

Determination of Structures of Proteins in Solution using Nuclear Magnetic Resonance

Siddhartha P Sarma

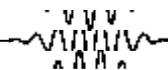
Nuclear magnetic resonance (NMR) is a manifestation of an intrinsic property of the nucleus, i.e. nuclear spin angular momentum. Spin angular momentum gives rise to magnetic moments. Thus, nuclei that possess net magnetic moments behave like very small bar magnets. NMR spectroscopy involves the study of the behaviour of these nuclei under suitable conditions. Analogous to other forms of spectroscopy, nuclear magnetic resonance (NMR) results from absorption of electromagnetic radiation that causes transitions between nuclear spin energy states. Absorption of radiation occurs in the radio frequency region of the electromagnetic spectrum. Not all isotopes of the elements in the periodic table possess nuclei that are 'NMR active'. What this means is that several nuclei possess zero nuclear magnetic moments in the ground state. This is a property that is difficult to predict. Certain empirical rules exist by which one can predict which elements possess net nuclear magnetic moments and which do not. For the present we will restrict ourselves to commonly occurring nuclei that possess a nuclear spin quantum number of $\frac{1}{2}$, viz., ^1H , ^{13}C , ^{15}N , ^{19}F and ^{31}P .

Nuclear magnetic moments are very weak.

In the presence of a strong magnetic field these spins align either in the direction of the applied magnetic field or in a direction opposite to the direction of the applied magnetic field, in a manner analogous to the behaviour of a bar magnet, which when suspended aligns in the direction of the earth's magnetic field. However, unlike the bar magnet, nuclei do not just align with the direction of the applied field and remain there, but precess around the applied field. The frequency of precession is given by the relationship;

$$\omega = -g B_0,$$

where ω (radians s^{-1}) is the angular frequency in radians per second, g ($\text{T}^{-1} \text{s}^{-1}$) is the gyromagnetic ratio and B_0 (T) is the strength of the applied or static magnetic field. This is known as the Larmor precessional frequency at a particular B_0 . The strength of the magnetic field is measured in units of Tesla. For the readers reference, $1\text{T} = 10000$ Gauss. The field strengths used in studying nuclear magnetic resonance vary from 1.5T to 21.14 T. For comparison the earth's magnetic field is 50mT . The sensitivity of a particular nucleus in an applied (external) magnetic field is determined by the gyromagnetic ratio (g) of the nucleus. Of the nuclei listed above, ^1H has the highest sensitivity. With a few exceptions, it is found in almost all molecules and hence is the most commonly studied nucleus by NMR methods. Absorption of energy at the characteristic precessional frequencies causes transitions between lower and higher energy states and results in the observance of the nuclear magnetic resonance



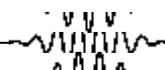
signal or spectrum. What is important to note is that the same nuclear species at different sites within a molecule can possess different precessional frequencies and thus give rise to unique signals. It is this property that has made NMR the most widely used physical tool to characterize structure of molecules. It is used to characterize organic, inorganic, natural and synthetic polymer molecules in the solid, liquid crystal and liquid state. In some instances, NMR has also been applied in the study of molecules in the gas phase. It is widely used in industries such as in pharmaceutical, paper, food, dyes, petrochemicals, pigments, etc., to identify molecules, to monitor reaction products and so on. One of the most spectacular contributions of NMR has been in the development of magnetic resonance imaging (MRI), a method that has today revolutionized diagnosis and treatment of diseases in the medical field. The application of NMR to the structure determination of biomacromolecules such as proteins and nucleic acids has been the most recent success story. Fundamental contributions towards the development of NMR as a method for determining the structure of proteins in solution were made by Kurt Wüthrich, for which he was awarded the Nobel Prize in Chemistry in 2002 (the Prize itself was shared with two other scientists for their contributions to large molecule mass spectrometry [1]). It must be noted that NMR is one of two methods used today to determine structures of biomolecules at atomic resolution, the other being the older and more established method of X-ray crystallography. In what

