ELSEVIER

Journal of Pharmaceutical and Biomedical Analysis
14 (1996) 233–246

JOURNAL OF
PHARMACEUTICAL
AND BIOMEDICAL
ANALYSIS

Review

Biomolecular structure and dynamics—experiment and theory[1]

S. Forsén[a,*], J. Kördel[b]

[a]*Department of Physical Chemistry 2, University of Lund, S-221 00 Lund, Sweden*
[b]*Department of Structural Biochemistry, Pharmacia Biopharmaceuticals, S-112 87 Stockholm, Sweden*

Received for review 15 June 1995; revised manuscript received 8 August 1995

Abstract

Biological macromolecules are complex systems and in order to understand their inner workings we need information from many sources. In this review we present some of the underlying principles for current methods of choice for structural and dynamical studies of biological macromolecules. Interplay between these disciplines—X-ray diffraction, nuclear magnetic resonance spectroscopy and theoretical calculations—has been extremely fruitful and our knowledge in this area of bioscience is rapidly increasing due to this cross-fertilization. While structural aspects of proteins are increasingly well studied and understood we do however still need to put more emphasis on their dynamical properties.

Keywords: NMR spectroscopy; X-ray crystallography; Protein structure; Protein dynamics; Molecular dynamics simulation

1. Introduction

The last decade has witnessed a remarkable upsurge in the numbers of published—or occasionally for commerical reasons unpublished—3D structures of biological macromolecules. X-ray crystal structure determinations have been greatly facilitated by modern molecular biology, synchrotron radiation sources, multiple wavelength anomalous dispersion (MAD), area detectors, charge coupled devices and proficient computer programs. NMR spectroscopy has, step by step, developed into a powerful method for the determination of solution structures. Through the imaginative combination of new radio frequency (r.f.) pulse sequences, isotope labelling (^{2}H, ^{13}C and ^{15}N), recombinant DNA techniques and bacterial, or even mammalian, expression systems the attainable molecular mass limit is presently pushing towards 35000 dalton. Even so, NMR structure determinations will always be limited to much smaller systems than can be approached by

[1] This review is based on a plenary lecture presented at the 5th International Symposium on Pharmaceutical and Biomedical Analysis, 21–24 September 1994, Stockholm, Sweden.

* Corresponding author.

SSDI 0731-7085(95)01640-6

X-ray techniques. NMR has, however, one important advantage over X-ray in that it can provide rather detailed dynamic information—information that for many problems will probably be as valuable as structures. At present comparably few detailed studies have been carried out but it has already been shown that some proteins exist in solution as equilibrium mixtures of two or more conformations. Also, the binding of a co-factor, an inhibitor or an ion to a protein may have only minor structural consequences, but may cause profound changes in dynamics even in regions distant from the site of binding. The rate of protein sequence determinations still outpaces by far the rate of experimental structure determinations by X-ray and NMR (see Table 1). The effort devoted to theoretical approaches to the structure problem therefore proceeds with some vigour at different levels of rigour. Although much progress has been made we are still far from able to predict detailed 3D structures. One of the problems is the fact that the potential energy functions presently widely used to describe inter- and intramolecular interactions in molecular dynamics (MD) calculations are very primitive. Much improved potential energy functions, derived from extended quantum chemical calculations, are now becoming available but have the drawback of calling for exceptional computer resources even for comparatively small molecules.

In this review we will make an attempt to discuss and compare the three main methods used to obtain structural information on biological macromolecules. The review is an attempt to report on some of the possibilities and limitations in these three fields and is by no means exhaustive. The description of X-ray structural studies will deal little with the underlying physical principles, partly because the technique has reached a significantly high level of maturity and partly because our views are expressed as non-members of that community.

Table 1
Well-documented published X-ray and NMR macromolecule structures [1] and submitted protein sequences in the SWISSPROT database

Year	X-ray structures	NMR structures	New sequences
1990	109	23	6059
1991	127	38	4290
1992	164	62	5500
1993	206	58	5174
1994	352	103	6963

2. Diffraction studies

X-ray diffraction has been used for studies of biological macromolecules for many decades and is now a well-established technique. Following the pioneering work on heme proteins by Perutz, Kendrew and co-workers in the 1950s [2] the technique has undergone remarkable development on many different levels. For one thing the intensities of the X-ray beams have become very much higher. The advent of synchrotron radiation sources has, on the one hand, allowed the use of significantly smaller cyrstals, while, on the other hand, rather dramatically shortening the exposure times necessary to obtain an adequate diffraction pattern. Shorter times of exposure of a sample in the X-ray beam have proven most beneficial for sensitive biological samples. This effect is largely due to the reduced absorption of the X-ray at the shorter wavelengths (< 1 Å) accessible with synchrotrons compared with conventional X-ray sources [3]. Shorter times per run also allow one to study a larger number of samples in a given time. Alternatively, the high intensities of the X-ray beams in modern synchrotron facilities may also be used to follow the time course of a biological process. Substrates have been allowed to diffuse into crystals of enzymes and the appearance of intermediates has been observed and their structure determined [4]. Characterization of transient structural intermediates can also be accomplished by cooling the samples to extremely low temperatures, 20–40 K, thereby prolonging the lifetime of such species beyond the time window of the X-ray crystallographic technique [5].

