Hypothesis

The direct determination of protein structure by NMR without assignment

R. Andrew Atkinson[a,*], Vladimír Saudek[b,1]

[a] *UPR 9004 du CNRS, Ecole Supérieure de Biotechnologie de Strasbourg, Boulevard Sébastien Brant, 67400 Illkirch, France*
[b] *7 Au Canal, 67300 Schiltigheim, France*

Received 25 October 2001; accepted 29 October 2001

First published online 30 November 2001

Edited by Thomas L. James

Abstract Assignment of the resonances in nuclear magnetic resonance spectra is considered a pre-requisite for the interpretation of spectra that yield structural information. The determination of the three-dimensional structure of a biological macromolecule may, however, be achieved directly without spectral assignment, using the same set of heteronuclear scalar and dipolar coupling experiments as normally used. A cross-peak in any of the spectra may be interpreted as a distance between atoms, yielding a set of distances between unassigned atoms that serves to define the tertiary structure of the molecule. The principle is illustrated using the 76 amino acid protein ubiquitin.

Key words: Assignment; Nuclear magnetic resonance; Structure; Ubiquitin

1. Assignment is a pre-requisite for structure determination

High-resolution solution-state nuclear magnetic resonance (NMR) spectroscopy of biological macromolecules lacks an elegant transform of experimental data to structure, such as that used in X-ray crystallography. Initial structural studies of proteins by NMR relied on scalar couplings to identify the sets of resonances of individual amino acids, and inter-residue nuclear Overhauser effects (NOEs) to assign these to specific residues in the sequence [1]. This 'assignment' procedure allows NOEs to be interpreted as distances between pairs of atoms, and these are used as restraints in algorithms designed to find structures compatible with both the chemical structure and the experimental observations [2]. It is, of course, now common to label molecules of interest with ^{13}C and ^{15}N: combinations of triple resonance experiments such as HNCA and HN(CO)CA are routinely used to establish sequential connectivities between amino acids; experiments such as HCCH-TOCSY allow full side-chain assignment and enable the interpretation of ^{13}C- and ^{15}N-edited NOESY spectra [3].

Assignment of resonances in the NMR spectrum is generally considered a pre-requisite for the interpretation of those spectra that yield structural information [1], such as NOEs, coupling constants, residual dipolar couplings [4] and hydrogen bonds [5,6]. It is also required, for example, for the detailed interpretation of relaxation measurements or chemical shift mapping experiments. Assignment is, however, an arduous, error-prone task that must be repeated for each new protein of interest, even when a crystallographic structure or a homology-based model is available. Attempts to automate the procedure have proven only partially successful [7]. Assignment of NOEs (an accurate but misleading term) to pairs of nuclei is an equally difficult and dangerous task, and although methods have been developed to remove the experimentalist's bias and to allow ambiguity [8–10], this remains a largely manual procedure.

2. Assignment is <u>not</u> a pre-requisite for structure determination

Methods for the direct determination of protein structure by NMR were first proposed by Malliavin et al. [11] who suggested recording a large number of heteronuclear 3D spectra to determine precisely the distances between backbone amide protons. These were to be used to place this subset of atoms in space using a distance geometry algorithm, and the backbone traced. Subsequent studies [12–14] proposed the interpretation of full sets of NOEs as distances between unassigned and unconnected hydrogen atoms that could be used to calculate a 'structure', or rather a proton cloud, into which the full covalent structure of the protein could be built. Spectral overlap has proven a stumbling block for the practical application of such an approach [15].

3. Heteronuclear scalar coupling experiments yield structural restraints

The determination of the three-dimensional structure of a biological macromolecule by NMR may, however, be achieved directly without prior spectral assignment, by exploiting the same set of heteronuclear scalar and dipolar coupling experiments as normally used. The assignment both of resonances and of NOEs may be entirely avoided simply by interpreting the experimental data in an alternative manner. A cross-peak in any of the spectra may be interpreted as a distance between atoms. The set of distances between unassigned atoms serves to define the tertiary structure of the molecule.

*Corresponding author. Fax: (33)-3-90 24 47 18.
E-mail addresses: andrew@esbs.u-strasbg.fr (R.A. Atkinson), vladimir@incyte.com (V. Saudek).

[1] Present address: Incyte Genomics, Botanic House, Cambridge CB2 1FF, UK.