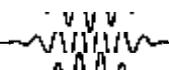
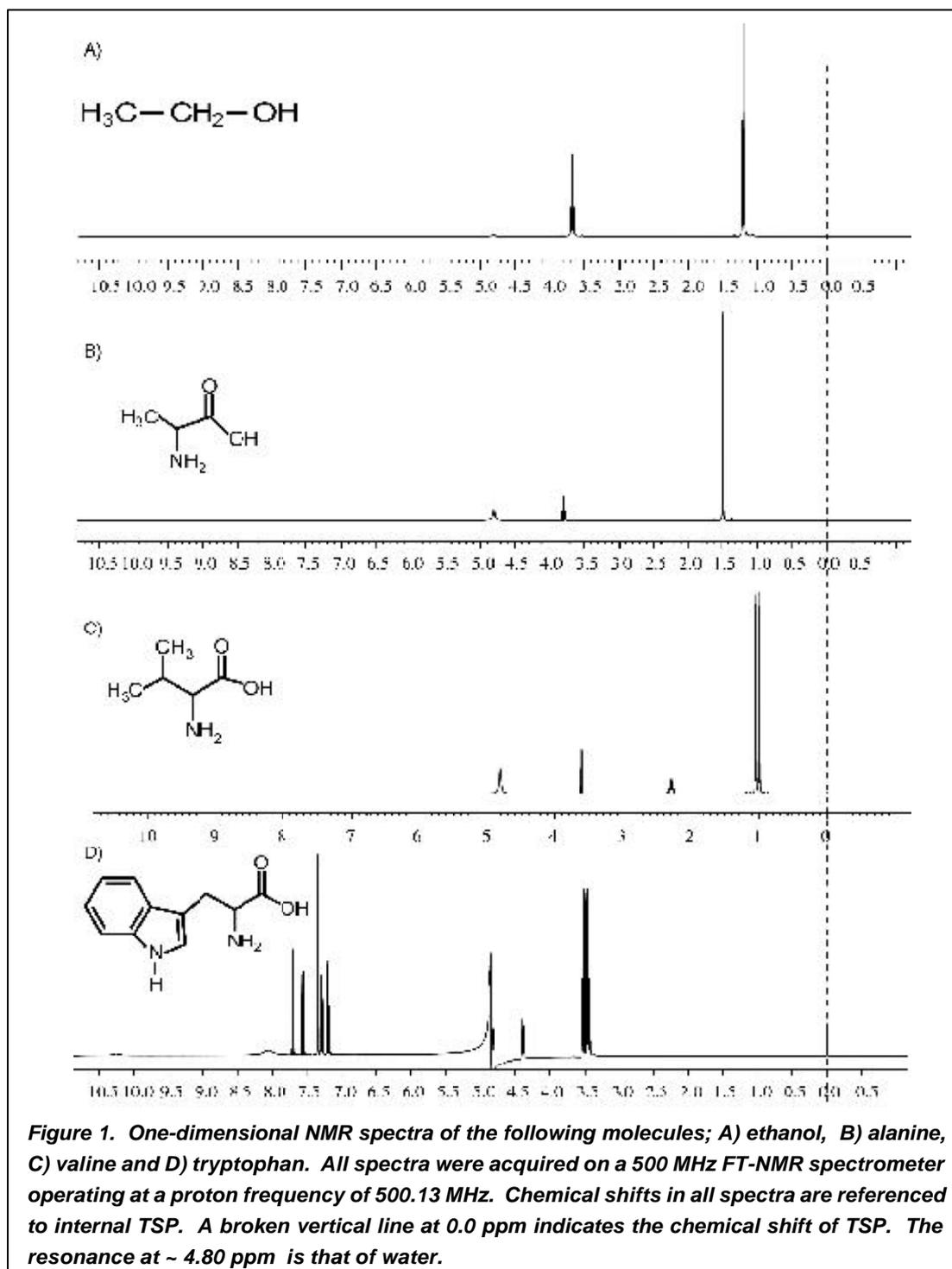
follows, I will attempt to describe NMR principles and methods used in the structure determination of proteins in solution.

The variation in precessional frequency of nuclei within a molecule is strongly dependant on the chemical environment in which the nuclei exist. Consider for example the ethanol molecule, which has the following chemical formula:



Figure 1 shows the one-dimensional proton (since the nucleus of the hydrogen atom contains a single proton, it is commonly referred to as proton in NMR literature and the same convention will be followed here) NMR spectra of ethanol and three amino acids. Let us first consider the spectrum of ethanol. There are essentially two sets of lines in the ethanol spectrum arising from the three protons in the methyl group and from the two protons in the methylene group (the hydroxyl proton is considered to exchange with bulk solvent, in this case water and hence is not observable as a distinct resonance in the spectrum). The unique feature of this spectrum is that the methyl and methylene protons resonate at different frequencies. It is the task of the researcher to determine which signal originates from which group of protons. The inductive effects of the oxygen atom are expected to deshield the methylene protons thereby causing them to shift 'downfield', i.e., to higher frequencies. The methyl protons on the other hand are less affected by the hydroxyl group. Thus the chemical environ-





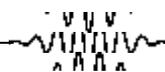
ments that protons find themselves in strongly influences the position at which they appear in the spectrum. The observed spectrum does not show single lines for each group of protons, but a multiplet fine structure because of a phenomenon known as 'scalar coupling' or '*J*-coupling'. What scalar coupling tells us is that the methylene protons are coupled to the methyl protons three bonds away, and thus make their presence felt mutually on each other. In general nuclei that are coupled by three bonds or less show such scalar coupling interactions. Thus the methylene resonance is split into four lines with an intensity ratio of 1:3:3:1, while the methyl resonance is split into three lines with an intensity ratio of 1:2:1. The separation between the lines in the multiplet gives an important structural parameter known as the coupling constant (*J*) and is represented in units of Hz. It is important to note here that the three methyl protons are chemically and magnetically equivalent, in that the three methyl protons resonate at an identical frequency. Same is the case with the two methylene protons. Another important feature is that the integrated area for the methylene resonance and the methyl resonance should be in the ratio of 2:3 as two protons contribute to the signal at ~ 3.6 ppm and three protons contribute to the signal at ~ 1.2 ppm.

A Note on the Appearance of the NMR Spectrum

The NMR spectrum essentially consists of a horizontal axis on which is represented the frequency at which the nuclei resonate. The

frequency is represented in units of Hertz or ppm (parts per million). The ppm scale removes spectrometer frequency dependence in comparing spectra acquired at different field strengths. For example, a spectrum acquired on an NMR machine operating at 7.05 T (300 MHz proton Larmor precessional frequency), 300 Hz corresponds to 1 ppm and for a spectrum acquired on an NMR machine operating at 11.74 T (500 MHz proton Larmor precessional frequency), 500 Hz corresponds to 1 ppm. The final spectrum is usually referenced to an internal standard that is ascribed a chemical shift of 0 ppm or 0 Hz. Thus a signal resonating 2 ppm higher than the internal standard would be at the same position in the spectrum acquired at 7.05 T or 11.74 T. The vertical axis represents the intensity of the observed signal and is usually not depicted in the final form of the spectrum. Multidimensional NMR spectra are viewed as contour plots. The system of referencing is the same as in one-dimensional NMR spectra.