Most X-ray structure determinations are of course carried out using monochromatic radiation. A monochromator is necessary if a synchrotron is the primary radiation source, and this will

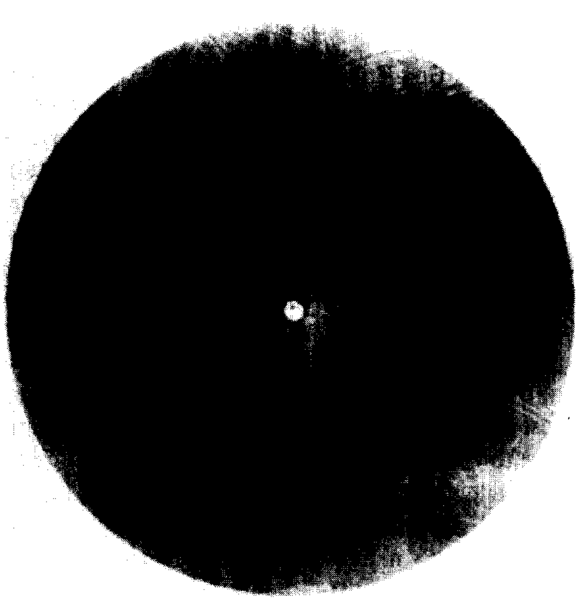

Fig. 1. Polychromatic (so-called Laue) diffraction pattern from a crystal of the enzyme ribulose–bisphosphate carboxylase/oxygenase (RuBisCo). The number of reflections predicted on the photograph are 126, 270. Data were collected and the Figure kindly provided by Dr. Inger Andersson, Uppsala.

by necessity reduce the intensity of the incoming polychromatic light. However, it is possible to do away with the monochromator and send the intense polychromatic beam directly into the crystal, generating a so-called Laue diffractogram. The diffraction pattern now becomes more complex, as can be seen in Fig. 1, but this diffractogram can also be analysed to yield a crystal structure. The striking opportunity is that a Laue diffractogram can be obtained in an amazingly short time—of the order of microseconds or less [4,6]. One can imagine that time-resolved Laue methods will be most useful in the study of time-dependent phenomena in biological systems—for example the detailed time course of an enzymatic reaction. There are however many constraints. For one thing, diffraction patterns are the result of the radiation field interacting with a large number of molecules and to get a clear diffraction pattern these molecules—or rather the unit cells—must all be in the same state. Thus the time dependent phenomenon must somehow by synchronised over a substantial part of the crystal. It is difficult to arrange this through diffusion methods as long as the effective diffusion rate is slower than the rate of the time dependent process one wants to study. An alternative that has been tried is to initiate a chemical transformation in the crystal through an intense laser flash [7]. Again, this ingenious approach is limited to macromolecular systems where a relevant molecular group—for example a substrate of a cofactor—has an absorption band that does not overlap with those parts of the system that should remain unaffected by the laser flash.

The so-called phase problem lies at the very centre of the diffraction method for cyrstal structure determinations. This is why crystallographers try to obtain heavy atom derivatives of their crystalline macromolecules—a heavy atom serves as a kink of internal diffraction reference point [3]. Synchrotron radiation sources have offered an attractive alternative, MAD. Atoms have slightly wavelength dependent interactions with the X-ray radiation field. This is most apparent when the energy of the X-ray photons exactly matches the energy required to excite an electron from a lower to a higher orbit. As outlined in a seminal paper by Hendrickson [8] it is possible to solve the phase problem if one can obtain a diffraction pattern at three X-ray wavelengths below, on top of and above such an absorption edge. Such closely spaced and well-defined X-ray wavelengths can presently only be obtained using a high quality monochromator attached to a synchrotron light source. But what absorption edge to choose? Those of carbon or oxygen are of no use—there are just too many atoms of that kind. Hendrickson came up with an unexpected solution: selenium! Hendrickson showed that it is possible to teach yeast cells—and more recently also Chinese hamster ovary cells—to accept selenomethionine in its proteins as a substitute for regular methionine without significant loss of biological activity [9]. There are usually not too many methionines in proteins and the selenium absorption edge is conveniently located far away from other edges. At present it has been used to solve quite a few structures, most recently that of human gonadotropin [10].

Development on the detector side of X-ray diffraction have been equally striking. Photographic films have been replaced by devices that in a more direct way detect the number of photons emitted in a certain direction. Many of these devices has been developed in high energy physics laboratories faced with the tremendous bookkeeping problem of tracking the subatomic debris after particle collisions. They have also been championed by astrophysicists faced with the problem of capturing nearly every photon reaching us from a distant astronomic object. Thus we now have area detectors, image plates and the charge-coupled device (CCD): make your choice if you have the necessary funding! The latter choice, CCD, is very attractive since it reduces the time between exposures from the present 3–4 min it takes to read off the image plates, to virtually no time at all. The CCDs will furthermore lead to a shorter exposure time due to their higher dynamic range and improved sensitivity as well as their improved resolution of measurement. The number of groups working with the fact generation of CCDs is slowly increasing and consequently their prices are likely to become more and more reasonable.

The final step of interpreting the diffraction pattern into a detailed crystal structure is still a challenge but gone are the transparent electron denstiy sheets, the Richard's boxes and the wire models. Powerful display systems and computer programs have confined the practitioners to secluded rooms filled with graphic wizardry. All in all, these advances have made life easier for the crystallographers and you find some of them bragging that they recently solved a protein structure in a few weeks. Nevertheless, problems remain. Making useful crystals is still very much of a gamble. The quality of the structure is to a large degree dependent on the resolution attainable, which in turn depends on how well-ordered the crystals are. At a resolution of about 2.0 Å the positioning of the individual residues is well-defined and mistakes are seldom made. However, at a resolution of around 3.0 Å it is possible to make serious errors in the interpretation of the electron density map if one is not very careful. It is, nevertheless, possible to make good use of such a dataset as long as great care has been taken in the determination of the initial phases.

Finally membrane proteins are still distressingly elusive—not the least annoying to molecular pharmacologists working with receptor related problems. This becomes especially true when one considers that more than 90% of prescription drugs are targeting membrane bound receptors. No crystal structures have so far been determined for full length receptor molecules. Protein crystallographers have therefore frequently focussed their work on the structures of the extracellular and intracellular parts of the molecules [11]. The light at the other end of the tunnel could be the present efforts to design molecules, such as amphipathic helices [12], that will help solubilize integral membrane proteins.

3. NMR structure determination in solution

As a result of a series of innovative developments in high resolution NMR spectroscopy it is presently possible to determine the 3D structure of biological macromolecules—in particular proteins and nucleic acids—in solution. In addition it is also possible to study dynamics as well as molecular interactions with NMR spectroscopy.

Before we go into a more detailed discussion of the characteristics of this new technique let us first briefly contrast the X-ray and NMR methods. A common feature is that they both permit studies at atomic resolution. While there is virtually no size limit in X-ray studies the situation is not as favourable in NMR. An NMR structure of a protein–DNA complex with a molecular mass of 37 kD was published in 1994 [13] and is presently the ultimate record. Although this is close to what at present is believed to be the size limit for high resolution structure determination with NMR spectroscopy it will probably not be the record for long.

