CHAPTER

11

An AI Approach to the Interpretation of the NMR Spectra of Proteins

Peter Edwards, Derek Sleeman,
Gordon C.K. Roberts & Lu Yun Lian

1 Introduction

The use of computers in chemistry and biochemistry has been widespread for many years, with machines performing many complex numerical calculations, e.g. solving quantum mechanical problems. However, some of the most interesting and challenging problems encountered in these domains are not numerical in nature. In particular, the interpretation or rationalization of many observed phenomena cannot be reduced to an equation or series of equations. Such problems are typically solved using intuition and experience and draw upon a great deal of empirical knowledge about the problem area. It is not surprising therefore, that these domains have proved such a fertile area for the application of artificial intelligence techniques. In this chapter we describe one such application, designed to assist a spectroscopist in the task of interpreting the Nuclear Magnetic Resonance (NMR) spectra of proteins [Edwards, 1989].

There are a number of scientific and medical applications of nuclear magnetic resonance spectroscopy and magnetic resonance imaging (MRI). The greatest impact of NMR in the chemical sciences has without doubt been in the elucidation of molecular structures. During the 1980s rapid developments in two-dimensional Fourier transform NMR made possible the determination of high quality structures of small proteins and nucleic acids. NMR spectrometers (in common with other laboratory experiments) invariably produce experimental data subject to noise, corrupted or missing data points, etc. User judgements in interactive processing of these data inevitably bias results, often unintentionally. The aim of the system currently under development is the automation of part of this task for NMR spectra of proteins. Our hope is that automation will limit the introduction of such user biases.

We now provide a brief introduction to proteins before describing the technique of nuclear magnetic resonance which can be used to elucidate the structure of such molecules.

2 Protein Chemistry

2.1 The Nature of Proteins

Proteins are probably the most diverse biological substances known. As enzymes and hormones, they catalyze 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, etc. Proteins come in all shapes and sizes and by the standard of most organic molecules, are of very high molecular weight. 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; their structures and names are shown in Figure 1. The amide linkages that join amino acids in proteins are commonly called peptide linkages and amino acid polymers are called polypeptides. Figure 2 shows a piece of protein backbone with the peptide linkages labeled.

2.2 Protein Structure

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; the secondary structure describes common structural arrangements of parts of the backbone; while the tertiary structure details the folding of these chains in three dimensional space.

Primary Structure The first stage in the process of protein structure pre-

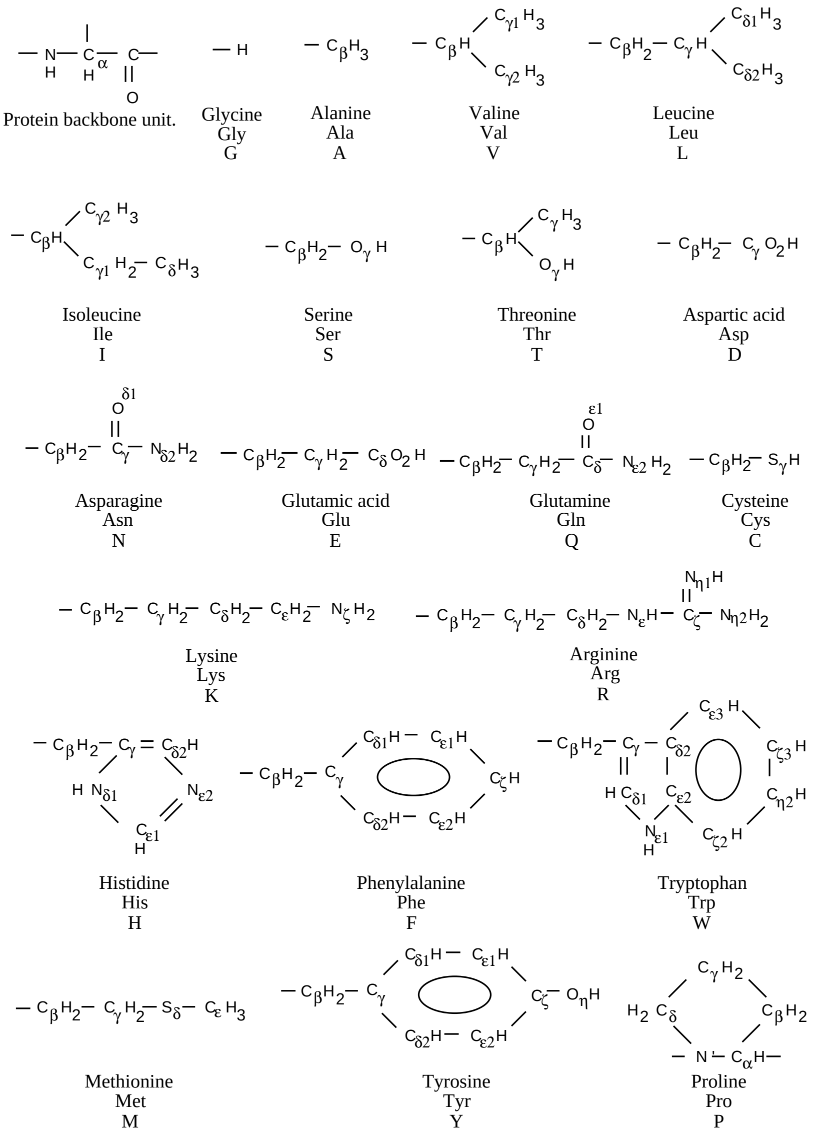

Figure 1 The protein backbone unit and the 20 amino acid side chains, shown with the three and one letter abbreviations for each. Proline is an imino acid, and its N and C_α backbone atoms are shown. Greek letters (α, β, γ, δ, ε, ζ, η) identify the distance (number of bonds) from the central (α) carbon atom. C=carbon, H=hydrogen, N=nitrogen, O=oxygen, S=sulphur atoms.

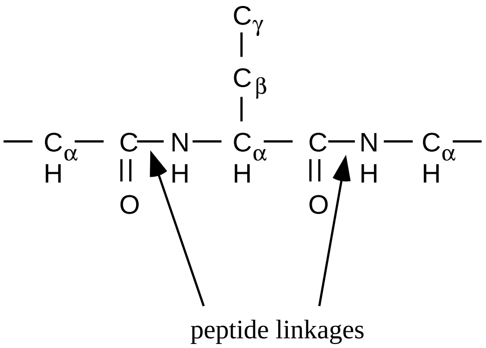

Figure 2. Protein backbone with α, β and γ carbons labeled. Peptide bonds link together adjacent amino acids. When an amino acid forms a peptide bond, two hydrogen atoms and one oxygen atom are released, and the remaining portion of the amino acid is called a residue.

diction is the determination of its primary structure, i.e., the linear arrangement of the amino acid residues within the protein. This is determined by chemical means.

Secondary Structure The major experimental technique that has been used in the elucidation of secondary structure of proteins is X-ray crystallographic analysis. When X-rays pass through a crystalline substance they produce diffraction patterns. Analysis of these patterns indicates a regular repetition of particular structural units with certain specific distances between them. The complete X-ray analysis of a molecule as complex as a protein can take many months. Many such analyses have been performed and they have revealed that the polypeptide chain of a natural protein can assume a number of regular conformations. Rotations of groups attached to the amide nitrogen and the carbonyl carbon are relatively free, and it is this property that allows peptide chains to adopt different conformations. Two major forms are the β sheet and α helix.

The β sheet consists of extended polypeptide chain with backbone residues forming hydrogen bonds between the chains. The sheet is not flat, but rather is pleated, in order to overcome repulsive interactions between groups on the side chains. The α helix is a right-handed helix with 3.6 amino acid residues per turn. Each NH group in the chain has a hydrogen bond to the carbonyl group at a distance of three amino acid residues. The side chain groups extend away from the helix. Certain peptide chains assume what is called random coil arrangement, a structure that is flexible, changing and statistically random. The presence of proline or hydroxyproline residues in polypeptide chains produces another striking effect. Because the nitrogen atoms of these residues are part of five-membered rings, the groups attached by the N - Cα bond cannot rotate enough to allow an α helical structure.

Tertiary Structure The tertiary structure of a protein is the three dimensional shape which arises from foldings of its polypeptide chains. Such foldings do not occur randomly: under normal environmental conditions, the tertiary structure that a protein assumes will be its most stable arrangement, the so-called "native conformation." Two major molecular shapes occur naturally, fibrous and globular. Fibrous molecules have a large helical content and are essentially rigid molecules of rod-like shape. Globular proteins have a polypeptide chain which consists partly of helical sections which are folded about the random coil sections to give a "spherical" shape.

A variety of forces are involved in stabilizing tertiary structures including the formation of disulphide bonds between elements of the primary structure. One characteristic of most proteins is that the folding takes place in such a way as to expose the maximum number of polar groups to the aqueous environment and enclose a maximum number of nonpolar groups within its interior. Myoglobin (1957) and haemoglobin (1959) were the first proteins whose tertiary structures were determined by X-ray analyzes.

3 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 realized 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 [Jardetzky, 1981; Wüthrich, 1986; Cooke, 1988].

NMR techniques are complementary to X-ray crystallography in several ways:

- NMR studies use non-crystalline samples e.g. solutions in aqueous or non-aqueous solvents. If NMR assignments and spatial structure determination can be obtained without reference to a corresponding crystal structure, a meaningful comparison of the conformations in single crystals and non-crystalline states can be obtained.
- NMR can be applied to molecules for which no single crystals are available.
- Solution conditions for NMR experiments (pH, temperature, etc.) can be varied over a wide range. This allows studies to be carried out on interactions with other molecules in solution.

We shall now define a number of terms commonly used by NMR spectroscopists.

Chemical shift defines the location of an NMR signal. It is measured relative to a reference compound. The chemical shift is normally quoted in parts per million (ppm) units and is primarily related to the magnetic environment of the nucleus giving rise to the resonance.

