Journal of Biomolecular NMR, **21:** 221–233, 2001.
KLUWER/ESCOM
© 2001 *Kluwer Academic Publishers. Printed in the Netherlands.*

Rapid protein fold determination using secondary chemical shifts and cross-hydrogen bond ^{15}N-^{13}C′ scalar couplings (^{3hb}J$_{NC'}$)

Alexandre M.J.J. Bonvin**, Klaartje Houben, Marc Guenneugues*, Robert Kaptein & Rolf Boelens
Bijvoet Center for Biomolecular Research, NMR Spectroscopy, Utrecht University, Padualaan 8, 3584 CH Utrecht, the Netherlands

Received 30 July 2001; Accepted 4 September 2001

Key words: chymotrypsin inhibitor 2, cross-hydrogen bond ^{15}N-^{13}C' scalar couplings, protein structure determination, secondary chemical shifts, TALOS

Abstract

The possibility of generating protein folds at the stage of backbone assignment using structural restraints derived from experimentally measured cross-hydrogen bond scalar couplings and secondary chemical shift information is investigated using as a test case the small α/β protein chymotrypsin inhibitor 2. Dihedral angle restraints for the ϕ and ψ angles of 32 out of 64 residues could be obtained from secondary chemical shift analysis with the TALOS program (Corneliscu et al., 1999a). This information was supplemented by 18 hydrogen-bond restraints derived from experimentally measured cross-hydrogen bond ^{3hb}J$_{NC'}$ coupling constants. These experimental data were sufficient to generate structures that are as close as 1.0 Å backbone rmsd from the crystal structure. The fold is, however, not uniquely defined and several solutions are generated that cannot be distinguished on the basis of violations or energetic considerations. Correct folds could be identified by combining clustering methods with knowledge-based potentials derived from structural databases.

Introduction

The number of NMR structures deposited into the Protein Data Bank (RCSB PDB, Berman et al., 2000) has strongly increased over the last years. This increase is, however, much slower than the recent increase in gene sequences. NMR structure determination is still a relatively slow process and speeding up this process is crucial. Once an NMR sample is available, the most lengthy step remains the interpretation and assignment of NOE data. Several automated NOE assignment and structure calculation methods have been developed to speed up this stage (Mumenthaler and Braun, 1995; Nilges and O'Donoghue, 1998; Duggan et al., 2001). These methods, however, are often not really used in a fully automated way, but rather iteratively by combining automated assignment and manual checking steps. Most methods are also quite sensitive to what happens in the initial iterations and conformations can sometimes get trapped in wrong minima. The reliability of such automated methods and the speed at which NMR could produce high resolution structures would certainly increase if low to medium resolution starting models or folds were available early on in the structure determination process.

The problem of quickly generating NMR folds has been addressed in various ways over the last years, typically combining secondary structure information derived from secondary chemical shifts with sparse NOE data and/or residual dipolar couplings. The accuracy of protein structures obtained from a limited number of NOE data and with various structure calculation protocols has been discussed (Clore et al. 1993; Karimi-Nejad et al., 1998). A great simplification of the NOE analysis and extension of NMR to larger

*Present address: ANTOMED S.A., Bd Sébastien Brant, 67400 Illkirch, France.

**To whom correspondence should be addressed. E-mail: abonvin@nmr.chem.uu.nl proteins could be achieved by the use of perdeuteration (Gardner and Kay, 1998) with, in some cases, selective reintroduction of methyl and/or aromatic protons (Mueller et al., 2000; Medek et al., 2000). Apart from the use of sparse NOE data, methods for fold generation have been developed based on residual dipolar couplings (Delaglio et al., 2000; Fowler et al., 2000; Hus et al., 2001). Residual dipolar couplings have the advantage that they can be measured very early on in the structure determination process at the stage of backbone assignment. These methods require, however, the measurement of several residual dipolar couplings per residue, preferably in several alignment media. This can be a prohibitive task if protein sample is a limitation. Next to the more 'classical' type of software used for NMR calculations such as for example CNS (Brünger et al., 1998) or DYANA (Güntert et al., 1997), methods developed for protein structure prediction have also been used such as for example MONSSTER (Skolnick et al., 1997) that makes use of lattice models, or more recently ROSETTA (Bowers et al., 2000) that builds structures from a library of fragments using a Monte Carlo procedure.