Since NMR studies are performed in solution, NMR structures should be free of artefacts due to crystallisation. While the differences in physical state should have little consequence for the interior of a protein they may have some consequences for the amino acids at the surface itself. Since surface groups are often the site of interaction with other molecules these differences may be of some importance. NMR will provide information on dynamics in different parts of the molecule at different time windows and is very good at exploring ligand interactions and rates of ligand exchange in solution. The mobility information that can be extracted from the crystallographic *B*-factors is hampered by the fact that it is not straightforward to disentangle contributions from static disorder and true mobility.

A fundamental difference between X-ray and NMR methods concerns the way the structural information is derived. The diffraction peaks in X-ray studies are directly related to distances in the periodic crystal lattice and to some extent contain information about the whole structure. In contrast the structural information in NMR spectroscopy—mostly cross peaks in multidimensional spectra—depends in a more convoluted way on intramolecular distances and dihedral angles. This dependence is not unrelated to the dynamic properties of the macromolecule. As we will discuss further below one of the consequences of these differences is that while we are used to seeing only one single refined X-ray structure of a protein molecule, the result of an NMR study is usually presented as an ensemble of structures.

Let us first briefly review the general strategy followed in NMR structure determinations. The first step is to obtain a well-resolved NMR spectrum—or rather NMR spectra—and then assign the individual spectral lines (^{1}H is implied). ^{1}H NMR spectra of a medium sized protein typically consist of more than 1000 partly overlapping lines. Higher spectrometer frequencies (500, 600, 750 MHz) do allow better dispersion due to chemical shifts [14], but even greater separation of lines is possible through multidimensional NMR spectroscopy. In addition, the spectrum may be further simplified by labelling the protein with isotopes such as ^{13}C, ^{15}N and also ^{2}H [15]. Methods based on the high yield expression of mammalian proteins in *E. coli* or other cell systems have been imperative for this latter approach. The pioneers of biological NMR working in the late 1960s could only solve the assignment of protein ^{1}H NMR spectra by selectively isotope-labelling the molecule at specific amino acids. A conceptual breakthrough was made around 1975 when it was demonstrated that once the signals from one unique amino acid in a peptide had been identified, the signals from neighbouring amino acids could be identified through a combination of spin coupling and relaxation data.

The advent of 2D NMR made this latter method—known as the sequential assignment method—much more simple to apply [16], as illustrated in the basic strategy outlined in Fig. 2A. Two types of 2D experiment are usually the minimum requirement to provide the necessary information. In correlated spectroscopy, or COSY, experiments the cross peaks only occur between protons that are connected by 2–3 covalent bonds. The nuclear Overhauser enhancement spectroscopy, or NOESY, experiments on the other hand produce cross peaks connecting protons that are close together (usually less than 4.5–5.0 Å) in space. The simplest form of the sequential assignment method—i.e. with no isotope labelling—tends to become more and more difficult as the size of the protein molecule increases due to extensive overlap of NMR signals. For proteins with more than about 100 amino acids it may then be necessary to label the protein

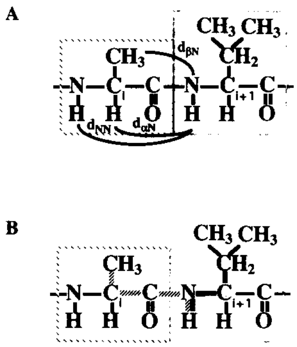

Fig. 2. The sequential assignment procedure using (A) ^{1}H NMR, (B) triple resonance (^{1}H, ^{13}C, ^{15}N) NMR. In (A) the crucial through-space connectivities d_{NN}, $d_{\alpha\mathrm{N}}$, $d_{\beta\mathrm{N}}$ are marked. These conectivities are found in a NOESY experiment and specify that the amide proton of residue $i+1$ is close in space to the N, α or β protons of residue i. In (B) the trough-bond connectivities found in a HBCBCA(CO)NNH experiment are shown hatched and those found in a HNCACB experiment are visualized as striped. The former experiment correlates the C$^\alpha$, C$^\beta$ nuclei of residue i with the N and NH nuclei of residue $i+1$ while the latter correlates the C$^\alpha$, C$^\beta$, N and NH nuclei of residue $i+1$ (and of course of residue i as well).

with ^{15}N and/or ^{13}C. Uniform ^{15}N labelling of recombinant proteins is today almost standard procedure and is most easily accomplished by feeding the producing cells ^{15}N-ammonium salts as the sole nitrogen source. ^{13}C labelling is much more expensive and is achieved by the use of ^{13}C-labelled compounds, such as ^{13}C glucose, as the sole carbon source.

We are now in a position to move from 2D to 3D spectroscopy. For a ^{15}N labelled protein it is possible to decompose a crowded 2D NMR NOESY spectrum into different subspectra, each subspectrum showing only those cross peaks related to protons that are bonded to a nitrogen atom with a particular ^{15}N resonance frequency. Should there still be problems with overlap we could also use uniform ^{13}C labelling and disperse each ^{15}N subspectrum into a set of new subspectra, each characterised by the resonance frequency of the attached carbons (see Fig. 3). Thus it is possible to carry out a 4D experiment!

In addition to using heteronuclei to move to higher dimensionality we have the added benefit of using them for the sequential assignment procedures. With a ^{13}C- and ^{15}N-labelled protein it is possible to base the sequential assignments on information derived only from (heteronuclear) correlated experiments [17], see Fig. 2B. This has several advantages, the main one being that the sequential assignment process now becomes independent of the conformation of the peptide backbone (see below).