0014-5793/02/$22.00
PII: S0014-5793(01)03208-2

Table 1
Numbers of synthetic distance restraints used for ubiquitin

Experiment	Number of restraints
^{15}N–^{1}H HSQC	72
HNHA	154
HNHB	214
^{13}C–^{1}H HSQC	365
HN(CO)CA/HNCA	288
HNCO/HN(CA)CO	288
CBCA(CO)NH/CBCANH	266
Inferred from above	1981
NOEs	2060
Inferred from NOEs	6089
$d_{H-H} > 4.0$ Å	92570

Thus, NOEs are again interpreted as distances between unassigned and unconnected atoms, but cross-peaks in all other spectra are also interpreted as distances instead of being used for assignment purposes. Thus, the ^{15}N–^{1}H HSQC yields a distance of 1.02 Å between nitrogen and hydrogen atoms (e.g. $N_{118.00}$–$H_{8.30}$); the HN(CO)CA spectrum yields distances of 2.40 Å and 2.49 Å between these nitrogen and hydrogen atoms respectively and a Cα atom (e.g. $N_{118.00}$–$C\alpha_{55.40}$; $H_{8.30}$–$C\alpha_{55.40}$); the HNCA spectrum yields distances of 1.46 Å and 2.23 Å between the same pair and a second Cα atom (e.g. $N_{118.00}$–$C\alpha_{58.30}$; $H_{8.30}$–$C\alpha_{58.30}$). Similar treatment of HNCO, HN(CA)CO, CBCANH, CBCA(CO)NH, HNHA and HNHB spectra yields a rich set of distances from which further interatomic distances may be inferred (e.g. Cα–C′, Cα–Cβ).

4. Testing the hypothesis

To validate the principle, synthetic data was produced for the protein ubiquitin, for which a high-resolution crystal structure and full backbone and side-chain ^{1}H, ^{13}C and ^{15}N assignments are available. The distance information that would, in practice, be extracted from ^{15}N–^{1}H HSQC, HNHA, HNHB, ^{13}C–^{1}H HSQC, HN(CO)CA, HNCA, HNCO, HN(CA)CO, CBCA(CO)NH and CBCANH spectra was generated. This was complemented by inferred distances between the following pairs of atoms: $C\alpha_i$–C'_i, $C\alpha_{i-1}$–$C\alpha_i$, C'_{i-1}–C'_i, $C\alpha_{i-1}$–C'_i, C'_{i-1}–$C\alpha_i$, $C\alpha_i$–$C\beta_i$, $C\beta_{i-1}$–$C\alpha_i$, $C\alpha_{i-1}$–$C\beta_i$, $C\beta_i$–C'_i, $C\beta_{i-1}$–C'_i, C'_{i-1}–$C\beta_i$, $C\beta_{i-1}$–$C\beta_i$, $C\beta_i$–$H\alpha_i$, C'_i–$H\alpha_i$, $C\alpha_{i-1}$–$H\alpha_i$, $C\beta_{i-1}$–$H\alpha_i$, C'_i–$H\alpha_i$, $H\alpha_i$–$H\beta_i$, $C\alpha_i$–$H\beta_i$, C'_i–$H\beta_i$, $C\alpha_{i-1}$–$H\beta_i$, $C\beta_{i-1}$–$H\beta_i$, C'_{i-1}–$H\beta_i$. Ranges were set for each distance, according to the minimum and maximum values possible in model polypeptides and a tolerance of 0.1 Å was added to each allowed range. As distance information concerning each atom was entered, the atom was added to a topology file for X-PLOR [16] using the full table of ^{1}H, ^{13}C and ^{15}N assignments. The 'residue ID' field was set to a sequential number according to a list ordered by ^{15}N chemical shifts, the 'residue name' field was used to code for the chemical shift of the atom and the 'charge' field was used to code for the chemical shift(s) of attached atoms. The distance information derived from a HN(CO)CA spectrum, for example, could then be specified using the sequential number (residue ID field) for the N and H atoms but the chemical shift (residue name field) for the $C\alpha_{i-1}$ atom, since the sequential number of the preceding residue is, at this stage, unknown.

To generate the set of inter-atomic distances that are expected to give rise to NOEs in 4D heteronuclear-edited NOESY spectra (< 4 Å), hydrogen atoms were added to the crystal structure of ubiquitin ([17]; PDB code: 1UBQ) and the structure was energy minimised. Distances were defined using both the chemical shifts of the proton (residue name field) and of the attached heteronucleus (charge field) and using a uniform upper limit of 4.0 Å (i.e. 1.8 Å $< d_o <$ 4.0 Å). Inferred upper limits between the pairs of attached heteronuclei and between each proton and the other heteronucleus were added. Finally, lower limits of 4.0 Å were set for all proton pairs not included in the NOE list (i.e. $d_o > 4.0$ Å).

