A Protein NMR Assistant

P Edwards D Sleeman

Department of Computing Science
King's College
University of Aberdeen
Aberdeen, Scotland
AB9 2UB

G C K Roberts L Y Lian

Biological NMR Centre
Medical Sciences Building
University of Leicester
Leicester, LE1 7RH

Introduction

The aim of this project is to develop a set of tools which will aid spectroscopists in the interpretation of nuclear magnetic resonance (NMR) spectra of proteins and polypeptides.

We begin by providing a brief overview of the nature and structure of proteins before describing the modern NMR techniques used to elucidate their structure. The design of a prototype tool is then discussed. A more detailed description of the background to this project can be found in Edwards et al 1989.

The Nature of Proteins

Proteins are probably the most diverse biological substances known. As enzymes and hormones, they catalyse and regulate the reactions that occur in the body; as muscles and tendons they provide the body with its means of movement; as skin and hair they give it an outer covering; in combination with other substances in bone they provide it with structural support. In spite of such diversity of size, shape and function, all proteins have common features that allow their structures to be deciphered and their properties understood. Proteins are biopolymers composed of amino acid building blocks or monomers. There are 20 common amino acids used to synthesize proteins.

The structure of a protein molecule is considered at three levels of detail: primary, secondary and tertiary structure. The primary structure describes the chemical composition of the protein in terms of the linear arrangement of amino acid residues within the molecule; the secondary structure describes common structural arrangements of parts of the backbone - two major forms are the β sheet and α helix; while the tertiary structure details the folding of these chains in three dimensional space. One of the experimental techniques used to investigate the structure of proteins is nuclear magnetic resonance spectroscopy (NMR).

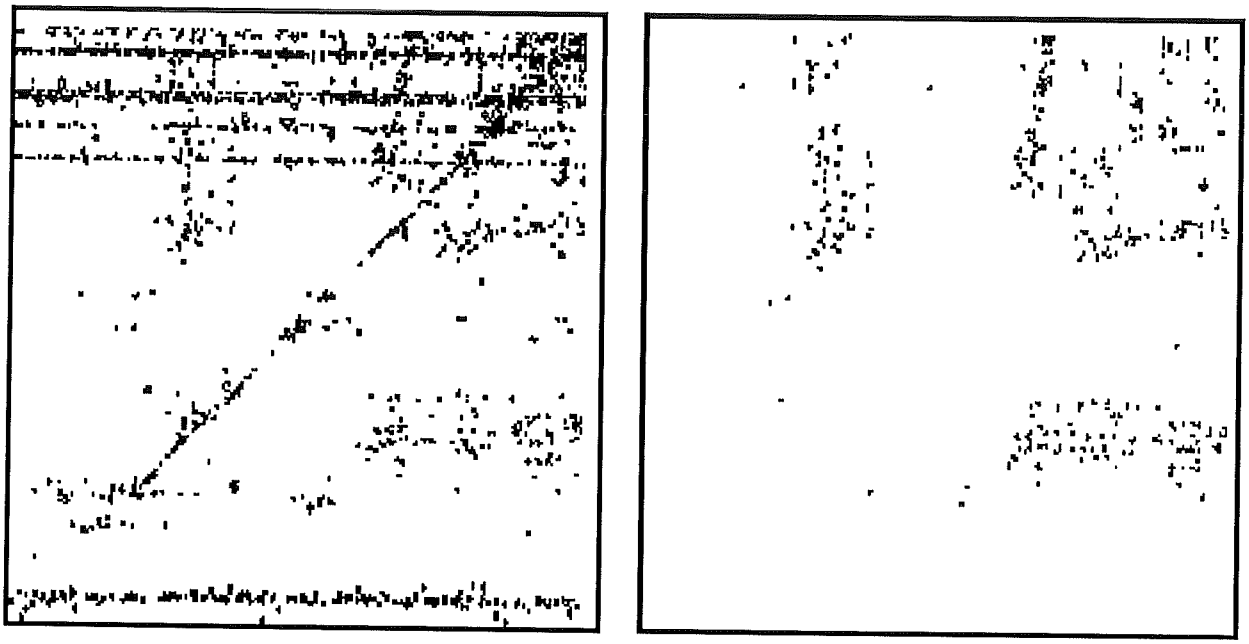
Protein NMR

The first NMR experiments with biopolymers were performed over thirty years ago. The potential of the method for structural studies of proteins was realised very early on. However, in practice, initial progress was slow because of limitations imposed by the instruments and the lack of suitable biological samples. In recent years there has been a huge increase in interest in the technique, primarily due to the development of two-dimensional NMR which makes the task of interpreting the data more straightforward (Cooke & Campbell 1988).