Assume that we have assigned all, or almost all, NOESY cross peaks and that we have also determined values of spin coupling constants between protons in the different amino acid residues. Both parameters contain structural information, in fact it is already now possible to draw a number of conclusions about the secondary structure of the protein. NOESY cross peak intensities—or "volumes"—depend to a first approximation on the interproton distance, and the spin couplings over three covalent bonds depend on the dihedral angle in the fragment. With these simple rules we can now identify β-sheets and α-helices. In the former the NH proton of each residue is close to the H^{α} proton of the preceding residue (≈ 2.2 Å), while the distance between successive NH protons is large (≈ 4.2 Å). In the α-helix the local conformation, by contrast, brings NH protons of successive residues close together (≈ 2.8 Å), while the NH proton is much further from the preceding H^{α} proton. If doubly labelled material is available then the secondary ^{13}C chemical shifts for the alpha and beta carbons can be used to determine the secondary structure [18]. The secondary chemical shift of C^{α} is positive for α-helical segments and negative for extended structures such as β-sheets, while the reverse is true for the C^{β} chemical shifts.

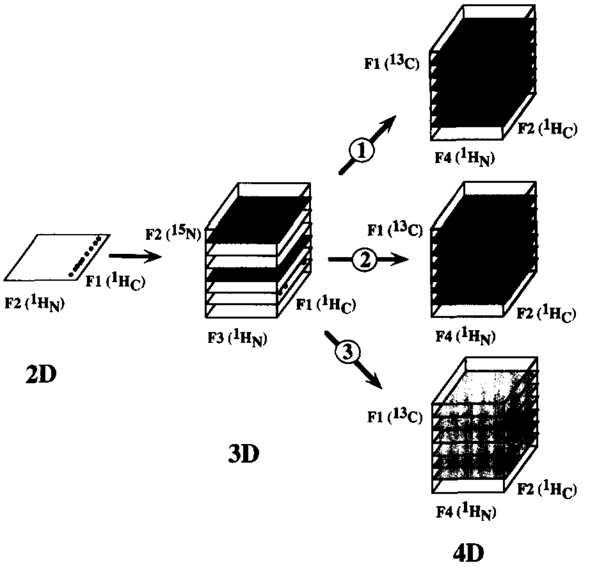

Fig. 3. Schematic illustration of the relationship between 2D, 3D and 4D heteronuclear-edited NMR spectroscopy. If a two-dimensional (2D) NOESY spectrum is recorded on a protein where aliphatic protons ($^{1}H_C$) are close in space to amide protons ($^{1}H_N$) that are degenerate in their chemical shift then all their NOE crosspeaks will line up in one column and no distinction can be made between the different amide protons. By going to a three-dimensional (3D) NOESY spectrum using the ^{15}N chemical shifts of the amide nitrogens to edit the information, then, assuming a non-degeneracy between the amide nitrogen shifts, the crosspeaks will be found at different 2D planes in the 3D cube depending on which amide proton the NOEs belong to. The NOEs in each of these planes can then be edited into yet another cube in a four-dimensional (4D) experiment where the ^{13}C chemical shift of the aliphatic carbon is used for the last editing.

We are not content though to have only secondary structure information. The NOESY cross peaks usually indicate that protons from different parts of the peptide chain are within a few Ångström of each other, so we would like to convert this information into a tertiary structure that agrees with these observations. Here we come to one of the crucial steps in our NMR method. Theory tells us that for a pair of isolated protons rotating randomly at a fixed internuclear distance the observed cross peak intensity is strictly proportional to the inverse of the sixth power of the internuclear distance (for short mixing times!). However, no interproton distances—not even those in the same molecular group—are rigidly fixed. No two protons in a protein molecule are isolated from their fellow protons. Also, there is no guarantee that the internuclear vector between two protons will rotate randomly and furthermore that all pair vectors in our protein will rotate with the same characteristic correlation time. Thus many caveats need to be made. As a consequence of this it has long been customary in biological NMR not to interpret NOESY cross peaks in terms of a single distance but in terms of a distance range. For example: strong = 1.8–2.7 Å; medium = 1.8–3.3 Å; weak = 1.8–5.0 Å; and very weak = 3.0–6.0 Å. The fact that no two protons are totally isolated results in NOESY cross peak intensities that are either too small—depending on the transfer of nuclear magnetization to neighbouring protons—or too large because two protons are not really close in space, but may have some neighbours that are. It takes time for this transfer to take place so NOESY spectra should ideally be recorded at short mixing times. However, this will make *all* cross peaks small so that the relevant information may be lost in the noise!

As mentioned above the COSY spectra contain information about dihedral angles through values of three bond spin coupling constants, ${}^3J_{ij}$. The most interesting for us on our attempt to determine the tertiary structure are the couplings between the C^α hydrogen and the amide hydrogen in the same amino acid residue (which tells us something about the ϕ angle) and between the C^α hydrogen and the C^β hydrogen (which tells us something about the orientation of the side chain C^β–C^γ bond with respect to the backbone amide nitrogen—the ψ^1 angle). Again we have to rely on some theoretical relationship between angle and spin coupling. It turns out that the relations have a built-in ambiguity, so that more than one value of the dihedral angle can fit a given coupling constant. For proteins with molecular masses above about 15 kD the increased line widths of the NMR lines tend to become so large as to obscure the small proton–proton spin coupling constants (mostly less than ≈ 12 Hz). Here the possibility to uniformly label proteins with ^{13}C and ^{15}N has opened up attractive alternatives—for example measuring spin couplings between amide ^{15}N and C^β hydrogen atoms—and there are also reports on side-chain orientations [19]. Since the origin of the ^{1}H NMR increase in line width is relaxation due to neighbouring ^{1}H nuclei, one way to reduce the line width is to reduce the number of proton neighbours either through specific or random fractional deuteration [20].

Early NMR solution structures were calculated, in a way that we will shortly discuss, using a kind of "pseudo atom" model for prochiral groups such as methylene groups and geminal CH_3 groups—simply because the correct stereochemical assignment of the protons could not be made. Progress in NMR methodology has recently facilitated such assignment and the NMR-derived structures have improved as a result.

Thus through the multidimensional NMR experiments we have arrived at a number of geometric constraints and our next problem is now to find that set of polypeptide chain conformations which satisfies these constraints. It should be remembered that these constraints actually involved *ranges* of allowed distances and angles. The conformation space available to a polypeptide chain is truly astronomical. For a 100 amino acid residue polypeptide chain there could be as many as 10^{90} possible conformations. How do we carry out a search among these? Already at the outset we may note that we can hardly hope to find a *single* conformation that satisfies our constraints, nor all conformations that do, because this is also likely to be a very large number. What we should aim for is a *representative* sample of the possible conformations.

