,4-DICHLORO	20.0	
3,4-DIBROMO	36.0	
1,2,3,6-TETRAHYDROPYRIDINE	28.0	
N-NITROSO-3-PYRROLINE	80.0	
N-NITROSO-HEXAMETHYLENEIMINE	28.0	
N-NITROSO-HEPTAMETHYLENEIMINE	25.0	
DIMETHYL-NITROSAMINE	25.0	
DIETHYL-NITROSAMINE	30.0	
BIS-(2-CHLORO)-	84.0	
BIS (2-CYANO) -	125.0	
BIS (2-METHOXY) -	63.0	

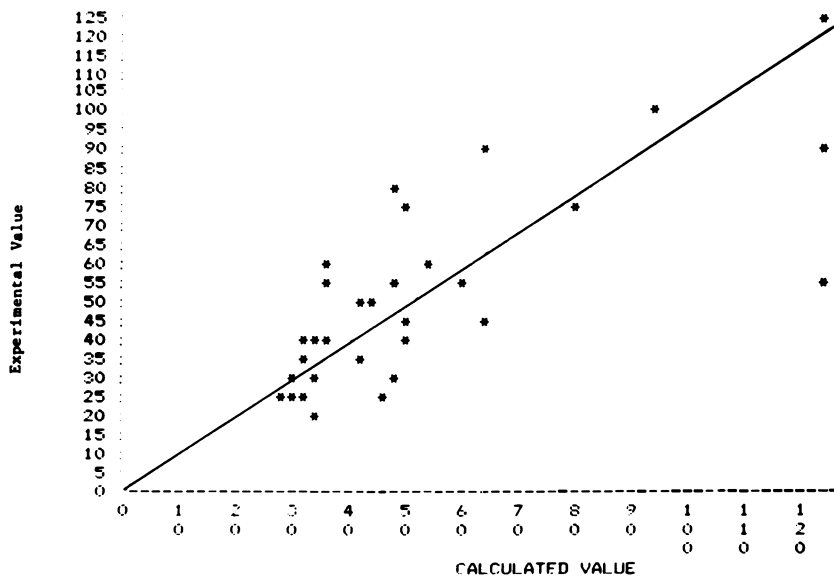


Figure 4. Plot of experimental TD_{50} (in weeks) vs. the calculated TD_{50} using the mechanism of action QSAR. $TD_{50} > 100$ are assigned a value of 125.

Discussion

1. QSAR Independent of Mechanism: The application of cluster analysis is an attractive method to use because relationships can be formulated independent of the biological data. The carcinogenic nitrosamines have been treated in such a manner. The molecular descriptors F_{Oct} , F_{H_2O} , $\log P$, and ΔE were cluster analyzed. The biological data was subsequently assigned to the clustered points in the descriptor space. It was then found that the potent carcinogenic nitrosamines cluster together, while the less carcinogenic compounds are scattered over descriptor space (see Fig. 2).

These findings both complement and support the significance of eq. (9) which is the optimum linear regression equation found using the data of Table 1. F_{Oct} is the dominant descriptor in the analysis, see Table 5b. It is not possible to give a truly reliable biochemical interpretation to this finding. However, one possible explanation is that the carcinogenic potency depends upon the bioaccumulation of the nitrosamine in a particular tissue having a specific nonpolar solubility. F_{Oct} , the free energy of interaction of a compound with a 1-octanol solution, is assumed to be a measure of the nonpolar solubility. $F_{Oct}(opt) = -3.9$, represents the solvation free energy of the "target" tissue. This conceptual model indirectly postulates that transport processes, as modelled through $\log P$, play only a minor role in the mechanism of action.

The second descriptor in eqn. (9), ΔE , is found to be of minimal significance according to the single variable ranking of Table 5b. F_{H_2O} is ranked as the second most significant single variable in a QSAR. However, the high correlation between F_{H_2O} and F_{Oct} , see Table 5a, eliminates this variable in multiple variable linear regression equations involving F_{Oct} .

It is difficult to give a physicochemical meaning to ΔE within the framework of an action mechanism. ΔE is the equilibrium energy difference between the most stable "boat" and "chair" conformer states. However, suppose the production of the ultimate carcinogenic metabolite depends upon two parallel metabolic reactions as depicted in Fig. 5a. The concentration, C_1 , of one metabolite increases with increasing ΔE , while the concentration of the second metabolite, C_2 , decreases with increasing ΔE , Fig. 5b. The concentration, C_3 , of the ultimate carcinogen is dependent upon the product C_1C_2 . In this case the ultimate dependence of C_3 upon ΔE can appear to be parabolic over some range of ΔE , see Fig. 5c. This explanation is not consistent with the action mechanism assumed in our modelling. Nor is there any experimental support for or against a pair of parallel metabolic reactions. However, parallel metabolic reactions can be used to account for the role of ΔE in eq. (9).

2. QSAR Using an Action Model: Eq. (12) indicates that $\log P$ is of only marginal significance in specifying activity. The

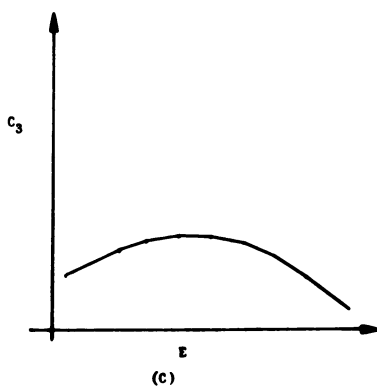
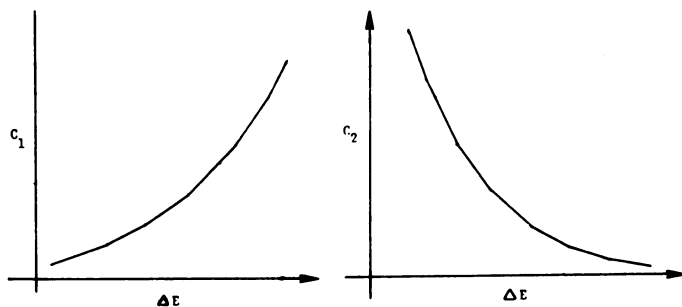
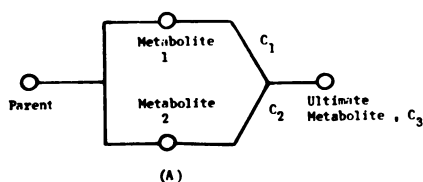
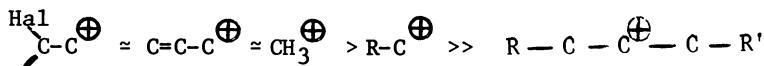


Figure 5. (A) Schematic of intermediate metabolite production leading to the production of the ultimate active metabolite; (B) Metabolite 1 concentration increases with increasing ΔE , opposite is true for Metabolite 2 concentration; (C) production (concentration) of the ultimate active metabolite exhibits parabolic dependence on ΔE .

optimum value for Log P is 1.32. This finding is consistent with the cluster and regression QSAR where Log P was not found to be an important descriptor. Overall, this suggests that transport processes may not be important in the carcinogenic potency of these cyclic nitrosamines. ΔE appears in eq. (12) as a minor descriptor whose increasing value increases carcinogenicity (TD_{50} decreases). We have not been able to explain the role of ΔE in this QSAR. However, ΔE is even less important in eq. (12) than in eq. (9).

The role of the carbonium ion metabolites of the nitrosamines in specifying carcinogenic potency can be estimated from an analysis of V_{ji} . The following conclusions have been made;

1. $V_{41} = 1.0$ implies that secondary carbonium ions are not carcinogenic.
2. $V_{11} \approx V_{21} \approx V_{31} \approx .5$ suggests that the nature of the side-chain of the carbonium ion is not important to specifying carcinogenic potency. Further, the common default option to each of these cases ($X_{1i} = X_{2i} = X_{3i} = 0$, $i = 1, 2$) corresponds to the formation of $-\text{CH}_3^+$. Thus $-\text{CH}_3^+$ is estimated to be twice as carcinogenic as each of the three alternate classes.
3. $V_{51} \approx V_{61} \approx 0$ implies that halogen substitution or formation of a double bond at an adjacent carbon to the carbonium ion enhances carcinogenic potency to about the same level as $-\text{CH}_3^+$ carcinogenic potency.
4. Qualitatively, carcinogenic potency can be ranked as,



3. Summary: The two QSAR's constructed in this current study of carcinogenic cyclic nitrosamines are complementary in that they suggest how to construct a better QSAR. F_{Oct} , formulated in terms of a bioaccumulation functional, along with the carbonium ion metabolite model should be coupled together to form a revised QSAR action model. Work is underway on this model presently in our laboratory.

Acknowledgements

We gratefully acknowledge the financial support of the National Cancer Institute (contract No. N01-CP-76927), the National Science Foundation (grant No. ENV 77-74061) and the National Institutes of Health (contract No. 217041).

References

1. The data base was provided by W. Lijinsky, Chemical Carcinogenesis Program, Frederick Cancer Research Center, Frederick, Md. 21701 (USA).
2. W. Lijinsky and H.W. Taylor, *Cancer Lett.*, 1, 359 (1976).
3. W. Lijinsky and H.W. Taylor, *Cancer Res.*, 35, 1270 (1975).
4. W. Lijinsky and H.W. Taylor, *Cancer Res.*, 36, 1988 (1976).
5. G.M. Singer, H.W. Taylor, and W. Lijinsky, *Chem.-Biol. Interactions*, 19, 133 (1977).
6. J.S. Wishnok, M.C. Archer, A.S. Edelman, and W.M. Rand, *Chem.-Biol. Interactions*, 20, 43 (1978).
7. See, E.J. Ariëns, *Drug Design*, Academic Press, New York Vols. 1-6 (1971-1976).
8. A.J. Hopfinger, *Conformational Properties of Macromolecules*, Academic Press, New (1973).
9. L.B. Kier, *Molecular Orbital Theory in Drug Research*, Academic Press, New York (1971).
10. R. Rekker, *The Hydrophobic Fragment Constant*, (Elsevier, New York), (1977).
11. A. Leo, Table of Common Fragment Constants, Dept. Chemistry, Pomona College, Claremont, CA (1976).
12. J. Hine and P.D. Mookerjee, *J. Org. Chem.*, 40, 292 (1975).
13. C. Hansch, A. Leo, S.H. Unger, K.H. Kim, D. Nikaitami, and E.J. Lien, *J. Med. Chem.*, 16, 1207 (1973).
14. A.J. Hopfinger and R.D. Battershell, *J. Med. Chem.*, 19, 569 (1976).
15. N.L. Allinger, J.T. Sprague and T. Kiljefors, *J. Am. Chem.* 96, 5100 (1974).
16. R. Potenzzone, Jr., E. Cavicchi, H.J.R. Weintraub, and A.J. Hopfinger, *Computers and Chem.*, 1, 187 (1977).
17. I.D. Blackburne, R.P. Duke, R.A.Y. Jones, A.R. Katritsky, and K.A.F. Record, *J. Chem. Soc. Perkins II*, 2, 332 (1972).
18. O. Gropen and P.N. Skancke, *Acta, Chem. Scand.*, 25, 1241 (1971).
19. a. R.K. Harris and R.A. Spragg, *J. Mol. Spect.* 23, 158 (1967).
b. Y.L. Chow and C.J. Colon, *Canad. J. Chem.*, 46, 2827 (1968).
c. R.K. Harris and R.A. Spragg, *J. Mol. Spect.*, 30, 77 (1969).
20. W. Lijinsky and H.W. Taylor, *Int. J. Cancer*, 16, 318 (1975).
21. J.A. Pople and D.C. Beveridge, "Approximate Molecular Orbital Theory" (McGraw-Hill, New York) (1970).
22. a. R.C. Bingham, M.J.S. Dewar, and D.H. Lo, *J. Am. Chem. Soc.*, 97, 1285 (1975).
b. R.C. Bingham, M.J.S. Dewar, and D.H. Lo, *J. Am. Chem. Soc.*, 97, 1302 (1975).
23. See, M.S. Tute, in *Advances in Drug Research* (N.J. Harper and A.B. Simmons, eds.) Academic Press, New York, Vol. 6, p. 1, (1971).
24. T.T. Young and T.W. Calvert, *Classification, Estimation, and Pattern Recognition*, Elsevier, New York (1974).