Consider next the proton NMR spectra of three amino acids alanine, valine and tryptophan (*Figure 1*) each of which has been dissolved in water. The chemical structures of the amino acids are also shown in the figure. The amino protons of each amino acid cannot be observed for the same reason that the hydroxyl proton cannot be observed in ethanol. The rest of the proton signals are observed. One can readily observe that the NMR spectrum of the individual amino acid becomes more complex as the size of the



amino acid increases and therefore requires careful interpretation for correct assignment of resonance signals. On the basis of the 'rules' presented above for the assignment of the ethanol spectrum, the assignments for individual amino acids too can be made in a straightforward manner. One can now expect that as the size of the molecule increases so will the complexity of the spectrum, due to an increase in number of nuclei that contribute to the observed spectrum. *Figure 2* shows the NMR spectrum acquired for different polypeptide molecules ranging from 5 amino acids to 230 amino acids. This makes unambiguous assignment of resonances in an one-dimensional spectrum to the various nuclei in larger molecules almost impossible.

One method by which the resonances in a spectrum can be resolved is by increasing the dimensionality of the NMR spectrum. By this it is meant that the resonances are separated over two or more frequency axes. How is this done? The exact spin physics behind this is beyond the scope of this article. Briefly, nuclear magnetic dipole moments interact with other neighboring nuclear magnetic dipole moments through the following mechanisms:

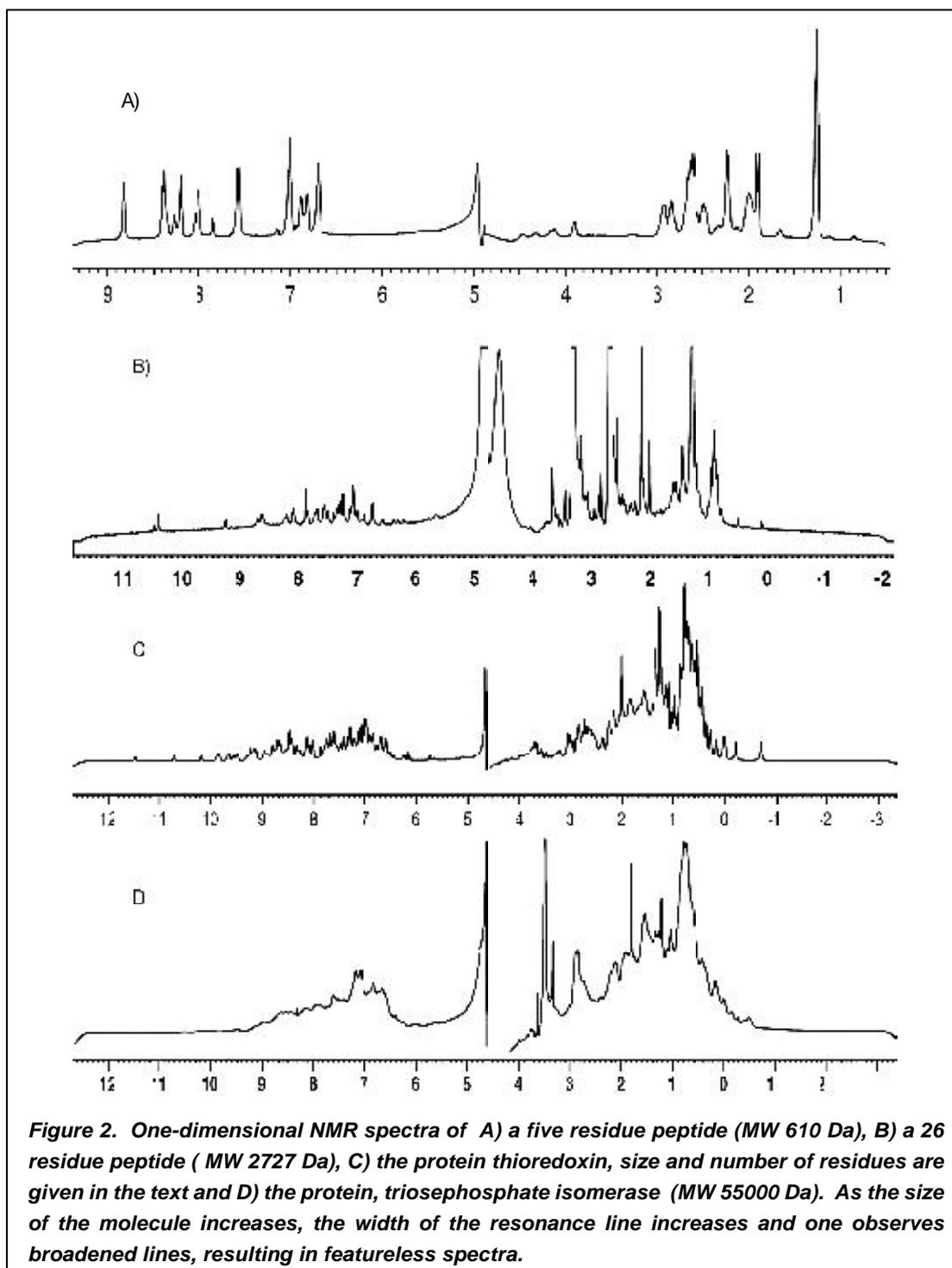
a) by indirect dipole-dipole coupling, also known as scalar coupling or *J*-coupling, through the involvement of bonding electrons.

