Journal of Biomolecular NMR, **16:** 197–208, 2000.
KLUWER/ESCOM
© 2000 *Kluwer Academic Publishers. Printed in the Netherlands.*

Simultaneous determination of disulphide bridge topology and three-dimensional structure using ambiguous intersulphur distance restraints: Possibilities and limitations

Jérôme Boisbouvier, Martin Blackledge, Albéric Sollier* & Dominique Marion**
Institut de Biologie Structurale, Jean-Pierre Ebel C.N.R.S.–C.E.A., 41 rue Jules Horowitz, F-38027 Grenoble Cedex 1, France

Received 29 October 1999; Accepted 21 December 1999

Key words: ambiguous intersulphur restraints, 3D structure refinement, disulphide bridge topology, identification of misassigned NOE, protein

Abstract

Knowledge of the native disulphide bridge topology allows the introduction of conformational restraints between remote parts of the peptide chain. This information is therefore of great importance for the successful determination of the three-dimensional structure of cysteine-rich proteins by NMR spectroscopy. In this paper we investigate the limitations of using ambiguous intersulphur restraints [Nilges, M. (1995) *J. Mol. Biol.,* **245**, 645–660] associated with NMR experimental information to determine the native disulphide bridge pattern. Using these restraints in a simulated annealing protocol we have determined the correct topology of numerous examples, including a protein with seven disulphide bridges (phospholipase A_2) and a protein in which 25% of the total number of residues are cysteines (μ-conotoxin GIIIB). We have also characterised the behaviour of the method when only limited experimental data is available, and find that the proposed protocol permits disulphide bridge determination even with a small number of restraints (around 5 NOEs – including a long-range restraint – per residue). In addition, we have shown that under these conditions the use of a reduced penalty function allows the identification of misassigned NOE restraints. These results indicate that the use of ambiguous intersulphur distances with the proposed simulated annealing protocol is a general method for the determination of disulphide bridge topology, particularly interesting in the first steps of NMR study of cysteine-rich proteins. Comparison with previously proposed protocols indicates that the presented method is more reliable and the interpretation of results is straightforward.

Abbreviations: (S-S), disulphur bridge; SA, simulated annealing.

Introduction

Proteins are linear chains of amino acids whose three-dimensional fold is mainly driven by long-range electrostatic, hydrophobic and van der Waals interactions between parts of the linear chain distant in the primary sequence, providing the molecule with an energetically favourable stable native fold. This fold can also be highly dependent on the presence of covalently bonded cofactors, or disulphide bridges (S-S), forming stable covalent links between remote parts of the chain. The formation of (S-S) between cysteine pairs is indeed a fundamental step in the folding pathway, which, in the case of incorrect matching, results in a misfolded, non-functional protein (Kortemme et al., 1996; Raina and Missiakas, 1997). In view of the high percentage of (S-S)-containing proteins, and the current interest in exploiting a few (S-S)-stabilised folds as scaffolds for medical (Daly et al., 1999), pharmaceutical (Vita et al., 1998) or agronomical applications (Fletcher et al., 1997; Oren et al., 1998), the development of specific methodology for the determination of

*Present address: European Synchrotron Radiation Facility, B.P. 220, F-38043 Grenoble Cedex, France
**To whom correspondence should be addressed. E-mail: marion@rmn.ibs.fr

correct (S-S) topology of proteins is of considerable importance.

The presence of (S-S) corresponds to the close proximity of the sulphur atoms (2.03 Å (Allen et al., 1987)), and a limited range of side chain dihedral angles (90° for the S_γ-S_γ bonds and 103° for the C_β-S_γ bonds (Heck et al., 1994)) . In terms of structure determination of solution state biomolecules using NMR spectroscopy, the correct identification of S-S pairs therefore also constitutes a powerful conformational restraint, considerably restricting the conformational space available for the protocol. The determination of a protein structure using NMR normally relies on the measurement of a large number of distances between protons: it is well established that the precision of each distance has a limited impact on the quality of the derived structures, whereas a single incorrect assignment can dramatically bias the result (Werner et al., 1997). Similarly, structure determination using an incorrect (S-S) topology could yield a completely erroneous fold. The number of possible (S-S) patterns (Θ) remains moderate as long as the number of cysteine residues remains small ($\Theta = 15$ for 6 cysteines), but increases factorially for a higher number of cysteine residues (2N) as depicted by Equation 1:

$$\Theta = (2N)!/(2^N \cdot N!) \qquad (1)$$

In order to establish the topology of disulphide bonds prior to an NMR investigation, several theoretical or experimental methods are accessible. The alignment of the sequence of the protein under study with analogous proteins can be used in the case of a high percentage of sequence identity or homology. Although this method is quick and straightforward, it should be used with caution: the patterns can differ for two proteins even with high sequence identity (>70%) and closely related biological activity (Smith et al., 1996). Biochemical methods provide direct experimental insight into the (S-S) topology: fragments are produced using specific endoproteases (Creighton, 1989) and then characterised by protein sequencing or mass spectrometry. An (S-S) bond is revealed as a link between two peptidic fragments. A small amount of material is needed for this strategy, which is however prone to misleading interpretation: lack of suitable cleavage products has been reported for small cysteine-rich proteins (Calvete et al., 1991) or for protease resistant proteins (Daly et al., 1999). For such cases, clever biochemical techniques have been devised (Gray, 1993), at the expenses of the experimental time and complexity. Moreover, in a few cases, which remain unexplained, unexpected side-reactions such as disulphide rearrangements (Lebrun et al., 1997) can occur, leading to misassigned disulphide bonds (Erlanson et al., 1974). These pitfalls warrant the development of an independent and reliable method in order to directly assign disulphide bonds by NMR using the native protein. Unfortunately for NMR spectroscopy, none of the sulphur isotopes exhibit a spin 1/2, which would enable the direct proof of the (S-S) formation by means of a J-correlation experiment. In contrast, ^{13}C NMR spectroscopy provides information on the redox state of each individual cysteine, as the Cys C_β chemical shift is shifted downfield by more than 10 ppm (Wishart et al., 1995) when an (S-S) bond is created. Once ^{1}H resonances are assigned, straightforward ^{13}C-^{1}H correlation experiments at natural abundance allow the identification of the Cys involved in a bridge (Boisbouvier et al., 1998). However, it should be emphasised that the assignment of the two partners in a given bridge cannot be established in this manner (Figure 1A). In order to pair the cysteines, a few experimental methods were proposed: Williamson et al. (1985) suggested the observation of H_β-H_β or H_β-H_α inter-cysteine NOEs. Unfortunately, these correlations, lying in a crowded spectral region of NOESY spectra, are generally difficult to resolve. Moreover, due to the through space nature of NOE, correlation peaks can also be observed between pairs of H_β belonging to two cysteines close in space but actually not linked together (Heitz et al., 1989; Delepierre et al., 1999). Indeed, Klaus et al. (1993) have analysed the number of cysteine residues close to another cysteine in X-ray structures of disulphide-rich proteins: up to 26% of the H_β-H_α distances shorter than 5 Å corresponds to pairs which do not belong to the same bridge and 11% of H_β-H_β distances. Such experimental probes are therefore not reliable enough to be converted into unambiguous S-S pair assignment for use during a structure refinement. An alternative strategy was proposed, where structure calculations are repeated for each possible topology in order to distinguish a single compatible restraint set (Heitz et al., 1989). Although this time-consuming method is nowadays conceivable due to the computing power available, it has been shown (Blanc et al., 1997) that the correct topology cannot always been picked out among a larger set.