Although computer model building has been used to search the conformational space the model mostly used is based on a procedure called distance geometry (DG) [21]. The basis is a mathematical theorem that essentially states that if all the distances between a number of atoms in space are known than the coordinates of the atoms can be calculated. That such a procedure may be applied to protein structure determinations from NMR constraints can perhaps be seen intuitively through the following analogy. Imagine we have available to table of train ticket prices between all cities connected by the Swedish State Railways. Just as our NOESY cross peaks are related to distances, the Swedish rail prices do indeed bear some crude relation to distance. Thus we may use the prices to construct a crude map of Sweden. Clearly the map would be more detailed the more detailed the price list is.

Obviously we do not know all interatomic distances in our protein, but first of all we can use standard bond lengths and bond angles for the residues in our polypeptide chain. Actually this assumption is also used in X-ray diffraction studies when the chain is traced through the electron density maps. To this we add the distance constraints from our NMR studies and these data form an input matrix for our distance geometry computer program. To give the program something to start from we also provide an initial coordinate matrix, which could in principle be a random arrangement of points. Then follows a step called "embedding" which transforms the total number of distances (based on the experimental NOESY data as well as covalent distances and geometry) into Cartesian coordinates. This coordinate set may actually represent a structure that violates constraints not initially considered—for example some atoms may have been placed within their respective van der Waals radii. Therefore a second optimisation step is necessary to eliminate violations as far as possible. Repeated calculation cycles hardly ever result in exactly the same structure. The standard procedure in NMR structure studies has therefore been to calculate a number of different structures—from 20 to 100—which are then overlaid in one picture. As an example the solution structure of an IgG binding domain of the bacterial protein Protein L has been recently solved by NMR, as shown in Fig. 4 [22].

Can we make use of additional information to check the validity of our NMR structure and can we improve the structure? First of all it may prove impossible to eliminate all violations in our NMR structures. This may be due to several factors. Some mistakes may have been made in the assignment and a critical reanalysis of data may eliminate these. More importantly, violations and inconsistencies may also be due to dynamic effects. The protein may undergo motional averaging between two or more well-defined conformations on a time scale comparable to that for NMR. Thus the resulting averaged NOEs and dihedral angles may actually represent a "pseudo conformation" that does not exist. Regions in an NMR structure, according to the DG calculations, come out as poorly defined, and are usually a consequence of there being very few observed NOEs. This could in turn be the result of local dynamic processes that tend to obliterate the NOESY cross peaks. Conformational flexibility in protein solutions is presently insufficiently

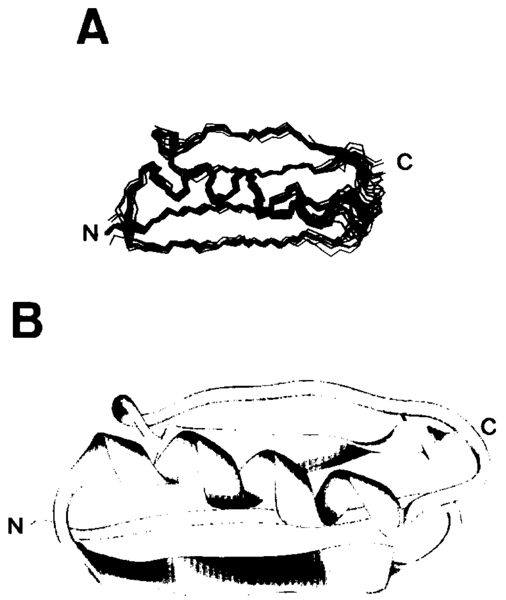

Fig. 4. NMR solution structure of the IgG binding domain B1 in protein L. The structure was calculated and the Figure kindly provided by Wikström [22]. (A) The backbone atoms (N, C^{α}, C′) of 21 refined structures are shown superimposed on their mean structure. (B) Ribbon diagram of the mean NMR solution structure calculated using the MidasPlus software (Computer Graphics Lab., UCSF, San Francisco, CA).

understood and not adequately quantitated and presents a major challenge to molecular biophysicists. Some progress has recently been made through relaxation studies of ^{13}C- and ^{15}N-labelled proteins [23].

Calculated structures can also be contrasted with experimentally determined chemical exchange rates of amide hydrogens—slow exchange is usually taken as an indication of hydrogen bond formation or structural integration in hydrophobic clusters. A major experimental NMR parameter that is easily available but largely unexplored is the chemical shift, of hydrogen as well as of ^{13}C and ^{15}N atoms. Chemical shifts can be incredibly sensitive to small changes in structure. It has been estimated that movements of a hydrogen atom 0.01 Å relative to an aromatic ring may be detected. While some progress has been made in the calculation of chemical shifts in a given structure the inverse problem of calculating structures from observed shifts is quite formidable and remains largely unsolved.

NMR solution structures may also be refined using energy minimisation methods—commonly MD. The garden-type MD method is an excerise in Newtonian mechanics applied to a molecular system. The equations of motion of the constituent atoms are solved in femtosecond time steps and the interactions between the atoms are described by some semi-empirical potential energy function. It is also possible and customary to introduce terms representing the NOE constraints as a kind of pseudo-potential that essentially tries to bring together protons to within the determined distance limits. Such restrained MD simulations tend to produce improved structures with fewer violations and better sampling of the available conformation space.

We now come to the important issue of the accuracy and precision of protein structures determined by NMR. We have already outlined some of the possible sources of error in the NMR method. How should we quantitate these? In this context accuracy denotes how well we can reproduce the "true" structure, while precision refers to how reproducible our individual structures are. We immediately see a problem here. Textbooks often explain the differences between accuracy and precision through a target-practice picture. In a way this picture is deceptive because it presumes that we *know* where the center of the target is located. But how do we know what a "true" protein structure is? The whole concept of a "true" structure may actually be something of a pipe-dream—the closer we try to define it, the more elusive it becomes. But let us assume that we have some kind of structural "gold standard"—it may for instance be a structure determined by some *independent* method—for example X-ray diffraction. Then we could illustrate the differences between accuracy and precision in a different way (Fig. 5). Researchers in the field argue over the accuracy of NMR structures. Some take the mean of a family of NMR structures as the "gold standard" and claim that the accuracy at best could be from 0.25 to 0.6 Å for backbone atoms and 1.1 Å for all atoms. Others find this choice of a "gold standard" questionable and tend to settle for values of 1.0–2.0 Å.