b) by direct through space interaction known as dipole-dipole coupling
In liquids, because of the rapid tumbling of

molecules the direct dipole-dipole couplings average to zero. This rapid tumbling does not cause the *J*-coupling to average to zero and it manifests as the multiplet fine structure observed in high resolution NMR spectra (cf. ethanol spectrum). The indirect dipole-dipole coupling also provides valuable structural information, in that the value of the three-bond ^1H - ^1H *J*-coupling in rigid molecule is dependant on the torsion angle around the central bond mediating the coupling. Thus the value of the three-bond coupling provides angular information and thus provides valuable local structural information. The direct dipole-dipole coupling between nuclear spins causes them to 'relax' each other, by a mechanism called the Nuclear Overhauser Effect (NOE). This mutual relaxation of nuclear spins is the most important parameter from which structural information is obtained. The rate at which the nuclei relax each other is dependant on the inverse of the sixth power of the distance separating the two nuclear dipoles and thus provides valuable distance information between pairs of interacting nuclei. The upper distance limit placed on such interaction is 5 Å.

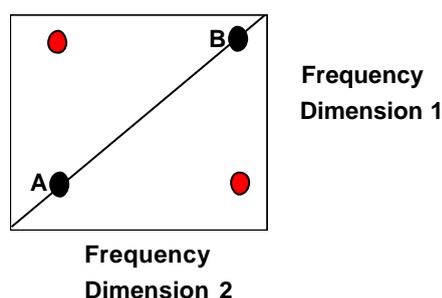
Many multidimensional NMR experiments (two or more dimensions) have been developed using the through bond or the through space interactions between nuclei to separate these interactions, thereby enabling the spectroscopist to identify nuclei that interact with each other through one or both these interactions. The NMR spectroscopist can thus es-





establish 'correlations' between interacting nuclei. As an illustration, a schematic representation of a two-dimensional NMR spectrum is shown below. In the spectrum, the black coloured circles correspond to peaks that one would observe in a normal one-dimensional NMR spectrum (see *Figure 2*) and they lie along the diagonal of the two-dimensional spectrum. The open circles are peaks that indicate that peaks A and B are connected in some manner, either through space (mechanism **a**) or through bond (mechanism **b**).

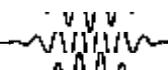
These experiments can and have been designed to study such correlations between



protons (homonuclear) or between protons and carbon-13 or nitrogen-15 (heteronuclear). At natural isotopic abundance, it is most efficient to study correlations between protons in a molecule. Homonuclear through bond correlations are established from homonuclear COSY and/or TOCSY experiments. The COSY (CORrelation SpectroscopY) experiment provides information on nuclei that are coupled by three bonds and TOCSY (TOtal CORrelation SpectroscopY) provides through bond correlations between all atoms

in a network of spins. Through-bond correlation type experiments do not provide information across the peptide bond in proton-proton correlated experiments of the type discussed above. Correlations between protons across the peptide bond are obtained via the through-space correlation experiments viz., the homonuclear NOESY (Nuclear Overhauser Enhancement and Exchange SpectroscopY) experiment. The information content in the spectrum is exactly as the name of the experiment suggests (if pronounced as NOSEY), in that each proton reports on all its' neighbours, which appear as crosspeaks (also known as NOEs), provided that they reside within a 5 Å radius.

How are these correlations established using two-dimensional NMR experiments and how is the structure of a protein calculated using NMR? Much of the methodology used today for the structure determination of proteins using nuclear magnetic resonance was developed by Kurt Wüthrich at the Department of Biophysics and Molecular Biology in ETH Zurich during the late 1970s and early 1980s. The first reported novel structure of a protein to be solved by NMR methods was that of Tendamistat, which was solved by Wuthrich and co-workers in 1986. Prior to this systematic methods for assignment of resonance signals in two dimensional spectra to individual protons or groups of protons within the protein and development of mathematical methods to compute structural models using NMR based structural restraints i.e., distance information between pairs of



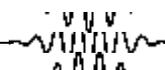
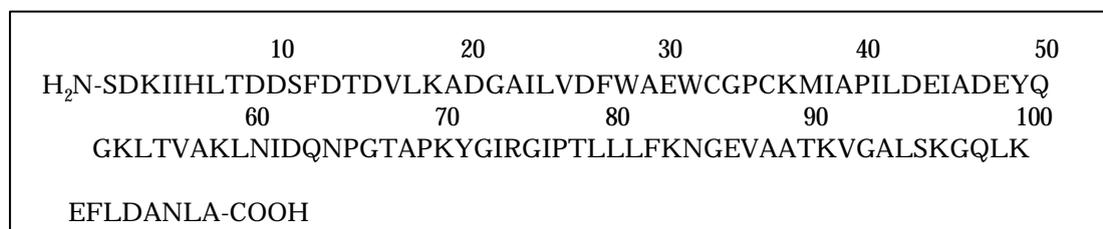
interacting protons from NOESY (Nuclear Overhauser Enhancement Spectroscopy) data and angular constraints from proton-proton scalar coupling constants measured across three bonds were developed by Kurt Wüthrich. More recently Wüthrich has developed NMR methods for the study of very large proteins in solution by making use of the TROSY (Transverse Relaxation Optimized Spectroscopy) principle. It is for these contributions that Wüthrich was awarded the Nobel Prize for the year 2002 in Chemistry, which he shared with John Fenn and Koichi Tanaka for their contributions to large molecule mass spectrometry.