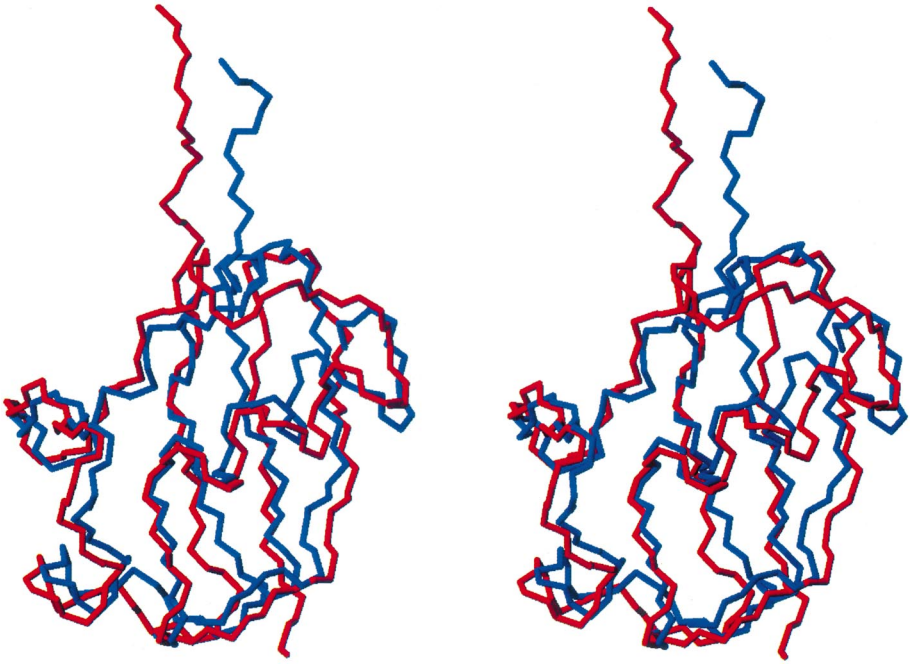

Fig. 1. Stereoview of the superimposed structures of ubiquitin, determined by X-ray crystallography (red) and calculated using synthetic NMR data but without assignment (blue). The C-terminus is located at the top of the picture. Breaks in the blue chain occur at proline residues.

Table 2
Data for runs with different input data sets

Data set	R.m.s.d. for lowest energy structure (Å)	Mean r.m.s.d. for four lowest energy structures (Å)	Correlation coefficient for r.m.s.d. vs. energy[a]
Complete set (as Table 1)	1.72	1.93	0.91
I – Upper limit set to 5 Å	2.05	1.85	0.82
II – 5 Å distances added	1.74	2.17	0.58
III – 10% data removed	2.15	2.32	0.84

[a]Calculated using a set of 16 structures.

A summary of the set of distance restraints used is given in Table 1. The two sets of input data, one derived from NOEs, the other from heteronuclear scalar coupling experiments, define distance restraints involving all observed nuclei.

A set of 16 starting structures was created with random coordinates for each atom in the topology file. Each starting structure was submitted to a standard X-PLOR simulated annealing protocol, in which only distance restraint and van der Waals terms were included. The resulting structures and their mirror images were compared to the structure from which data were derived. Since synthetic data sets were used, the correspondence between atoms in the two structures is known. The r.m.s.d. value for the lowest energy structure was 1.72 Å when superimposing the Cα atoms of the elements of secondary structure (as defined for the crystal structure [17]). R.m.s.d. values were found to correlate well with the energies of the structures ($r^2 = 0.91$), that is, the lowest r.m.s.d. values were found for the structures with the lowest energies. The set of four structures with the lowest energies gave a mean r.m.s.d. with ubiquitin of 1.93 Å (again superimposing secondary structure). The lowest energy structures closely resemble that of ubiquitin determined by X-ray crystallography (Fig. 1). It should be re-iterated at this point that the structures were determined with no prior assignment of any spectral resonance or cross-peak, but that every hydrogen atom in the structure is labelled by both its own chemical shift and that of the attached heavy atom, and vice versa.