Conventional (one dimensional) NMR of proteins are densely crowded with resonance lines. There is no straightforward correlation between the NMR spectrum of the simple, constituent amino acids and the macromolecules. This makes it difficult to detect individual residues within the spectrum. As a consequence of the difficulties involved in interpreting such data, spectroscopists choose to produce two dimensional spectra of proteins and other biopolymers.

With 2D NMR the natural limitations of 1D NMR can largely be overcome. The main advantages of 2D NMR relative to 1D NMR for proteins are that connectivities between distinct individual spins are delineated, and the resonance peaks are spread out in two dimensions leading to a substantial improvement in peak separation, thus making the spectra far easier to interpret. The selection of techniques for the visualisation of the data from a 2D experiment is of considerable practical importance. Spectral analysis relies primarily on contour plots. Such plots are suitable for extracting resonance frequencies and for delineating connectivities via cross peaks, but care must be taken when attempting to extract quantitative information from such a plot.

Two main types of 2D experiment are important for proteins. One records through-bond interactions between H nuclei (COSY, HOHAHA) while the other detects through-space interactions (NOESY). We shall not go into the details of how these different experiments are performed, suffice it to say that the first pair of techniques allow one to study interactions occurring within amino acid residues while the second illustrates longer range interactions occurring between amino acid residues. An example HOHAHA spectrum for the 34 amino-acid polypeptide Nisin is shown in Figure 1.



A B

Figure 1 The two-dimensional HOHAHA spectrum of Nisin.
A - raw spectral data after processing by peak picking algorithm.
B - extent of the peak information used during interpretation.

We shall now describe how such techniques may be used to determine protein structure.

Protein Structure Prediction

The process of determining the structure of a protein by NMR relies on a chemical sequence for the protein (assumed to be correct) being available. Each residue in the protein will give rise to a characteristic set of peaks in the HOHAHA and COSY spectra and interactions between residues will lead to cross peaks in the NOESY spectrum. The interpretation of these spectra involves detection of the residue spin-systems in the HOHAHA and COSY, followed by analysis of the NOESY in order to link these spin-systems together.

The steps involved are:

i. The spin systems of individual amino acid residues are identified using through-bond ^1H - ^1H connectivities. Each spin system produces a pattern of signals within the HOHAHA and COSY spectra that is characteristic of one or more amino-acid residue. Often it is only possible to identify something as belonging to a class of residues. For small proteins it is usually possible to pick out all the spin systems despite there being many hundreds of protons contributing to the spectrum.

ii. Residues which are sequential neighbours are identified from observation of signals in the NOESY spectrum indicating sequential connectivities αN, NN and possibly βN.

iii. Steps i. and ii. attempt to identify groups of peaks corresponding to peptide segments that are sufficiently large to be unique in the primary structure (sequence) of the protein. Sequence specific assignments are then obtained by matching the segments thus identified with the corresponding segments in the chemically determined amino acid sequence. Note that for larger proteins, crystallographic data may also be used here.

iv. The occurrence of certain patterns of NMR parameters along the polypeptide chain is indicative of particular features of secondary structure. NOESY signals are used to detect interactions between residues in the protein.

For certain residues, the chemical shift values of the α, β, γ and other protons can be very similar, leading to the ordering of signals becoming confused. From the HOHAHA spectrum it is impossible to say which signals are due to which protons and in such a situation it is necessary to resort to a COSY spectrum as this makes explicit the "adjacent" protons. Thus, in order to perform a complete spin-system assignment, it is necessary to have both the HOHAHA and COSY spectra of the protein. The HOHAHA spectrum is used to identify the spin-systems while the COSY spectrum is used to identify trouble-

some α and β protons prior to the sequential assignment process.

The region of the spectrum displaying peaks due to interaction between the N and Cα protons in residues (approx. 7 - 9 ppm) is termed the "fingerprint" region, and all interpretations begin in this area.