RECEIVED June 8, 1979.

Computer-Aided Selection of Novel Antitumor Drugs for Animal Screening

LOUIS HODES

National Cancer Institute, 8300 Colesville Road, Silver Spring, MD 20910

An earlier paper (1) described a method for estimating the probability of activity of compounds. In that report the method was applied to a small set of data for the purpose of comparing the performance to that of more sophisticated pattern recognition methods. The intended application was the large volume of data from the antitumor screening program of the National Cancer Institute (NCI).

Early in 1976 a new panel of animal models for screening was adopted. These include mouse colon, breast, and lung tumors; and corresponding human tumor xenografts in athymic mice. Most compounds are required to show activity in an in vivo pre-screen in order to receive the more extensive testing. P388 mouse lymphocytic leukemia was chosen as the pre-screen and, thus, P388 became the logical choice for a trial of the statistical-heuristic method. There were already substantial amounts of P388 data from earlier tests, and roughly 15,000 compounds a year are currently being screened in P388.

From this data two trials of the method were designed to approximate its use in the operating environment. These are described later as experiments 1 and 2. A series of enhancements followed. Some are reported on here, while others are still being worked on.

There are two aspects of this method in the selection of compounds for antitumor screening. The first can be considered an enrichment in the yield of actives. For example, suppose we can achieve an enrichment factor of 1.2 by taking arbitrarily just those compounds scoring in the top 50%. In that case, instead of screening 15,000 compounds, NCI can obtain the same number of P388 actives by screening 12,500 compounds which were derived from an initial evaluation of 25,000 compounds by this computer method.

Such a model is envisioned as part of a preselection module attached at the beginning of the Drug Research and Development automated chemical information system. This module will evaluate chemical structures prior to acquisition. It will include the currently performed search for duplicates, as well as a comparison

This chapter not subject to U.S. copyright.
Published 1979 American Chemical Society

with an extensive list of skeleton structures to detect analogs. A comprehensive report would then be examined by the chemist responsible for acquisition, who will decide whether to acquire the compound and, if so, whether special consideration or testing is called for.

The second aspect of the method is selection by surveillance. This would involve running even larger numbers of compounds through the model, perhaps all new Chemical Abstracts registrations. Under these conditions, only the high scoring compounds would be reviewed manually. In contrast, under the first aspect, even the lowest scoring can be reviewed.

The main requirement under both aspects is the ability to detect new active classes of compounds. Therefore, the first enhancement of the method was the elimination of familiar drugs from the training set, following the original version of the method (2). It will be seen that the resulting model gives higher scores to compounds with relatively unfamiliar structures which have appeared in active P388 tests. These compounds are more in demand for further development than are analogs of known drugs.

Also, certain other properties of the method point it toward novel structures. Since each feature is treated independently, new combinations of features which have not appeared together before are more likely to be selected in this method than in other more elaborate methods. Moreover, as will be seen later, a modification of the method gives more weight to lower incidence features. No feature is discarded because of low incidence alone. In addition, a separate program prints out, for each candidate compound, its lowest incidence feature in P388 testing as well as all features not yet having occurred in P388 testing.

It must be emphasized that although this method runs on a computer, it is not designed to automatically pass or reject compounds. Rather, it is proposed as a tool to aid the medicinal chemist in selecting compounds. Although its assumption of feature independence is a strong limitation, the unbiased use of much data should make the scores useful.

Review of the Method.

The method was described previously (1) so only a brief review will be attempted here. Some modifications are discussed in later sections of this paper.

A model is based on a collection of compounds of known activity, called the training set. A set of features pertaining to these compounds are produced; we use molecular substructure descriptors. These features must, of course, be generally relevant to activity for the method to work.

An activity weight is assigned to each feature independently. This weight is based on circumstantial evidence, that is, on how often the feature occurs in the active compounds of the training set relative to how often it is expected to occur assuming that

the feature is not relevant to activity. In computing the expected incidence, we use as reference the incidence of the feature in the entire Drug Research and Development collection. The weight is expressed as a number of standard deviations using the statistical distribution derived from considering the active compounds as a random selection from the file. This procedure compensates for the wide range of incidences among structure features.

The features currently are those routinely generated as keys for the substructure inquiry system. (3) This system incorporates an open-ended feature set as opposed to a dictionary. The main types of keys are the augmented atom (AA), the ganglia augmented atom (GAA) the ring key and two kinds of nucleus key, in addition to individual element keys. The number of features appearing in the P388 data are roughly 2000 without the GAA keys increasing to roughly 8000 when the GAA keys are substituted for the AA keys. Other features being considered are keys used in the BASIC search system (4) and an algorithmic approximation to functional groups. The different classes of features will be covered later.

An activity score is produced for a compound of presumably unknown activity by adding the weights of the features it presents. The score is not intended to estimate the strength of activity, but only some measure of the likelihood that the compound is active.

Experiments with a Large Data Set.

Several experiments were performed to establish the feasibility of this method on large amounts of data. These experiments required a few searches of the Drug Research and Development biological file to find compounds that met various criteria as a result of P388 testing. These criteria are usually stated in terms of the T/C ratio, which is the median life-span of the treated animals divided by that of the controls.

The output of each biological search is a list of NSC numbers. These numbers are assigned successively to each compound upon registration of that compound in the chemical information system (CIS). This occurs upon acceptance of the compound for screening. NSC number 1 was assigned in 1955 and current (1979) NSC numbers are over 300,000. The results of the biological search are then sent to the CIS for extraction of the structure keys to be used as features.

This section describes the experiments that were performed before the well known classes were removed. They showed that the large number of compounds and features could easily be handled. It will be seen later that removing the classes of compounds led to some basic changes in the model. Also, it will be interesting to compare the results of the modified model.

Preliminary Experiments. In November, 1975 a search of the biological files produced a list of active and a list of inactive

compounds with respect to P388 according to the following criteria. The actives were picked from a file of roughly 6000 selected agents and amounted to 489 compounds with a manually assigned code indicating significant activity in P388 (this was based on a confirmed T/C of 175% or greater.) The inactive compounds, numbering 4174, were collected from two sources. (1) Compounds from the selected agents file which had been assigned a manual code signifying negative in P388. (2) Compounds within the latest 100,000 NSC number range which had definitive P388 testing with T/C always less than 175%. Definitive testing requires a regimen of at least three dose levels with the bottom two non-toxic. Toxicity is defined as a T/C less than 85% or less than 2/3 survivors on initial toxicity day. An additional weight loss criterion was not used.

The preliminary experiments used these compounds mostly as training sets. At first, every tenth compound in the NSC sequence beginning with the first was selected to be included in a test set, leaving the remaining 90% as the training set. A second cut consisted of every tenth compound starting with the second. The actives and inactives were cut separately to equalize their relative incidence in the test and training sets. Five such cuts were run.

In summary, these five cuts contained 236 active compounds. (Of the 489 active compounds, several were disqualified because they lacked chemical structure keys, so that 236 are about half of those remaining.) Only 5 of the 236 compounds ranked in the lower 50% of the activity scores in their respective cuts. This remarkable performance is due in large measure to heavily populated classes of similar compounds occurring in the file of selected agents. Nevertheless, these results were considered sufficiently encouraging to try data closer to current input.

This trial also revealed properties of the method as applied to the P388 prescreen that warrant further attention. First, the inactivity score was not useful, so that the results were equally good when it was ignored. Second, conspicuous in the small number of low scoring P388 active compounds were 5-fluorouracil, platinum and ellipticines, since they are of great interest in the treatment of human cancer.

Also, these failures represented the three main categories of failure. The platinum scores increased upon better data as described soon. More powerful features are required for compounds like 5-fluorouracil, whereas ellipticines are more intractable since there are many inactives with the same basic structure. All examples of these three compounds were later removed as analogs of well known classes.

Test Set 1. A more realistic trial was the selection, as a test set, of compounds as they had been input into the program. The 5,000 compounds from NSC number 260,001 through 265,000 were selected as having a fair amount of P388 testing at the time. (October, 1976.) These were categorized by manual review of the

screening data summaries as follows according to their activity in P388.

'A'--T/C at least 175% in two separate screening tests.

'C'--T/C at least 120% in two separate screening tests but not in category A. Two tests in P388 with T/C of 120% or greater is sufficient to be considered a candidate for testing in the tumor panel.

'D'--T/C at least 120% in one screening test and not yet retested.

'E'--T/C at least 120% in one screening test and T/C less than 120% in a separate screening test and not in category C or A.

'N'--T/C less than 120% in all screening tests. Must have a test with at least three dose levels, the bottom two non-toxic.

'M'--insufficient testing in P388.

'T'--not tested at all in P388.

The training set. There were reasons that performance on the C's could be improved by beginning again with a new training set. First, there were a few false negative platinum compounds, and it was noticed that the platinum key was not weighted highly simply because of a time lag in assigning compounds to the selected agents file with the appropriate manual code. Second, it was felt that a training set containing C's as well as A's should perform better on incoming selections which would yield T/C between 120% and 175%. Third, a search of the entire file rather than the most recent 100,000 compounds, and a T/C cutoff of 120% might make the inactive training set more useful.

Thus, a more comprehensive training set was collected from the biological file. The search took place early in 1977 and covered NSC numbers 1 through 260,000. NSC number 260,000 corresponds chronologically to near the end of 1975, so biological testing should have been fairly complete.

Five lists of NSC numbers were collected according to degree of activity in P388. The search of the biological file was specified according to the categories A, C, D, E and N established earlier. From the truncated file of 260,000 compounds, category A yielded 880 compounds, category C yielded 1916 compounds, category D yielded 1402 compounds, category E yielded 1787 compounds, category N yielded 15524 compounds, for a training set total of 21509 compounds in P388.

Experiment 1. The test set for this experiment consisted of the 5000 compounds from NSC number 260001 to 265000, of which 2322 satisfied the biological testing and chemical structure data requirements. Various combinations of weights and scores derived from the sets of A's, C's and N's of the training set were tried with the following results.

First, the 880 A's alone were used as the training set and the 2322 compounds were scored and ranked, As shown in Table I, all 26 A's scored in the upper half but there were 11 C's out of 85 in the lower half.