In addition to residual dipolar couplings and NOE data, the cross-hydrogen bond scalar coupling (Dingley and Grzesiek, 1998; Pervushin et al., 1998) can provide a valuable source of structural information that can also be obtained at the stage of backbone assignment. The observation of cross-hydrogen bond scalar couplings gives direct evidence of the existence of hydrogen bonds in biomolecules. These couplings were first observed in proteins for backbone hydrogen bonds (Cordier and Grzesiek, 1999; Cornilescu et al., 1999b) but have since also been detected for side-chain backbone (Liu et al., 2000a) and side-chain side-chain hydrogen bonds (Liu et al., 2000b). Cross hydrogen bond couplings are typically measured by performing HNCO-type experiments (Cordier and Grzesiek, 1999; Cornilescu et al., 1999b). The rather long evolution periods needed to observe these very weak couplings might put limitations on the protein size for which such data can be obtained. Cross-hydrogen bond scalar couplings could, however, be detected for subtilisin PB92 (269 amino acids; Boelens, unpublished data) and a perdeuterated 30 kDa protein (Wang et al., 1999).

In this article we study the use of cross-hydrogen bond scalar couplings for defining three-dimensional protein folds. Early theoretical studies have previously indicated the great potential of hydrogen-bonds for defining a protein fold (Levitt, 1983). Here, results of ^{15}N-^{13}C$'$ cross-hydrogen bond scalar coupling ($^{3hb}J_{NC'}$) measurements are presented for the small 64 residue protein Chymotrypsin inhibitor 2 (CI2). This small protein, for which both solution (3CI2) (Ludvigsen et al., 1991) and crystal (2CI2) (McPhalen et al., 1987) structures are available, possesses an α/β fold. The experimentally identified hydrogen bonds in CI2 are combined with backbone dihedral angle restraints derived from secondary chemical shifts to calculate initial folds and assess their quality. This extremely sparse information obtained at the stage of backbone assignment is not sufficient to unambiguously determine the fold. We will, however, show that it is possible to identify correct folds from a combination of clustering methods and empirical potentials.

Material and methods

Sample preparation

The 64 residue CI2 sequence used in this work corresponds to residues 20-83 in the crystal structure (2CI2) (McPhalen et al., 1987). All NMR experiments were performed at 300 or 303 K on a 0.5 ml ^{15}N and ^{13}C labeled sample containing 2 mM CI2 in 90% H_2O/ 10% D_2O (v/v) and 50 mM of an acetate buffer with a pH of 4.6.

Chemical shift assignments

Although proton chemical shift assignments of CI2 have been published previously (Mogens et al., 1987)(BMRB accession numbers 1869–1872), a complete set of resonance assignments including those of ^{15}N and ^{13}C are reported here. The sequential connectivities were identified using HNCO, HN(CA)CO, HNCA, HN(CO)CA, CBCANH and CBCA(CO)NH spectra (Grzesiek et al., 1992; Grzesiek and Bax, 1993; Muhandiram et al. 1994). ^{1}H resonances were assigned using NOESY, TOCSY, NOESY-^{15}N-HSQC and TOCSY-^{15}N-HSQC spectra (Cavanagh et al., 1993). Side-chain ^{13}C resonances were obtained from C(CO)NH and HCCH-TOCSY spectra (Grzesiek et al., 1993; Kay et al., 1993). Side-chain NH_2 protons from Asn and Gln were identified using ^{15}N-HSQC and NOESY-^{15}N-HSQC. Prolines were not ^{13}C and ^{15}N labeled since additional unlabeled proline had been added to the expression system. All spectra were recorded on a Bruker 600 MHz spectrometer, with the exception of the C(CO)NH experiment that was recorded on a Bruker 500 MHz spectrometer. The spectra were processed using the NMRPipe software package (Delaglio et al., 1995) and analyzed using our in-house program REGINE (Kleywegt et al., 1993).