These drawbacks have led to the development of alternative protocols with greater speed and reliability. When the (S-S) pattern is to be determined by NMR, a preliminary run without any input on this pattern is first carried out. Then, in favourable cases, the native topology can be inferred from a conservative comparison of the S_γ-S_γ or C_β-C_β distances with ideal disulphide geometry (Cooke et al., 1992; Johnson and Sugg, 1992; Klaus et al., 1993). More recently, Nilges has proposed a new strategy which combines *ambiguous* distances between sulphurs with NMR restraints and aims at picking out the real topology (Nilges, 1995). This protocol assumes that a given Cys is part of a bridge (as supported experimentally by the chemical shift of its C_β) and does not overinterpret the data by supposing a particular partner. During the calculations, each disulphide bond is allowed to float freely and the protocol is driven to the most compatible (S-S) topology under the influence of the 'standard' NMR restraint set. The applicability of this method was first demonstrated on a theoretical example simulated using BPTI (Nilges, 1995). We successfully applied this method on a snake venom protein (MIT1) (Boisbouvier et al., 1998), clearly determining the (S-S) pattern from 945 possibilities, which could not be established using simple biochemical methods due to the exceptional resistance of MIT1 to endoproteases. The purpose of the present study is to investigate the general viability of this method in the case of other cysteine-rich proteins for which the structure has been solved by NMR. Finally, the advantages and limitations of this method are discussed.

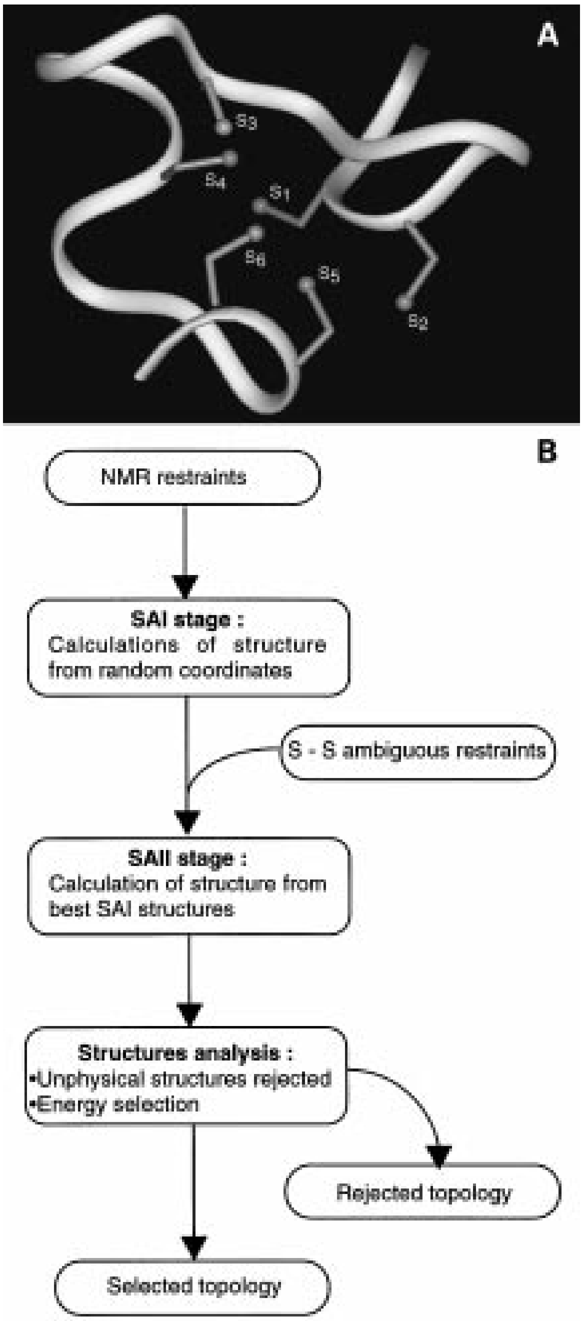
A
S3
S4
S1
S6
S5
S2
B
NMR restraints
SAI stage :
Calculations of structure from random coordinates
S - S ambiguous restraints
SAII stage :
Calculation of structure from best SAI structures
Structures analysis :
•Unphysical structures rejected
•Energy selection
Rejected topology
Selected topology

Figure 1. Ambiguous intersulphur distance restraints for (S-S) pattern determination. (A) Backbone and cysteine side chain representation of best SAI structure calculated for μ-Conotoxin GIIIB. Without (S-S) ambiguous restraints, the proximity of all sulphur atoms renders the direct determination of native intersulphur connectivity virtually impossible. (B) Summary of the different steps used in the proposed protocol for the determination of (S-S) topology from experimental NMR restraints.

Materials and methods

Determination of the protein fold

Structure calculations were performed on SGI computers (R10000) using DISCOVER interfaced to the InsightII program for visualisation and result analysis (version 6/97, MSI). For the simulated annealing (SA) protocols, the AMBER4 force field was used for the covalent terms, whereas the non-covalent terms were replaced by simple quartic non-bonded terms (Blackledge et al., 1995). All peptide ω dihedrals were forced to *trans*, except the one preceding *cis*-Pro, when strong evidence from the NOE data indicated that this peptide bond occurs in *cis* conformation. All distance, backbone ϕ angles and χ_1 angle restraint sets used in this work were previously measured with classical NMR experiments (Wüthrich, 1986). Distance restraint lower bonds were fixed to 1.7 Å. Restraints involving groups of equivalent hydrogen atoms (methyl, aromatic, ...) were taken into account as a $\left(\sum r^{-6}\right)^{-1/6}$ sum (Nilges, 1993). A floating chirality was used for non-stereoassigned methylene protons (Folmer et al., 1997).

The global fold of proteins was determined using a 60 ps SA protocol (called SAI) at a nominal temperature of 1000 K starting from randomised initial co-ordinates and containing 10 ps of slow cooling to 300 K. During this stage only NMR data (with no information concerning cysteine sulphur atoms) were used, distance restraints were scaled to a maximum of 50 kcal mol^{-1} Å^{-2} (200 kcal mol^{-1} rad^{-2} for dihedral restraints) during the first 30 ps. The molecule was then minimised with the same force field as used in the previous SA; the electrostatic interaction between neighbour cysteine sulphur atoms were suppressed to avoid repulsion. It should be kept in mind that in a regular protein structure calculation the repulsion between two S atoms in a bridge is not taken into account due to the covalent bond.