The calculated structures only include those atoms that are observable by NMR. Breaks in the chain occur at proline residues. At present, no distance information from heteronuclear scalar coupling spectra connects the atoms of the side-chains beyond Cβ to the chain comprised of the backbone and the Cβ and Hβ atoms. The HCCH-TOCSY is the obvious experiment to achieve this, but implementation in a distance-based approach is not straightforward and has not been attempted here. The more remote atoms of the side-chains are, however, present as C–H pairs and it is the NOEs from these to atoms of the chain that serve to position side-chain atoms close to the residue in the chain to which they belong.

While the fold of the chain is close to that of ubiquitin, side-chain atoms remain unconnected, degenerate resonances unaccounted for (e.g. phenylalanine Cδ–Hδ atoms are represented by a single C-H pair), and unobserved atoms are absent. In a refinement stage, side-chains may be completed, covalent geometry and Cα chirality imposed, peak intensities calibrated, pseudo-atom corrections included and tools such as TALOS [18], that require sequence specific assignment, applied.

5. Results with less perfect data sets

Data sets derived from experimental spectra are expected to be less ideal than those used in the calculations described above. To test the effects of such imperfections, a number of additional data sets were generated and calculations performed as above. Occasionally, lower r.m.s.d. values were found for structures other than those with the lowest energy. In practice, however, no criterion would allow selection of these structures, so only the values for the lowest energy structures are reported here (Table 2). The results suggest that imperfections in input data sets are tolerated in this approach.

Set I – To simulate overestimation of the maximal distance observed in the NOESY experiments, the upper limit for the set of NOE restraints (corresponding to distances in the structure of ubiquitin less than 4.0 Å) and the lower limit for all other inter-proton distances were set to 5 Å. The lowest energy structure gave a r.m.s.d. value of 2.05 Å.

Set II – To allow for the possible observation of some NOEs where the distance between protons is greater than 4.0 Å, 5% of the set of inter-proton distances between 4.0 and 5.0 Å were added to the data set, but defined in the same manner as all other NOE restraints (i.e. 1.8 Å $< d_o <$ 4.0 Å). The lowest energy structure gave a r.m.s.d. value of 1.74 Å.

Set III – To account for incomplete data sets, 10% of the input data was removed at random. The lowest energy structure gave a r.m.s.d. value of 2.15 Å.

6. Overlap and other practical considerations

It was noted above that resonance overlap represented a major difficulty in applying 'no assignment' strategies. Indeed, two resonances from nuclei that are far apart in the structure with identical chemical shifts but distinct sets of neighbours would be represented by a single atom with one set of neighbours, leading to gross distortion of the calculated structure. Here, however, use of heteronuclear-edited NOESY spectra drastically reduces the likelihood of overlap: the assignment table for ubiquitin contains only one occurrence (a ^{15}N–^{1}H pair) of precise overlap of both proton and heteronucleus chemical shifts. This particular case of overlap may be readily identified during data preparation. Overlap of two peaks in the ^{15}N–^{1}H HSQC is resolved during peak picking of the HNCOCA spectrum: a strip that contains cross-peaks to two $C\alpha_{i-1}$ resonances serves to identify overlap of the ^{15}N and ^{1}H resonances. Precise overlap would require the ambiguity to be propagated to restraints involving the $C\alpha_i$, $C\beta_i$ and $C\beta_{i-1}$ atoms.

Other practical issues remain to be addressed for the application of this method to experimental data: most importantly, uncertainties in the chemical shifts of picked peaks must be incorporated. The resonance of a Cα nucleus, for example, may not be picked at exactly the same chemical shift (to two decimal places) in both HNCOCA and HNCA spectra. Here,

aspects of the methodology used in ARIA [9] might be applicable. Since only distance information is used, a structure and its mirror image satisfy the data equally well. The choice between the two may be made in the refinement stage, as is currently done in standard distance geometry methods. Tools for generating input files from peak lists and for coupling to refinement steps also need to be developed.