Chemical shifts-derived restraints

The chemical shifts of 91% of all nuclei in CI2 (excluding four unlabeled prolines) were determined (BMRB accession number 4974). The C_α,C_β,C',H_α and N chemical shifts of 60 residues served as input for the TALOS program (Cornilescu et al., 1999a). TALOS derives information on the ϕ and ψ backbone dihedral angles from a comparison of secondary chemical shifts patterns of amino-acid triplets against a database of secondary chemical shifts corresponding to known conformations. The predictions are classified as 'good' if 9 out the 10 best matches agree with each other, depending on the values of the ϕ and ψ backbone dihedral angles. Here, a more conservative approach was chosen requiring that all 10 best matches agree for a prediction to be accepted. No outlier should thus be present. This conservative approach was chosen after noticing a few cases where the outlier actually corresponded to the correct conformation. The TALOS 'good' predictions were converted into dihedral angle restraints as the average ϕ and ψ angles ± twice the standard deviation with a minimum of ±10 deg.

The C_α,C_β,C' and H_α chemical shifts were also analyzed according to the Chemical Shift Index (CSI) method (Wishart et al., 1994) and the results were combined with the TALOS predictions to define elements of secondary structure along the sequence.

Cross-hydrogen bond ^{15}N-$^{13}C'$ scalar couplings ($^{3hb}J_{NC'}$)

Cross-hydrogen bond ^{15}N-^{13}C$'$ scalar coupling constants ($^{3hb}J_{NC'}$) were measured from 2D-CT-HNCO spectra (Grzesiek and Bax, 1992) on a Bruker 600 MHz spectrometer using the constant time long range HNCO pulse sequence of Cordier and Grzesiek (1999). In both experiments the evolution time T was 64.5 ms. For the reference spectrum the delay between the two ^{15}N and ^{13}C 180° pulses was set to 16.6 ms to maximize the one bond coupling $^1J_{NC'}$ evolution. This delay was set to zero in the long range spectrum. The number of complex points and the spectral widths in both experiments were 112 × 512 (F1 × F2) and 11.0 ppm (F1 C$'$) and 18 ppm (F2 H_N), respectively. The total number of scans were 32 and 640 for the reference and long range experiments, respectively, resulting in a total experiment time of 6 and 60 hours.

For t_2 a 0.4π shifted sine-bell function was used and zero-filling was applied up to 2048 points. Mirror-image linear prediction was performed to double the number of points in t_1. This could be done because of the constant time evolution. A squared sine-bell function shifted by 0.45π was used as a window function for t_1. The spectra were analyzed with the NMR analysis program FELIX (Molecular Simulation Inc.) on a Silicon Graphics workstation.

Although a correlation has been demonstrated between $^{3hb}J_{NC'}$ and the inverse of the hydrogen bond distance (Dingley and Grzesiek, 1998; Cordier and Grzesiek, 1999; Cornilescu *et al.*, 1999c) the identified hydrogen bonds were uniformly converted into distance restraints with upper and lower limits of 2.5–1.7 Å between proton and acceptor (H^N-O) and 3.5–2.3 Å between donor and acceptor (N-O).

Structure calculations

All calculations were performed with CNS (Brünger et al., 1998) using the regular ARIA parameters and protocols (Nilges and O' Donoghue, 1998; Linge and Nilges, 1999).

A soft-square restraining potential was used for the hydrogen bond restraints with a force constant of 50 kcal mol^{-1} Å^{-2}. The TALOS-derived dihedral angles were restrained with a harmonic potential using a force constant of 200 kcal mol^{-1} rad^{-2}.