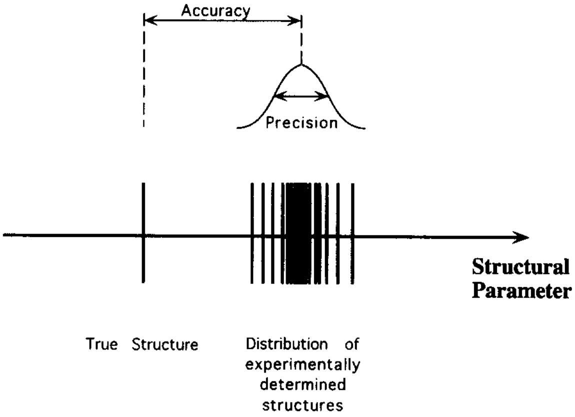

Fig. 5. One representation of the differences between accuracy and precision.

However, the precision of backbone atoms can today certainly be somewhere in the vicinity of 0.3–0.5 Å. In particular, during the early days of NMR structure determination the question of how NMR structures and X-ray structures did compare was very much debated. Certainly this is an interesting issue that some scientists still love to bring up. As is well discussed in a review by Wagner et al. [24] comparison of NMR and X-ray data can be made in at least three ways. First, one may compare the calculated ensemble of NMR structures with the crystal structure to see whether or not the crystal structure will quantify as another member of the ensemble. In this vein one will often see pictures in which the crystal structure is superimposed on the NMR structure. The possible differences tend to be readily visible to the eye, but can of course also be analysed in a more detailed way. Secondly one may compare the crystal structure with the geometric constraints obtained from the NMR data. Such a comparison can be made even before any NMR structures have been calculated. For example if we find a strong NOESY cross peak between two protons—indicating close contact in space—that in the crystal are more than 5 A away, this indicates that significant differences are at hand. Finally we may use the crystal structure to back-calculate the experimental NMR parameters. The problem with this approach—which otherwise has some attractive features—is that a number of simplifying assumptions have to be made in carrying out the calculations. Thus we must be aware that observed differences could be due to such shortcomings rather than representing true structural differences.

A brief trip through the published data will take us to early cases like the α-amylase inhibitor tendamistat, for which parallel studies with NMR and X-ray were made [25]. Here the average root mean square deviation (RMSD) between the mean NMR structure and the X-ray structure was ≈ 1 Å while the RMSD among the NMR structures was 0.85 Å. There are many other examples where solution structures and crystal structures agree to a similar extent. There have also been a few cases where significant differences have been found. One concerns the homologous "inflammatory" proteins C5a and C3a. Here the NMR structure [26] has defined an N-terminal α-helix that is absent in the crystal structure [27]. However, the crystal structure shows a C-terminal helix where NMR does not. Another case is interleukin 8—a homodimer of 2 × 8 kD. The NMR structure has been determined by Clore et al. [28]. The crystal structure was later solved by Baldwin et al. [29] using the NMR structure as an input to solve the phase problem. The RMSD between the X-ray structure and the NMR mean structure is 1.1 Å for all atoms in the well-defined core of the monomer (72 amino acid residues). However, the RMS distances in some parts of the core are sometimes as large as 3-4 Å. The quaternary structures obtained also showed significant differences. Two α-helices on top of an extended β-sheet structure are 14.8 Å apart in the NMR structure, but only 11.1 Å apart in the crystal structure.

If there is a "take home" message to be gathered from structural studies published so far, it would probably be that crystallisation can either destabilize or stabilize parts of a protein structure. As pointed out by Wagner et al. [24] the active sites—or rather interaction sites—of many protease inhibitors, hormones and growth factors often appear disordered or ill-defined in NMR studies while they may be seen as ordered in X-ray structures. Since some mobility in the interaction sites probably facilitates the recognition event, the NMR results are not unreasonable. The ordered structures seen in the crystal may represent a conformation that deviates from the conformation that is most important for the interaction.

4. Theoretical attempts to calculate structures of proteins

As early as 1929 one of the founding fathers of quantum mechanics, Paul Dirac, made a bold and much cited declaration that "the underlying physical laws for the mathematical theory of a large part of physics and the whole of chemistry are completely known". Thus it seemed that chemistry was merely a subdiscipline of physics and that the only sensible thing left for chemists to do was to resort to theoretical calculations. In some senses Dirac was of course right—and his somewhat arrogant statement has certainly cast its spell over chemistry. But the complexity, immensity and practical intractability of the approach did somehow escape him—and perhaps also some of his followers—since not even the hydrogen molecule can be treated exactly! It would be most attractive if we could do away with tedious exper-

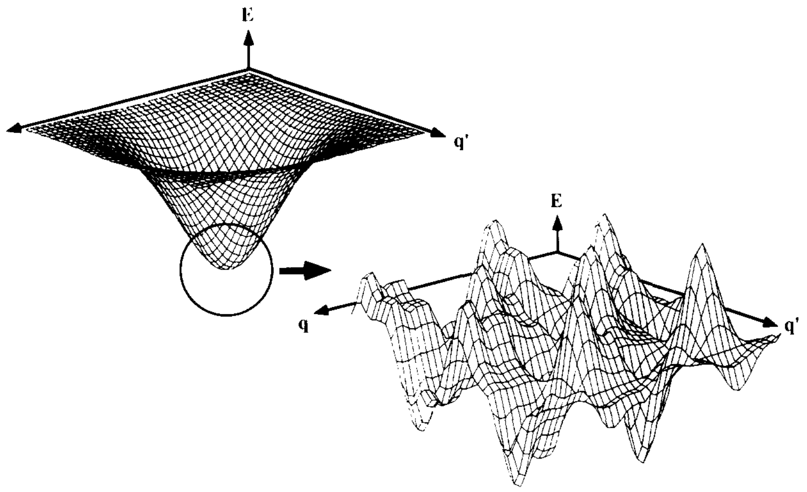

Fig. 6. The energy landscape of a protein molecule. The expansion shows, contrary to widespread blief, that the energy surface (described in two dimensions with the coordinates q and q') will have a very jagged appearance with many local minima of similar energy (E).

iments altogether and accurately predict the structure of a protein simply from knowledge of its amino acid sequence. After all, Mother Nature does this trick—the folding trick—within a second or less, repeatedly in our cells.