Table I. Ranking of Actives in Test Set 1. Training set A.

RANK (Deciles)	NUMBER OF A COMPOUNDS	CUM. PERCENT	NUMBER OF C COMPOUNDS	CUM. PERCENT
10	18	69	41	48
9	5	88	19	71
8	1	92	5	76
7	2	100	3	80
6	0		6	87
5	0		5	93
4	0		1	94
3	0		4	99
2	0		1	100
1	0		0	

A new training set was formed by combining the A's and C's so that there were now $880 + 1916 = 2796$ compounds in the training set. Actually, each of the A compounds was counted as appearing twice in the training set, so the weightings resulted in what may be considered a 2A+C model. Now, 1 A out of 26 in the test set and 7 C's of 85 fell into the lower half. See Table II. Again, the weights derived from the rather large set of inactives did not help the performance.

Table II. Ranking of Actives in Test Set 1. Training set 2A+C.

RANK (Deciles)	NUMBER OF A COMPOUNDS	CUM. PERCENT	NUMBER OF C COMPOUNDS	CUM. PERCENT
10	15	58	41	48
9	8	88	20	72
8	0		8	81
7	2	92	5	87
6	0		4	92
5	0		2	94
4	1	100	1	95
3	0		2	98
2	0		2	100
1	0		0	

Expressed as an enrichment of the yield of actives, these results show that selecting the upper half under the 2A+C model would have changed the yield of actives from 111 out of 2322 or 4.8% to 103 out of 1161 or 8.9%

Experiment 2. An attempt was now made to select a sequence

of compounds which seemed relatively free of analogs. The new test set chosen consisted of compounds with NSC numbers 268001 through 272000. This test set, like the earlier one, was rated into biological categories A,C,D,E,N,M,T by manual review of the screening data summaries.

Excluding M's and T's and compounds with no chemical structure keys, there were 3239 compounds left totally, of which 32 were A's and 145 were C's. Upon running the same 2A+C model, all 32 A's scored in the upper half and 45 out of 145 C's scored in the lower half. The results are summarized in Table III. Expressed as an enrichment, the yield went from 177 out of 3239, or 5.5%, to 132 out of 1620 for the upper half or 8.1%. This time all the A's scored in the upper 50% with respect to the 2A+C model.

Table III. Ranking of Actives in Test Set 2. Training set 2A+C.

RANK (Deciles)	NUMBER OF A COMPOUNDS	CUM. PERCENT	NUMBER OF C COMPOUNDS	CUM. PERCENT
10	27	84	41	28
9	5	100	24	45
8	0		15	55
7	0		13	64
6	0		7	69
5	0		10	76
4	0		10	83
3	0		11	90
2	0		7	95
1	0		9	100

Removal of Analogs.

At this point the model could be considered extremely successful from the point of view of enrichment of the yield of actives. But, its performance depends to a large extent on the continued testing of analogs of well known active compounds. The main purpose of P388 testing and also of the model remains the discovery of new active classes. The proficiency of the model toward this end must be enhanced and its effectiveness measured.

Analogs of the well known active compounds, which will now be referred to simply as analogs, make up a large part of the training set. More than 85% of the highly active A's and over one third of the C's belong to these classes. Therefore, any new training set formed without them should produce substantially different results.

It is clear that the new model will not perform as well on

any further analogs of those removed. Many of the previously most highly weighted features will now be decreased in weight or else be completely absent from the model. Hence, the new model will tend to show decreased predictability on a test set to the extent that it contains analogs. Recall that a provision for detecting analogs is included in the selection process.

On the other hand, the presence of large numbers of analogs in the training set can deteriorate performance of the model on those compounds which may provide new leads. This occurs when a known active group such as an alkylating function is combined with otherwise inactive moieties, producing an active compound. The weights of all its keys is raised. When such keys occur in inactive compounds they can raise their score sufficiently to yield many false positives. False positives, of course, will lower the relative rank of the true positives which score below them.

Removal of analogs from the training set was first performed on the highly active A's by Ken Paull, then with Starks C. P., the acquisitions contractor. He presented Table IV, which classifies analogs among the 846 A's with defined structure. Familiar alkylating agents, which include mustards, methanesulfonates, epoxides, aziridines and nitrosoureas, make up 52% of the 846. Other, more specialized classes reduced the A's further until they were cut by 87%.

Table IV. Active Classes in 'A' Training Set.
846 Structures with P388 T/C % \geq 175%

CLASS	NUMBER	CUM. PERCENT
Alkylating Agents	440	52%
Anthracyclines	50	
Wander	50	64%
Antifols	32	
Actinomycins	23	70%
Nucleocides	22	
9-Aminoacridine	20	
Platinum	19	
Planar Quaternary Polycyclics	18	
Triyl cysteines	16	82%
Quinoliniums	14	
hiosemicarbizones	9	
Hydrazonoyl halides	9	
Thioxanthenones	8	
Diazo compounds	7	87%

The reduction on the C's was less drastic, the number of compounds going from 1916 to 1182. Several additional classes were removed as shown in Table V. The selection was performed

by Ken Paull in coordination with Mike Hazard and Bob Ing of NCI.

Table V. Additional Active Classes Removed
from Training Sets.

Bruceantin	Podophyllotoxin
Halacanthane	Phenazine
Colchicine	Cycloleucine
Antimony	Mithramycin
Hydroxyurea	Charged Nitrogen
Phosphonium	Sulfonium
Triazeno	Ellipticine
Emetine	Vincristine
Methyl GAG	Styryl Quinoline

Of course, many compounds in most of these active classes tested negative in P388. A great many others were not tested at all in P388.

Ideally, one would desire a new file with all these compounds eliminated. Recall that the method uses the incidence of a feature in the entire file as a reference. The analogs in the file would distort the statistics for the same reasons given earlier. However, it was impractical to search all those classes over the entire file. Machine searches would consume too much computer time.

It was practical to remove the analogs from the set of P388 inactives. Machine searches for all the classes were performed by Mike Hazard and Frank Sordyl, reducing the inactives from 15524 to 14357 compounds.

The method must now be revised so that the training set rather than the entire file provides the reference for the expected incidence of a feature among the actives. Thus, a feature would have positive weight if its proportional incidence were greater in the actives than in the inactives.

Correction of the Method.

A correction to a more precise model became more imperative when the removal of analogs induced the universe of compounds to be limited to the training set, as depicted in the preceding section.

The method as described in (1) assumes, under the null hypothesis, that the active compounds are a random selection of compounds from the file. The probability of feature incidence in the actives was said to follow the binomial distribution.

Robert Tarone of NCI advised me as follows. 1) The binomial distribution holds only when the random selection is done with replacement. 2) A more precise model should assume random selection without replacement. 3) Feature incidence in the

actives then follows the hypergeometric distribution. 4) The standard deviation, S , is now given by

$$S^2 = (F(T-F)N(T-N)) / (T^2(T-1)).$$

Here F is the number of compounds with the feature and T is the total number of compounds, now the number of compounds in the training set. N is the number of actives, so now $T-N$ is the number of inactives.

Under these conditions there is effectively one weight for each feature, since the active and inactive weights become equal in magnitude and opposite in sign. The same holds true for the scores for each predicted compound. Thus, when the training set is the universe of compounds, there is only one activity score. The inactivity score which had not proved useful anyway, being redundant, can be ignored.

If N is much less than T then the hypergeometric distribution becomes indistinguishable from the binomial distribution where $\mathcal{L} = P(1-P)N$ and P is simply F/T . This condition certainly holds when the entire file is used as a reference. It is even a good approximation when only the training set is used since the actives are much less numerous than the inactives.

However, the correction was necessary upon consideration of weights for multiple occurrences of features. Here, for example, when computing the weight for two or more occurrences of a feature the above formula for the standard deviation becomes

$$S^2 = (F_2(F-F_2)N_1(F-N_1)) / (F^2(F-1)).$$

Here F is, as before, the number of compounds with one or more occurrences of the feature, F_2 is the number of compounds with two or more occurrences and N_1 is the number of actives with the feature.

Now, if the feature happens to be more prominent in the actives than in the inactives, then N_1 can be significant compared with F . Thus, in some cases the corrected version will show a significantly different weight. These differences are much more likely when the training set becomes the reference universe.

Modification of the Method.

The removal of analogs accentuated a weakness of the method which allowed some highly prevalent features to receive weights of much larger magnitude than seems reasonable. For example, keys like benzene and six-carbon sugars showed incidences in the actives that were between five and ten standard deviations away from their expected incidences. Before the removal of analogs, these numbers were dominated by the higher weighted analog keys such as aziridine which ranged into the twenties and higher. In the new model the presumably innocuous keys exerted a dominant

influence and became more obviously troublesome. It was necessary to modify the method so that such features are automatically deemphasized.

The anomalous effect is due to the fact that a relatively minor deviation from the expected incidence in the actives of a high incidence feature has a large statistical significance. For example, a feature with 60% incidence, which is expected to occur in 1000 out of 1667 actives has standard deviation approximately 20. Thus, the chance that it occurs in more than 1100 or less than 900 actives is five standard deviations away. The model is based on random selection, but of course the selection of compounds is not really random.

The large deviations are due to selection parameters which may or may not be relevant to activity, but which tend to give high incidence features increased effect.

A heuristic for reducing the magnitude of these weights is simply the division of the computed weight of each feature by the logarithm of its incidence. We use the incidence in the training set of about 15000 compounds, after the removal of analogs. To illustrate, we can divide the weight by the logarithm to the base ten of the incidence for every feature with incidence greater than ten, so that no weight is increased. Recall that $\log 10 = 1$. So, for example, a feature that occurs 100 times will have its weight divided by two; if it occurs 1000 times its weight is divided by 3; etc.

This modification of the weights has the serendipitous effect of emphasizing low incidence features, which helps point the model toward the discovery of new classes.

Results without Analogs.

A new model consisted of weights generated from the training set without analogs. This model required the correction described earlier. Results are to be compared with those from the old model including analogs. A further model incorporated the modification based on incidence.

Since the ratio of C's to A's after the removal of analogs increased to more than ten to one, a 5A+C model, presenting each A five times, replaced the 2A+C model used earlier. The five to one ratio was a heuristic compromise; something like ten to one giving too much influence to individual compounds.

Results on the Test Set. Results on test set 2, which contained fewer analogs, are more comparable to the earlier results. The new results are shown in Table VI. Compared with Table III, performance on the A's deteriorates somewhat, but results on the C's are about the same. Overall, the top 50% contained 70% of the actives, compared with 75% earlier.

Table VI. Ranking of Actives in Test Set 2. Training set 5A+C.

RANK (Deciles)	NUMBER OF A COMPOUNDS	CUM. PERCENT	NUMBER OF C COMPOUNDS	CUM. PERCENT
10	19	59	40	28
9	3	68	15	40
8	2	75	11	46
7	1	78	21	60
6	1	81	12	68
5	0		13	77
4	1	84	10	84
3	3	94	9	90
2	2	100	11	98
1	0		3	100

The results after the incidence modification are shown in Table VII. Performance on the A compounds is effectively the same. Performance on the C's is improved by the incidence modification, but not significantly.

Table VII. Ranking of Actives in Test Set 2. Training set 5A+C. After Incidence Modification.