Covalent interactions were calculated with the 5.2 version of the PARALLHDG parameter file (Linge and Nilges, 1999) based on the CSDX parameter set (Engh and Huber, 1991). In addition to the bonded energy terms typically used in NMR structure calculations (bond, angle and improper energy terms), the dihedral angle energy term describing torsions around rotatable bonds ('dihe' flag in CNS) was turned on with a force constant of 5 kcal mol^{-1} rad^{-2} in order to improves the quality of side-chain χ_1 and χ_2 rotamers as assessed by PROCHECK (Laskowski et al., 1993). Non-bonded interactions were calculated with the repel function using the PROLSQ parameters (Hendrickson, 1985) as implemented in the new PARALLHDG parameter file. The OPLS non-bonded parameters (Jorgensen and Tirado-Rives, 1988) were used for the final water refinement including full van der Waals and electrostatic energy terms. The non-bonded pair list was generated with a 9.5 Å cut-off

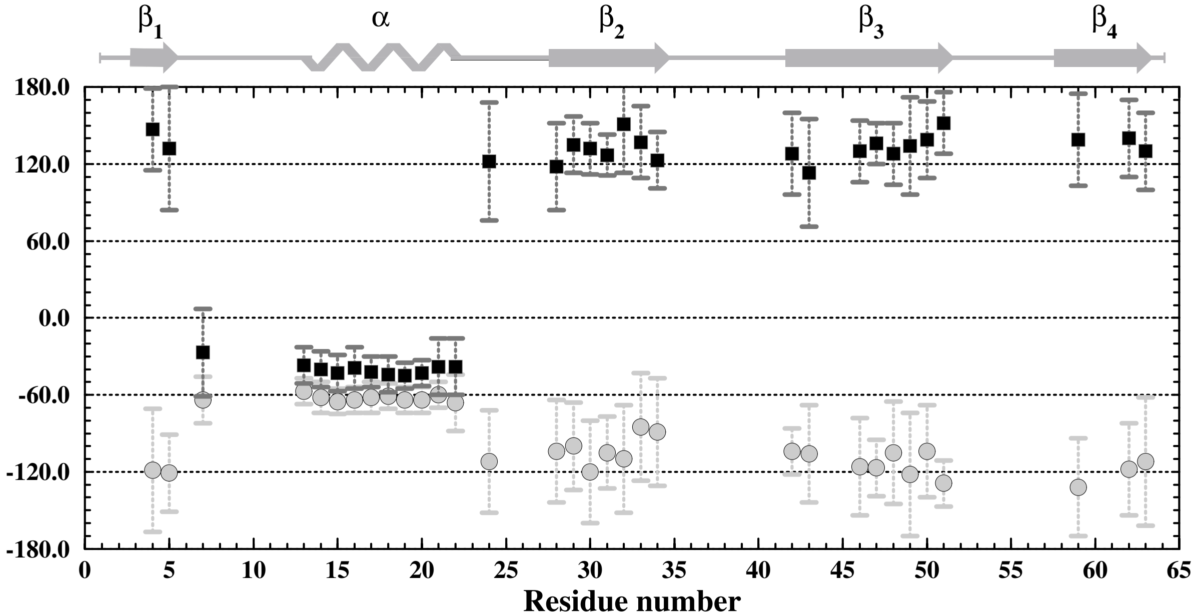

Figure 1. TALOS-derived φ (gray circles) and ψ (black squares) dihedral angle predictions as function of the residue sequence. The average angles with their respective error ranges used in the structure calculations are indicated. The secondary structure cartoon on top corresponds to the predictions obtained from TALOS (Cornilescu et al., 1999a) and the chemical shift index analysis (Wishart et al., 1994).

and the non-bonded interaction were calculated with a 8.5 Å cut-off using a shifting function.

A simulated annealing protocol was used starting from an extended conformation with a combination of torsion angle dynamics (TAD) (Stein et al., 1997) and Cartesian dynamics. Force constants were scaled throughout the protocol following the default ARIA/CNS setup. The atomic masses were set uniformly to 100 amu and the friction coefficient f_β for the coupling to the external temperature bath to 20 ps^{-1}. The simulated annealing protocol, which is similar to the one described in Linge and Nilges (1999), consisted of five stages: (i) high temperature TAD stage (10 000 steps, 10000 K), (ii) a TAD cooling phase in 50 K steps from 10 000 to 50 K in 10 000 steps, (iii) a first cooling phase in Cartesian space from 2000 to 1000 K in 10 000 steps, (iv) a second cooling phase in Cartesian space from 1000 to 50 K in 5000 steps, followed finally by (v) 200 steps of energy minimization. The time step for the integration was set to 0.003 ps. This was increased by a factor 9 during the TAD calculations while reducing the number of steps by a factor 9.