Spin-spin coupling constants characterize through-bond interactions between nuclei linked by a small number of covalent bonds in a chemical structure.

NOEs (Nuclear Overhauser Enhancement/Effect) are due to through-space interactions between different nuclei and are correlated with the inverse sixth power of the internuclear distance.

3.1 Two Dimensional NMR

Conventional (one dimensional) NMR spectra 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. There are a number of reasons for this, including the spatial folding of proteins, which has an effect on chemical shift values; and physical side-effects due to the size of proteins. As a consequence of the difficulties involved in interpreting such data, spectroscopists choose to produce two dimensional spectra of proteins and other biopolymers[1].

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 that resonance peaks are spread out in two dimensions leading to a substantial improvement in peak separation, thus making the spectra far easier to interpret.

Two main types of 2D experiment are important for proteins. One records through-bond interactions between ^{1}H nuclei (HOHAHA, COSY) 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 acids. Figure 3 shows a piece of protein backbone with selected through-bond and through-space interactions labeled.

The selection of techniques for the visualization of the data from a 2D experiment is of considerable practical importance. Spectral analysis relies primarily on contour plots of the type shown in Figure 4. Contour 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.

Limitations for the analysis of 2D NMR spectra may arise from a phe-

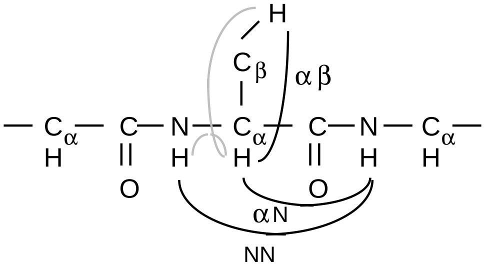

Figure 3. Protein backbone illustrating through-bond (dotted line) and through-space (solid line) ^1H - ^1H interactions.

nomenon termed “t_1 noise”, i.e. bands of spurious signals running parallel to the ω_1 axis at the position of strong, sharp diagonal peaks. These signals may arise due to spectrometer instability or other sources of thermal noise. They may also be an artifact of inadequate data handling during the Fourier transform. Ideally, NOESY or HOHAHA spectra should be symmetrical with respect to the main diagonal. In practice, however, noise, instrumental artifacts and insufficient digitization tend to destroy perfect symmetry. A number of 2D NMR experiments, including COSY and NOESY are described by Morris [1986].

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

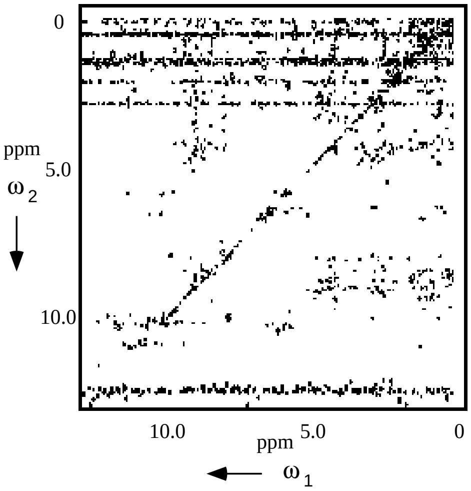

Figure 4. The two-dimensional HOHAHA spectrum of Nisin (a 34 amino-acid polypeptide)

4 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:

1. 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. (Section 4.1)

2. Residues which are sequential neighbors are identified from observation of signals in the NOESY spectrum indicating sequential connectivities[2] αN, NN and possibly βN. (Section 4.2)

3. Steps (1) and (2) attempt to identify groups of peaks corresponding to peptide segments that are sufficiently long 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.[3] Note that for larger proteins, crystallographic data may also be used here. (Section 4.2)

4. 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. (Section 4.3)

4.1 Assignment of Spin Systems

HOHAHA & COSY techniques A COSY spectrum consists of the conventional NMR signal along the diagonal and off-diagonal peaks (cross peaks) corresponding to ^1H - ^1H interactions. The peaks along the diagonal represent a normal spectrum of the system. Figure 5 is a schematic 2D plot showing the approximate positions of different types of protons along the diagonal.

With few exceptions COSY cross peaks are only observed between protons separated by three or less covalent bonds and thus are restricted to protons within individual amino-acid residues. Some of the 20 residues found in proteins give rise to unique patterns in the COSY spectrum. Not all residues

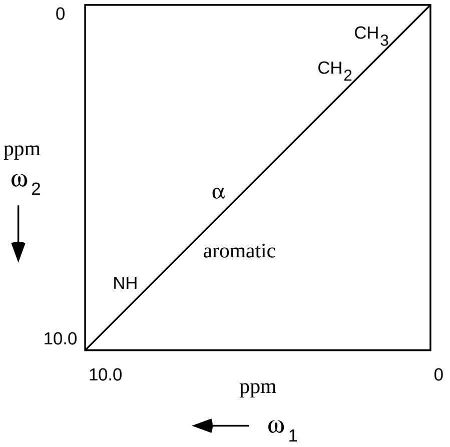

Figure 5. Positions of the different types of protons along the diagonal of a HOHAHA or COSY spectrum.

produce unique patterns as a number of them have similar structures and thus give rise to very similar COSY cross peak patterns. Often one can only identify something as belonging to a class of residues. NH protons and aromatic protons are relatively easy to identify. However, multiple methylene (CH_2) groups often cause problems as it becomes difficult to ascertain their ordering.

In order to make the signals easier to analyze, a variation on the standard COSY technique is employed, HOHAHA. Whereas COSY only shows interactions occurring between neighboring protons, such as αβ, βγ and so on, HOHAHA provides in principle, an overall picture by showing all ^1H - ^1H interactions occurring for each proton within the residue. Thus for a residue containing N, α, β and γ protons, the HOHAHA spectrum will contain a peak for each interaction with the N proton: Nα, Nβ, Nγ; a peak for each interaction with the α proton: αβ, αγ and so on. This technique has only been in routine use relatively recently, and has largely superseded COSY as it provides *additional* information. However, as noted below in point (3) it is often necessary to use these two techniques together. Thus, in the NH region of the HOHAHA spectrum one sees cross peaks due to each of the protons in the residue. In the Cα region one sees peaks caused by the Cβ, Cγ protons, etc. and in the Cβ region peaks resulting from Cγ, Cδ, etc. Figure 6a shows the HOHAHA spectrum for the threonine residue. Even with this technique we find that not all residues can be uniquely identified.

Correlations to δ protons are often not observed from the amide (NH) protons. The δ protons can however be observed in the Cα and Cβ regions and it is therefore quite common for different regions of the spectrum to be examined in order to detect the differing signals belonging to a spin system. In

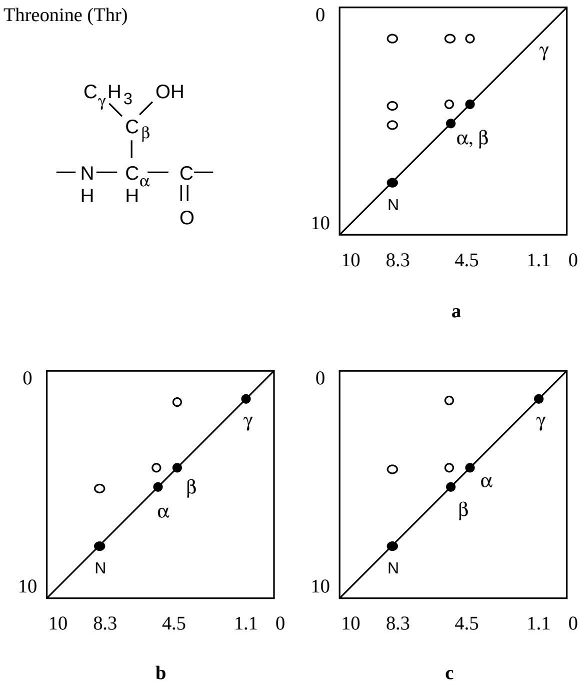

Figure 6. The HOHAHA spectrum (a) of threonine together with two possible forms of its COSY spectrum (b & c).

order to make the δ protons observable in the NH region it is necessary to adjust the experimental parameter known as the mixing time. Unfortunately, as this parameter is increased information starts to be lost from the spectrum due to relaxation processes.

For small proteins it is usually possible to pick out all of the spin systems despite there being many hundreds of protons contributing to the spectrum. The interpretation process begins with an attempt to assign the individual spin-systems within the HOHAHA spectrum. The region of the spectrum displaying peaks due to interaction between the N and Cα protons (approximately 3.8 - 5.5 / 7.6 - 9 ppm) is termed the “fingerprint” region and all interpretations begin in this area. Study of the HOHAHA spectrum of Nisin[4] shown in Figure 4 serves to illustrate the reason for this decision, as the resolution in this region is a great deal better than in the Cα or Cβ regions. Spin systems are detected by the following procedure:

1. Find a group of peaks which are aligned[5] along a vertical in the NH region.

2. As it is possible to have more than one set of spin system peaks on the

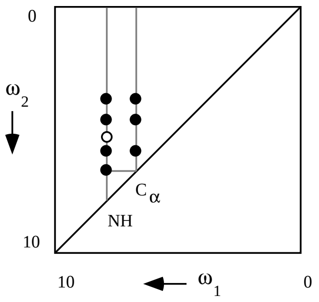

Figure 7. Detection of peaks belonging to the same spin-system. (The white circle indicates a peak that does not belong to the same spin-system as the others.)

same vertical, it is often necessary to resolve such overlapping systems. This is accomplished by the choice of a Cα signal in the NH region (Figure 7); a horizontal line is then traced to the diagonal; looking up the vertical from this point, all the Cβ, Cγ, etc. signals belonging to the same residue as the Cα are observed. If these signals are compared with the signals in the original part of the NH region, the group of peaks in that region belonging to the same spin-system should become obvious, as they will have a constant chemical shift value in the vertical direction.