RANK (Deciles)	NUMBER OF A COMPOUNDS	CUM. PERCENT	NUMBER OF C COMPOUNDS	CUM. PERCENT
10	18	56	42	29
9	4	69	17	41
8	2	75	17	52
7	2	81	16	63
6	0		12	72
5	0		9	78
4	1	84	15	88
3	3	94	4	91
2	2	100	10	98
1	0		3	100

As expected, the removal of analogs lowers the overall performance of the model. However, it is also possible to get a better indication of its performance on novel compounds as follows.

The actives (A+C) in the test sets were examined to flag analogs according to the criteria used in removing them from the training set. There were only twenty active non-analogs in the 2322 compounds of test set 1 and 72 in the 3239 compounds of test

set 2. Further, these compounds tend to come in small clusters of similar structures so the diversity is quite limited, especially on test set 1. The results on the 68 non-analogs that rated C in test set 2 are shown in Table VIII for three versions of the model. The two versions of the 5A+C-analogs model, before and after the incidence modification, did not show much difference, but the difference from the 2A+C model may be significant.

Table VIII. Ranking of Non-Analogs in Test Set 2.

RANK (Deciles)	2A+C MODEL		5A+C MODEL		WITH INCID. MOD.	
	NO.	CUM.%	NO.	CUM.%	NO.	CUM.%
10	11	16	18	26	20	29
9	10	31	9	40	8	41
8	8	43	6	49	6	50
7	7	53	6	57	6	59
6	5	60	5	65	4	65
5	4	66	7	75	6	74
4	8	78	5	82	7	84
3	6	87	3	87	2	87
2	3	91	7	97	7	97
1	6	100	2	100	2	100

The non-analogs do not do quite as well as the total C's in their respective models. At least in this test set, the analogs tend to score fairly high, even when they have been removed from the training set.

The Training Set as a Test Set. It has just been shown that the results on the test set do not provide a completely satisfactory indication of any effect of removing analogs from the training set. However, the training set minus analogs does provide a diverse body of novel compounds to compare both models. Since these compounds were used in the construction of the models, the performance should be better than on a new test set; but since they were used in both models one should obtain a direct indication of any improvement.

The training set of A's, C's and N's minus analogs was combined and run through the same three versions of the model as applied to test set 2. The graph in Figure 1 shows the performance on all the actives, as they rank among the N's. The A's and the C's are graphed separately. Recall that the C's are more than ten times as numerous as the A's. Again, the 5A+C model was run before and after the modification for incidence.

Figure 1 shows a definite improvement on the novel compounds upon the removal of analogs. Although there are two basic differences in the construction of the models, i. e., the A/C ratio, and the reference incidence, the results are comparable

Figure 1. Cumulative percent active for the training set without analogs vs. decile. Percentages are plotted at the lower end of each decile and joined linearly. (---) A's; (—) C's; (○) original 2A + C model; (□) 5A + C-analogs; and (△) 5A + C-analogs with incidence modification.

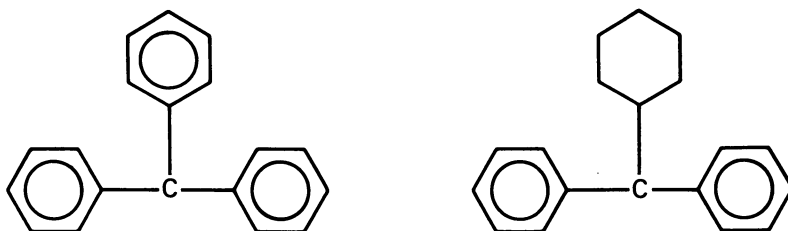
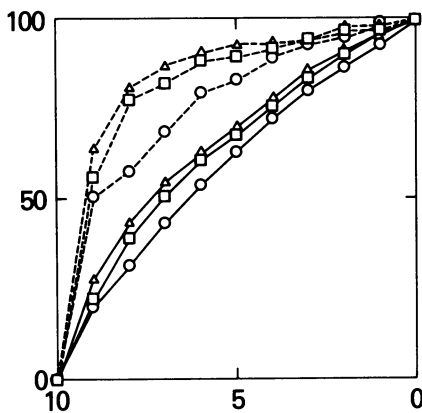


Figure 2. The branching carbon atom yields exactly the same AA keys as the two structures shown, but different GAA keys (AA keys cannot distinguish carbon central atom in these structures; GAA keys can)

because each model was suited to the conditions under which it was constructed. What is most surprising is the effect of the incidence modification. By this change, the features whose weights are diminished are those most prevalent in the training set, so one would think these features would have induced greater separation at their original higher weight, at least on the training set.

In order to achieve the improvement shown in Figure 1, this effect must be more than balanced by the effect due to relatively increased weight of low incidence features. Sufficiently novel compounds will often have one or more low incidence feature. It was gratifying to see that this influenced enough compounds to outweigh the loss due to diminished weight of the high incidence features even on the training set.

The GAA Model.

Thus far the features used were as described in (1). These are augmented atom (AA) keys and certain kinds of ring keys. The AA keys consist of a central atom, its bonds, and the neighboring atoms attached to the bonds. All combinations of attachments are permitted.

There was also available an alternate set of keys for searching which consists of features slightly larger than the AA keys. These were called the ganglia augmented atom (GAA) keys. They include the AA key and all the bonds attached to all the atoms. The GAA keys and not the AA keys are capable of distinguishing the central carbon atom in Figures 2a and 2b.

Examination of some of the failures in the preceding models indicated that the GAA keys would improve performance, although there were other examples which were less tractable. The easy availability of these features became the determining factor in deciding to use them. To avoid questions of redundancy, the GAA keys were used in place of the AA keys.

The results on test set 2 shown in Table IX should be compared to Tables VII and VIII. There appears to be a significant improvement, especially in the central column which lists the total C rated compounds.

Table IX. Ranking of Actives in Test Set 2.
GAA Key 5A+C Model with Incidence Modification.

RANK (Deciles)	A COMPOUNDS		C COMPOUNDS		C NON-ANALOGS	
	NO.	CUM.%	NO.	CUM.%	NO.	CUM.%
10	19	58	60	40	20	29
9	1	61	25	57	14	50
8	1	64	11	64	4	56
7	3	73	9	70	3	60
6	2	79	13	79	6	69
5	0		11	86	8	81
4	4	91	7	91	5	88
3	2	97	5	94	2	91
2	1	100	4	97	4	97
1	0		5	100	2	100

The results on the training set itself are compared with the AA models in Figure 3. The GAA model is 5A+C-analogs with incidence modification, as is the best AA model. On the training set, the main improvement lies in performance on the A's.

The Rarest Key Report.

Since the emphasis of the screening program remains the testing of novel compounds, a crude approximation to a measure of novelty was attempted. A program was written to simply print out, for each compound its key which has least incidence in P388 testing and also any new keys, which have not yet appeared in P388.

The rarest key also shows significant differences in switching from AA keys to GAA keys as seen from the example in Figure 4. The rarest AA key has occurred in 107 compounds which have been tested in P388, but the rarest GAA key, with a different central atom, has occurred only once.

Work in Progress.

Operational Model. During the past year an operational model has been installed at Chemical Abstract Service, where the NCI Chemical Information System is maintained. This model will yield reports on activity score and rarest key beginning April 1, 1979.

The installed model is based on an extension of the training set without analogs as described earlier. A search of the biological file, NSC numbers 260001 and above, was performed May, 1978, followed by a machine search to remove analogs similar to the search performed on the earlier training set. The augmented training set has the A's increased from 83 to 115, the C's from

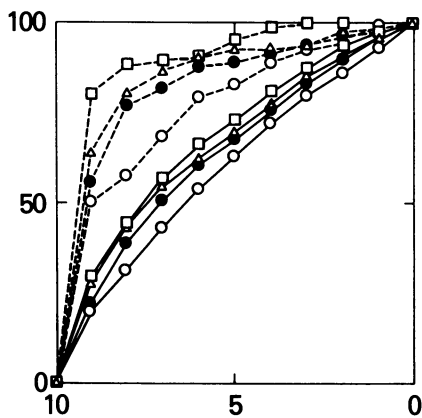
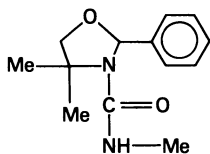
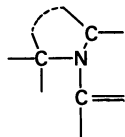


Figure 3. The performance of the GAA model on the same training set is shown in (□) together with the graphs of Figure 1. Note that the 5A + C-analogs model is now represented by (●) instead of (□).

IS THIS COMPOUND NOVEL?



NO — AA KEY 107 OCCURRENCES



YES — GAA KEY 1 OCCURRENCE

Figure 4. The rarest key often turns out to be quite different in the AA key set from that of the GAA key set

1182 to 1950 and the N's from 14357 to 33828.

The model was run first with AA, then with GAA keys. At this time a new version of the GAA key model was generated; here keys which had ganglia on the central atom were eliminated. Thus only the largest GAA keys were kept. This amounts to a decrease in redundancy which is easier to perform on the GAA than the AA keys.

Performance was again compared by running the extended training set through the respective models. The model using the largest GAA key was decidedly the best, with about 45% of the actives in the top 10%. A more appropriate indication of performance would involve setting aside ten or twenty percent cuts as test sets. This is being worked on now and will be reported soon.

The Test of 1000. An assortment of 1000 potential acquisitions was provided by Starks C. P., the acquisitions contractor as a test of the model. These compounds were first reviewed by Ken Paull and were assigned on the basis of their chemical structure only to one of the following three categories: active, novel, or inactive. Then they were run through the models at CAS. In addition, the entire 1000 have been put into P388 testing.

This experiment serves several purposes. It was undertaken mainly to familiarize Starks C. P. with the model. Also, it shows how unselected compounds rank against the NCI file. If one uses the scale of the training set, then 23-30 of the 1000 fall in the highest decile and 350-500 in the lowest decile, depending on the version of the model.

Another purpose of the test of 1000 is to compare the relative performance of an experienced chemist with the performance of the model, not to see which is better (undoubtedly the chemist) but to see if there are some things the model will catch which are not obvious to the chemist. In fact, review by the chemist of the 30 compounds which scored in the top decile, led him to revise his opinion of activity from originally 15 to almost all, i. e., 28 as possibly active. A second chemist, Frank Williams, performed the review.

This experiment should be completed in about a year, when the P388 testing is finished. It will be reported at that time.

Other Structure Features. In this type of work, more specific features tend to yield better results, as is evidenced by the use of the GAA keys.

With permission from the Basel Information Center for Chemistry (BASIC), we have been experimenting with their keys on our data at CAS. The BASIC keys include linear sequences, which are chains of atoms of length four to six with bonds specified only as to their ring or non-ring character.

The BASIC has for some time been performing substructure search on the entire CAS file. Thus, BASIC keys are routinely generated in large volume for compounds newly registered at CAS and can be used for surveillance at low cost.

Performance on test set 2 is shown in Table X. The A's and C's are combined in these statistics for a comparison of performance on all the actives. So far the BASIC keys are not as discriminating as the NCI keys. Improvement was expected by the use of conditional probabilities among linear sequences to cut down redundancy, but early results show no such improvement. This is being worked on by Paul Blower of CAS, who is also investigating the use of co-occurrences.