Ambiguous restraints for disulphide bridge topology assignment

In a second protocol, the information of (S-S) is incorporated in the form of ambiguous restraints (from 0 to 2.03 Å, with 100 kcal mol^{-1} Å^{-2} as force constant) between any Cys-S$_\gamma$ and all other Cys-S$_\gamma$ (Nilges, 1995). We have used the implementation available in DISCOVER for handling overlapped NOE cross peaks: each Cys-S$_\gamma$ atom is forced to get close to at least one among the 2N − 1 other Cys-S$_\gamma$ atoms according to Equation 2:

$$\begin{aligned} r_{\text{effective}} &= \left(\sum_{i\neq j} \frac{1}{\left(r_{S_i-S_j}\right)^6}\right)^{-\frac{1}{6}} && \\ E_{\text{SSamb}} &= 0 && \text{if } r_{\text{effective}} < r_u \\ \text{or } E_{\text{SSamb}} &= (r_{\text{effective}} - r_u)^2 && \text{if } r_{\text{effective}} > r_u \end{aligned} \qquad (2)$$

As DISCOVER allows the use of ambiguous restraints only between H atoms, we have modified the AMBER4 force field library to include a pseudo cysteine residue, with an H$_\gamma$ proton overlapping on the S$_\gamma$ atom: note that apart from the distance restraint the proton is transparent to the calculation. Note that in the DISCOVER implementation $r_{\text{effective}}^{-6}$ is divided by the number of ambiguous partners. The lower distance for the ambiguous restraints is set to 0, but steric repulsion between two Cys-S$_\gamma$ atoms prevents them from getting closer than 1.67 Å without incurring an energy penalty. Such restraints can be satisfied by more than two Cys-S$_\gamma$ atoms, especially for small Cys-rich proteins or poorly defined structures (few NOEs). Should such a physically unlikely situation arise, these structures can be identified a posteriori and rejected (Nilges, 1995). Note that in the case of a starting SAII structure with incorrect topology, the convergence towards the native topology needs to go through an intermediate unphysical situation, where more than two sulphur atoms are in close proximity.

The best SAI structures selected on the basis of their experimental energy were used as starting coordinates for a second SA protocol (SAII) in which we introduced this ambiguous information for each cysteine sulphur atom in addition to other NMR restraints (Figure 1B). This SAII protocol starts with a 5 ps scaling period at 2000 K followed by 2 ps of sampling before smooth cooling to 100 K over 13 ps, then a further 2 ps of dynamics was performed. Structures were then minimised as in SAI. This second SA protocol at high temperature allows side chain reorientation, but does not permit refolding of the molecule (Blackledge et al., 1995): note that dynamic pairing of Cys residues involves side chain adjustment.

Efficiency and limits of the protocol

For the present study, the NMR restraint sets were taken from the Brookhaven Protein Data Bank: only proteins with three or more (S-S) per chain were considered. An (S-S) is considered assigned as soon as one sulphur atom is closer than 2.1 Å from another isolated cysteine sulphur atom. All final SAII structural ensembles were analysed and the determined disulphide topologies were compared with those determined in the original study. Half of the structures from each ensemble were selected for further analysis using the experimental (target function comprising distance and dihedral restraints) and physical (covalent geometry and quartic non-bonded terms) energy terms (Figure 1B) as criteria. In a first step, the two-stage SA protocol has been applied on six different examples (Figure 2) including Dendrotoxin (1dtk), Phospholipase A2 (1sfv), Kistrin (1kst), Flavoridin (1fvl), μ-Conotoxin GIIIB (1gib) and epidermal growth factor-like module of human complement protease C1r (1apq).

The case of Dendrotoxin was used to analyse how the quality of the NMR data affects the efficiency of the (S-S) assignment: this very complete data set includes a large number of correctly assigned NOE information (Berndt et al., 1993).

(a) Initially the number of NMR restraints was examined. When the signal-to-noise ratio of NOESY spectra drops, then the cross peaks corresponding to larger distances become harder to observe and as-

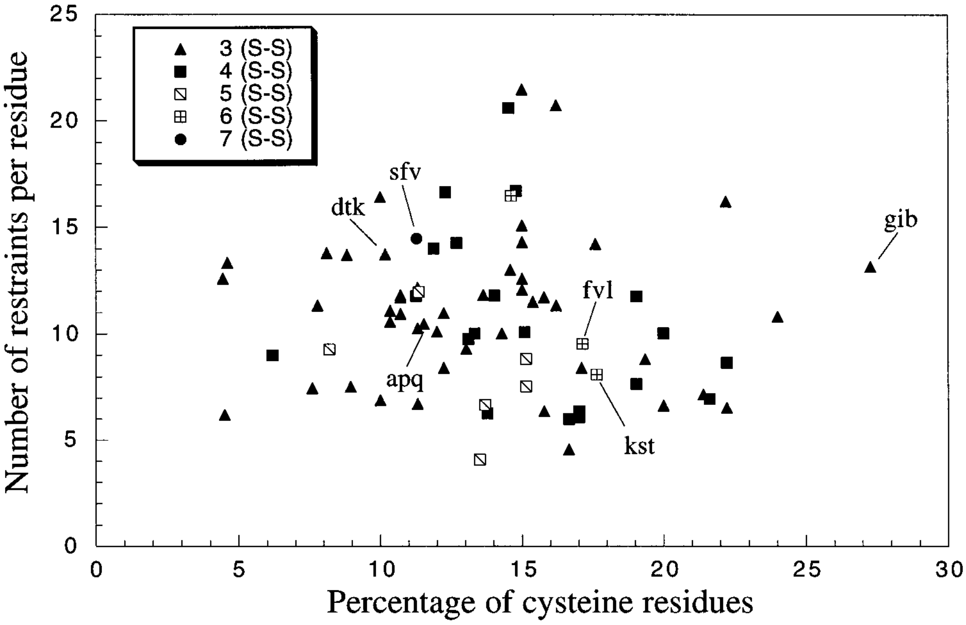

Figure 2. Distribution of NMR structure of cysteine-rich protein. The x-axis corresponds to the proportion of cysteine residues in a sequence and the y-axis to the number of restraints per residue. Only structures resolved after 1990 for which the NMR restraints file has been deposited in the PDB are taken into account. Label points correspond to studied examples. The notation is the same as in Table 1.

Table 1. Statistical analysis of (S-S) determination

Protein PDB code (AA)	Number of (S-S)	Restraints total (LR)	Calculated structures	Non-physical structures	Non-native (S-S)	Selected structures	Native (S-S)	Convergency rates (%)
Non-modified PDB file								
1dtk(57)	3	761(287)	96	0	0	48	48	100
1gib(22)	3	302(22)	100	15	2	43	43	100
1apq(52)	3	543(180)	105	9	25	48	41	85.4
1fvl(70)	6	666(219)	100	31	1	35	35	100
1kst(68)	6	549(216)	70	26	2	22	21	95.5
1sfv(124)	7	1727(549)	100	1	1	50	50	100
Modified PDB file								
1dtk-2(57)	3	298(57)	100	0	0	50	50	100
1tdk-3(57)	3	268(43)	103	0	1	52	52	100
1dtk-4(57)	3	249(35)	96	1	13	48	42	87.5
1dtk-5(57)	3	222(28)	100	20	22	40	30	75
1dtk-A(57)	3	769(296)	120	4	7	58	56	96.5
1dtk-B(57)	3	769(296)	100	2	10	49	43	87.8

Non-physical structures and non-native (S-S) topology are related to all SAII calculated structures. Native (S-S) topology and convergency rates are related to the selected SAII structures (see Materials and methods and Figure 2). LR: long range distance restraints; AA: number of residues.

sign. On this basis, we have used the upper bound distance as a selection criterion, and have progressively reduced the number of restraints from 761 to 222 (this latter case corresponds to a maximum upper bond of 3.3 Å). As χ_1 side chain rotamer identification depends both on J coupling measurements and intra-residue NOE data (H_N-H_β and H_α-H_β), we have removed χ_1 restraints as soon as one of these NOEs was eliminated from the restraint input file.