3. 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 (Table 1). 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. Figures 6b and 6c show the COSY spectra for threonine with the chemical shift values of the α and β protons occurring in slightly different positions. As COSY only shows cross peaks for “one step” interactions it is quite easy to differentiate between the α and β protons, regardless of their chemical shift values; the α interacts with N and β, while the β interacts with α and γ. From the HOHAHA spectra in Figure 6a it is impossible to distinguish between these protons.

4. Often, all the signals for a particular spin system are not present, due to peak overlap and other effects. In such a situation the spectroscopist will often resort to “intelligent” guesswork based upon his knowledge of the technique to “fill the gaps.” This knowledge is used to provide a plausible NMR reason why signals do not appear and may, for example, involve decisions based upon the similarities of chemical shift values for individual protons.

5. Once a pattern of signals has been detected within the spectrum, it is la-

Residue	Protons
Arg	H_β 1.63 (.43) H_β' 1.79 (.34) H_γ 1.52 (.34) H_γ' 1.56 (.34)
Gln	H_β 1.92 (.27) H_β' 2.10 (.20) H_γ 2.29 (.25) H_γ' 2.35 (.20)
Glu	H_β 1.97 (.20) H_β' 2.04 (.18) H_γ 2.27 (.20) H_γ' 2.34 (.21)
Ile	H_β 1.74 (.37) H_γ' 1.30 (.32)
Leu	H_β 1.60 (.37) H_β' 1.71 (.31) H_γ 1.51 (.30)
Lys	H_β 1.74 (.38) H_β' 1.84 (.34) H_γ 1.30 (.39) H_γ' 1.36 (.37) H_δ 1.54 (.24) H_δ' 1.57 (.23)
Met	H_β 1.89 (.19) H_β' 2.03 (.21) $H_{\varepsilon 3}$ 1.98 (.21)
Pro	H_β 1.88 (.35) H_β' 2.18 (.40) H_γ 1.92 (.50) H_γ' 2.02 (.45)
Ser	H_α 4.50 (.47) H_β 3.72 (.44) H_β' 3.89 (.43)
Thr	H_α 4.53 (.43) H_β 4.17 (.31)
Trp	H_α 4.29 (.80) H_β' 3.42 (.22)

Table 1. Residues with protons which are difficult to identify. Each proton is followed by a mean chemical shift value, determined from a study performed on 20 proteins. The figure in parentheses is the standard deviation.

beled as having been produced by one or more of the 20 amino-acid residues. In the case of some patterns the spin system may only be labeled as belonging to a group of residues with similar structure, such as those with long side chains or aromatic groups.

Chemical shift values could be used to distinguish between the different residue types, but in practice such values are regarded as being too unreliable and are little used.

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 troublesome α and sidechain protons prior to the sequential assignment process.

This entire process is currently performed using a ruler and pencil (to link signals in the spectrum together) and can take several days of a spectroscopists time.

4.2 Connecting the Spin Systems

NOESY technique Depending on the actual settings used during the experiment, NOESY cross peak signals (off-diagonal peaks) can be obtained for pairs of protons at varying distances apart. Figure 3 illustrates the interactions that can occur between 2 adjacent residues.

Which through-space interaction is prevalent will depend upon the geometric shape of the protein. It is possible to get non-sequential NOE interactions due to hydrogen-bonded interactions between adjacent sheets, etc. The NMR experiment may be "fine-tuned" to indicate only those interactions occurring within a certain distance. For example, those occurring between adjacent residues. This is achieved by use of the experimental parameter, mixing time (τ_m). It is usual to set τ_m initially to exclude all but the shortest range NOEs which are due to sequential interactions and very short through-space interactions[6]. This type of experiment is used during the sequential assignment process. For the determination of secondary structure it becomes necessary to alter τ_m to allow the longer range NOEs to give rise to signals. The region 3.8 - 5.5 / 7.6 - 9.0 ppm is the "fingerprint" region of a NOESY spectrum (c.f. HOHAHA).

It is possible to set τ_m in order to exclude all but *one* sequential neighbor of each residue and the shortest through-space interactions. This technique is particularly useful for sequence confirmation experiments when segments of polypeptide chain can be constructed based on the spectrum and checked against the chemically derived sequence.

The sequential assignment process requires that the chemical sequence of the protein be available.

Sequential assignment Using the three sequential connectivities αN, NN, βN it is possible to "walk" the entire length of the residue chain. Using just one of these types of connectivity is often not sufficient, due to absent or overlapping peaks, etc. The HOHAHA and NOESY spectra both possess diagonal peaks corresponding to correlations between protons from the residue. As these peaks occur in the same positions in both spectra, this gives us a means of relating cross peaks in the NOESY to the spin systems identified in the HOHAHA. Figure 8 illustrates the process of sequential assignment using these techniques. One begins by selecting a diagonal peak, such as a Cα peak, in the HOHAHA spectrum which belongs to a known spin system (*d1*). *A* is an off-diagonal peak within that spin-system. The corresponding diagonal peak in the NOESY is then detected (*d1'*) and a horizontal line drawn away from the diagonal to find the NOESY off-diagonal peak. (If more than one peak is present along the horizontal, then they are all treated as possibly being due to sequential connectivity.) A vertical line is then drawn back to the diagonal (*d2*). The corresponding diagonal peak in the HOHAHA spectrum is then identified (*d2'*) and the fact that the two residues are adjacent is noted. *B* is an off-diagonal peak in the adjacent spin-system[7]. This process is repeated until a peptide segment of perhaps five or six residues has been detected, e.g.

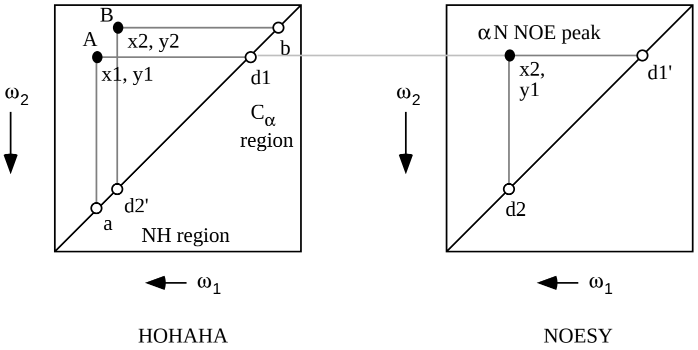

*Figure 8. Use of the HOHAHA and NOESY spectra to perform sequential assignments. HOHAHA peaks **a**, **d1** and **A** are due to residue R1, while peaks **b**, **d2'** and **B** are due to residue R2. NOESY peak **d1'** is due to R1 and peak **d2** to R2.*

GCA∗L , where ∗ indicates a residue which cannot be identified with absolute certainty from the HOHAHA/COSY spectra

The chemically determined amino acid sequence is then searched for peptide segments that will match the partial sequences identified by sequential NMR assignments. The sequence is needed at this stage to eliminate erroneous sequential assignment pathways, which may have resulted because non-sequential NOE connectivities have been interpreted as sequential ones[8]. In principle, all the information missing in incomplete spin system identifications can be obtained during the sequence specific assignment process. Patterns in the HOHAHA spectrum that have been labeled as one of a group of residues can be uniquely identified once their sequence positions are known. Once all the spin systems in the HOHAHA have been fully identified, they are labeled with their residue name and sequence position.

The sequence specific assignment technique described here works well for small proteins up to approximately 100 residues. If there are too many residues, signal overlap becomes a major problem. For larger proteins it is often necessary to use crystallographic data to help with NMR assignments.

4.3 Secondary Structure Prediction

We have already seen that the short range NOESY interactions allow us to determine which of the residues detected in the HOHAHA spectrum are adjacent. This same information can also be used to indicate some features of secondary structure. Non-sequential interactions also indicate secondary structure, e.g. interaction between the *ith* and *ith+3* residues is seen in α-helices. Accurate identification of the ends of a helix can be difficult. Table 2 summarizes the type of interactions seen in the NOESY spectrum for partic-

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

Table 2. Common secondary structure NOESY interactions.

ular secondary structures, together with a guide to the peak intensity.

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

4.4 3D Structure Determination

As we have seen, the NOESY spectrum can be used to indicate features of secondary structure. They 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. Tables containing all internuclear distances in the protein are constructed with the spectroscopic information used to provide some of the entries and the geometry of common structural features used to provide others. Upper and lower limits are recorded for each distance. It should be noted that this is an approximate technique as there is no straightforward mapping from NOE to distance (as the general environment complicates the signal). Strong, medium and weak NOEs are taken to indicate upper distance limits of 2.5, 3.5 and 4.5Å respectively. Known molecular bond lengths, bond angles and standard geometries are used to provide interatomic distances for atoms separated by one or two bonds. Peptide dihedral angles obtained from coupling constants can also be recast as limits on distances between atoms separated by three bonds. A lower limit on interatomic distances is normally set as the sum of the van der Waals radii.

Generating a three dimensional structure from this data is not straightforward and a number of different approaches exist including Distance Geometry algorithms [Havel, 1983], Molecular Dynamics [Hermans, 1985] and systems employing geometric constraint satisfaction, such as the PROTEAN system (see below). Distance Geometry and Molecular Dynamics are examples of methods within the *adjustment paradigm* for interpretation of NMR data [Altman, 1988], i.e. they generate starting structures, usually at random, and then search the neighboring conformational space until the mismatch between the data predicted from the adjusted structure and the experimental data is minimized in terms of some chosen function. By contrast, methods within the *exclusion paradigm* generate starting structures in a systematic manner and test them for agreement with the given data set. All structures compatible with the data are retained as possible solutions, and all incompatible structures are excluded from further consideration.