From a theoretical point of view, what is the meaning of the "structure" of a protein molecule? As already noted above, a polypeptide chain can in principle take up an astronomical number of conformations. Even with only two allowed conformations per residue we will have at least 10^{30} allowed conformations for a moderate 100 residue polypeptide chain. We are interested in that, or rather those, with the lowest free energy—the *global minimum*. We should not forget that we usually mean the global energy minimum in aqueous solution! There have been a number of attempts to calculate the global free energy minimum of polypeptide chains. An accurate calculation of the complete multidimensional free enregy surface from first principles is way beyond our present capabilities. Therefore many scientists have settled for enthalpy calculations that are much more accessible than free energy calculations. The energy of a protein is a function of the topological arrangment of all the constituting atoms. The energy of even a small protein is dependent on thousands of coordinates—it constitutes what mathematicians call a hypersurface. What will it look like? It will most likely not have a very smooth and well-behaved appearance with a single deep minimum. On the contrary it will have a very rugged look with lots of minima of nearly the same energy, separated by small and large ridges (Fig. 6). Each minimum represents a different conformation—if you like, a different "structure" or "substate" (as coined by Frauenfelder et al. [30]). Most theoretical calculations performed up to now have not aimed for a calculation of the energy "hypersurface" in the minimum energy region but have rather tried to find the "global energy minimum". The primary tool for studies of polypeptide thermodynamics and dynamics is MD simulations, already mentioned above. The protein molecule is treated as a classical system with a potential energy that depends on the coordinates of all atoms. The potential energy function includes terms that depend on bond lengths, bond and torsion angles, van der Waals and Coulombic interactions. To make extensive computer simulations possible the force fields are considerably simplified and different parametrization schemes are implemented, according to the preference of the different labs. There are several problems with MD simulation as a method. Early simulations ignored the surrounding solvent water. In prolonged simulations—starting from known X-ray structures—the proteins tended to literally explode unless the Coulomb interactions were scaled down. Present day simulations of proteins mostly include a solvent shell—the computational demand is considerably increased as a consequence. Predictions of the global minimum 3D structure of a protein-based MD simulation without any previous knowledge of the global fold have been unobtainable in practice up to now, although inroads have been made with a related theoretical method—Monte Carlo simulation [31]. One problem is the extensive computer time required even on a powerful supercomputer. Another is the problem of local free energy minima—how to differentiate local minima from the global minimum and how to avoid getting stuck in a local minimum and thus come up with an erroneous structure. Many schemes to solve this latter problem have been suggested [32] but the general applicability of such schemes is presently unknown.

The parametrization of the potential energy functions—"the force field"—used in MD calculations of proteins is an issue of some importance. As mentioned above a number of different parameter sets are presently being used. Which set gives the most reliable result? It will most likely depend on the type of information desired. Let us illustrate some of the limitations of some of the currently used semi-empricial potential energy functions, taking results obtained on a simple model system of polypeptides—the "alanine dipeptide". Assuming planarity of the amide bonds and neglecting the barrier to rotation of the methyl group, the conformations of this molecule can be specified by only two parameters, namely the backbone dihedral angles, ϕ and ψ. Despite this apparent simplicity the potential energy surface turns out to be fairly complicated with several minima and barriers. The alanine dipeptide is sufficiently simple to allow not only semi-empirical energy calculations but also ab initio quantum mechanical (QM) calculations of some quality (using an extended basis set) and detailed quantum mechanical energy surfaces have been calculated [33]. Let us now compare the energy difference between two conformations, represented by the two energy minima denoted C_7^{eq} and C_7^{ax}, as calculated by QM and miscellaneous potential functions (Table 2). We note that while the QM results tend to put the energy difference in the range 2.0–2.5 kcal mol^{-1} the force fields used in various models predict this difference to be from 0.3 to 8.8 kcal mol^{-1}. Some of the commonly used computer programs do fairly well, in particular CHARMM versions. The results just presented have been analysed by Brooks and Case [33] and the differences may be traced back to the peculiarities of individual potentials and assumptions. The results may seem both discouraging and encouraging depending on the expectations of the observer. At least they indicate that caution should be exercised in theoretical studies of protein stability and energetics.

5. Concluding remarks

X-ray diffraction is a mature scientific field. It has attracted a number of brilliant scientists that are not only masters of the technique but, most importantly, are also on top of the biological problems. Thus X-ray diffraction has, for instance, been indispensible for advancing our knowledge on the workings of the MHC proteins and the recognition of peptide antigens. It has provided us with the only available detailed structures of complex membrane proteins. The molecular structure of several viruses has been revealed.

Table 2
Relative energies in the alanine dipeptide in vacuum as a function of the forcefield used [33]

Method/potential energy function	$\Delta E(C_7^{ax}-C_7^{eq})$ (kcal mol^{-1})
"Model 4"	0.3
AMBER (united atom model)	1.1
AMBER (all atom model)	1.3
CHARMM-19	2.0
CHARMM-22	2.2
Quantum mechanical Hartree–Fock	2.2
ECEPP/2	7.3
ECEPP	8.8

The interactions of transcription factors with their target DNA sequences have been detailed. The list could easily be extended.

Structure determination by NMR is still in a developmental phase. New ingenious pulse sequences, spectral editing and simplifications are still emerging at a brisk pace. Like teenyboppers, many practitioners in this field are actively exploring the limits of their universe. A driving force is to attack ever bigger systems and to push the accuracy even further. This is not a futile activity; on the contrary it is probably a most necessary period that serves to advance the field. Concurrently, the biological problems are certainly gradually coming to the forefront in the NMR community. As we have seen, NMR offers many unique opportunities but it will never be considered to be on a par with X-ray diffraction techniques when it comes to the size and complexity of the structures to be determined. Our personal belief is that the dynamic information that NMR can provide—ligand binding and exchange as well as dynamics of macromolecules—will become of ever increasing importance in the future.