In what follows, the steps involved in determining the structures of proteins using NMR methods will be briefly outlined using the protein thioredoxin as an example. Thioredoxin is a protein that contains 108 amino acids, has a molecular weight of 11677 Daltons and has approximately 850 protons. The first step is to assign each resonance in the two-dimensional NMR spectrum to a specific proton within the protein, in other words to make 'sequence specific' assignments. This is the most important step in the analysis of NMR data if any conclusions about molecular structure at the atomic level are to be made. The next step is to establish the spa-

tial proximity of each proton to other protons within the protein. In practice both types of spin-spin interactions, viz., through bond and through space interactions are simultaneously examined and sequence specific assignments made.

Shown below is a two-dimensional TOCSY spectrum (Figure 3). In this spectrum only through-bond correlations are manifest. At first glance it may seem that assignment of individual resonances to specific protons within the protein is an impossible task. Fortunately, it is not as bleak as it appears. We have at our disposal a wealth of chemical information about the molecule. For instance, we know that the molecule is a protein and that proteins are made up of the 20 standard amino acids. At the next level of information the chemical structures and stereochemistry of these amino acids are known. The final piece of information comes from the number of amino acids present and the sequence in which these amino acids are arranged in the molecule. The primary amino acid sequence of thioredoxin is given below in the one letter amino acid code.

Once this is known our knowledge of the primary structure of the protein is complete. Furthermore, from accumulated knowledge



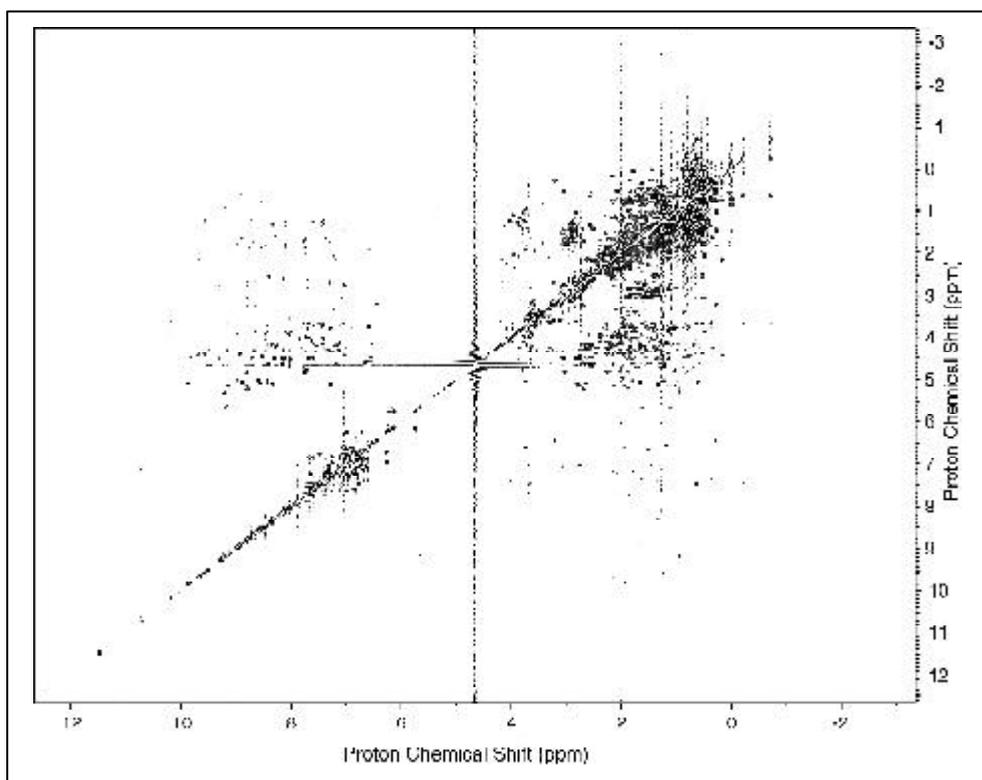
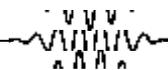