Distance Geometry algorithms work with distances between points rather than Cartesian coordinates. They allow the choice of three mutually perpendicular axes to be made such that a "best fit" emerges as a 3D description of the structure. This fit usually contains some small incompatibilities with the distance information. These are minimized according to user supplied criteria, often some kind of energy relaxation calculation is required to relieve strain in the structure. Alternative solutions are generated by repeating the calculation with a random choice for the distances, each somewhere within its limits. The effect is to sample the conformation space. Confidence in a solution grows if repeated calculations arrive at a similar end point.

The input to Molecular Dynamics programs consists of the covalent structure of the molecule and a number of energetic terms, e.g. energy to stretch bonds, energy for van der Waal's repulsion. Other energy terms are linked to distance constraints. The program then solves Newton's equations of motion using the energy terms. Balance between the energy terms is important. There is no known test for uniqueness, but confidence increases if repeated simulations from different starting points converge to give a similar final result.

5 Computational Aspects of NMR

The earliest use of computers in NMR was for time-averaging of multiple scans in the mid-1960s. Systems for performing the first Fourier transforms in commercial NMR instrumentation appeared in 1969. From the early 1970s the majority of NMR instruments were interfaced to minicomputers which controlled data acquisition and performed FFT (Fast Fourier Transform) and standard post-FT processing. By the mid-1970s NMR instrumentation was designed around 16-20 bit word minicomputers with low resolution colour graphics and digital plotters used for output. In the early 1980s NMR instrument computers began to be replaced by modern microcomputer and minicomputer systems, augmented by array processors. The current generation of NMR instruments incorporate microcomputers performing tasks ranging from sample temperature control to data acquisition and supervision and control of the user interface. Most NMR instruments make use of high resolution graphics for data display. The current trend is to perform data reduction away from the spectrometer using general-purpose commercial workstations. Levy [1986] reviews some of the computational aspects of NMR.

The application of computers to NMR can be separated into two areas: 1) data acquisition and experiment control and 2) data reduction. The task of acquiring data and performing control over the spectrometer is handled by computers embedded in the instrumentation, usually through proprietary software that is not available to the user. The timing of experimental events, pulse programming and so on occurs on a rapid time scale. The data reduction task, on the other hand, has relatively light real-time constraints. Data reduction is usually performed using ex-spectrometer computers, which facilitate the use of new data reduction techniques and which remove lengthy processing from the instrument and thus lead to an increase in spectrometer throughput. The most common language in scientific computing remains FORTRAN, although recently C has begun to be widely used also. Artificial intelligence languages such as LISP, PROLOG and POP-11 are finding use in scientific software, but as yet only on a very small scale.

5.1 AI Applications & NMR

Chemistry was one of the first disciplines, aside from Computer Science, to actively engage in research involving AI techniques. The Dendral project [Carhart, 1977; Lindsay, 1980; Smith, 1981; Djerassi, 1982] is almost certainly the most well-known of these attempts to use AI for chemical applications, and aimed to develop computer programs to assist structural organic chemists in the process of structure elucidation. Dendral was the first major application of heuristic programming to experimental analysis in an empirical science, a practical problem of some importance. It was the first large scale program to embody the strategy of using detailed, task-specific knowledge about the problem domain as a source of heuristics, and to seek generality through automating the acquisition of such knowledge. The structure elucidation process involves a number of steps. First, chemical and spectroscopic data (including NMR data) are interpreted to provide a number of structural constraints. These constraints are substructures that must either be present, or absent from the molecule under investigation. All possible candidate structures consistent with these constraints are then generated. Additional discriminating experiments are then planned so that the one correct structure can be determined.

Heuristic Dendral was constructed from a simple acyclic structure generator and a planning module (the preliminary inference maker) that performed a classification based on mass spectral data. The early version of the system dealt only with ketone molecules while subsequent versions of the system were extended to handle additional classes of molecules such as ethers and amines [Schroll, 1969]. At the same time other spectral data were incorporated into the system in the form of one dimensional ^{1}H NMR spectra.

Another aspect of the Dendral project was the Meta-Dendral system [Buchanan,1971; Buchanan, 1973]. This arm of the project was concerned with the production of useful tools for chemists at a lower level than the complete structure elucidation system. The system was originally devised for the analysis of mass spectral data although it was extended by Mitchell and Schwenzer [Schwenzer, 1977; Mitchell, 1978] to the analysis of ^{13}C NMR data. The principles governing ^{13}C NMR are similar to those of ^{1}H spectroscopy, although the scale of observed shifts is greater for the former. Again, as in ^{1}H NMR, the precise chemical shift of a nucleus depends on the atom or atoms attached to it. The system generated rules which relate precise ^{13}C shift ranges to specific environments for the resonating carbon atom. The chemical shift range associated with a particular environment is found by matching the generated structure against a training set of molecules and their spectra. The minimum and maximum values of the shift corresponding to that environment are recorded and form the range used in the rules. Goal states can be characterized by various criteria such as requiring a rule to have a sufficiently narrow range or to be supported by a minimum number of examples in the training data. The system begins with a very primitive substructure (e.g. a simple carbon atom) and a correspondingly vague chemical shift range ($-\infty \rightarrow +\infty$). Operators modify this structure by adding hydrogen, carbon, and so on The generated rules are used to predict spectra for a set of candidate molecules and the structures ranked by comparison of the predicted spectrum with that of the unknown.

The use of a database of ^{1}H NMR data to eliminate incompatible candidates from the list of structures produced by exhaustive generation of isomers is described by Egli [1982]. Structures obtained by a generator program are evaluated by prediction of their ^{1}H NMR spectra. The predicted and observed spectra are then compared and the candidates ordered based on such comparisons. The approach to spectrum prediction is strictly empirical and involves the derivation of a set of expected chemical shifts for the protons in each candidate. Egli describes a suite of programs which allow a user to build and maintain a ^{1}H NMR database that correlates substructural environments with observed proton resonances; to predict the spectrum of one or more candidate structures for an unknown compound; to compare the predicted and observed spectra of the molecule and to order the candidates based upon this comparison.

A similar database of ^{13}C NMR correlations containing 10,350 distinct substructure/chemical shift pairs is described by Gray [1981]. This database is also used for prediction of spectra for generated structures. It is also used to perform the interpretation of the ^{13}C NMR spectra of unknown molecules (to arrive at a set of substructural fragments). This interpretation is performed so as to arrive at the minimal, internally consistent set of substructures.

The use of structural constraints provided by two-dimensional NMR is described by Lindley [1983]. Partial structures obtained from the two-dimensional NMR spectrum are combined with other spectral data in an effort to elucidate the correct structure of an unknown molecule. All the constraints are provided by a chemist, who is required to interpret the spectroscopic data.

A number of workers (other than those involved in the Dendral project) have addressed the problem of constructing computer programs to automate or semi-automate the task of structure elucidation. The use of a number of different techniques for computer-assisted structure elucidation is described in Hippe [1985]. These include library-search algorithms which perform the comparison of an unknown spectrum with those in a standard collection stored on disc. Such algorithms typically return a list of spectra and their associated structures ranked according to some matching function. Hippe also describes integrated methods of structure elucidation. Three major components are common to all systems which attempt the structure elucidation task. First, some interpretation of the chemical and spectral data is performed, in order to derive structural fragments. The next step involves molecule assembly, i.e. the generation of complete structures compatible with the fragments and constraints provided by the first phase. Finally, spectra of the generated structures are simulated and compared with the observed data. This allows structures to be ranked on the basis of the quality of the fit between predicted and observed data.

The CASE system [Shelley, 1977; Shelley, 1981; Munk, 1982] is a suite of programs designed to accelerate and make more reliable the entire process of structure elucidation.The task of reducing chemical and spectroscopic data to structural information is currently shared by the chemist and the system. Interpreters capable of detecting the presence of structural fragments based on infrared and ^{13}C NMR data [Shelley, 1982] exist. Two-dimensional NMR data may be used to provide information about the connectivity of atoms in a molecule. The INTERPRET2D module of the system [Christie, 1985] accepts 2D-NMR data input by the chemist and generates the structural conclusions consistent with this information as a set of alternative fragment sets. These sets describe all possible carbon-carbon atom connections consistent with the data and may also be used as input to a structure generator program.

CHEMICS [Sasaki, 1971; Yamasaki, 1977; Sasaki, 1981] uses ^{1}H NMR, ^{13}C NMR, infrared and ultra-violet data to decide which of a set of 150 structural fragments are present in an unknown molecule. The fragments believed to be present are arranged into sets which satisfy the molecular formula and ^{1}H and ^{13}C NMR spectra. A structure generator uses these sets as input to create molecular structures. CHEMICS analyzes the ^{1}H NMR data by first calculating the area of each group of signals in the spectrum. The number of protons associated with each group is thus assigned. Recognizable spin-system patterns are then identified. The most probable structural fragments are then inferred based on chemical shift values. The number of each of these fragments is estimated based on the peak area values. ^{13}C NMR data is interpreted as follows: first, the number of carbons associated with each peak in the spectrum is computed, based on signal intensities; next, the splitting of each peak is examined and individual signals are labeled as arising due to protonated or non-protonated carbons and the number of protons on each carbon recorded. Finally, based on the information already extracted together with the chemical shifts of the signals, a set of structural fragments consistent with the information are obtained.

The STREC system [Gribov, 1977; Gribov, 1980; Elyashberg, 1987] also uses the plan, generate and test approach. During the plan phase, infrared and ^{1}H NMR data are examined and a set of plausible fragments computed. A generator uses these fragments to generate all possible structural isomers. Each structure is then checked against a library of structural fragments for which spectroscopic data are available. The fragments detected have their characteristic spectral information compared with the experimental data for the unknown. If the experimental data do not confirm the presence of the fragment, analysis of that structure is terminated. Each fragment in the library has data for infrared, ^{1}H NMR, ultra-violet and mass spectra. STREC2 [Gribov, 1983] is an enhanced version of the original STREC system, capable of handling larger structures, which makes use of ^{13}C NMR data in addition to the techniques described above.