Table X. Cumulative Percent Active in Test Set 2.
BASIC and NCI models.

DECILE	BASIC	NCI-AA	NCI-GAA
10	30.5	33.9	43.5
9	40.1	45.8	58.7

Another set of structure features is given by an algorithmic definition of a functional group. This is essentially any connected subset of atoms which does not contain a carbon-carbon single bond or a carbon-carbon ring alternating bond.

A program for generating such groups has been presented to the author by Paul Blower and this will be worked on soon. An intriguing combination would be linear sequences with internal carbon atoms and functional groups.

Summary.

Current animal screening methods at the National Cancer Institute include the use of a standard pre-screen (P388) to test roughly 15,000 new compounds a year. Results from P388 testing have been used to create an experimental structure-activity model to aid in selecting compounds likely to be active in P388. In order to better detect new leads, well-known active classes of compounds have been eliminated from the training data. There were still approximately 1300 active compounds and 14,000 inactives so that a statistical model for prediction of activity was the method of choice.

The predictors are molecular structure features previously used in searching the chemical structure file. The method assumes independence of features, so new combinations can be easily detected. Emphasis on low incidence features also helps point toward novel compounds. In this vein, for each compound the feature occurring least often and all features not yet in P388 testing are flagged.

At this time there is some general agreement regarding the usefulness of the methods. The results on test data are quite encouraging, especially with the GAA model. The methods are being put into operational use so more concrete results are expected.

Meanwhile, improvements are planned for the models.

Acknowledgement.

Besides the people mentioned in the paper I owe thanks to Sid Richman and Ruth Geran and several others at NCI and CAS.

Literature Cited.

1. Hodes, Louis et al, J. Med. Chem., (1977), 20, 469.
2. Cramer, Richard D. et al, J. Med. Chem. (1974), 17, 533.
3. Richman, Sidney et al, in Retrieval of Medicinal Chemical Information, ACS Symposium Series, (1978).
4. Schenk, H.R. and Wegmuller, F., J. Chem. Inf. Comput. Sci., (1976), 16, 153.

RECEIVED June 8, 1979.

INDEX

A	
AA keys (augmented atom)	585
<i>Ab initio</i> (AI)	
calculation of acetylcholine (ACh)	166
drug design	150, 445
MODPOT/VRDDO calculations	
on DNA bases	422
quantum chemical calculations 4,	415–428
Absorption	
of acids, buccal	511, 512 <i>t</i> , 514 <i>t</i>
colonic	508, 510
studies	508
at various pH's, buccal	513
Acetates, acid hydrolysis of	35
Acetic acid	302
Acetylcholine (ACh)	301
<i>ab initio</i> calculation of	166
AChE catalyzed hydrolysis of	302
acylation	
intermediates for	307 <i>t</i> , 310 <i>t</i>
mechanism of	303
enzymic hydrolysis of	301–314
intermediates	306
PCILO calculations of	166
valence energy for enzymic	
acylation of	305 <i>t</i>
Acetylcholinesterase (AChE)	301
active site of	301
catalyzed hydrolysis of ACh	302
catalyzed hydrolysis of ASCh	302
<i>N</i> -Acetyl phenylalanyl tyrosine	90
crystal structure of	89
Acetylthiocholine (ASCh)	301
AChE catalyzed hydrolysis of	302
acylation intermediates for	307 <i>t</i> , 310 <i>t</i>
acylation mechanism of	303
enzymic hydrolysis of	301–314
intermediates	306
valence energy for enzymic	
acylation of	305 <i>t</i>
ACh (<i>see</i> Acetylcholine)	
AChE (<i>see</i> Acetylcholinesterase)	
ASCh (<i>see</i> Acetylthiocholine)	
Active analog approach (AAA)	205–223
Active site–substrate complexes,	
computer-generated	192, 194
Activity, biological	21
Activity relations, structure–(SAR)	103–125
Acylation mechanism of ACh	303
Acylation mechanism of ASCh	303
Acyl enzyme	306
Adamantane(s)	155
analogues, structures and inhibi-	
tory activities of	157
-1-carboxylic acid	155
ADAPT	
computer software system	106
descriptor manipulation routines in	112
descriptors used for <i>N</i> -nitroso data	124 <i>t</i>
descriptors used for PAH study	119 <i>t</i>
environment descriptors in	109
fragment descriptors in	109
geometric descriptors in	112
linear discriminant function in	115
linear discriminant function analysis	
(LDFA) in	114
molecular connectivity descriptors	
in	109
<i>N</i> -nitroso compound study in	122
nonparametric recognition pro-	
grams in	114
PAH study	120 <i>f</i>
parametric pattern recognition	
programs in	113
path and path environment	
descriptors in	109
pattern recognition analysis in	112
sigma charge descriptor in	110
substructure descriptors in	109
topological descriptors in	109
training set/prediction set genera-	
tion routine in	113
variance feature selection in	116
weight vector utility routine in	113
Additivity postulate	22
<i>S</i> -Adenosylmethionine	215
<i>S</i> -Adenosylmethionine synthetase	214
Adenylate cyclase	441
ADH-NAD (<i>see</i> Alcohol-dehydro-	
genase-nicotinamide adenine	
dinucleotide)	
Adrenergic agents, active	442
Agonist(s)	219
/antagonist potency	244
receptor	227
AI (<i>see ab initio</i>)	
ALCHEM	543
Alcohol dehydrogenase, horse liver	191
Alcohol-dehydrogenase–nicotinamide	
adenine dinucleotide (ADH-	
NAD)	
active site of	192
complex with alkyl-substituted	
cyclohexanols, interaction of	192
Alcohol dehydrogenase, stereoview of	
the active site of	193 <i>f</i>
Algorithm in CAMSEQ, bond	
rotation	355
Algorithm in CAMSEQ, geometry	
determination	356

Aliphatic systems, π constants from	29t	Antibiotics, structure-activity relationships of tetracycline	142
Allosteric effects in enzymes	206	Antitumor drugs for animal screening, computer-aided selection of	583-601
Alloxan	215, 217, 218f	Antitumor screening program of the National Cancer Institute (NCI)	583
Amine(s)		(-)-Amorphine	228
biogenic	215	ARCANA	302, 313
receptor, biogenic	221	calculations, orbital parameters in	314t
sympathomimetic	439	Aromatic	
structure-activity relationships (SAR) in	441	binding site	231
Amino acid residues of glycyl	407	hydrocarbons, polycyclic (PAH)	111
Amino acid residues of valyl	407	carcinogens	116
2-Amino-6,7-dihydroxy-1,2,3,4-tetrahydronaphthalene (6,7-ADTN)	219	substitutions, electrophilic systems, π constants from	29t
endo-2-Amino isomer of benzobicyclo [2.2.2] octene	459-481	N-Arylanthranilic acids	520
exo-2-Amino isomer benzobicyclo [2.2.2] octene	459-481	α -Aryl- α -cyanocarbonylphenylhydrazones	521
γ -Aminobutyric acid	206	2-Aryl-1,3-indanediones	521
2-Aminonornbornane-2-carboxylic acid	214	Atlas of Steroid Structure	81
Ammonium methylphosphate (AMP)	245f, 246	Atomic charge densities	562
complexes of	251-257, 253f	ATPase (Na ⁺ ,K ⁺)	441
H-bonded	248	inhibition data	272t-273t
Ammonium methylsulphate (AMS)	245f, 246	inhibition studies of genins	270
complexes	251-257, 254f	Augmented atom (AA) keys	585
H-bonded	248		
AMP (see Ammonium methylphosphate)		B	
Amphetamine(s)	451, 459	Barbiturates	323f
analogs	478	commercial	322
decalin	454	Basel Information Center for Chemistry (BASIC)	600
CNDA/2 calculations on	471	BASIC (Basel Information Center for Chemistry)	600
conformational energy map of	474f	Bayesian classification surfaces	104
energy surfaces for	472	Bayesian theorem	113
endo-isomers of	459	Bacteriostatic action	152
exo-isomers of	459	of carboxylic acids	522
AMS (see Ammonium methylsulphate)		of nitrophenols	518t
Analgesic(s)	94	of phenols	517
activity, opiate	244	1,4-Benzenediamine, synthesis of	344
pharmacophore	221	7,8-Benzflavone	412f
structures, opiate	94	Benzobicyclo [2.2.1] heptenes	459
Anhydrobutaclamol	237	Benzobicyclo [2.2.2] octene(s)	459
Anilines, dissociation of	33	endo-2-amino isomer of	459-481
Anionic receptor sites	244	endo-2-methylamino isomer of	459-481
model	245f	exo-2-amino isomer of	459-481
opiate	243-257	exo-2-methylamino isomer of	459-481
Anion interaction, cation-	246	system	459
Antagonist(s)	219	Benzoic acids	32
dopamine receptor	236	Benzomorphans	243
narcotic	417	antagonist potency and binding affinity of	252t
opiate	243	Bicyclic ring structures, ring strain in the	464
potencies	255	Bicyclic systems, CNDO/2 calculations on	476
agonist/	244	Binding	
and binding affinity of benzomorphans	252t	interactions	197
receptor	227	sites	214
H ₂ -Antagonist activity	181		

- Binding (*continued*)
 sites (*continued*)
 aromatic 231
 on dopamine receptor
 lipophilic accessory 238*f*
 primary 235*f*, 238*f*
 lipophilic accessory 237
 opiate 243
 primary 228
 substrate 320
 Bioactive synthesis 319–339
 binding parameters in 320
 reaction parameters in 320
 transport parameters in 320
 Biodata 320
 Bioenergetic models 320
 Biososterism 207
 Biological activity 21, 39
 influence of electronic effects on 40
 influence of steric effects on 45
 relations using pattern recognition,
 chemical structure— 103–125
 Biologically active compounds, design
 of 383–413
 Biologically active substance, confor-
 mationally-restricted analogs of a 454
 Biological
 mechanisms, theoretical models for 165
 potency of a drug 67
 systems, interaction of drugs with .. 53
 Biomolecular interactions 161
 Bocek–Kopecky interaction model 64
 Bond
 angle deformation energy 84
 distance deformation energy 84
 formation, hydrogen 234
 length shortening, thermal 82
 overlap population 309
 rotation algorithm in CAMSEQ 355
 Boronate derivatives 199
 Boronate inhibitor to subtilisin,
 stereoview of 202*f*
 Brain membranes, LSD binding to 177
 B. subtilis by carboxylic acids,
 inhibition of 525*t*
 Buccal absorption 514*t*
 of acids 511
 of organic acids 512*t*
 at various pH's 513
 Butaclamol 219, 227–239
 enantiomers of 228
 (+)-Butaclamol 228
 conformers, Dreiding models of
 nuclei of 232*f*
 hydrobromide 231
 crystal structure of 230*f*
- C**
- Cambridge Crystallographic Data
 Centre 81
- CAMSEQ
 batch 364
 bond rotation algorithm in 355
 geometry determination algorithms
 in 356
 potential functions in 356
 Software System 353
 CAMSEQ-II 356, 562
 CAMSEQ/M (Microprocessor-Based
 Conformational Analysis Sys-
 tem) 353–369
 computer hardware for 357
 data analysis options 366
 graphics model builder in 359, 360*f*
 isoenergy contour maps (CMAPS)
 in 366
 structure of phenethylamine by 360*f*
 Cancer Institute (NCI), antitumor
 screening program of the
 National 583
 Carbaldoximes, configurational analy-
 sis, inversion and reduction of
 pyridine 489–501
 Carbonium ion metabolites of cyclic
 nitrosamines, role of 580
 Carbonium ion metabolites of
 nitrosamines 564
 Carboxylic acids
 bacteriostatic activity of 522
 inhibition of B. subtilis by 525*t*
 2-Carboxy-nitrosopyrrolidine 571
 Carboxypeptidase 53
 active site of 173*f*
 enzymatic mechanisms in 165
 mechanism of action of 172
 molecular electrostatic potential of
 the active site of 174*f*
 Carcinogenic
 activity 105
 nitrosamines 105
 data base, hierarchal QSAR
 molecular structure calcula-
 tor applied to a 553–580
 potency 116, 553, 578
 of a compound 554
 Carcinogens 105
 polycyclic aromatic hydrocarbon
 (PAH) 116
 Cardenolide(s)
 backbones, saturated 262
 8(14)-ene D-ring structure of 265, 267
 14-ene D-ring structures of 265, 267
 energy diagrams of 267
 functional receptor mapping for
 modified 259–277
 genins 260
 saturated D-ring structures of 265
 17 β -side group orientation of 265
 structural studies 260
 toxicity 259