It is jokingly said that when an experimentalist presents his new results at a meeting everyone in the room, except himself, tends to believe them. But when a theoretician presents his new data to his peers he is the only one in the room to be convinced! Surely an exaggeration—theoretical results tend to be rather seductive, in particular to people outside the field. We seem to have an almost inborn disposition to believe written words and numbers. Even with all their present shortcomings computer simulations could be most useful. It is important to understand the limitations of the method and to ask the relevant types of questions. After all, even simple hard sphere models of molecules have considerably advanced our understaning of the thermodynamic properties of liquids.

7. Acknowledgements

We thank Drs. Anders Liljas, Thomas Lundqvist and Gunnar Karlström for critical reading of sections of this paper. We also thank Dr. Inger Andersson for providing Fig. 1 and Dr. Mats Wilkström for providing Fig. 4.

References

[1] W.A. Hendrickson and K. Wüthrich (Eds.), Macromolecular Structures, Current Biology Ltd., London, 1991.

[2] D.W. Green, V.M. Ingram and M.F. Perutz, Proc. R. Soc., Ser. A, 225 (1954) 287.
J. C. Kendrew, R.E. Dickerson, B.E. Strandberg, R.G. Hart, D.R. Davies, D.C. Phillips and V.C. Shore, Nature, 185 (1960) 422.

[3] J. Drenth, Principles of Protein X-ray Crystallography, Springer-Verlag, New York, 1994.

[4] D.W.J. Cruickshank, J.R. Helliwell and L.N. Johnson, Time-resolved Macromolecular Crystallography, Oxford Science Publications, Oxford, 1992.

[5] I. Schlichting, J. Berendzen, G.N. Phillips, Jr. and R.M. Sweet, Nature, 371 (1994) 808.

[6] J. Hajdu and L.N. Johnson, Biochemistry, 29 (1990) 1669.

[7] I. Schlichting, S.C. Almo, G. Rapp, K. Wilson, K. Petratos, A. Lentfer, A. Wittinghofer, W. Kabsch, E.F. Pai, G.A. Petsko and R.S. Goody, Nature, 345 (1990) 309.

[8] W. Hendrickson, Trans. Am. Crystallogr. Assoc., 21 (1985) 11.

[9] W.A. Hendrickson, J.R. Horton and D.M. LeMaster, EMBO J., 9 (1990) 1665.

[10] H. Wu, J.W. Lustbader, Y. Liu, R.E. Canfield and W.A. Hendrickson, Structure, 2 (1994) 545.

[11] S.S. Hall, Science, 267 (1995) 620.

[12] C.E. Schafmeister, L.J.W. Miercke and R.M. Stroud, Science, 262 (1993) 734.

[13] H. Zhang, D. Zhao, M. Revington, W. Lee, X. Jia, C. Arrowsmith and O. Jardetzky, J. Mol. Biol., 238 (1994) 592.

[14] J. Boyd, N. Soffe and I. Campbell, Structure, 2 (1994) 253.

[15] A. Bax and S. Grzesiek, Acc. Chem. Res., 26 (1993) 131.
G.M. Clore and A.M. Gronenborn, Prot. Sci., 3 (1994) 372.

[16] K. Wütrich, NMR of Proteins and Nucleic Acids, John Wiley & Sons, New York, 1986.

[17] G. Montelione and G. Wagner, J. Magn. Reson., 87 (1990) 183.
M. Ikura, L.E. Kay and A. Bax, Biochemistry, 29 (1990) 4659.

[18] S. Sper and A. Bax, J. Am. Chem. Soc., 113 (1991) 5490.

[19] G. Montelione and G. Wagner, J. Am. Chem. Soc., 111 (1989) 5474.

[20] D.M. Lemaster and F.M. Richards, Biochemistry, 27 (1988) 142.
S. Grzesiek, J. Anglister, H. Ren and A. Bax, J. Am. Chem. Soc., 115 (1993) 4369.

[21] T. Havel and K. Wüthrich, J. Mol. Biol., 182 (1985) 281.

[22] M. Wikström, T. Drakenberg, S. Forsén, U. Sjöbring and L. Björck, Biochemistry, 33 (1994) 14011.

[23] G. Wagner, Curr. Opin. Struct. Biol., 3 (1993) 748.

[24] G. Wagner, S. Hyberts and T.F. Havel, Ann. Rev. Biophys. Biomol. Struct., 21 (1992) 167.

[25] M. Billeter, A.D. Kline, W. Braun, R. Huber and K. Wüthrich, J. Mol. Biol., 206 (1989) 677.

[26] M.P. Williamson and V.S. Madison, Biochemistry, 29 (1990) 2895.

[27] R. Huber, H. Scholze, E.P. Paques and J. Deisenhofer, Hoppe-Seyler's Z. Physiol. Chem., 361 (1988) 1389.

[28] G.M. Clore, E. Appella, M. Yamada, K. Matsushima and A.M. Groneneborn, Biochemistry, 29 (1990) 1689.

[29] E.T. Baldwin, I.T. Weber, R. St. Charles, J.C. Xuan, E. Appella, M. Yamada, K. Matsushima, B.F. Edwards, G.M. Clore, A.M. Gronenborn and A. Wlodawer, Proc. Natl. Acad. Sci. USA, 88 (1991) 502.

[30] H. Frauenfelder, F. Parak and R.D. Young, Ann. Rev. Biophys. Biophys. Chem., 17 (1988) 451.

[31] A. Kolinski, A. Godzik and J. Skolnik, J. Chem. Phys., 98 (1993) 7420.

[32] D.R. Ripoll, L. Piela, M. Vasques and H.A. Scheraga, Proteins, 10 (1991) 188 (and references cited therein).

[33] C.L. Brooks, III and D.A. Case, Chem. Rev., 93 (1993) 2487.