SEAC (Structure Elucidation Aided by Computer) uses infrared, ^{1}H NMR and ultraviolet data to infer the structure of an unknown molecule [Debska, 1981]. A system for the interpretation of infrared, ^{13}C NMR and mass spectral data, based on the idea of intersecting the interpretations of each of these techniques has been developed by Moldoveanu [1987]. Each of the three spectra is interpreted to generate three sets of plausible fragments; the intersection of these sets is then found and the resulting group of fragments is output to the user. The output also indicates the number of each of these functional groups present in the molecule and the possible positions of substitution of these groups in the unknown molecule.

Knowledge-based techniques have also been applied to the interpretation of other kinds of spectroscopy, including gamma ray activation spectra [Barstow, 1979; Barstow, 1980], ESCA (Electron Spectroscopy for Chemical Analysis) [Yamazaki, 1979], X-ray fluorescence spectroscopy [Janssens, 1986] and X-ray diffraction spectra [Ennis, 1982].

5.2 Computational Aids for Protein NMR

A number of attempts have been made to automate part of the protein structure determination process. One of these systems [Billeter, 1988] starts from a well-defined list of spin-systems which have been identified by the user. The program considers all possible assignments that are consistent with the data currently available. If new data are provided the program eliminates assignments that are inconsistent. It performs logical decision-making and bookkeeping functions and avoids making ambiguous decisions when multiple assignments are possible. Uncertain decisions, i.e. decisions based on NMR data that do not allow a unique interpretation are left to the user. Another system, developed by Cieslar [1988], identifies potential spin-systems within the HOHAHA spectrum by locating aligned peaks. However, attaching residue labels to these spin-systems is left to the user. Once spin-systems have been labeled, the program endeavors, through the use of NOESY signals, to identify sequential connectivities. Partial sequences are identified in this manner and then located within the chemical sequence. The system then constructs all possible assignments for all partial sequences that are consistent with the input data. In order to achieve consistency the partial sequences must not contain overlaps and no particular spin-system should be used in more than one position. All solutions for the assignment of the complete sequence are then generated and checked by the system. Eads [1989] describes a suite of programs which assist in the sequential assignment process and which use peak coordinates and intensity values directly as input. The programs trace spin-systems out to the β protons, look for NOESY cross peaks between relevant protons and create lists of sequential spin-systems. Tracing the spin-systems beyond the β protons and establishing correspondence with the primary sequence is left to the user.

The ABC system [Brugge, 1988] automates the process of determining secondary structure from NMR data. The program is able to identify α helical and β strand segments of chain by means of a set of qualitative criteria that are used in analyzing data derived from the NMR spectra. ABC is implemented within the BB1 architecture [Hayes-Roth, 1988]. Input to ABC consists of the primary sequence of the protein, lists of observed NOEs and residue information. The output of the program is a set of secondary structure elements, defined by their extent over the primary sequence. Each structure is also labeled with the evidence used to derive it and pointers to partial structures from which it was constructed. The output of ABC can be used as part of the input to programs for determining the tertiary structure of proteins. ABC has been tested using published data on nine different proteins and its ability to locate regions of secondary structure, and its precision in defining the extent of these regions have been measured. The system performs well and comes close to reproducing the results of expert analysis of NMR data.

The PROTEAN system [Lichtarge, 1986; Altman, 1988; Altman, 1989] is based on the exclusion paradigm described earlier. Its purpose is to sample the conformational space of a protein systematically and to determine the entire set of positions for each atom that is compatible with the given set of constraints. To maintain computational feasibility, PROTEAN solves the protein structure problem in a hierarchical fashion. The program uses knowledge of the protein sequence together with NMR data to determine the secondary structure. It next defines the coarse topology of the folded structure and then specifies the spatial distribution of atomic positions using a description of accessible volumes. From these values, the original data are predicted to verify the resulting family of structures.

The secondary structure of the protein is determined using the ABC system described above. The units of secondary structure and a set of experimental constraints (primary structure, NOE distances, surface and volume information) are then passed to the SOLID GS module. This computes the accessible volume for the units of secondary structure. SOLID GS uses abstract representations to reduce the number of objects whose positions need to be sampled. For example, helices are represented by cylinders. The next module, ATOMIC GS, refines the secondary structures and coils using discrete sampling for atoms. The output of this module is then processed by another (KALMAN) which employs a probabilistic refinement method[9] for determination of the uncertainty in each atom. The final component of the system, BLOCH, calculates NMR data and evaluates the match between observed and predicted values. The system has been used to investigate the tertiary structure of the *lac-repressor headpiece*, a protein with 51 amino acid residues. The structural solution proposed by PROTEAN closely matches that proposed by a manual interpretation of the data performed by an expert protein spectroscopist.

6 The Protein NMR Assistant

We are in the process of developing a Protein NMR Assistant (PNA) 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 addressed this problem, such as those described earlier, have tackled only part of the task and have left much of the interpretation to the spectroscopist. PNA aims to provide a complete system for the identification and assignment of spin-systems, leading to the prediction of secondary structure.

Two previous attempts at inferring protein structure using AI techniques are CRYSALIS [Engelmore, 1979; Terry, 1983] and PROTEAN [Hayes-Roth, 1986]. CRYSALIS attempted to infer the structure of a protein of known composition but unknown conformation using X-ray diffraction data. Both these systems made 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.

6.1 The Blackboard Architecture

A blackboard system consists of three main components: the blackboard, a set of problem-solving knowledge sources and a control mechanism. The blackboard serves to partition the solution space of the problem domain into one or more domain-specific hierarchies, representing partial solutions. Each level in the hierarchy possesses a unique vocabulary that serves to describe the information at that level. Objects on the blackboard can be input data, partial solutions as well as final solutions and possibly control information. Relationships between objects are denoted by named links. Domain knowledge is partitioned into separate modules which transform information at one level into information on the same or different levels. These modules are termed knowledge sources (KSs) and perform transformations using rules or procedures. The KSs are separate and independent. Each KS is responsible for knowing the conditions under which it can contribute to the solution and thus has a precondition which indicates the conditions on the blackboard that must exist before the main part of the KS can be activated. The choice of which KS to use is based on the state of the solution, the latest additions and modifications to the solution and on the existence of KSs capable of improving the state of the solution. A controller monitors the changes to the blackboard and decides what to do next. The solution evolves one step at a time with any type of reasoning step (data-driven, goal-driven and so on) being applied at each stage.

For a particular application it is necessary to define a solution space and the knowledge needed to find the solution. This space is divided into levels of analysis corresponding to partial solutions and the domain knowledge is divided into specialized KSs that perform the subtasks necessary to arrive at a final solution. How the problem is partitioned into subproblems makes a great deal of difference to the clarity of the approach, the resources required and even the ability to solve the problem at all. This discussion has been necessarily brief; a number of excellent articles on this subject exist [Hayes-Roth, 1983; Nii, 1986a; Nii, 1986b], together with two books [Engelmore, 1988; Jagannathan, 1989].

A blackboard system can serve as a powerful research tool, allowing the solution space and domain knowledge of an application problem to be partitioned in different ways and a variety of reasoning strategies to be evaluated. The robustness of blackboard systems stems primarily from the way in which they are organized which tends to localize changes. The answer produced by a blackboard system is often a complex datastructure, different parts of which may have been computed through different reasoning paths. A trace of the system's execution history is unlikely to prove very useful to the user. We are addressing the problem of visualization of results as part of the development of the current system.

The blackboard architecture has been used in a wide variety of applications, including speech understanding [Erman, 1980]; submarine detection [Nii, 1982]; image understanding [Nagao, 1979]; and computer controlled manufacturing [Ayel, 1988]. A number of generalized architectures have also been developed to allow blackboard application systems to be constructed more easily. Examples of such tools include: AGE [Nii, 1979], Hearsay-III [Balzer, 1980], BB1 [Hayes-Roth, 1988], GBB [Corkill, 1986], PCB [Edwards, 1990]. The PCB system is a problem-solving architecture designed to ease the construction of complex knowledge-based systems in chemical domains. Although the system we shall describe below is not built within this framework, its design and implementation owe much to the PCB system.

We shall discuss the Protein NMR Assistant in terms of the components of the blackboard architecture described above, i.e. the blackboard (its levels and objects), the knowledge sources (structure and function) and control. The system architecture is shown in Figure 9.

6.2 The PNA Blackboard

The blackboard is divided into five levels (as shown in Figure 9): data, spin-system, segment, labeled residue and secondary structure. The contents of each level are as follows:

data: Spectroscopic data (HOHAHA, COSY and NOESY) plus the chemical sequence.

spin-system: Hypotheses describing the identification of residue spin-systems within the HOHAHA spectrum.

segment: Partial sequences of 5 or 6 residues assembled from the spin- system hypotheses.

labeled residue: Fully labeled residue hypotheses each of which describes the sequence position of a residue, together with the spectroscopic data used to identify it.

secondary structure: Units of secondary structure identified through examination of the NOESY spectrum.