The NOESY experiment may be "fine-tuned" to indicate only those interactions occurring within a certain distance, e.g. those occurring between adjacent residues. This type of experiment is used during the sequential assignment process. For the determination of secondary structure it becomes necessary to allow the longer range interactions to give rise to signals.

Secondary Structure Prediction

Non-sequential interactions also indicate secondary structure, e.g. interaction between the ith and ith+3 residues occurs in α-helices. Some of the short range NOESY interactions can also be used to indicate features of secondary structure. Accurate identification of the ends of a helix can be difficult. Table 1 is a summary of the type of interactions seen in the NOESY spectrum for particular secondary structures, together with a rough estimate of the peak intensity.

α helices	
αN (i, i+4)	weak
αβ (i, i+3)	strong
αN (i, i+3)	medium
NN (i, i+2)	weak
αN (i, i+2)	NONE
NN (i, i+1)	strong
αN (i, i+1)	medium
β sheet	
NN (i, i+1)	weak
α N (i, i+1)	very strong
extended form	
αN (i, i+1)	strong

Table 1 Common secondary structure NOESY interactions.

Coupling constant values (determined from the spectrum) can also be used to provide support for a particular structure.

3D Structure Determination

As we have seen, the NOESY spectrum can be used to indicate features of secondary structure. It can also be used to determine a tertiary (3D) structure for the protein. The NOESY data is converted into a set of limits on the distances between pairs of interacting protons. Generating a three dimensional structure from this data is not straightforward and a number of different approaches exist including distance geometry algorithms, molecular dynamics programs and systems employing geometric constraint satisfaction, such as the PROTEAN system (Hayes-Roth et al 1986).

System Design

We are in the process of developing a protein NMR assistant which will aid a spectroscopist in the identification of residue spin systems and the prediction of secondary structure. (We are not currently interested in the problem of tertiary structure prediction.)

Previous systems which have attempted to tackle this problem (Billeter, Basus & Kuntz 1988, Cieslar, Clore & Gronenborn 1988, Eads & Kuntz 1989) have relied upon the protein chemist to perform a large part of the process and make little or no use of chemical shift information.

Two previous attempts at inferring protein structure using AI techniques are CRYSALIS (Terry 1983) and PROTEAN (Hayes-Roth et al 1986). CRYSALIS attempted to infer the structure of a protein of known composition but unknown conformation using X-ray diffraction data. PROTEAN aimed to derive the conformation of proteins in solution by using NMR data to provide distance constraints. Both these systems make use of the blackboard architecture to integrate diverse sources of problem-solving knowledge and to partition the problem into manageable "chunks".

We are currently investigating whether such an approach would be appropriate for the task of interpreting 2D NMR of proteins. The characteristics of this task are: a large solution space; noisy data; likelihood of multiple, competing solutions; and the use of a number of cooperating sources of knowledge. This would seem to make it suitable for the blackboard approach.

The system currently consists of six main modules: SID (Spin-system IDentifier), CSA (Chemical Shift Analyzer), COSI (COSy Interpreter), SAM (Sequential Assignment Module), SLOC (Sequence LOCator) and STAN (STructure ANalyzer). We shall now describe each of these in turn.

SID uses the coordinate representation of the HOHAHA spectrum and attempts to identify residue spin-systems. Each peak in the spectrum is represented by an x, y coordinate indicating the peak centre and two other values indicating its size in the x and y directions. SID also uses the chemical sequence of the protein, represented by a list of the usual one letter residue abbreviations, to prevent residues absent from the protein being proposed.

SID uses a knowledge base containing a description of each of the twenty common amino acid residues and the approximate chemical shift values of each of their protons. Each of the

residues is represented by a frame containing a description of the protons found in that residue, represented by a list. For example, isoleucine is represented by the list [N Ca Cb Cg1 Cg1 Cg2 Cg2 Cg2 Cd1 Cd1 Cd1], i.e. 1 amide proton, 1 Cα, 1 Cβ, etc. Another slot contains a list of the approximate chemical shift values of each proton. Thus, for isoleucine, the chemical shift list is: [8.26 4.13 1.74 1.30 1.01 0.78 0.78 0.78 0.69 0.69 0.69], i.e. the amide proton has a value of approximately 8.26, the Cβ a value of 1.74, etc.