The structures were subjected to a final refinement protocol in explicit water by solvating them with a 8 Å layer of TIP3P waters (Jorgensen et al., 1992). The water refinement consisted of a heating period (50 MD steps at 100, 200, 300, 400 and 500 K, time step 0.005 ps) with harmonic position restraints on the C_α atoms (k_{harm} = 10 kcal mol^{-1} $Å^2$) followed by 2500 MD steps at 500 K without any position restraints and a final cooling stage from 500 to 100 K in 100 K steps (1000 MD steps per temperature step). The resulting structures were energy minimized with 100 steps of Powell steepest descent minimization.

Structure analysis

The stereochemical quality of the structures was analyzed with PROCHECK (Laskowski et al., 1993), their packing quality with PROVE (Pontius et al., 1996). Hydrophobic/hydrophilic solvent accessible surface areas were calculated with NACCESS (Hubbard and Thornton, 1993) using a 1.4 Å radius water probe. The quality of the generated models was further assessed by calculating their mean force potentials (average z-score over all residues) with ProsaII (Sippl, 1993) and by calculating their three-dimensional profile score with Profiler_3D (Bowie et al.,1991; Lüthy et al., 1992).

A simple clustering of the 50 best final structures selected based on their restraint energy was performed. For this, the pairwise positional root mean square deviations (rmsd) matrix was calculated for