Objects on each of these levels are represented using a frame-based representation. The chemical sequence is represented by a frame containing a number of slots, the first of which contains the full sequence represented as a

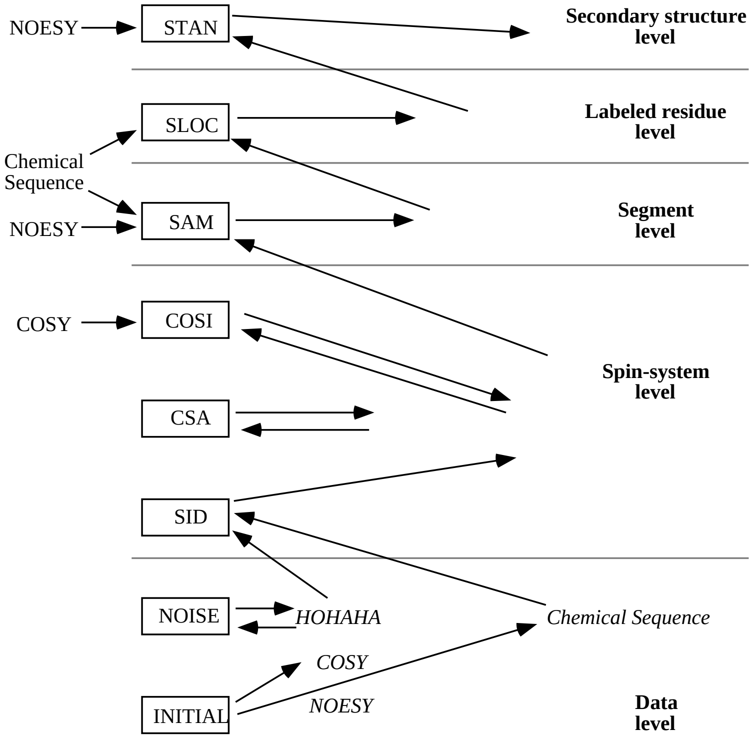

Figure 9. The Protein NMR Assistant system blackboard architecture.

list of the usual one letter abbreviations. Other slots contain the length of the sequence and the number of each of the twenty amino-acid residues present within the sequence. The spectroscopic data is represented on a peak-by-peak basis. Each peak in the HOHAHA spectrum is represented using the following slots: *id* (unique identification number for the peak); *xcoord* (x co-ordinate of the peak center); *ycoord* (y coordinate of the peak center); *xsize* (“width” of the peak in the x direction); *ysize* (“width” of the peak in the y direction); *peak-type* (label indicating whether or not the peak is noise); *infers* (list of the spin-system hypotheses which the peak is associated with).

Objects on the spin-system level define the nature of the residue spin-systems identified from the HOHAHA spectrum. Each hypothesis contains the following slots: *infers* (list of segment hypotheses supported by the spin-system); *supported-by* (list of HOHAHA peaks which make up the spin-system); *residue-type* (name of the amino-acid residue giving rise to the spin-system); *peak-list* (identification numbers and coordinates of the HOHAHA peaks which make up the spin-system); *diagonal-peaks* (positions of each of the diagonal peaks involved in the spin-system); *protons* (label indicating whether the COSY spectrum should be used o distinguish between peaks in the spin-system). If the system is unable to uniquely assign a spin-system to a particular residue, the *residue-type* slot contains a list of the possible residues associated with that spin-system, instead of an individual residue name.

Segment level hypotheses also contain *infers* and *supported-by* slots. *infers* is used to indicate which of the objects on the labeled residue level the segment hypothesis has provided evidence for, while *supported-by* lists those spin-systems which were connected to form the segment. Other slots within the segment hypotheses are: *segment-sequence* (the partial sequence stored as a list); *noesy-links* (peak data indicating the sequential connectivities for each of the residue pairs in the segment). In the event that a residue is not uniquely identified on the spin-system level, the *segment-sequence* slot will contain a list of possible residues in place of a single residue.

A fully labeled residue hypothesis contains all the information associated with the identification of a residue. As well as *infers* and *supported-by* slots (which indicate which secondary structure unit the residue is involved in and which segment supports it), objects on this level also contain the following: *residue-type* (residue name); *sequence-position* (position of the residue within the chemical sequence); *peak-list* (identification numbers and coordinates of the HOHAHA peaks which comprise the residue spin-system); *diagonal-peaks* (positions of each of the diagonal peaks involved in the residue spin-system); *noesy-links* (NOESY peak data used to assemble the segment in which the residue occurs).

The final level of the PNA blackboard contains the secondary structure hypotheses. These objects detail the exact nature and extent of any secondary structure unit identified within the protein. Structural hypotheses contain the following: *supported-by* (list of labeled residues which make up this unit); *structure-unit* (type of unit, i.e. α helix, β sheet); *start* (position in the chemical sequence at which the unit commences); *finish* (position in the sequence where the unit terminates); *spatial-noesy* (NOESY interactions used to infer the presence of the unit). In cases were there is uncertainty as to the exact point in the sequence where the structural unit begins or ends, the *start* and *finish* slots contain lists of residues, indicating a region of the protein sequence.

6.3 The PNA Knowledge Sources

The system currently consists of eight knowledge sources: INITIAL (Initialization) NOISE (Noise removal), 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 KSs in turn.

INITIAL The first of the PNA KSs deals with the initialization of the blackboard and with the loading of spectroscopic and chemical sequence data. The coordinate data representing the HOHAHA, COSY and NOESY spectra are held in text files which are compiled by this KS into the internal representation described above. The chemical sequence is also compiled from a file containing the one letter residue symbols. In addition to loading the sequence, INITIAL also calculates its length and the number of each of the amino-acid residues that are present within it.

NOISE As described earlier (Section 3.1), the HOHAHA spectrum contains bands of noise (t_1 noise) which run parallel to the ω_1 axis. Before the system attempts to identify spin-systems within the spectrum, it first uses the NOISE KS to identify peaks which may be due to noise. NOISE examines the spectroscopic data and searches for groups of peaks which run parallel to the ω_1 axis, i.e. peaks which possess approximately the same *y* coordinate value. These peaks, once identified, have *noise* written to their peak-type slot. This information is then used by the other KSs during analysis of the spectrum.

SID This KS uses the coordinate representation of the HOHAHA spectrum[10] together with the chemical sequence of the protein and attempts to identify residue spin-systems. The chemical sequence is used in order to prevent residues absent from the protein being proposed. SID contains knowledge describing 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, 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 (9.0 ppm), 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 labeled as possibly belonging to the same spin-system. This list is then processed to remove all but one peak with any

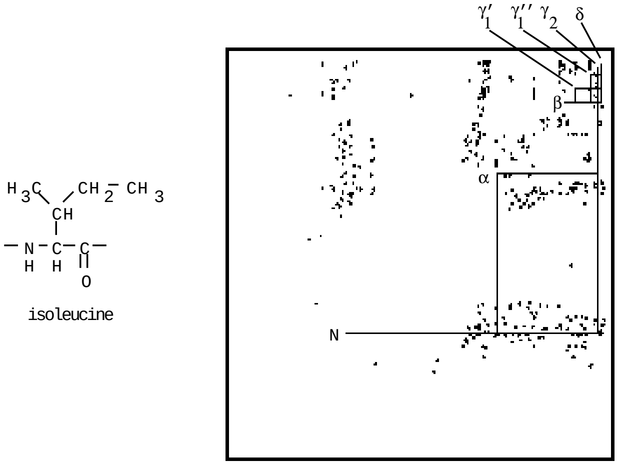

Figure 10. The identification of an isoleucine spin-system.

y coordinate value. The contents of this list correspond to the protons in an individual spin-system. This list is compared against the list of chemical shifts held by SID for each residue 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 labeled 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. As we have already seen (Table 1), it can be difficult to detect the Cα and Cβ protons of certain residues due to very similar chemical shift values for different protons. It is important that the N, Cα and Cβ protons are clearly labeled as it is the positions of these protons that are used by the sequential assignment module (SAM). Figure 10 shows the alignment of cross-peaks in the Nisin spectrum corresponding to an isoleucine residue, while Figure 11 contains the spin-system hypothesis created by PNA to describe the identified 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.

CSA This KS examines spin-system hypotheses which have been created

```
spin-system hypothesis 23:

     infers :             [ ]
     supported-by :       [ peak835    peak1076   peak1082   peak1085   peak1086
                            peak1153   peak1154   peak1285 ]
     knowledge-source :
                          [ spin-system-id ]
     residue type :       ile
     protons :            [ Cb  Cg1 ]
     peak list :          [ 835  310.8  712.34 ]        [ 1076  889.76  920.13 ]
                          [ 1154  923.58  984.78 ]      [ 1085  890.7  974 ]
                          [ 1285  310.7  976.6 ]        [ 1153  923.51  974.18 ]
                          [ 1086  890.65  984.05 ]      [ 1082  890.6  957.56 ]
     diagonals peaks:     [ 310.8  N    712.34  a    890.7  b    923.58  g1    957.56  g1
                            974  g2     974  g2      974  g2     984.78  d     984.78  d
                            984.78  d ]
```

Figure 11. The spin-system hypothesis corresponding to the isoleucine spin-system shown in Figure 10.

by the SID KS and uses knowledge of chemical shift data in order to check whether the residue has Cα and Cβ protons which may be confused with other protons. From an examination of Table 1, it is obvious that the residues listed there are likely to lead to just this kind of assignment difficulty. If such confusion occurs, the hypothesis is labeled accordingly. For example, in isoleucine residues, the Cβ and Cγ1' protons may easily be confused and thus [Cb Cg1] is placed in the protons slot of the spin-system hypothesis. If protons with difficult to assign resonances are not believed to occur in the residue, CSA writes *complete* to the protons slot.

COSI Using the COSY coordinate data, this KS attempts to distinguish between protons within spin-system hypotheses which have been labeled by the CSA KS. It performs this task using the list of coordinates of the protons within the spin-system together with the information provided by the CSA label. 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 labeled within the spin-system hypothesis and the value of the protons slot set as *complete*.

To illustrate the solution adopted by PNA to these problem assignments, consider Figure 6. From the HOHAHA spectrum of Threonine (6a) it is impossible to distinguish between the Cα and Cβ proton. However, as COSY only has cross peaks due to adjacent interactions (6b, 6c), the Cα - Cβ interaction can be seen, as can the N - Cα and Cβ - Cγ interactions. As the Cα proton gives rise to two cross peaks, one with N and the other with Cβ, while the Cβ proton is involved with Cα and Cγ, it is quite straightforward to differentiate between the α and β protons.