Catalytic activity	165	Cocaine	441
Catecholamines	458, 467	Collander's equation	31, 32
Cation-anion interaction	246	Colonic absorption	508
Cell membranes, phenol inhibition of chloride passage across	519	of acids and bases	510t
Chance correlations	131	Complete Neglect of Differential Overlap (CNDO)	445
Charge densities, atomic	562	calculations	
Charge transfer interactions	246	on amphetamines	471
Chemical Abstract Service	598	on bicyclic systems	476
Chemical Information System (CIS) NIH-EPA-	356	and X-ray crystallographic analy- sis to the design of confor- mationally defined analogs of methamphetamine	439-481
Chemical reactivity	163	method	564
Chemical searches, computer-based ..	155	wave functions	427
Chemical structure-biological activity relations using pattern recog- nition	103-125	Computer -aided	
Chirality	234	design of inhibitors	194
Chloride passage across red cell mem- branes, phenol inhibition of	519	design of substrates	194
Chloromethylketones	201	selection of antitumor drugs for animal screening	583-601
peptide	199	-assisted synthesis	321
<i>p</i> -Chloro-phenoxyacetic ester	373	-assisted synthetic analysis	527-549
Chlorpromazine	210f	-based chemical searches	155
conformers of	221, 222f	-controlled diffractometers	81
phenothiazine neuroleptic	211	-generated active site-substrate complexes	194
α 3OH-5 β cholanic acid, stereoview of model of	195f	graphic(s)	99, 384
β 3OH-5 β cholanoic acid, stereoview of model of	195f	interactive	189
Choline	302, 306	graphic techniques for molecular manipulations	151
Chorismate	155	hardware	190
Chorismate mutase, inhibition of	152	for CAMSEQ/M	357
Chorismate mutase-prephenate dehydrogenase	154	program	
Chorismate to prephenate, Claisen rearrangement of	152	synthesis of Depoprovera pro- posed by	345f
Chorismate to prephenate, isomeriza- tion of	153	syntheses of drugs by a	341-352
Chronotropic activity	259	synthesis of PGF _{2α} proposed by ..	347f
Chymotrypsin, enzymatic mechanisms in	165	software	190
Cimetidine	181	system, ADAPT	106
CIS (Chemical Information System), NIH-EPA-	356	system, NIH PROPHET	260
Claisen rearrangement of chorismate to prephenate	152	Computing systems, crystallographic	81
Classification surfaces, Bayesian	104	Comsub Data	338
CLINORIL	538	Configuration interaction calculations	424
synthesis of	539	Conformation-activity relationships ..	166
Clofibrate	373	Conformation(s)	166
chemical structures of	375f	analysis of module MOLY (CONFOR)	386
Clonidine, QSAR in centrally acting imidazolines structurally related to	143	crystal	87
Cluster analysis	571, 578	energy	387
CMAPS (isoenergy contour maps) in CAMSEQ/M	366	molecular	84
CNDO (<i>see</i> Complete Neglect of Differential Overlap)		receptor-bound	205
		receptor-site	231
		Conformational analysis	383-413
		in drug design, applications of	367
		module in MOLY, parameteriza- tion of the	390
		in molecular modeling, applica- tions of	367

- Conformational (*continued*)
 characteristics of molecules 9
 hyperspace 366
 parameter in drug design 205-223
 search, systematic 209
 studies of phenylethylamines 445
- Conformationally-defined phenyl-ethylamines 471
- Conformationally-restricted analogs of a biologically active substance 454
- CONFOR (Conformation analysis module of MOLY) 386
 comparison of PCILO and 390
- Conjugation 388
- Connectivity indices 36
- π Constants from aliphatic systems 29t
- π Constants from aromatic systems 29t
- Correlation equations 131
- Coulombic interactions 192
- Coulomb's law-type function 354
- Crystallographic computing systems 81
- Crystallography, x-ray 443
- Crystal
 conformation 87
 structure
 of *N*-acetyl-phenylalanyl tyrosine analysis in drug design 79-100
 thermal vibration in 82
- Cyclohexanols, interaction of ADH-NAD complex with alkyl-substituted 192
- Cysteine proteinase 11
- Cytochrome P-450 564
- D**
- Data base
 hierarchical QSAR molecular structure calculator applied to a carcinogenic nitrosamine 553-580
 for multiple regression 321
 on-line 341
- Data storage 543
- Deamino acids 286
- Decalin amphetamine analogs 454
- Decalin compounds, *trans*- 454
- Deformation energy
 bond angle 84
 bond distance 84
 out-of-plane 85
- Dehydrogenases 327
 fungicides inhibiting 328f
- Del Re sigma charge calculation 110
- Density map, pseudoelection 214
- Depoprovera 344
 proposed by computer program, synthesis of 345f
- Descriptors
 in ADAPT
 environment 109
 fragment 109
- Descriptors (*continued*)
 in ADAPT (*continued*)
 geometric 112
 manipulation routines in ADAPT 112
 molecular connectivity 109
 path and path environment 109
 sigma charge 110
 substructure 109
 topological 109
 electronic
 group additive molecular 554
 molecular
 mechanics 555
 substructure 584
 used for *N*-nitroso data, ADAPT 124t
 used for PAH study, ADAPT 119t
 quantum mechanical 562
- Desipramine 441, 479
- Desoxybutaclamol 237
- Diastereoisomers 125
- 5,11-Diazaditwistanes 541
- exo*-3-Dicarboxylic acid 156
- Differential solvation approach 168
- Diffractionometers, computer-controlled 81
- Digitalis 259
 genin- Na^+ , K^+ -ATPase inhibition 258-277
 glycosides 259
 receptor 259
- Digitoxigenin 259-277
 molecular structure of 261f
- Digitoxin 260
- Dimaprit 181
- Dimethyl 2'-hydroxy 6,7 benzomorphan analogues, *N*-substituted 5,9 255
 2',3'-Dimethyl-3,5-diiodothyronine 283
 2',5'-Dimethyl-3,5-diiodothyronine 283
- Dinucleoside monophosphate (CpG) crystalline complex, ethidium/ 358f
- Diphenyl methane structure, generation of 362
- Dispersion-type interaction 246
- Dissociation, process of 59
- DNA
 bases, *ab initio* MODPOT/VRDDO calculations on 422
 double helix 12
 -drug models 12
 -hydrocarbon binding 122
 intercalator 12
- ^3H -Dopamine (DA) 478
- Dopamine
 receptor 215, 227-239
 antagonists 236
 lipophilic accessory binding site on 238f
 model of 238f
 primary binding sites on 235f, 238f
- Double helix, DNA 12

Dreiding models of nuclei of (+)-butaclamol conformers	232f
Dreiding models of nuclei of (+)-isobutaclamol conformers	232f
Drug	
action	4
equation	568
molecular determinants for	169
activity	169
for animal screening, computer-aided selection of antitumor	583-601
biological potency of a	67
with biological systems, interaction of	53
by a computer program, syntheses of	341-352
conformation, solvent effects on	168
design	3
<i>ab initio</i>	150
applications of conformational analysis in	367
conformational parameter in	205-223
crystal structure analysis in	79-100
molecular mechanics in	79-100
quantum mechanical methods in	60
distribution phenomena	59
intercalation	12
models, DNA-	12
reactive sites of	169
-receptor interactions	161
activation in	175
quantum chemical recognition mechanisms in	161-183
quantum chemical studies on	169
recognition in	175
schematic	210f
specificity in	54
receptor site	190

E

EHT (Extended Hückel Theory)	445
Electron density fitting in protein crystallography	214
Electronic	
charge distribution	44
descriptor	110
effects	39, 59
on biological activity, influence of	40
substituent constants for	23
factors in QSAR using distribution coefficients	507-526
mechanisms in substrate-enzyme interactions	175
properties of nitrosamines	125
substituent constant	40
Electrophilic aromatic substitutions ..	33
Electrostatic	
interactions	85, 354
molecular potential	374

Electrostatic (<i>continued</i>)	
contour maps	427
energy contours of the	374
potential	
of the active site of carboxypeptidase, molecular	174f
field	170
of 5-hydroxytryptamine, molecular	180f
of D-LSD	180f
maps	9
of 5-hydroxytryptamine	178
of LSD	178
molecular	170
Ellipticines	586
Enkephalin	221
Enthalpy	42
Entropy	42
Enantiomers of butaclamol	228
Enantiospecificity	228
Energy	
bond angle deformation	84
bond distance deformation	84
calculations of opiates, empirical potential	243-257
calculations of opiates, PCILO potential	243-257
of complex formation, intermolecular	246
conformation	387
contours of the electrostatic molecular potential	374
decomposition, SCF	426
diagrams of cardenolides	267
electrostatic interaction	85
free	28, 87
function, classical-type empirical potential	355
nonbonded interaction	84
out-of-plane deformation	85
solvation	168
strain	84, 85, 87, 108
total interaction	247
total overlap	311
valence	304
Enzymatic mechanisms in carboxypeptidase	165
Enzymatic mechanisms in chymotrypsin	165
Enzyme(s)	
acyl	306
allosteric effects in	206
inhibitors	149
structure-activity studies of	194
Enzyme-substrate complex	168
Enzyme-substrate interactions	10
Ephedrine	
conformations of	451
energy contours for	403f