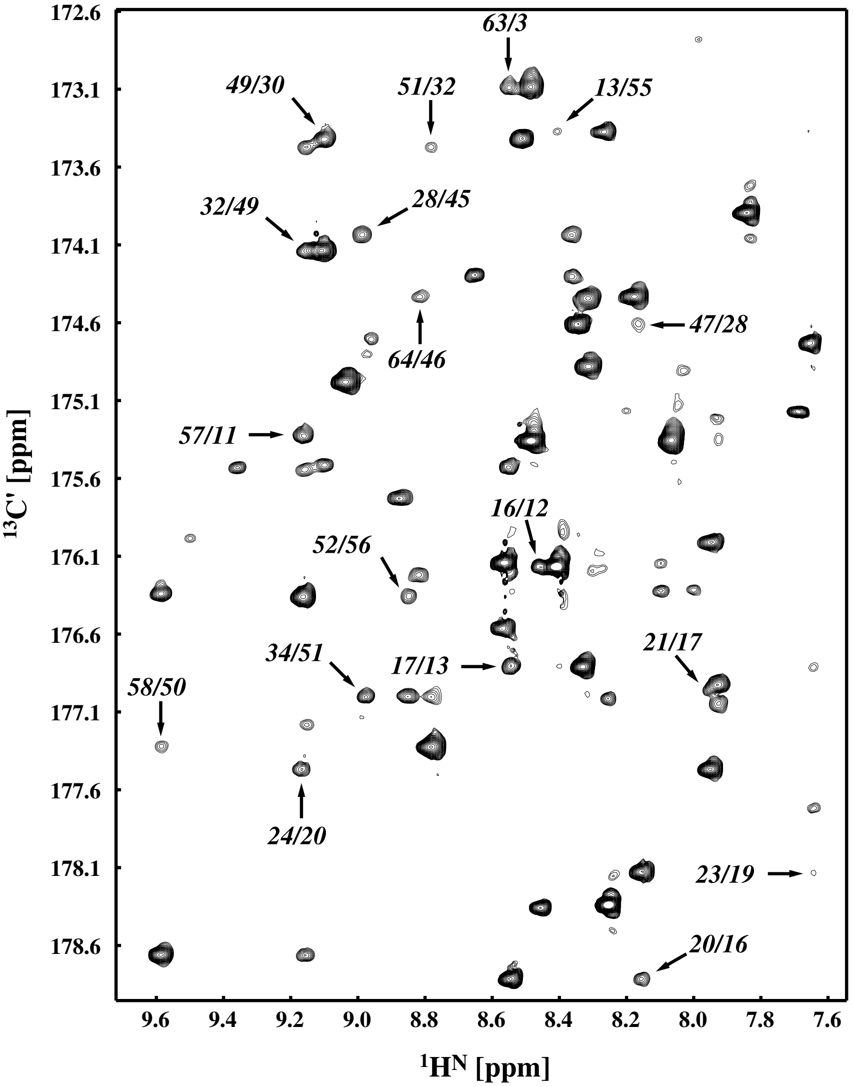

Figure 2. 600 MHz 2D long range CT-HNCO spectrum of a 2 mM ^{15}N-^{13}C CI2 sample. The spectrum was recorded with a transfer time of 2 × 64.5 ms (see Material and methods). The peaks originating from cross-hydrogen bond long range scalar couplings are labelled.

backbone atoms of the secondary structure elements identified from the CSI index. Clustering was done for various distance cut-offs ranging from 1.5 Å to 3 Å. A cluster was accepted only if it contained a least four members.

Results and discussion

Chemical shifts-derived restraints

The C_α,C_β,C′,H_α and N chemical shifts (deposited under accession number 4974 at the BMRB) of 60 out of the 64 residues (the four unlabeled prolines were excluded) were analyzed with TALOS (Cornilescu et al., 1999a). TALOS resulted in 41 ‘Good’ predictions, corresponding to 68% of the 60 residues analyzed. 9 predictions were, however, rejected because of the presence of one outlier in the 10 best matches; ϕ/ψ dihedral angle restraints were generated for the remaining 32 residues. The resulting average ϕ/ψ angles with their corresponding error ranges are given in Figure 1. From this analysis backbone dihedral angle restraints could be obtained for 50% of the residues. The TALOS predictions were combined with the results of the CSI analysis (Wishart et al., 1994) (data

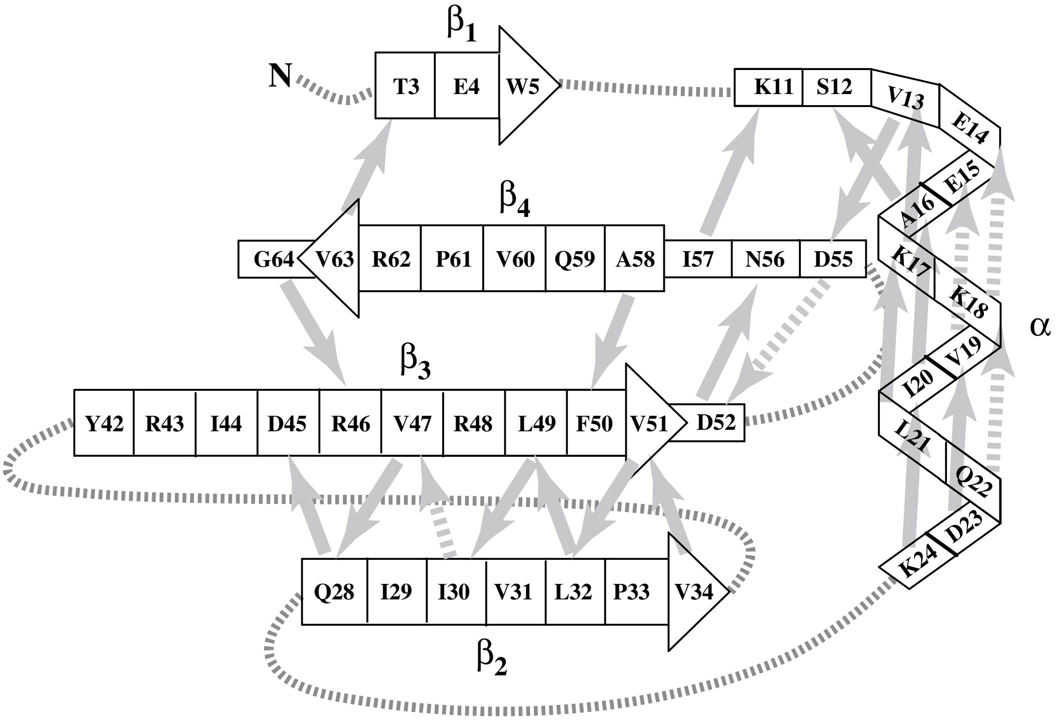

Figure 3. Schematic representation of the hydrogen bonds detected from the long range HNCO experiment. The arrows point toward the acceptor. Dashed arrows indicate hydrogen bonds identified from very weak peaks. These were not included in the structure calculations. The secondary structure elements correspond to those identified from chemical shift data.