7. A flexible approach

The approach may be tailored to the problem in hand. For example, if no structural information (NOEs, residual dipolar couplings, hydrogen bonds) is available, the use of heteronuclear scalar coupling spectra alone yields enough data to produce an unfolded chain that may be inspected to obtain the assignment. This chain is expected to be broken at proline residues and where an amide proton is, for some reason, not observed. Hydrogen bonding data [5,6] will fold this chain into its elements of secondary structure. It has been shown recently that good folds may be determined with no NOE data, but using instead a rich set of residual dipolar couplings [19] – adapting this approach to unassigned data will yield structures without assignment and without NOEs. Finally, addition of NOE data yields the protein structure. If, however, a crystal structure or a reliable homology-based model is available, a molecular replacement strategy might be used in which the unfolded chain is placed onto the template structure to yield both the assignment and a structure that may be refined using additional available information. The investigation by NMR (e.g. chemical shift mapping, changes in dynamics) of the interactions of a protein of known structure may thus be greatly accelerated.

8. A way forward?

The results presented here were obtained using synthetic data sets but lay the foundation for a new approach to the structural characterisation of proteins in solution that entirely avoids the pre-requisite of assignment. The approach uses the same information as assignment-based structure determination, but in an alternative manner. As such, it is no more nor less susceptible to the problems of resonance overlap, or incomplete data: indeed, it is only the use of a full covalent structure in conventional methods that creates an impression of completeness. Development of this approach will allow the assignment stage to be by-passed and will allow NMR to perform optimally in solving large numbers of structures and in characterising larger systems, such as labelled components of complexes, that are becoming amenable to structure determination by NMR through higher fields, the TROSY experiment, deuteration and protein splicing (for reviews, see [20,21]).

Acknowledgements: We thank Dr A. Bax for kindly providing a full set of assignments for ubiquitin and Drs G. Kelly, B. Kieffer and G. Travé for numerous discussions. We also thank the Institut National de la Santé et de la Recherche Médicale, the Centre National de la Recherche Scientifique and the Hôpital Universitaire de Strasbourg for financial support.

References

[1] Wüthrich, K. (1986) NMR of Proteins and Nucleic Acids, John Wiley and Sons, New York.
[2] Güntert, P. (1998) Q. Rev. Biophys. 31, 145–237.
[3] Cavanagh, J., Fairbrother, W.J., Palmer, A.G., III and Skelton, N.J. (1996) Protein NMR Spectroscopy: Principles and Practice, Academic Press, San Diego, CA.
[4] Tjandra, N. and Bax, A. (1997) Science 278, 1111–1114.
[5] Cordier, F. and Grzesiek, S. (1999) J. Am. Chem. Soc. 121, 1601–1602.
[6] Cornilescu, G., Hu, J.-S. and Bax, A. (1999) J. Am. Chem. Soc. 121, 2949–2950.
[7] Moseley, H.N.B. and Montelione, G.T. (1999) Curr. Opin. Struct. Biol. 9, 635–642.
[8] Mumenthaler, C. and Braun, W. (1995) J. Mol. Biol. 254, 465–480.
[9] Nilges, M. (1995) J. Mol. Biol. 245, 645–660.
[10] Mumenthaler, C., Güntert, P., Braun, W. and Wüthrich, K. (1997) J. Biomol. NMR 10, 351–362.
[11] Malliavin, T.E., Rouh, A., Delsuc, M.A. and Lallemand, J.-Y. (1992) C. R. Acad. Sci. Paris II 315, 653–659.
[12] Oshiro, C.M. and Kuntz, I.D. (1993) Biopolymers 33, 107–115.
[13] Kraulis, P. (1994) J. Mol. Biol. 243, 696–718.
[14] Atkinson, R.A. and Saudek, V. (1996), in: Dynamics and the Problem of Recognition in Biological Macromolecules (Jardetzky, O. and Lefèvre, J.-F., Eds.), pp. 49–55, Plenum Press, New York.
[15] Atkinson, R.A. and Saudek, V. (1997) J. Chem. Soc. Faraday Trans. 93, 3319–3323.
[16] Brünger, A. (1992) X-PLOR Software Manual, version 3.1, Yale University Press, New Haven, CT.
[17] Vijay-Kumar, S., Bugg, C.E. and Cook, W.J. (1987) J. Mol. Biol. 194, 531–544.
[18] Cornilescu, G., Delaglio, F. and Bax, A. (1999) J. Biomol. NMR 13, 289–302.
[19] Hus, J.-C., Marion, D. and Blackledge, M. (2001) J. Am. Chem. Soc. 123, 1541–1542.
[20] Kay, L.E. and Gardner, K.H. (1997) Curr. Opin. Struct. Biol. 7, 722–731.
[21] Wider, G. and Wüthrich, K. (1999) Curr. Opin. Struct. Biol. 9, 594–601.