SAM This KS uses the chemical sequence and the spin-system hypotheses, together with a coordinate representation of the short τ_m 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, SAM notes that the two residues are adjacent and the process is repeated using the spin-system hypothesis for the second residue. This continues until a 5 or 6 residue segment has been assembled at which point SAM creates a segment hypothesis. This hypothesis contains the partial sequence and the NOESY peak data used to construct it. SAM then selects another spin-system hypothesis and attempts to generate another 5 or 6 residue segment.

SLOC Once a peptide segment has been created by the SAM KS, the SLOC KS may be invoked. This KS attempts to locate the partial sequence defined by the segment hypothesis within the overall chemical sequence of the protein. The sequence is searched for a matching segment and the sequence position numbers of each of the residues are noted. At this stage, uncertainties as to the exact nature of a residue spin-system are resolved using the sequence. Each of the spin-system hypotheses used to generate the segment are then examined and the appropriate fully labeled residue hypotheses created.

STAN This KS uses the fully labeled residue hypotheses and a coordinate representation of the NOESY spectrum with an intensity descriptor for each peak. It contains information on the type of interactions expected for each secondary structure unit. 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. For example, 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]. STAN examines the NOESY data for cross peaks indicating particular secondary structure units and creates structure hypotheses (described above) detailing the nature and extent of these structures. Table 3 contains a summary of the function of each of the PNA knowledge sources.

6.4 Control

The control component of PNA must integrate the performance of each of the domain knowledge sources described above with intervention by the spectroscopist during problem-solving. Unlike the KSs, each of which is a specialized problem-solving entity dealing with a small part of the overall task, the user is able to contribute at any stage of the process. The user may choose to interrupt the performance of the system and may, for example, create a new hypothesis or modify an existing one on any level of the blackboard. The control task faced by PNA is therefore a dynamic constantly changing one, with the system requiring a flexible control structure.

Rather than encoding a fixed control strategy into the system we are implementing a control framework which will allow the user to intervene during problem-solving. However, we have restricted the amount of user interaction which is allowed during the analysis of the data. For example, the SID knowledge source generates **all** potential spin-system hypotheses without any interruption by the user. One the spin-system identification process is complete the user is free to intervene and to inspect the hypotheses and if necessary to modify or even delete some of them. Other KSs, such as COSI or SAM modify or create only one hypothesis before allowing the user to intervene. This approach is, we feel, a useful compromise between no user interaction during problem-solving and allowing the user to intervene at any point during problem-solving - with all its inherent difficulties.

It should be noted that although Figure 9 indicates the flow of reasoning moving upwards from the data level, that the system also supports top-down reasoning. For example, the identification of a segment hypothesis within the chemical sequence may remove an uncertainty as to the nature of a residue spin-system, which will result in modifications to lower-level hypotheses.

7 Discussion

The Protein NMR Assistant aims to provide a spectroscopist with a powerful tool for the analysis of nuclear magnetic resonance spectra of proteins. Currently, much of this task is performed by hand and is extremely time consuming. By providing an interactive environment for the analysis of HOHA-

Knowledge Source	Function
INITIAL	Initialises the PNA blackboard by loading spectroscopic and sequence data.
NOISE	Identifies noise bands and other spurious peaks in the HOHAHA spectrum prior to the spin-system identification process.
SID	Attempts to identify residue spin-systems within the coordinate representation of the HOHAHA spectrum.
CSA	Using knowledge of approximate residue chemical shifts, examines spin-system hypotheses and labels those which contain "troublesome" protons.
COSI	Examines the COSY coordinate map in an effort to distinguish between troublesome signals identified by CSA.
SAM	Links spin-system hypotheses together to form segment hypotheses.
SLOC	Searches the chemical sequence for a segment generated by SAM and generates a residue hypothesis labeled with its sequence position.
STAN	Infers the presence of secondary structure units using residue hypotheses and the NOESY spectrum.

Table 3 Summary of the Protein NMR Assistant Knowledge Sources.

HA, COSY and NOESY spectra, it is hoped that the time required to perform the analysis of such data will be reduced and that the reliability of results will be increased.

A user interface, consisting of a series of windows displaying the spectra and allowing the user to interact during the interpretation process is under development. Such an interface is, we feel, a vital part of the overall architecture. We are investigating how the partial solutions created on the blackboard can be displayed in such a way that they are meaningful and assist the user in comprehending the actions of the system. Once fully implemented, PNA will be used to examine a number of proteins for which NMR data are available and the results and performance of the system evaluated.

We are currently investigating the application of machine learning techniques to the 2D NMR of Carbohydrates [Metaxas, 1991]. This study aims to generate an empirical theory relating the structural form of a molecule with its 2D NMR spectrum. It is hoped that the experience gained through this project will allow us to investigate the applicability of such methods to 2D NMR of Proteins. Empirical rules relating spectral features to protein structure could be used to assist in secondary structure prediction and perhaps during the sequential assignment process. The existence of some rules, such as those relating peaks in the NOESY spectrum to secondary structure, gives us confidence that this domain will prove suitable for the application of machine learning techniques. Any knowledge obtained using such techniques could be tested within the problem-solving environment provided by the Protein NMR Assistant.

Notes

[1] It should be noted that three dimensional NMR experiments are also now possible.

[2] Figure 3 illustrates these sequential NOE connectivities.

[3] Proline residues can present a problem during the interpretation process, as they do not possess an amide proton and thus any residue adjacent to a proline will appear to be a terminal residue.

[4] A 34 amino-acid peptide with molecular formula $C_{143}H_{230}N_{42}O_{37}$

[5] By aligned we mean that the peak centers lie along the same vertical line allowing for some scatter value.

[6] As some NOEs due to secondary structure features may appear in this experiment, it is necessary to refer to the chemical sequence during sequential assignment of spin systems.

[7] If the NOE cross peak occurs between two protons within the same residue it is ignored.

[8] We are of course assuming that the sequence is correct.

[9] The double-iterated Kalman filter.

[10] The transformation from the original HOHAHA spectrum to this coordinate representation is performed using a commercial 2D “peak picking” program.

References

R.B. Altman, B.S. Duncan, J.F. Brinkley, B.G. Buchanan and O. Jardetzky, Determination of the Spatial Distribution of Protein Structure Using Solution Data, in J. W. Jaroszewski, K. Schaumburg and H. Kofod (Eds.), *NMR Spectroscopy in Drug Research,* Munksgaard, Copenhagen, 1988, 209-232

R.B. Altman and O. Jardetzky, Heuristic Refinement Method for Determination of Solution Structure of Proteins from Nuclear Magnetic Resonance Data, *Methods in Enzymology,* 1989, 177, 218-246

J. Ayel, A Conceptual Supervision Model in Computer Integrated Manufacturing, in *Proceedings of the Eighth European Conference on Artificial Intelligence (ECAI88),* Munich, FRG, August 1-5, 1988, 427-432

R. Balzer, L.D. Erman, P.E. London and C. Williams, Hearsay-III : A Domain-Independent

Framework for Expert Systems, in *Proceedings of the First National Conference on Artificial Intelligence (AAAI80),* Stanford, California, USA, August 18-21, 1980, 108-110

D.R. Barstow, Knowledge Engineering in Nuclear Physics, in *Proceedings of the Sixth International Joint Conference on Artificial Intelligence (IJCAI79),* Tokyo, Japan, August 20-23, 1979, 1, 34-36

D.R. Barstow, Exploiting a Domain Model in an Expert Spectral Analysis Program, in *Proceedings of the First National Conference on Artificial Intelligence (AAAI80),* Stanford, CA, USA, August 18-21, 1980, 276-279

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 Resonance,* 1988, 76: 400-415

J.A. Brugge, B.G. Buchanan and O. Jardetzky, Toward Automating the Process of Determining Polypeptide Secondary Structure from ^{1}H NMR Data, *Journal of Comput. Chemistry,* 188, 9 (6): 662-673

B.G. Buchanan, E.A. Feigenbaum and J. Lederberg, A Heuristic Programming Study of Theory Formation in Science, in *Advance Papers of the Second International Joint Conference on Artificial Intelligence (IJCAI71),* London, September 1-3, 1971, 40-50

B.G. Buchanan and N.S. Sridharan, Analysis of Behaviour of Chemical Molecules: Rule Formation of Non-Homogeneous Classes of Objects, in *Advance Papers of the Third International Joint Conference on Artificial Intelligence (IJCAI73),* Stanford, CA, USA, August 20-23, 1973, 67-76

R.E. Carhart, T.H. Varkony and D.H. Smith, Computer Assistance for the Structural Chemist, in D.H. Smith (Ed.), *Computer-Assisted Structure Elucidation,* ACS Symposium Series no. 54, ACS, Washington D.C., 1977, 126-145

B.D. Christie, Personal Communication, September 1985

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

D.D. Corkill, K.Q. Gallagher and K.E. Murray, GBB: A Generic Blackboard Development System, in *Proceedings of the Fifth National Conference on Artificial Intelligence (AAAI86),* Philadelphia, PA, USA, August 11-15, 1986, 2, 1008-114

B. Debska, J. Duliban, B. Guzowska-Swider and Z. Hippe, Computer-Aided Structural Analysis of Organic Compounds by an Artificial Intelligence System, *Anal. Chim. Acta.,* 1981, 133: 303-318

C. Djerassi, D.H. Smith, C.W. Crandell, N.A.B. Gray, J.G. Nourse and M.R. Lindley, The Dendral Project: Computational Aids to Natural Products Structure Elucidation, *Pure & Applied Chem.,* 1982, 54 (12), 2425-2442