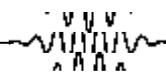
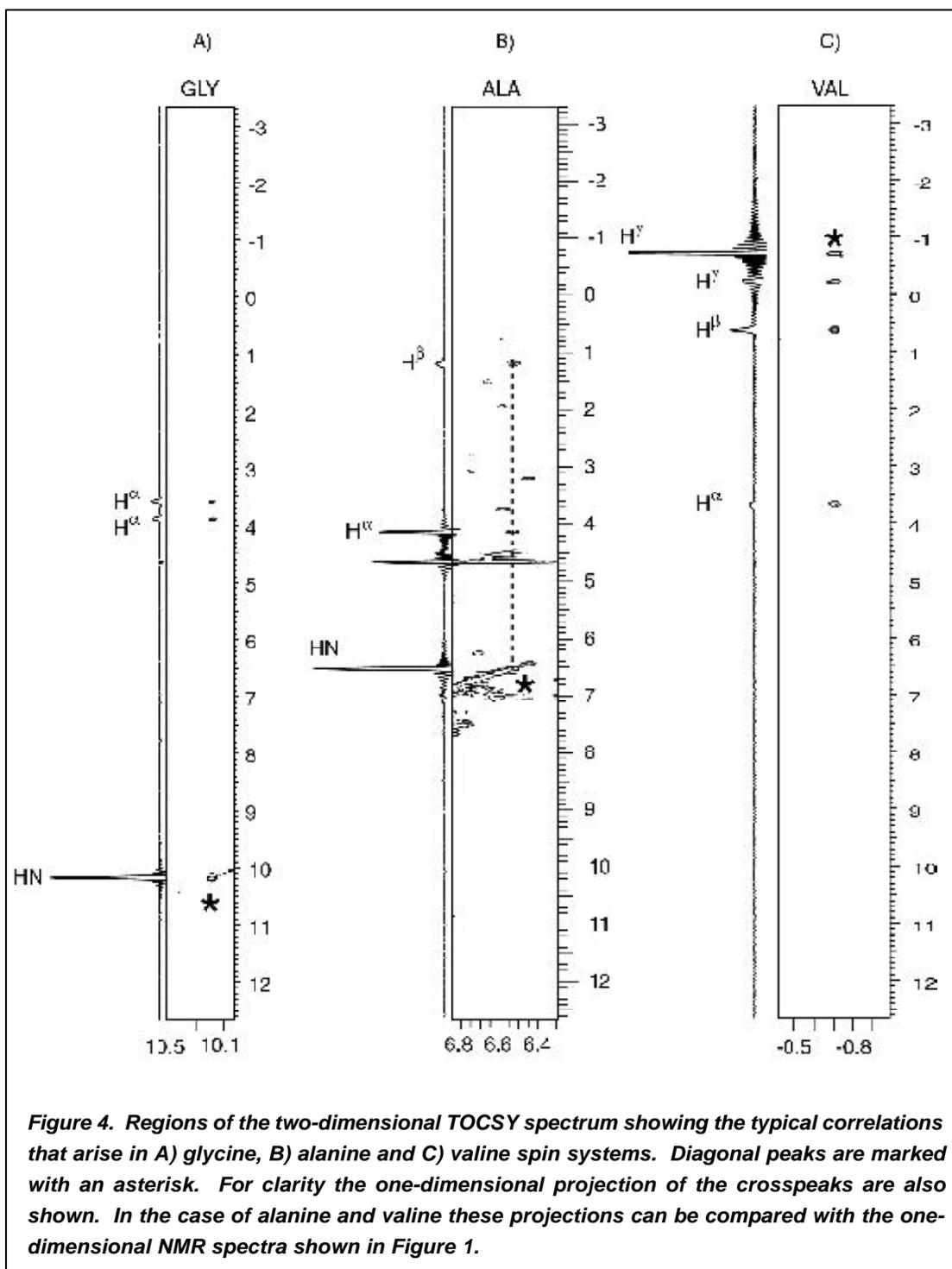
Figure 3. Two-dimensional TOCSY spectrum of thioredoxin. This spectrum shows through-bond correlations, i.e., crosspeaks between protons within an amino acid. Two dimensional spectra are normally viewed as two-dimensional contour plots as is the case here.

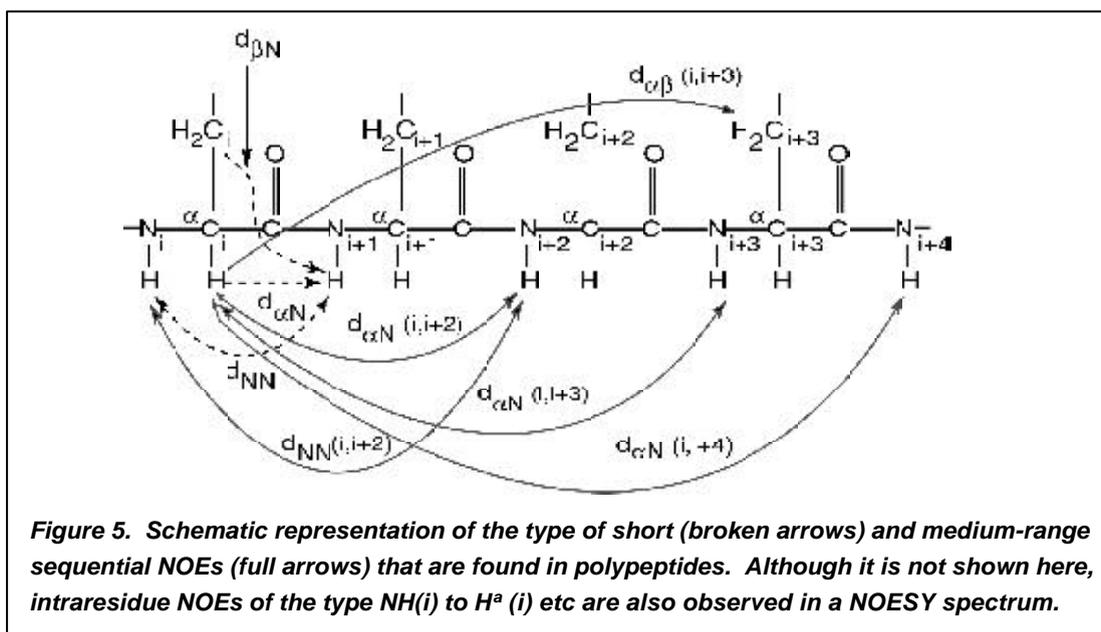
on protein structures (obtained by X-ray diffraction methods), we know that regions of the protein backbone adopt one of the following secondary structures, i.e., helical, *b*-extended or *b*-turn-type. From our NMR data we must now define which regions of the protein adopt what type of secondary structure and how these secondary structure elements are spatially distributed in the molecule. With this in mind we can take another look at the spectrum with less trepidation.