(b) The influence of NOE assignment errors was then examined by introducing eight randomly misassigned long-range NOE cross peaks. Distance upper limits were set to 5 Å between H_N and H_β of the two randomly chosen residues. With such modified restraints, the above two-stage protocol was repeated twice. Whereas the regular protocol is applied in the first case (1dtk-B), the penalty function was modified in the second run to limit to 50 kcal Å^{-1} the force brought on by any violated NOE. This modification was introduced to take into account the fact that a restraint file is usually self-consistent if the protein does not undergo large conformational transitions. By limiting the penalty to pay due to a misassigned NOE, one limits the consequences of a single mistake, which may otherwise be resolved by overall distortion of the structure. Final Dendrotoxin structures calculated with these modified restraint files have been analysed and (S-S) topologies assigned in each case.

Results and discussion

Among the 569 NMR structures deposited in the PDB in May 1999, 92 restraint sets correspond to a Cys-rich protein structure with at least 3 (S-S). These Cys-rich proteins belong to 26 different 3D folds (Murzin et al., 1995), the highest populated of which corresponds to a knottin fold (Heitz et al., 1989). Not surprisingly, the number of entries decreases with the number of (S-S) bonds – a single example with seven bridges (Phospholipase A2 – 124 residues) is available. In order to thoroughly evaluate the efficiency of our protocol, examples of increasing complexity were chosen. The success or failure in assigning the bridges in a protein depends not only on the intricacy of the bridge pattern (number of Cys and spatial proximity) but also on the quality of the NMR data. The first feature is intrinsic to each protein and can be roughly characterised by the proportion of cysteine residues in a sequence. In contrast, many parameters influence the quality of the data: some are related to the protein (its solubility, its shape, its internal flexibility) while others are user-determined (amount of purified protein, magnetic field strength, experimental time). A standard measure of the NMR data quality is provided by the number of restraints per residue, although the relative proportions of redundant and useful restraints are not always separated. The distribution of the Cys-rich proteins as a function of these two criteria (number of bridges and of restraints per residue) is shown in Figure 2. We have included in our study the three proteins with the highest number of (S-S): Phospholipase A2 (Van den Berg et al., 1995; PDB-code: 1sfv) with seven bridges, Flavoridin (Senn and Klaus, 1993; PDB-code: 1fvl) and Kistrin (Adler et al., 1991; PDB-code: 1kst) with six bridges. μ-Conotoxin GIIIB (Hill et al., 1996; PDB-code: 1gib), with 3 bridges for only 22 amino acids, has been also considered as an extreme case.

The protocol presented in Figure 1 was implemented for the six molecules with different (S-S) topological complexity. The results are presented in Table 1. The (S-S) distance selection is illustrated in Figure 3 for the Dendrotoxin (Berndt et al., 1993; PDB-code: 1dtk), a member of the BPTI-fold family with three (S-S) for 57 residues. Ninety-six structures were calculated. With this very complete data set (761 NOEs including more than one third of long-range restraints), the SA protocol leads to a single and correct (S-S) topology (5–55, 14–38, 30–51). The physical or experimental energy terms at the end of this computation are comparable to those obtained without the ambiguous restraints (data not shown), demonstrating the consistency of the NOE data set with the (S-S) topology. Thus, the present study confirms that the proposed method is fully appropriate for assigning the (S-S) topology from real experimental NMR information.

The case of the μ-Conotoxin GIIIB, a 22-residue peptide with three bonds and a small NOE data set (22 long-range restraints), is more demanding. In the initial structures calculated using SAI, all six sulphur atoms are clustered together within the polypeptide core (Figure 1A). Nevertheless, among the 85 structures obtained after the SAII stage, only two do not have the native topology. Moreover, these two exceptions correspond to the highest experimental and physical energies of the ensemble (Figure 4) and are therefore not retained for further analysis following the selection criterion described in the Methods section.

As the number of cysteines increases, the possible pairing topologies increases factorially: for instance,

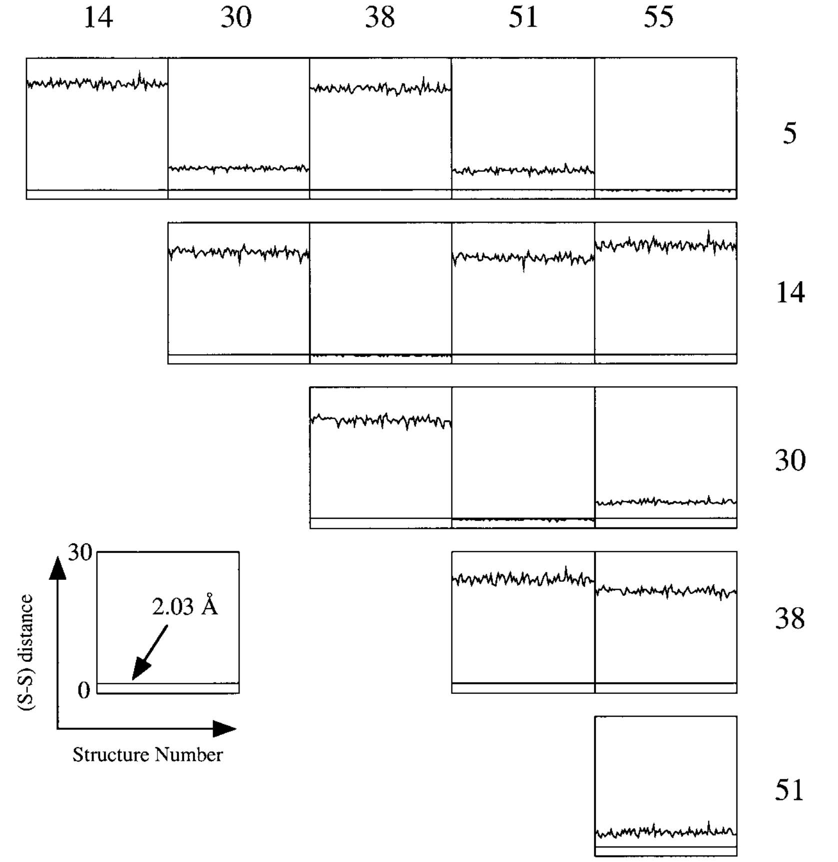

Figure 3. Distribution of the distances between cysteine sulphur atoms of Dendrotoxin SAII structures. The x-axis corresponds to each of the conformers of the ensemble. The y-axis corresponds to the distance (from 0 to 30 Å) between S atoms. The normal S-S distance for a disulphide bridge (2.03 Å) is shown as a horizontal line. Each column (and row) corresponds to one cysteine residue. In each box are plotted intersulphur distances of cysteine residues corresponding to a row and column of this box.