- Esophagus, tumors of the 554
- Ethidium 12
- Ethidium/dinucleoside monophosphate (Cpg) crystal complex 358f
- N*-Ethylamphetamine 476
- conformational energy map of 475f
- energy surfaces for 472
- EXBIOCOR (Extrapolated Biocorrelation Program) 336
- Excluded volume map 214
- Extended Hückel Theory (EHT) 445
- Extrapolated Biocorrelation Program (EXBIOCOR) 336
- Extrathermodynamic derivation of QSAR parameters 22
- Extrathermodynamic parameters 61
- Extrathermodynamic relationships 21
- F**
- Fentanyl
- citrate 83
- conformations of 98
- molecule 94
- FITMOL, PROPHET procedure 270
- Flavin coenzymes 165
- 5-Fluorouracil 586
- Fock matrix elements 424
- Formamide 175
- FORTRAN 386, 531
- Multiple Regression Analysis Program 132
- Free energy 28
- Free-Wilson method for QSAR 63
- Frontier densities 177
- Functional groups 163
- Functions, mathematical form of potential 389t
- Function, nonbonded potential 395t-400t
- Fundus, rat 177
- Fungicidal activity 330
- Fungicides 327
- inhibiting dehydrogenases 328f
- Furan compounds 247
- G**
- GAA (ganglia augmented atom) 585
- GAA (ganglia augmented atom) keys 597
- Ganglia augmented atom (GAA) 585
- Ganglia augmented atom (GAA) keys 597
- Gaussian integral programs 418
- Gaussian Integral Program System (GIPSY) 418
- Genin(s) 259
- cardenolide 260
- Na⁺,K⁺-ATPase inhibition studies of 270
- Geometry determination algorithm in CAMSEQ 356
- Geometry optimization, intermolecular 251
- GIPSY (Gaussian Integral Program System) 418
- Globular proteins, glycol residues in 407
- D-glucose 215
- Glucose receptor 215
- Glycosides, digitalis 259
- Glycol
- amino acid residues 407
- residue, energy contours for 408f
- residues in globular proteins 407
- Graphics model builder in CAMSEQ/M 360f
- Graphics model builder option in CAMSEQ/M 359
- Graphics system, MMS-X 214
- Guanidines 496
- H**
- Hallucinogens, action of 169
- Hammett equation 23
- Hammett parameters 23
- Hansch methods for QSAR 62
- Hansch equation 320
- Hardware for CAMSEQ/M computer 357
- Hartree-Fock approximation 446
- Herbicides 332f
- Herbicides, thiocarbamate 325, 326f
- Herbicides, urea 331
- trans*-Hexahydrocarbazole 219
- Hexanoic acid-4-aza-4-methyl-6-(3-pyrrolyl)-*N,N*-diethylamide 181
- Hexanoic acid-4-aza-4,5-dimethyl-6-(3-indolyl)-*N,N*-diethylamide 181
- Hexapeptide, stereoview of model of 198f
- Hexoses 215, 217f, 218f
- Highest Occupied Molecular Orbitals (HOMO) 177, 178
- Hildebrand-Scott solubility parameter 38
- Hill reaction 331
- inhibition, QSAR in 142
- inhibitors 332f
- Histimine 181
- H₂-receptors 181
- Histidine 199
- HOMO (Highest Occupied Molecular Orbits) 177, 178
- Hormones
- peptide 212
- thyroid 281, 282f
- cisoid/transoid conformation of 283
- diphenyl ether conformation of 286
- distal/proximal 3'-substituents of 283
- and metabolites 294t
- protein binding characteristics of 292
- receptor interactions, binding models 281-297
- Horse liver alcohol dehydrogenase 191

Hückel		Inotropic activity	259
calculations, extended	415	Insulin release	215, 217f
method	313	Interaction Field Modified Hamil-	
molecular orbital calculations	374	tonian method (IFMH)	172
Theory, Extended (EHT)	445	Interaction pharmacophore	170, 171
Hydrocarbon binding, DNA-	122	of LSD	178
Hydrocarbon (PAH) carcinogens,		<i>syn-anti</i> Interconversion in oximes	497f
polycyclic aromatic	116	Intermediate Neglect of Differential	
Hydrocarbons, polycyclic aromatic	111	Overlap (INDO)	445
Hydrogen		wave functions	427
bond formation	234	Intermolecular interactions	52
-bonding effect	354	Intramolecular interactions	51
bonding sites	170	Iodines, diortho	281
Hydrophilic molecules	57	5-Iodocytidylyl (3'-5') guanosine	12
Hydrophobic		Ionophores	515
binding	57	Iodothyronines	286
effect(s)	39, 53	5-Iodouridylyl (3'-5') adenosine	12
interaction(s)	38	Ionization fraction	58
model	53	(+)-Isobutclamol	228
substituent constant	38, 50, 59	conformers, Dreiding models of	
Hydrophobicity	112, 161	nuclei of	232f
6-Hydroxyadamantane-1,3-dicar-		hydrobromide	229, 231
boxylic acid	154, 156	crystal structure of	230f
<i>exo</i> -6-Hydroxybicyclo [3.3.1] nonane-1	156	Isoenergy contour maps (CMAPS)	
<i>exo</i> -6-Hydroxybicyclo [3.3.1] nonane-		in CAMSEQ/M	366
1, <i>endo</i> -3-dicarboxylic acid	154, 156	Isokinetic relationship	42
<i>exo</i> -6-Hydroxybicyclo [3.3.1] nonane-		Isolipophilic groups	321
1, <i>exo</i> -3-dicarboxylic acid	154, 156	Isolipophilic thiolcarbamate	332
3 α -Hydroxy-5 β -cholanoic acid	194	Isomers, optical	47
3 β -Hydroxy-5 β -cholanoic acid	192		
α -Hydroxy fatty acids	522	K	
5-Hydroxytryptamine (5-HT)	163, 177	K ⁺ -ATPase inhibition studies of	
electrostatic potential maps of	178	genins, Na ⁺	270
molecular electrostatic potential of	180f		
Hyperconjugation constant	46	L	
Hypolipidemic agent	373	Langerhans, islets of	215
		LCAO (Linear combination of atomic	
I		orbitals)	61
IBM RANDU subroutine	132	LCAO-MO-SCF calculations	415
IFMH method (Interaction Field		LEMO (Lowest energy empty	
Modified Hamiltonian)	172	molecular orbital)	308
Imidazol	302, 304	Lennard-Jones 6-12 potential	390, 391f
Imidazolines structurally related to		functions	353
clonidine, QSAR in centrally		Leukemia, P388 mouse lymphocytic	583
acting	143	Ligand complexes, receptor-	227
Imidazolium	304	Linear	
Incumbrance area	122	combination of atomic orbitals	
Indicator variables	65	(LCAO)	61
Indole derivatives	177	discriminant function in ADAPT	115
Indoles, methylated	177	discriminant function analysis	
INDO (<i>see</i> Intermediate Neglect of		(LDFA) in ADAPT	114
Differential Overlap)	445	regression analysis	571
Inductive effects	27, 34	LINPLOT (Linear energy vs. rotation	
Inductive substituent constants	35	angle plots)	366
Inhibitors		Lipophilic accessory binding site	237
computer-aided design of	194	on dopamine receptor	238f
enzyme	149	Lipophilic molecules	57
SH	328		

- LSD 169
 binding to brain membranes 177
 discriminant molecular determinants for the actions of 178
 electrostatic potential maps of 178
 interaction pharmacophore of 178
 reactive sites in 177
 D-LSD, molecular electrostatic potential of 180f
 D-LSD molecular structures of 182f
 Lysine 197
- M**
- Macrodeformations 10
 Macromolecular modeling 190
 D-Mannose 215
 Map, excluded volume 214
 Mapping, receptor 227
 for modified cardenolides, functional 259-277
 Master Data parameter table 319-339
 Mechanics, molecular 84
 Mechanics, statistical 9
 Membranes 56
 Meperidine 10
 hydrobromide 97
 hydrochloride 97
 Mercks, implementation of SECS at 528
 MERGE option 423
 Mescaline hydrobromide 445
 Mesomeric effect 27
 Methamphetamine 459
 analogs 478
 application of CNDO/2 calculations and x-ray crystallographic analysis to the design of conformationally defined analogs of 439-481
 endo-isomers of 459
 exo-isomers of 459
 Methanol 302
 Methionine 214
 Methylation, biological 215
 Methylcyclopropyl compounds 247
 N-Methyl-2-pyridiniumaldoxime (2-PAM) 489
 Methyl *cis*-4-*t*-butyl cyclohexane carboxylates 47
 Methyl *trans*-4-*t*-butyl cyclohexane carboxylates 47
 1-Methyl-2-pyridinium carbaldoxime isomers 500t
 2-Methylamino-1,2,3,4-tetrahydro-1,4-ethanonaphthalene hydrochlorides 463
 2-Methylnitrosopiperidine 564
 mechanism of metabolic activation of 565f
 4-Methyl-N-nitrosopiperidine 125
 6- α -Methyl-17- α -acetoxyprogesterone.. 344
 5-Methyl benz(a)anthracene 116
 12-Methyl benz(a)anthracenes 111
 7- α -Methyl-19Norandrostenediol (7- α -Me) 80, 89
 anti-1-Methyl-1,4-dihydropyridine-2-carbaldoxime 502f
 anti-1-Methyl-1,4-dihydropyridine-2-carbaldoxime ion, N-protonated 505f
 anti-1-Methyl-1,6-dihydropyridine-2-carbaldoxime 503f
 anti-1-Methyl-3,6-dihydropyridine-2-carbaldoxime, C-protonated 503f
 anti-1-Methyl-5,6-dihydropyridine-2-carbaldoxime ion, C-protonated 504f
 anti-1-Methyl-2-pyridinium-carbaldoxime 494f
 endo-2-Methylamino isomer of benzobicyclo [2.2.2] octene 459-481
 exo-2-Methylamino isomer of benzobicyclo [2.2.2] octene 459-481
 syn-1-Methyl-2-pyridiniumcarbaldoxime 495f
 Metamide 181
 Michael addition 328
 activators 329f
 Michael reaction, intramolecular 542
 Michaelis complex 301
 Microdeformations 10
 Mimetics 372, 374
 MIMETIC-GEMO 372
 MINDO/3 molecular orbital calculations 150, 153
 MINDO/3 SCF MO program 490
 Minimal steric difference (MSD) 37
 Mitochondrial monoamine oxidase (MAO) 441
 Mitochondrial protein, binding of phenols to 519
 Molecular
 basis of structure-activity relationships 161-183
 conformations 84, 166
 descriptors, group additive 554
 determinants for drug action 169
 electrostatic potential map 170
 geometries 372
 interactions 425
 perturbation procedures for 425
 three dimensional specificity of .. 371
 manipulation and superimposition.. 151
 mechanics 84
 descriptors 555
 in drug design 79-100
 modeling 108
 applications of conformational analysis in 367
 interactive 189
 of site-specific pharmacophores .. 190