In the TOCSY spectrum the residues glycine, alanine and valine are relatively easy to

identify. In *Figure 4*, regions of the 2D TOCSY spectrum highlighting the spin system identification of glycines, alanines and valines are shown. Also shown is the one-dimensional projection for each of the residues. Compare these 1D spectra with those of the individual amino acids shown in *Figure 1*. Looking back at the primary sequence, we find that thioredoxin contains 9 glycine residues, 12 alanines and 5 valines. Which of these alanines, glycines and valines give rise to the respective signals that we have just identified? In order to determine this, we must identify the preceding and following



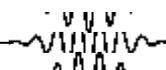
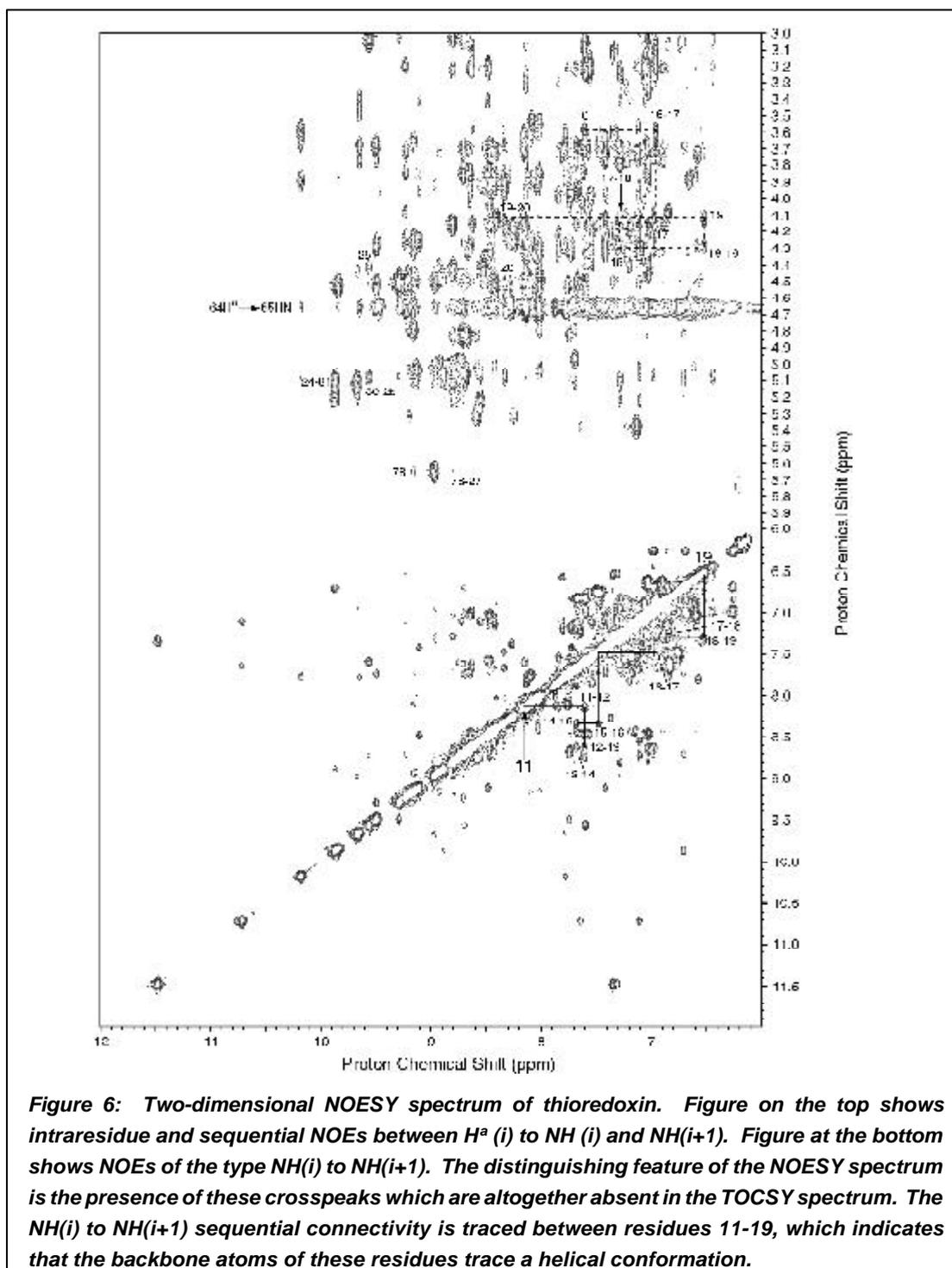




residue in the sequence. This is best done by examining the NOESY spectrum to establish 'sequential connectivities'. We must look for correlations involving backbone H^a(i) to NH(i+1) or NH(i) to NH(i+1). Figure 5 shows a schematic representation of the sequential NOEs that would be observed in the NOESY spectrum of a polypeptide. Figure 6 shows sequential connectivities to the residues that we have just identified. In the case of glycine the preceding residue is a proline and the succeeding residue is a threonine. Looking back at the sequence it is safe to conclude that the glycine we have identified is G65 in the sequence. Similarly valine is preceded by a leucine and succeeded by aspartic acid, which would identify the valine as V16. In the case of alanine it is preceded by isoleucine and succeeded by a proline, which would identify the alanine as

A19. Thus by identifying V16 and A19 by their preceding and succeeding residues in the sequence, we have thus identified residues 15-20 in the sequence. Similar logic should be applied and the assignment process be extended for the rest of the resonances in the spectrum.