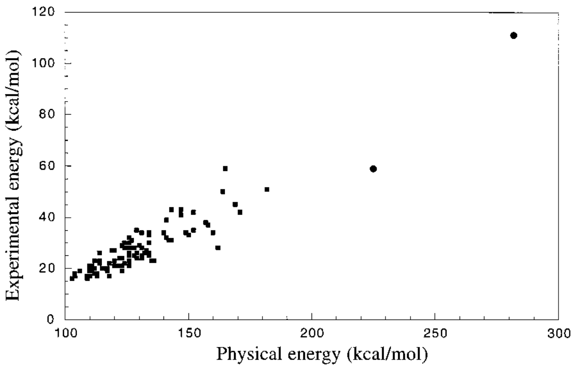

Figure 4. Distribution of energy of μ-Conotoxin GIIIB SAII structures. The x-axis and y-axis correspond to physical and experimental energies, respectively (in kcal/mol). Calculated structures with native (S-S) topology are represented by squares and others by a point. Non-physical structures containing clusters with more than two sulphurs have been discarded from the structure ensemble.

135135 possibilities are available for Phospholipase A_2, with seven bridges. Once the topologies bringing more than two Cys sulphurs in close proximity are discarded, all selected structures correspond to native topologies for Phospholipase A2 (seven bridges) and Flavoridin (six bridges). For Kistrin (high homology with Flavoridin) the success rate is reduced to 95.5% (see Table 1); note that the latter has a slightly less complete NOE data set than Flavoridin. It is also interesting to note that the rate of unphysical structures obtained is higher for the Flavoridin and Kistrin (about 35%) and for μ-Conotoxin (15%) than for both Dendrotoxin and Phospholipase A2, supporting the intuitive assumption that the challenge of the (S-S) assignment is increased by the density of the cysteines in a protein and facilitated by good quality of the NOE data set (see Figure 2).

These examples demonstrate that the native (S-S) topology can be established with a high level of confidence for a variety of proteins, for a variety of chain lengths (22 to 124 residues), number of (S-S) (3 to 7) or overall fold. No correlation was found between the initial (S-S) distances present in the selected SAI ensemble and the final (S-S) topologies present in SAII. The second part of the algorithm (SAII) therefore sufficiently samples conformational space to avoid any detectable dependence on initial fold. The topology found matches the best compromise between all experimental information and is not derived from a few NOEs, as is the case when the assignment is done manually. It is of interest to investigate whether this protocol remains reliable, when the NOE information is not longer equally distributed along the peptidic chain, for example under the conditions when a part of a protein undergoes conformational averaging and signals become broad, reducing the number of observable NOEs.

Rapid conformational averaging of this kind occurs for the epidermal growth factor (EGF)-like module of human complement protease C1r (Bersch et al., 1998; PDB-code: 1apq). This domain shares the common EGF module (S-S) pattern, i.e. bridges between C129 and C148, C159 and C172 and C144 and C157, but has an unusually large loop connecting C129 and C144. The statistical distribution of the NOEs (Figure 5A) is clearly non-uniform and very few long range restraints have been observed and assigned between C129 and C144. We have applied our protocol to this case to analyse the consequence of a local lack of restraints on the (S-S) topology. Most of the non-native topologies display much higher energy than the native one (see Table 1) and can thus be rejected. As far as the pairing of the cysteine residues is concerned, there is a clear contrast between one bridge (C159–C172) and the two others (C129–C148) and (C144–C157). All 48 retained structures show the correct topology between C159 and C172 due to the numerous NOEs in this part of the protein. However, a small number of structures show an interchange of the pairing of C144 and C148, although the native (S-S) topology remains statistically favoured. In comparison to the calculation without intersulphur ambiguous restraints (Figure 5B), the rmsd to the mean structure is similar for the well-structured regions (the two-third near the C-terminus), but a lower rmsd (about 1 Å) is observed around the flexible loop. In conclusion, this protocol not only provides information on the (S-S) topology, but can also increase the precision of the derived structure.

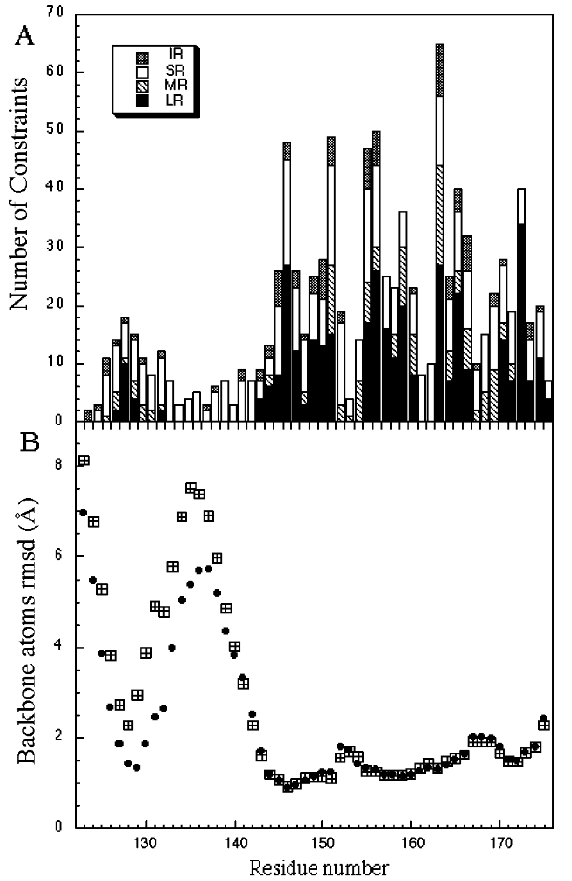

Figure 5. Experimental statistics for C1r-EGF module. (A) Sequential distribution and range of NOE restraints: black, long-range; hatched, medium-range (i≤4); white, sequential; and grey, intraresidual restraints. (B) Positional backbone rmsd of C1r-EGF module along the protein sequence. SAII structures have been superposed on (N, C_α and C′) of the mean structure. Dashed squares represent rmsd of SAII structures obtained without (S-S) ambiguous restraints. Filled circles represent the rmsd of SAII structures obtained with (S-S) ambiguous restraints (only structures corresponding to the statically favoured topology are taken into account).