The approximate chemical shift values we are using were obtained from a statistical analysis of water soluble polypeptides and proteins (Groß & Kalbitzer 1988). It should be noted that the values are only approximate and are merely used as a guide to the likely nature of the spin-system.

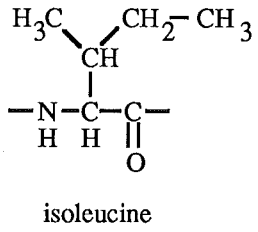
The spin-system identification process proceeds as follows. Beginning at the limit of the amide proton region of the HOHAHA spectrum (9.0ppm), a peak is selected that is close to the diagonal. All peaks with the same *x* coordinate as this peak (+/- some threshold value) are detected. SID then examines the spectrum for peaks in other regions with the same *y* coordinate as the peaks in this list. The set of peaks which are aligned in the NH region of the spectrum and which have companion peaks in other regions which are also aligned along a vertical, are then labelled as possibly belonging to the same spin-system. This list is then processed to remove all but one peak with any *y* coordinate value. The contents of this list correspond to the protons in an individual spin-system. This list is compared against the chemical shifts held in the SID knowledge base and all those residues which match the pattern of peaks are retained. By match here, we mean that the shift values for the spin-system peaks are equal to those in the residue chemical shift lists +/- some scatter parameter. We are assuming (for the moment) that the spectral data is complete, i.e. that each residue in the protein gives rise to the correct number of cross peaks and that there are no missing or extraneous peaks.

All peaks in the spectrum that have been assigned to a spin-system are labelled as such and a spin-system hypothesis created. This hypothesis holds the identification numbers and coordinates of each peak involved in a spin-system together with the name of the residue. A set of hypotheses are the output of the SID module. A typical hypothesis (for a serine spin-system) contains the following:

ser [57 305.9 661.7] [29 305.7 734.8] [8 660.7 735.6]

It is important that the N, Cα and Cβ protons are clearly labelled as it is the positions of these protons that are used by the sequential assignment module (SAM). Figure 2 shows the alignment of cross-peaks in the Nisin spectrum corresponding to an isoleucine residue.

One of the problems to be solved within SID is a means of resolving peak overlap, i.e. how to distinguish between a number of peaks which occur in very close proximity. It is obvious that for spin-system identification to be successful, such peaks must be differentiated. As we have already described above, SID currently assumes that no overlap occurs and that we have a "perfect" spectrum. Later versions of the



isoleucine

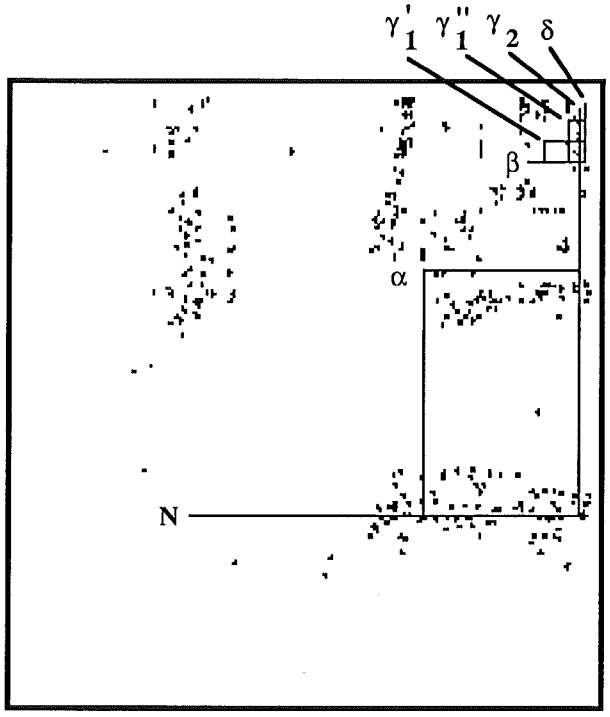


Figure 2 An isoleucine spin-system.

system will endeavour to cope with such messy data.

CSA examines spin-system hypotheses and uses the chemical shift data in its knowledge base in order to check whether the residue has Cα and Cβ protons which may be confused with other protons. If such confusion occurs, the hypothesis is labelled accordingly. e.g. for isoleucine residues, the Cβ and Cγ1' protons may easily be confused.