The intensities of the 'sequential crosspeaks' and the presence/absence of a certain type of sequential NOE provides valuable information with regard to the type of secondary structure that the residue lies in. For instance, if a stretch of sequential NOEs are identified as being H^a(i) to NH(i+1) with the concomitant absence of NH(i) to NH(i+1) NOEs then it can be safely assumed that the residues in that stretch are present in a *b*-strand like conformation. In the case of residues that lie in *a*-helical regions, both



$H^a(i)$ to $NH(i+1)$ and $NH(i)$ to $NH(i+1)$ NOEs are found. While sequential connectivities ($i - j = 4$; short-range) give secondary structural information, what are needed to determine the structure of a protein are 'long-range' NOEs i.e., correlations between nuclei that are separated in sequence by at least 5 residues ($|i - j| > 4$). These NOEs tell us that while the nuclei may be well separated in primary sequence, they are close together in space. This information is critical for determining the tertiary structure of the protein. Examples of such corre-

lations are also shown in the homonuclear NOESY spectrum. Distances between pairs of protons that contribute to a particular NOE peak in a NOESY spectrum can be obtained from the intensity ($\propto 1/r^6$) of the peak. This should then be repeated for all assigned NOE peaks and using these distances the tertiary structures can be calculated. The algorithm used to calculate structures is based on distance geometry. To invoke an analogy, it is quite similar to constructing the map of India from a knowledge of the distances between various cities and

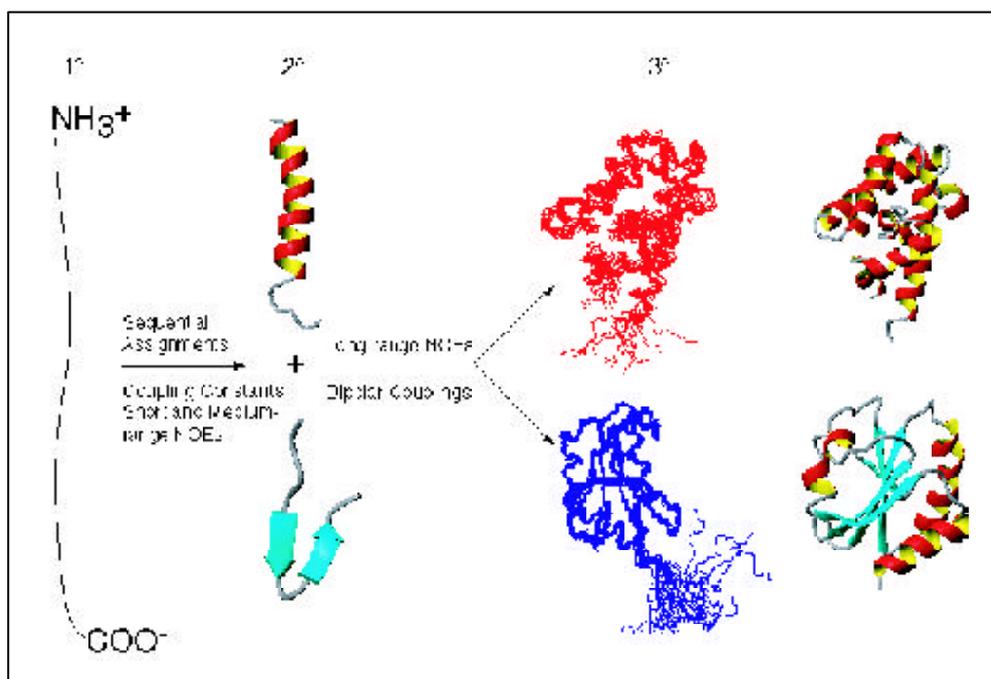


Figure 7. Schematic representation of the steps involved in protein structure determination using nuclear magnetic resonance methods and the information content vis-à-vis the structure of the molecule obtainable at each step. 1°, 2° and 3° refer to the primary, secondary and tertiary structure of proteins. An ensemble of closely related structures are obtained using NMR structure determination methods. Shown to the extreme right is a representative structure determine using NMR parameters.

towns. The greater the number of distances we can define between cities and towns in the hinterland to those in border areas, the more accurate the map.

Figure 7 summarizes the steps involved in protein structure determination using NMR methods. Unlike X-ray diffraction methods, no unique structure is obtained by NMR methods. Instead, an ensemble of closely related conformers, in which all conformers display the same global architecture but differ in local aspects of the protein fold, are obtained. For structures that are determined from a reliable set of experimental data the RMSD (root mean square deviation) should be small. The minimum energy conformers are those that satisfy all experimental data and also the Ramachandran torsion angle energy space. Several recent advances have been made in NMR methods which have shortened the assignment process significantly and thereby speeded up structure calculation and has concurrently improved the

accuracy of the determined structures. Principal among this is the introduction of heteronuclear NMR methods involving nuclei such as carbon-13 and nitrogen-15. Today high resolution structures of proteins with molecular masses upto 30 kDa are being determined on a routine basis. The future holds promise for much larger molecules to be explored by NMR methods.

Suggested Reading

- [1] M Vairamani, **Chemistry Nobel Prize 2002 – Mass Spectrometry**, *Resonance*, Vol.8, No.4, pp.69-76, 2003.
- [2] Kurt Wüthrich, **NMR of Proteins and Nucleic Acids**, *John Wiley and Sons, New York, 1986*.
- [3] Donald Voet and Judith Voet, **Biochemistry**, *John Wiley and Sons, New York. Chapters 6-8, 1995*.
- [4] [HTTP://www.cryst.bbk.ac.uk/PPS2/projects/schirra/html/2dnmr.htm](http://www.cryst.bbk.ac.uk/PPS2/projects/schirra/html/2dnmr.htm)

Siddhartha P Sarma, Assistant Professor, Molecular Biophysics Unit, Indian Institute of Science, Bangalore 560012, India, Email:sidd@mbu.iisc.ernet.in



“I have never wanted to direct the work of others. I was always afraid to waste their time. I don’t mind wasting my own time, but I don’t want to waste the time of others”

– Dirac