Disulphide bridges assignment during preliminary steps of NMR study

NMR structure determination is often performed as an iterative process in which previous structures are used to correct and complete the NOE assignment: this is based on the assumption that initial structures are of sufficient quality to reliably assign hitherto ambiguous NOEs. As the knowledge of (S-S) connectivities improves the precision of the calculated structures, there is a strong motivation to determine their topology as soon as possible. The efficiency of our protocol obviously depends upon the number of NOEs and their

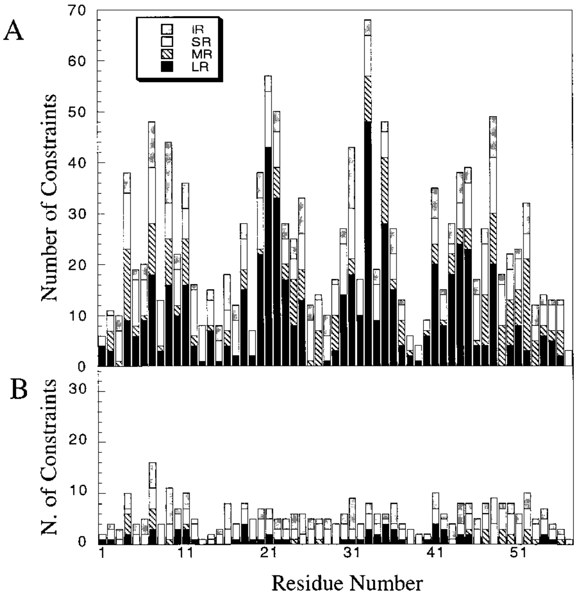

Figure 6. Sequential distribution and range of reduced and full NOE restraint files for Dendrotoxin. Black, long-range; hatched, medium-range (i$\leq$4); white, sequential; and grey, intraresidual restraints. The restraint distributions for the 1dtk and 1dtk-5 reduced restraint files are shown in parts A and B, respectively.

Table 2. Identification of misassigned NOEs

Restraint assignment	1dtk-A average violation	1dtk-B average violation	Rank in 1dtk-B violation list
3_H$_{\beta 2}$/18_H$_{\mathrm{N}}$ #	6.77	2.2	1
52_H$_{\beta 2}$/40_H$_{\mathrm{N}}$ #	6.52	1.84	2
15_H$_{\beta *}$/31_H$_{\mathrm{N}}$ #	6.28	1.58	3
38_H$_{\beta *}$/25_H$_{\mathrm{N}}$ #	5.17	0.96	10
8_H$_{\beta *}$/20_H$_{\mathrm{N}}$ #	5.07	1.37	5
46_H$_{\beta 1}$/28_H$_{\mathrm{N}}$ #	3.86	0.73	12
49_H$_{\beta *}$/3_H$_{\mathrm{N}}$ #	2.92	0.62	15
53_H$_{\beta *}$/15_H$_{\mathrm{N}}$ #	2.20	0.12	81
9_H$_{\beta *}$/22_H$_{\beta 2}$	1.40	1.37	4
9_H$_{\beta *}$/22_H$_{\beta 1}$	1.30	0.98	9
4_H$_{\beta 1 *}$/7_H$_{\delta 1 *}$	0.94	1.33	6

In this table statistics are given for each distance restraint with a violation $\geq$3 Å during calculation 1dtk-A. In the first column, assignments of these violated restraints have been reported in decreasing order of average violation in 1dtk-A (# indicates the misassigned NOE). For each restraint, the average violation in computation 1dtk-A (modified penalty function) and 1dtk-B (classical penalty function) is reported in the second and third column. The last column reports the ranking of the same violation in the 1dtk-B calculation.

correct assignment. We have extensively analysed the limits of successful assignment for the case of Dendrotoxin (Berndt et al., 1993). To mimic the poorer quality spectra, the original distance restraint set was progressively reduced by removal of more and more restraints corresponding to the weakest NOEs (Table 1), which would be expected to disappear earlier in the case of lower sensitivity.

When more than five NOEs per residue (and one long-range NOE per residue) are still present, only the native topology is compatible with structural restraints. The smaller the NOE set, the larger the number of non-physical structures, and the lower the difference in experimental and physical energy between structures with correct and incorrect topology. At a ratio of four NOEs per residue (Figure 6), 75% of the topologies in the retained ensemble are correct and 25% correspond to misassignment of the (S-S), although the backbone rmsd is higher than 5 Å compared to the known structure. This result, along with that for the μ-Conotoxin discussed earlier, implies that preliminary NOE data should contain at least one long-range NOE per residue for convergence. This requirement (five NOEs per residue, including at least one long-range NOE per residue) can easily be met for well-behaved proteins, as a final goal of 15 NOEs per residue is not unreasonable at currently available B_0 magnetic field strengths (Clore and Gronenborn, 1998).

For obvious reasons, deposited NMR restraint files (as used here) generally contain no significant violation of the accompanying structure. However, this is not always true at intermediate stages of a structure calculation, when (S-S) topology is of most use. We have thus checked whether our protocol remains stable when wrongly assigned NOEs are incorporated into the calculation, by adding eight misassigned long-range restraints (see Table 2) to the Dendrotoxin restraint file. The robustness of the algorithm is illustrated by the fact that the native topology remains statistically favoured (calculation 1dtk-B), although the correct topology drops from 100% to 90%. With the standard distance restraint potential (1dtk-B), any misassignment leads to a huge energy penalty which the calculation tries to compensate by an overall distortion, inducing a large number of smaller violations. In contrast, when a boundary is set for each individual violation at 50 kcal mol^{-1} Å^{-1} (see Materials and methods), the convergence rate greatly improves: the percentage of misassigned (S-S) topology is scaled down by a factor of 3.