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

P. Edwards, A Cooperative Expert System for Spectra Interpretation, PhD thesis, School of Chemistry, University of Leeds, 1990

H. Egli, D.H. Smith and C. Djerassi, Computer-Assisted Structural Interpretation of ^{1}H

NMR Spectral Data, *Helvetica Chimica Acta,* 1982, 65: 1898-1920

M.E. Elyashberg, V.V. Serov and L.A. Gribov, Artificial Intelligence Systems for Molecular Spectral Analysis, *Talanta,* 1987, 34 (1), 21-30

R.S. Engelmore and A. Terry, Structure and Function of the CRYSALIS System, in *Proceedings of the Sixth International Joint Conference on Artificial Intelligence (IJCAI79),* Tokyo, Japan, August 20-23,1979, 1: 250- 256

R.S. Engelmore and A.J. Morgan, *Blackboard Systems,* Addison-Wesley, Wokingham, England, 1988

S.P. Ennis, Expert Systems : A User's Perspective of Some Current Tools, in *Proceedings of the Second National Conference on Artificial Intelligence (AAAI82),* Carnegie-Mellon University, Pittsburgh, PA, USA, August 18-20 1982, 319-321

L.D. Erman, F. Hayes-Roth, V.R. Lesser and D.R. Reddy, The Hearsay-II Speech Understanding System: Integrating Knowledge to Resolve Uncertainty, *ACM Computing Surveys,* 1980, 12 (2), 213-253

N.A.B. Gray, C.W. Crandell, J.G. Nourse, D.H. Smith, M.L. Dageforde and C. Djerassi, Computer-Assisted Structural Interpretation of Carbon-13 Spectral Data, *J. Org. Chem.,* 1981, 46: 703-715

L.A. Gribov, M.E. Elyashberg and V.V. Serov, Computer System for Structure Recognition of Polyatomic Molecules by IR, NMR, UV and MS Methods, *Anal. Chim. Acta.,* 1977, 95: 75-96

L.A. Gribov, Application of Artificial Intelligence Systems in Molecular Spectroscopy, *Anal. Chim. Acta.,* 1980, 122: 249-256

L.A. Gribov, M.E. Elyashberg, V.N. Koldashov and I.V. Pletnjov, A Dialogue Computer Program System for Structure Recognition of Complex Molecules by Spectroscopic Methods, *Anal. Chim. Acta.,* 1983, 148: 159-170

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

T.F. Havel, I.D. Kuntz and G.M. Crippen, The Theory and Practice of Distance Geometry, *Bulletin of Mathematical Biology,* 1983, 45 (5), 665-720

B. Hayes-Roth, *The Blackboard Architecture: A General Framework for Problem Solving ?*, Stanford University, Computer Science Department, Heuristic Programming Project Report No. HPP-83-30, 1983

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

B. Hayes-Roth and M. Hewett, BB1: An Implementation of the Blackboard Control Architecture, in *Blackboard Systems,* Addison-Wesley, Wokingham, England, R.S. Engelmore and A.J. Morgan (Eds.), 1988, 297-313

J. Hermans (Ed.), *Molecular Dynamics and Protein Structure,* Polycrystal Book Service, 1985

Z. Hippe, Problem-Solving Methods in Computer-Aided Organic Structure Determination, *J. Chem. Inf. Comput. Sci.,* 1985, 25: 344-350

V. Jagannathan, R. Dodhiawala and L.S. Baum (Eds.), *Blackboard Architectures and Applications,* Academic Press: San Diego, CA, 1989

K. Janssens and P. Van Espen, Evaluation of Energy-Dispersive X-Ray Spectra with the Aid of Expert Systems, *Anal. Chim. Acta.*, 1986, 191: 169-180

O. Jardetzky and G.C.K. Roberts, *NMR in Molecular Biology*, Academic Press, New York, 1981

G.C. Levy, Current Trends in Computing: Hardware, Software and Nuclear Magnetic Resonance Research, *J. Molec. Graphics,* 1986, 4 (3),170-177

O. Lichtarge, C.W. Cornelius, B.G. Buchanan and O. Jardetzky, *Validation of the First Step of the Heuristic Refinement Method for the Derivation of Solution Structures of Proteins from NMR Data*, Knowledge Systems Laboratory, Computer Science Department, Stanford University, Report No. KSL-86-12

M.R. Lindley, J.N. Shoolery, D.H. Smith and C. Djerassi, Application of the Computer Program GENOA and Two-Dimensional NMR Spectroscopy to Structure Elucidation, *Organic Magnetic Resonance,* 1983, 21 (7), 405- 411

R.K. Lindsay, B.G. Buchanan, E.A. Feigenbaum and J. Lederberg, *Applications of Artificial Intelligence for Organic Chemistry: The Dendral Project*, McGraw-Hill, New York, 1980

S. Metaxas, P.Edwards and D. Sleeman, *The Interpretation of COSY ^{1}H NMR Spectra for Small Sugar Molecules*, Aberdeen University Computing Science Department Technical Report, AUCS/TR9109

T.M. Mitchell and G.M. Schwenzer, Application of Artificial Intelligence for Chemical Inference XXV. A Computer Program for Automated Empirical ^{13}C NMR Rule Formation, *Org. Mag. Res.,* 1978, 11 (8), 378-384

S. Moldoveanu and C.A. Rapson, Spectral Interpretation for Organic Analysis Using an Expert System, *Anal. Chem.*, 1987, 59: 1207-1212

G.A. Morris, Modern NMR Techniques for Structure Elucidation, *Magnetic Resonance in Chemistry,* 1986, 24: 371-403

M.E. Munk, C.A. Shelley, H.B. Woodruff and M.O. Trulson, Computer- Assisted Structure Elucidation, Z. *Anal. Chem.,* 1982, 313: 473-479

M. Nagao, T. Matsuyama and H. Mori, Structural Analysis of Complex Aerial Photographs, in *Proceedings of the Sixth International Joint Conference on Artificial Intelligence (IJCAI79),* Tokyo, Japan, August 20-23,1979, 2, 610-616

H.P. Nii and N. Aiello, AGE (Attempt to Generalize) : A Knowledge-Based Program for Building Knowledge-Based Programs, in *Proceedings of the Sixth International Joint Conference on Artificial Intelligence (IJCAI79),* Tokyo, Japan, August 20-23,1979, 2, 645-655

H.P. Nii, E.A. Feigenbaum, J.J. Anton and A.J. Rockmore, Signal-to-Symbol Transformation: HASP/SIAP Case Study, *AI Magazine,* 1982, 3 (2), 23-35

H.P. Nii, Blackboard Systems: The Blackboard Model of Problem Solving and the Evolution of Blackboard Architectures, *AI Magazine,* 1986, 7 (2), 38-53

H.P. Nii, Blackboard Systems: Blackboard Application Systems, Blackboard Systems from a Knowledge Engineering Perspective, *AI Magazine,* 1986, 7: (3), 82-106

S. Sasaki, Y. Kudo, S. Ochiai and H. Abe, Automated Chemical Structure Analysis of Organic Compounds: An Attempt to Structure Determination by the Use of NMR, Mikrochimica *Acta.,* 1971, 726-742

S. Sasaki, H. Abe, I. Fujiwara and T. Yamasaki, *The Application of ^{13}C NMR in CHEMICS, The Computer Program System for Structure Elucidation,* in Z. Hippe (Ed.), *Data Processing in Chemistry* (Studies in Physical and Theoretical Chemistry 16), Elsevier, 1981, 186-204

G. Schroll, A.M. Duffield, C. Djerassi, B.G. Buchanan, G.L. Sutherland, E.A. Feigenbaum and J. Lederberg, Applications of Artificial Intelligence for Chemical Inference III. Aliphatic

Ethers Diagnosed by their Low-Resolution Mass Spectra and Nuclear Magnetic Resonance Data, *Journal American Chemical. Society,* 1969, 91, 7440-7445

G.M. Schwenzer and T.M. Mitchell, Computer-Assisted Structure Elucidation Using Automatically Acquired ^{13}C NMR Rules, in D.H. Smith (Ed.), *Computer-Assisted Structure Elucidation,* ACS Symposium Series no. 54, ACS, Washington D.C., 1977, 58-76

C.A. Shelley, H.B. Woodruff, C.R. Snelling and M.E. Munk, Interactive Structure Elucidation, in D.H. Smith (Ed.), *Computer-Assisted Structure Elucidation,* ACS Symposium Series no. 54, ACS, Washington D.C., 1977, 92-107

C.A. Shelley and M.E. Munk, CASE, A Computer Model of the Structure Elucidation Process, *Anal. Chim. Acta.,* 1981, 133: 507-516

C.A. Shelley and M.E. Munk, Computer Prediction of Substructures from Carbon-13 Nuclear Magnetic Resonance Spectra, *Anal. Chem.*, 1982, 54: 516-521

D.H. Smith, N.A.B. Gray, J.G. Nourse and C.W. Crandell, The Dendral Project: Recent Advances in Computer-Assisted Structure Elucidation, *Anal. Chim. Acta.,* 1981, 133, 471-497

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

K. Wüthrich, *NMR of Proteins and Nucleic Acids*, Wiley, New York, 1986

T. Yamasaki, H. Abe, Y. Kudo and S. Sasaki, CHEMICS: A Computer Program System for Structure Elucidation of Organic Compounds, in D.H. Smith (Ed.), *Computer- Assisted Structure Elucidation,* ACS Symposium Series no. 54, ACS, Washington D.C., 1977, 108-125

M. Yamazaki and H. Ihara, Knowledge-Driven Interpretation of ESCA Spectra, in *Proceedings of the Sixth International Joint Conference on Artificial Intelligence (IJCAI79),* Tokyo, Japan, August 20 - 23 1979, 2, 995 - 997