COSI examines the COSY coordinate map in an effort to distinguish between two such proton signals.It performs this task by using the list of coordinates of the protons in a spin-system. The *y* coordinates are used to detect the appropriate diagonal peaks in the COSY coordinate map. Cross peaks which occur between these diagonal peaks are then traced. The representation of the structure of the residue (described above) is then called upon and the system determines (based on knowledge about COSY interactions) which of the COSY cross peaks is due to each of the one step interactions. Thus, each of the important α, β and N protons is correctly labelled.

SAM uses the chemical sequence and the labelled spin-system hypotheses, together with a coordinate representation of the NOESY spectrum with an additional descriptor for each peak to provide intensity information. The sequential assignment process then proceeds as follows. The chemical sequence is examined and either a unique residue, or unique dipeptide segment (pair of adjacent residues) is detected. In the case of a unique residue, the system then looks through the spin-system hypotheses for a hypothesis corresponding to this residue. For dipeptides, one of the residues in the pair is selected and the appropriate hypothesis retrieved. The coordinates of the Cα peak are extracted and the NOESY spectrum examined for a cross peak with the same *y* coordinate. The *x* coordinate of this peak is then retrieved and the spin-system hypotheses examined for a N proton with the same *x* coordinate. This group of connected peaks corresponds to a αN short range interaction. If the search for an interaction is unsuccessful, then the coordinates of the N proton peak in the starting residue are used and if this fails, the Cβ peak is used. If such an interaction is detected, the two residues are labelled as adjacent and the process repeated using the spin-system hypothesis for the second residue. This continues until a 4 or 5 residue segment has been assembled.

SAM then selects another spin-system hypothesis and attempts to generate another 4 or 5 residue segment. This process continues until either all the spin-systems are labelled with a sequence position number or no further NOESY connectivities can be found.

SLOC When a peptide segment of 4 or 5 residues in length has been pieced together using the technique above, SLOC turns to the chemical sequence. The sequence is searched for a matching segment and the spin-system hypotheses used by SAM to generate the segment are labelled with the appropriate sequence position number. At this stage, uncertainties as to the exact nature of a residue spin-system are resolved using the sequence.

STAN uses the fully labelled spin-systems and a coordinate representation of the NOESY spectrum with an intensity descriptor for each peak. This module has access to a knowledge base containing information on the type of interactions expected for each secondary structure. This information is represented as a series of frames containing details of the type of protons involved, their relative positions in the sequence, the intensity of the signal and the secondary structure. e.g. to represent that an αN (i, i+4) interaction with weak intensity indicates an α-helix, a frame would contain the following: [a n 4 weak alpha].

References

M. Billeter, V.J. Basus & I.D. Kuntz, A Program for Semi-Automatic Sequential Resonance Assignments in Protein ^{1}H Nuclear Magnetic Resonance Spectra, Journal of Magnetic Resonace, 1988, **76**, 400-415

C. Cieslar, G.M. Clore & A.M. Gronenborn, Computer-Aided Sequential Assignment of Protein ^{1}H NMR Spectra, Journal of Magnetic Resonance, 1988, **76**, 119-127

R.M. Cooke & I.D. Campbell, Protein Structure Determination by Nuclear Magnetic Resonance, BioEssays, 1988, 8 (2), 52-56

C.D. Eads & I.D. Kuntz, Programs for Computer-Assisted Sequential Assignment of Proteins, Journal of Magnetic Resonance, 1989, **82**, 467-482

P. Edwards, D. Sleeman, G.C.K. Roberts & L.Y. Lian, An Intelligent Assistant for Protein NMR, Aberdeen University Computing Science Department Technical Report, AUCS/TR8910, 1989

K.H. Groß & H.R. Kalbitzer, Distribution of Chemical Shifts in ^{1}H Nuclear Magnetic Resonance Spectra of Proteins, Journal of Magnetic Resonace, 1988, **76**, 87-99

B. Hayes-Roth, B. Buchanan, O. Lichtarge, M. Hewett, R. Altman, J. Brinkley, C. Cornelius, B. Duncan & O. Jardetzky, PROTEAN: Deriving Protein Structure from Constraints, in Proceedings of the Fifth National Conference on AI (AAAI86), August 11-15, 1986, **2**, 904-909

A. Terry, The CRYSALIS Project: Hierarchical Control of Production Systems, Stanford University Technical Report, HPP-83-19, 1983