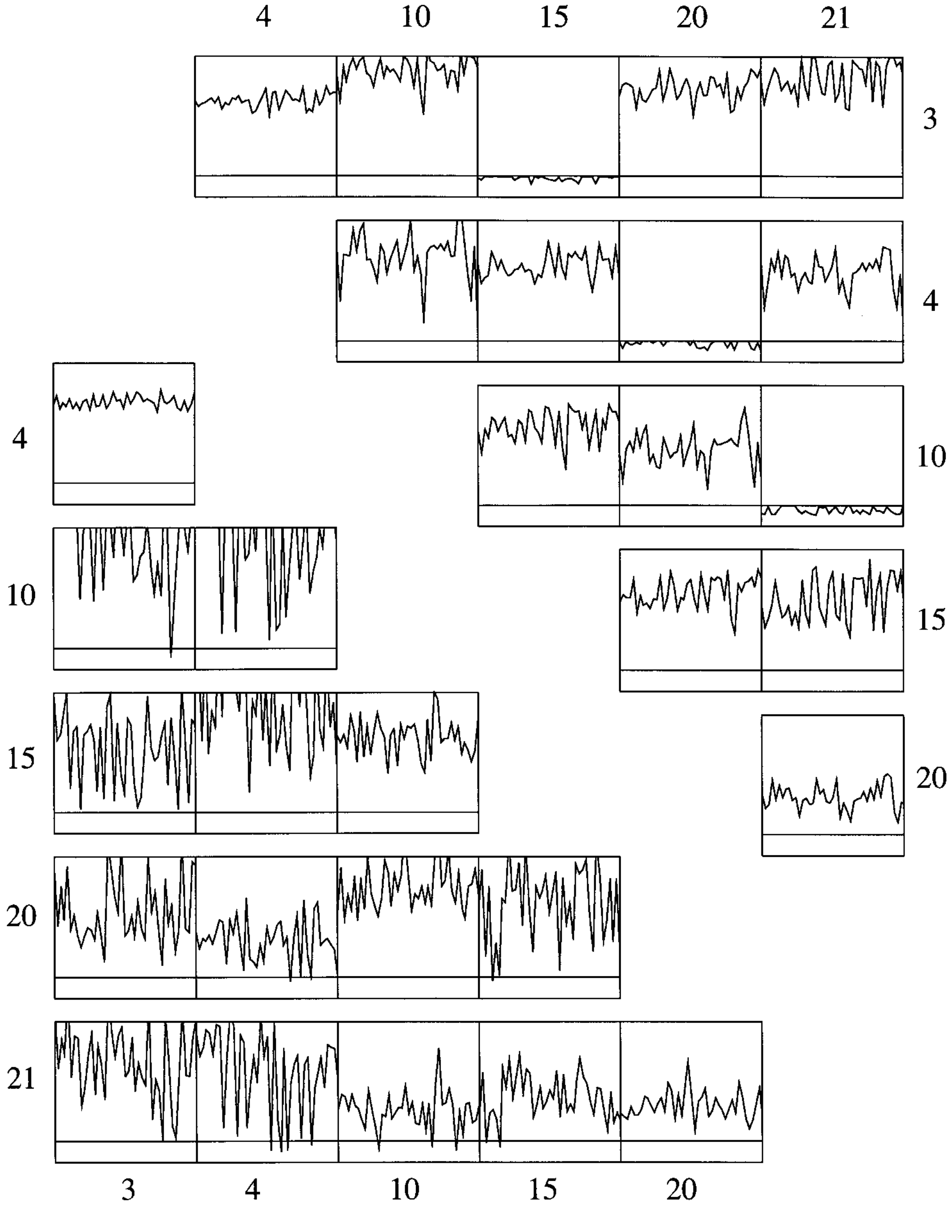

Figure 7. Comparison of distribution of the distances between cysteine sulphur atoms of μ-Conotoxin GIIIB structures. Results of SAII calculation without (left) and with (right) (S-S) ambiguous restraints. In both cases non-physical structures (see Figure 4) have been discarded and only half of the remaining structures with lower energy have been taken into account for this comparison. Each column (and row) corresponds to one cysteine residue. In each box are plotted intersulphur distances of cysteine residues corresponding to a row and column of this box. The x-axis corresponds to each of the conformers of the ensembles. The y-axis corresponds to the distance (from 1 to 8 Å) between S atoms. The normal S-S distance for a disulphide bridge (2.03 Å) is shown as a horizontal line (see Figure 3).

An interesting feature appears during the analysis of restraint violations for both runs incorporating misassigned NOEs. When there is a boundary to the penalty, the eight largest distance violations correspond to the misassigned NOEs and can thus be unambiguously identified (see Table 2). In contrast, when the classical penalty is used, only four out of the eight misassigned NOEs are among the most common violations. Using the boundary penalty during the SA protocol, these inconsistent restraints remain significantly violated, while the self-consistent restraint set leads the floating topology towards the native bridges.

Comparisons with previously proposed protocols

The only information supplied by NMR on each cysteine sulphur atom describes its oxidation state, but does not identify its (S-S) partner. The 10 ppm difference in ^{13}C chemical shifts between a reduced and an oxidised cysteine (as monitored by means of a simple ^{13}C-^{1}H correlation) allows the assignment of cysteines involved in disulphide bridges (Boisbouvier et al., 1998). The NMR spectroscopist has the choice of neglecting this information (as far as bond connectivity is concerned, a cysteine is not different than any other residue) or attempting to assign it using a local set of NOEs to create a permanent covalent bond between pairs of cysteines. The former is clearly unsatisfactory, while the latter runs a high risk of overinterpretation. The method investigated here, initially proposed by Nilges (1995), can be seen as an intermediate method which has neither of these disadvantages.

To further test the importance of ambiguous intersulphur restraints we have compared two runs of computations using the μ-Conotoxin GIIIB presented above. Our protocol was repeated with and without ambiguous (S-S) restraints; this latter case corresponds to the first method mentioned above (Cooke et al., 1992; Johnson and Sugg, 1992; Klaus et al., 1993). The distribution of intersulphur distances for both runs is displayed in Figure 7. Note that without ambiguous (S-S) restraints (lower left part of Figure 7), most of the bridge assignments remain uncertain; even though some possibilities can be rejected (3–4, 10–15, 10–20), the most probable among 15 possible topologies would have to be chosen on the basis of the average deviation from ideal intersulphur distances (Klaus et al., 1993; Hill et al., 1996). This is very difficult, as even for the true S-S bonds the mean intersulphur distance deviates considerably from the known S-S bond length (Daly et al., 1999). Such ambiguities, when combined with experimental errors, make any decision delicate. In contrast, with ambiguous intersulphur restraints added to the same set of restraints as before, the topology assignment is straightforward and reliable as shown in the upper right part of Figure 7.

An efficient NMR refinement protocol should meet two apparently conflicting requirements: it should yield structures with a reasonable precision but remain able to detect possible misassignment in the course of the refinement. We have demonstrated that our protocol achieves a correct balance between these two goals: taking into account the ambiguous restraint leads to an improvement of the resolution but errors in the restraint file can still be identified. A classical strategy without the use of (S-S) ambiguous restraints would require more computing time and may result in incorrect assignment. In contrast, the protocol proposed here allows the identification of misassigned NOEs in a single iteration. This modification can easily be implemented in most software used for NMR structure determination, and should significantly improve and shorten the refinement protocol of large proteins.

Conclusions

We have investigated the robustness of structure calculation using a simulated annealing protocol proposed for determining the protein (S-S) topology (Nilges, 1995), with experimental NMR information associated with ambiguous intersulphur restraints. We find that this protocol is generally applicable, and is able to determine the correct topology of most cysteine-rich proteins, for a variety of global folds. In most cases, the floating (S-S) topology converges towards a single topology, i.e. the native one. In less well behaved proteins, the native topology remains statistically and energetically favoured. As the (S-S) topology can be established automatically and on the basis of objective criteria, our protocol can be used in the early stages of the NMR protein structure determination. The optimisation of the (S-S) topology is not gained at the expense of other important aspects of the NMR refinement: misassigned NOEs can still be identified and an increased resolution can be achieved using this protocol.

Due to the reductive environment of the cytosol, most (S-S) bonds appear as post-translational modifications of the secreted polypeptide (Raines, 1997). The folding pathways of small disulphide-rich proteins remain uncertain: it is therefore advisable to assign the disulphide bonds directly using the native protein in the NMR sample. For such proteins, sample purified directly from the host organisms should be preferred to that obtained by chemical synthesis or bacterial overexpression, as one may be unaware of disulphide bond swapping, which may be biochemically important. When a protein is first obtained in a reduced form, it has been shown that the ratio between different (S-S) topology depends upon the oxidation conditions (air versus glutathione oxidation) (see Hernandez et al. (1997) for an example). When possible, it is advisable to carry out a functional test of the protein, to make sure that the NMR studied form is the active one and has thus the native fold.

Recent advances in NMR methodology (nano-NMR probe (Varian), ^{1}H cryo-probe (Bruker), Olson et al., 1995) enable the comparison of the fingerprint region of 2D spectra of small amounts of the natural product (Delepierre et al., 1999) with engineered samples. Once the native (S-S) topology has been ascertained, a new sample could then be engineered to obtain enough material for a classical solution structure determination. We have shown in this paper that the correct topology of the (S-S) bridges can be obtained without external information and at the preliminary stage of the NMR study.