Molecular (<i>continued</i>)	
orbital	
calculations	149, 445
Hückel	374
on Large Systems (MOLASYS)	418
MINDO/3	153
semiempirical	302
lowest energy empty (LEMO)	308
MOSES program	416
potential, energy contours of the	
electrostatic	374, 427
reactivity	60
recognition	205
structure calculator applied to a	
carcinogenic nitrosamine data	
base, hierarchal QSAR	553-580
structure, theoretical calculations	
on	167
substructure descriptors	584
wavefunction	61
Molecules, conformational characteristics of	9
MODPOT (model potential)	419
/VRDDO calculations on DNA	
bases, <i>ab initio</i>	422
Molar attraction constant	38
Molar refractivity	38
MOLASYS (Molecular Orbital Calculations on Large Systems)	418
MOLMEC	108
MOLY program	384
architecture of	385f
conformation analysis module of (CONFOR)	386, 395
MOSES program (Molecular orbitals)	416
Monte Carlo approach	168
Morphinans	243
Morphine	9, 243
molecule	94
monohydrate	88, 90
structure and atom numbering	8f
MSD (Minimal steric difference)	37
MULFIT	568
Mulliken net atomic charges	306
Multiple	
linear regression analysis in	
QSAR	66
perturbation expansion	178
regression analysis	131, 133
Multiplicative Congruential Method	132
Muscimol	213f
N	
Na ⁺ ,K ⁺ -ATPase	441
enzyme	259
inhibition data	272t-273t
inhibition studies of genins	270
Naloxone	243
Narcotic(s)	417
antagonists	417
opiate	243
National Cancer Institute (NCI), antitumor screening program of the	583
NCI (National Cancer Institute), antitumor screening program of the	583
Nerve ending, noradrenergic	440f
Neuroleptic(s)	219, 417
activity	219
pharmacophore	221
Neuronal transmitter	441
Neurotransmitter(s)	442
receptors for	227
NIH-EPA-Chemical Information System (CIS)	356
NIH PROPHET computer system	260
Ninhydrin	215, 217f
Nitrophenols, bacteriostatic activity of	518t
Nitrosamine(s)	553, 556t-560t
action model for cyclic	575
bioactivation	125
carbonium ion metabolites of	564
carcinogenic	105
cyclic	554
data base, hierarchal QSAR model	
structure calculator applied to a carcinogenic	553-580
electronic properties of	125
partition coefficients of	125
role of carbonium ion metabolites of cyclic	580
used in action model QSAR, cyclic	576t
N-Nitroso compounds	123t
N-Nitroso compound study in	
ADAPT	122
N-Nitroso data, ADAPT descriptors used for	124t
Nitrosopiperidine	563f
conformers of	562
Nitrosopyrrolidines	562
NMR (Nuclear magnetic resonance)	447
Nonbonded interaction energy	84
19-Nor (19-Norandrostenediol)	80, 89
Noradrenergic synapse	439
Noradrenergic nerve ending	440f
19-Norandrostenediol (19-Nor)	80, 89
Norephedrine	394, 406f
energy contours for	404f
Norepinephrine	439
³ H-Norepinephrine (NE)	478
Nuclear Magnetic Resonance (NMR)	447
Nuclear proteins	293
Nucleophilic reactions	46
Nucleotide-serotonin complex	163

O

Octanol/water partition coefficients ..	55
Octoclotheptin ..	219
(+)-Octoclotheptin ..	228
Opiate(s)	
analgesic	
activity ..	244
structures ..	94
anionic receptor sites, model ..	243-257
antagonist ..	243
binding site ..	243
empirical potential energy calculations of ..	243-257
fused-ring ..	244
narcotics ..	243
PCILO potential energy calculations of ..	243-257
protonated amine in ..	246
receptor ..	243
-receptor interactions ..	9
modeling ..	243-257
rigid ..	9, 243-257
N-substituents, PCILO-calculated energies in H-bonded complexes of ..	249t
N-substituents, PCILO-calculated energies in complexes of ..	250t
N-substituents in rigid ..	244
Optical isomers ..	47
Organic synthesis, software development for ..	341-352
Orientation map ..	207
Oripavines/thebaines ..	243
Ortho effect ..	47
Oxidative phosphorylation, uncoupling of ..	520
Oximes ..	498t-499t
<i>syn-anti</i> interconversion of ..	496, 497f
Oxygen, formation of the activated ..	165
Oxyheme geometry ..	165

P

PAH (<i>see</i> Polycyclic aromatic hydrocarbon)	
2-PAM (<i>N</i> -methyl-2-pyridinium-aldoxime) ..	489
reduction of ..	496
<i>syn</i> and <i>anti</i> forms of ..	493
Pancreatic trypsin inhibitor (PTI) ..	194
Papain ..	11
Parachor ..	38
Parathion ..	337

Pargyline-treated rat atria ..	479
Partition ..	53
coefficients ..	27, 50, 59
estimation ..	111
membrane/buffer ..	55
of nitrosamines ..	125
octanol/water ..	55
water/octanol ..	553
process of ..	59
Pattern recognition	
analysis in ADAPT ..	112
chemical structure-biological activity relations using ..	103-125
programs in ADAPT, nonparametric ..	114
programs in ADAPT, parametric ..	113
in QSAR ..	62
SAR applications of ..	104
techniques ..	107f
PCILO (<i>see</i> Perturbative Configuration using Localized Orbitals)	
Peptide	
bond, sissile ..	194
chloromethylketones ..	199
hormones ..	212
Perturbative Configuration Interaction using Localized Orbitals (PCILO) ..	246, 394, 445
-calculated energies in H-bonded complexes of opiate <i>N</i> -substituents ..	249t
-calculated energies in $n \rightarrow \pi$ complexes of opiate <i>N</i> -substituents ..	250t
calculations of acetylcholine ..	166
and CONFOR, comparison of ..	390
contour maps ..	394
potential energy calculations of opiates ..	243-257
PGF _{2α} , syntheses of ..	346
proposed by computer program ..	347f
Pharmacologicals, theoretical studies on ..	5t-7t
Pharmacology, quantum ..	3-12
Pharmacophore(s) ..	205-223, 371
analgesic ..	221
chiral ..	228
interaction ..	170, 171
of LSD, interaction ..	178
molecular modeling of site-specific neuroleptic ..	190
neuroleptic ..	221
Phosphorylation, uncoupling oxidative ..	515, 520
Physicochemical properties, prediction of ..	164
Phenethylamine ..	394
by CAMSEQ/M, structure of ..	360f
derivatives ..	367
energy contours for ..	401f
moiety ..	236

- Quantitative structure-activity relationships (QSAR) (*continued*)
 pattern recognition in 62
 in pyrimidinones 142
 quantum mechanical indices of reactivity in 61
 quantum mechanical parameters in 60
 statistical methods in 66
 studies, chance factor in 131-145
 -using an action model 575
 using distribution coefficients, electronic factors in 507-526
- Quantum
 chemical recognition mechanisms in drug-receptor interactions 161-183
 chemical studies on drug-receptor interactions 169
 chemistry 60
 Chemistry Program Exchange 150
 mechanical
 descriptors 562
 indices of reactivity in QSAR 61
 methods 164
 in drug design 60
 semiempirical 35
 parameters in QSAR 60
 techniques, *ab initio* 4
 pharmacology 3-12
- QSAR (*see* Quantitative Structure-Activity Relationships)
- R**
- Random number generator 132
 RANDU subroutine, IBM 132
 RANGE Program 321, 327, 331, 335
 modifiers 324f
- Reaction(s)
 building 342
 coordinate 150
 facilitating 342
 pathways 150
- Reactive sites of drugs 169
- Receptor
 agonists 227
 antagonists 227
 dopamine 236
 -binding assays 223
 biogenic amine 221
 -bound conformation 205
 digitalis 259
 dopamine 215, 227-239
 essential volume 215
 glucose 215
 interactions
 binding models thyroid hormones- 281-297
 modeling opiate 243-257
 opiate- 9
 specificity in drug- 54
- Receptor (*continued*)
 -ligand complexes 227
 lipophilic accessory binding site on dopamine 238f
 mapping 227, 369
 for modified cardenolides, functional 259-277
 model 99
 of dopamine 238f
 for neurotransmitters 227
 opiate 243
 primary binding sites on dopamine 235f, 238f
 protein 292
 recognition 161
 -site conformations 231
 sites
 anionic 244
 model anionic 245f
 model opiate anionic 243-257
 and substrate interactions, modeling 189-201
- Regression analysis, linear 571
 Regression, data base for multiple 321
 Reserpine-treated rat atria 479
 Resonance effects 27, 34
 Ribonuclease A 206
 Ribonuclease S 206
- Ring strain in the bicyclic ring structures 464
- Routine in ADAPT, training set/prediction set generation 113
- Routine in ADAPT, weight vector utility 113
- S**
- SA (Serum albumin) 281
 SAR (Structure-activity relations) 103-125
 SCF energy decomposition 426
 SECS at Merck, implementation of 528
 SECS program modules 530f
 SECS (Simulation and Evaluation of Chemical Synthesis) program 527
 Separability postulate 23
 Serine proteases 191, 194
 Serum albumin (SA) 281
 Serum proteins 293
 binding 283
 SH inhibitors 328
 Sigma charge calculation 110
 Del Re 110
 Simulation and Evaluation of Chemical Synthesis (SECS) program .. 527
 Sissile peptide bond 194
 Software development for organic synthesis 341-352
 Solubility parameter, Hildebrand-Scott 38
 Solubility of water in apolar solvents .. 55f

- Thyroid (*continued*)
 hormones (*continued*)
 protein binding characteristics of 292
 -receptor interactions, binding
 models 281-297
 Thyronine nucleus 286
 Thyronine structures 295
 Thyroxine 281-297, 282*f*, 284*f*
 -binding globulin (TBC) 281, 295
 binding affinity of 297
 -binding prealbumen
 (TBPA) 281, 293, 295
 binding affinity of 297
 -receptor-interaction, model of 296*f*
 diphenyl ether conformation of 282*f*
 Topographical similarity 417
 Torsion angles 472
 Torsional barriers 388
 Torsional interactions 386
 Torsional strain energy 85
 Toxicity 586
 cardenolide 259
 Transform 543
 Transition state
 analogues 149-158
 of alternative 154
 calculation 152
 structure, determination of 149
 Treloxinate 373
 chemical structures of 376*f*-377*f*
 Triiodothyronine 281-297, 282*f*
 Trypsin
 bound to the trypsin inhibitor
 complex, stereoview of the
 active site region of 196*f*
 inhibitor complex, stereoview of
 the active site region of trypsin
 bound to the 196*f*
 inhibitor, pancreatic (PTI) 194
 substrates 194
 -trypsin inhibitor complex 191
 + trypsin inhibitor, stereoview of .. 198*f*
 Tryptamine(s) 163, 177
 congeners, reactive sites in 177
 derivatives 163, 169
 Tumors of the esophagus 554
- U**
- Uncoupling 521
 activity 516*t*, 523*t*
 Ureas, herbicidal 332*f*
 Urea herbicides 331
- V**
- Valence energy 304
 Valyl amino acid residues 407
 Valyl residue 410*f*
 contour map for 411*f*
 energy contours for 409*f*
 van der Waals' interactions 192
 Variable retention of diatomic differ-
 ential overlap (VRDDO)
 approximation 418
 Variance feature selection in ADAPT 116
 Vibration in crystals, thermal 82
 VRDDO (Variable retention of di-
 atomic differential overlap)
 approximation 418
- W**
- Water
 in apolar solvents, solubility of 55*t*
 desolvation of 54
 /hexane partition coefficient 553
 /octanol partition coefficient 553
 Wave functions, CNDO/2 427
 Wave functions, INDO 427
- X**
- X-ray
 crystallographic analysis to the de-
 sign of conformationally de-
 fined analogs of methampheta-
 mine, application of CNDO/2
 calculations and 439-481
 crystallography 443
 diffraction 79
 studies of phenylethylamines 467
- Z**
- Zipper hypothesis 206