Acknowledgements

We thank Beate Bersch for providing, before publication, the NMR-restraint set corresponding to the epidermal growth factor (EGF)-like module of human complement protease C1r. We thank Muriel Delepierre, Michel Jaquinod and Ada Prochnika-Chalufour for stimulating discussions. This work was supported by the Commissariat à l'Energie Atomique, the Centre National de la Recherche Scientifique, and Molecular Simulations Inc. (San Diego, CA). This is publication no. 734 of the *Institut de Biologie Structurale Jean-Pierre Ebel.*

References

Adler, M., Lazarus, R.A., Dennis, M.S. and Wagner, G. (1991) *Science,* **253**, 445–448.

Allen, F.H., Kennard, O., Watson, D.G., Brammer, L., Orpen, A.G. and Taylor, R. (1987) *J. Chem. Soc. Perkin Trans. II,* S1–S19.

Berndt, K.D., Güntert, P. and Wüthrich, K. (1993) *J. Mol. Biol.,* **234**, 735–750.

Bersch, B., Hernandez, J.-F., Marion, D. and Arlaud, G. (1998) *Biochemistry*, **37**, 1204–1214.

Blackledge, M.J., Medvedeva, S., Poncin, M., Guerlesquin, F., Bruschi, M. and Marion, D. (1995) *J. Mol. Biol.*, **245**, 661–681.

Blanc, E., Sabatier, J.M., Kharrat, R., Meunier, S., El Ayeb, M., Van Rietschoten, J. and Darbon, H. (1997) *Proteins Struct. Funct. Genet.*, **29**, 321–333.

Boisbouvier, J., Albrand, J.-P., Blackledge, M., Jaquinod, M., Schweitz, H., Lazdunski, M. and Marion, D. (1998) *J. Mol. Biol.*, **283**, 205–219.

Calvete, J.J., Schäfer, W., Soszka, T., Lu, W., Cook, J.J., Jameson, B.A. and Niewiarowski, S. (1991) *Biochemistry*, **30**, 5225–5229.

Clore, G.M. and Gronenborn, A.M. (1998) *Proc. Natl. Acad. Sci. USA*, **95**, 5891–5898.

Cooke, R.M., Carter, B.G., Murray-Rust, P., Hatshorn, M.J., Herzyk, P. and Hubbard, R.E. (1992) *Protein Eng.*, **5**, 473–477.

Creighton, T.E., In *Protein Structure, a Practical Approach* (Ed., Creighton, T.E.), IRL Press, Oxford, pp. 155–167.

Daly, N.L., Koltay, A., Gustafson, K.R., Boyd, M.R., Casas-Finet, J.R. and Craik, D.J. (1999) *J. Mol. Biol.*, **285**, 333–345.

Delepierre, M., Prochnicka-Chalufour, A., Boisbouvier, J. and Possani, L. (1999) *Biochemistry*, **38**, 16756–16765.

Erlanson, C., Charles, M., Astier, M. and Desnuelle, P. (1974) *Biochim. Biophys. Acta*, **359**, 198–203.

Fletcher, J.I., Smith, R., O'Donoghue, S.I., Nilges, M., Connor, M., Howden, M.E.H., Christie, M.J. and King, G.F. (1997) *Nat. Struct. Biol.*, **4**, 559–566.

Folmer, R.H.A., Hilbers, C.W., Konings, R.N.H. and Nilges, M. (1997) *J. Biomol. NMR*, **9**, 245–258.

Gray, W.R. (1993) *Protein Sci.*, **2**, 1749–1755.

Heck, S.D., Kelbaugh, P.R., Kelly, M.E., Thadeio, P.F., Saccomano, N.A., Stroh, J.G. and Volkmann, R.A. (1994) *J. Am. Chem. Soc.*, **116**, 10426–10436.

Heitz, A., Chiche, L., Le-Nguyen, D. and Castro, B. (1989) *Biochemistry*, **28**, 2392–2398.

Hernandez, J.-F., Bersch, B., Pétillot, Y., Gagnon, J. and Arlaud, G.J. (1997) *J. Pept. Res.*, **49**, 221–231.

Hill, J.M., Alewood, P.F. and Craik, D.J. (1996) *Biochemistry*, **35**, 8824–8835.

Johnson, B.A. and Sugg, E.E. (1992) *Biochemistry*, **31**, 8151–8159.

Klaus, W., Broger, C., Gerber, P. and Senn, H. (1993) *J. Mol. Biol.*, **232**, 897–906.

Kortemme, T., Hollecker, M., Kemmink, J. and Creighton, T.E. (1996) *J. Mol. Biol.*, **257**, 188–198.

Lebrun, B., Romi-Lebrun, R., Martin-Eauclaire, M.-F., Yasuda, A., Ishiguro, M., Oyama, Y., Pongs, O. and Nakajima, T. (1997) *Biochem. J.*, **328**, 321–327.

Murzin, A.G., Brenner, S.E., Hubbard, T. and Chothia, C. (1995) *J. Mol. Biol.*, **247**, 536–540.

Nilges, M. (1993) *Proteins Struct. Funct. Genet.*, **17**, 297–309.

Nilges, M. (1995) *J. Mol. Biol.*, **245**, 645–660.

Olson, D.L., Peck, T.L., Webb, A.G., Magin, R.L. and Sweedler, J.V. (1995) *Science*, **270**, 1967–1970.

Oren, D.A., Froy, O., Amit, E., Kleinberger-Doron, N., Gurevitz, M. and Shaanan, B. (1998) *Structure,* **6**, 1095–1103.

Raina, S. and Missiakas, D. (1997) *Annu. Rev. Microbiol.*, **51**, 179–202.

Raines, R.T. (1997) *Nat. Struct. Biol.*, **4**, 424–427.

Senn, H. and Klaus, W. (1993) *J. Mol. Biol.*, **232**, 907–925.

Smith, K.J., Jaseja, M., Lu, X., Williams, J.A., Hyde, E.I. and Trayer, I.P. (1996) *Int. J. Pept. Protein Res.*, **48**, 220–228.

Van Den Berg, B., Tessari, M., de Haas, G.H., Verheij, H.M., Boelens, R. and Kaptein, R. (1995) *EMBO J.*, **14**, 4123–4131.

Vita, C., Vizzavona, J., Drakopoulou, E., Zinn-Justin, S., Gilquin, B. and Ménez, A. (1998) *Biopolymers*, **47**, 93–100.

Werner, M.H., Clore, G.M., Fisher, C.L., Fisher, R.J., Trinh, L., Shiloach, J. and Gronenborn, A.M. (1997) *J. Biomol. NMR*, **10**, 317–328.

Williamson, M.P., Havel, T.F. and Wüthrich, K. (1985) *J. Mol. Biol.*, **182**, 295–315.

Wishart, D.S., Bigam, C.G., Holm, A., Hodges, R.S. and Sykes, B.D. (1995) *J. Biomol. NMR*, **5**, 67–81.

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