not shown) to define elements of secondary structure along the sequence. From this, four β-sheet fragments were identified spanning residues 3 to 5, 28 to 34, 42 to 51 and 58 to 63, and one α-helix from residue 13 to 22. This information could, in principle, also be translated into dihedral angle restraints by using typical ϕ/ψ angle values for α-helix and β-sheets. In this particular case, adding those restraints did not affect the results of the structure calculations significantly (data not shown) and only TALOS-derived restraints were used. The identified secondary structure elements were, however, used for rmsd calculations in the analysis (see below).

Cross-hydrogen bond ^{15}N-$^{13}C'$ scalar couplings ($^{3hb}J_{NC'}$)

From the long-range HNCO spectrum (Figure 2) 18 peaks originating from cross-hydrogen bond ^{15}N-^{13}C′ scalar couplings could be unambiguously identified. Five additional peaks originating from cross-hydrogen bond couplings could be detected, but these are very weak and close to the noise level. The connectivities established by those hydrogen bonds are presented schematically in Figure 3. From the 18 observed hydrogen bonds (plain arrows in Figure 3) six are within the α-helix, six connect β-strands 2 and 3, well establishing the parallel orientation of those two strands. Strands 3 and 4 are connected in an antiparallel fashion by three hydrogen bonds, mainly found at the extremities of the β strands. This observation is consistent with the presence of water molecules bridging those two strands as observed in the crystal structure (McPhalen and James,1987) and by solution NMR (Melacini et al., 1999). Of the remaining three observed hydrogen-bonds, one ties the C- and N-termini together and the other two connect residues at the N-terminal side of the α-helix to the loop preceding strand 4. Three of the five weak peaks (dashed arrows in Figure 3) correspond to hydrogen bonds within the α-helix, the other two to hydrogen bonds between strand 2 and 3 and in the loop between strands 3 and 4. A number of hydrogen bonds that were observed in the NMR (3CI2) and crystal (2CI2) structures of CI2 could not be detected because of the unlabeled prolines (W5-P61) or because of peak overlap (G10-I57, R46-R43, R48-R62). A few others could not be detected (L8-W5, K11-L8, A27-K24 and R43-G64).

Seventeen cross-hydrogen bond couplings have been reported previously for CI2 (Meissner and Sørensen, 2000a, b). Four of those were not observed in this work because of the unlabeled prolines or peak overlap (see above) while 10 additional couplings could be detected.

The 23 hydrogen bonds detected by NMR define a β-sheet plane composed of four strands. Although the α-helix is tied to the residues preceding strand 4 and connected by a four residue loop to strand 2, nothing can be said about its orientation with respect to the β-sheet plane (above or below). Choosing a conservative approach, only the 18 hydrogen bonds identified from strong peaks were included as distance restraints in the structure calculations. The exclusion of the remaining 5 weak hydrogen bonds should not affect the result of the structure calculations too much: three of them are in the α-helix that is already well-defined from the TALOS prediction, one is between strand 2 and 3 which are already connected by 6 strong hydrogen bonds and the last one is in a tight turn which is already closed by one hydrogen bond (D52-N56).

Structure calculations and analysis

Using 36 distance restraints (H^N-O and N-O) derived from the 18 hydrogen bonds and the 64 ϕ/ψ dihedral angle restraints derived from the TALOS predictions for 32 residues, 200 structures were calculated following a simulated annealing protocol followed by refinement in explicit water. The best 50 models were selected for analysis after sorting the structures according to the sum of their restraint energies (Figure 4a). 30 structures out of the selected 50 have no distance violation larger than 0.5 Å and no dihedral angle violation larger than 5 Å. These are indicated by black dots in Figure 4a. The 50 structures were compared to the crystal conformation. The backbone positional rms deviations for those range from 1.0 to 9.0 Å and from 2.25 to 11.5 Å for secondary structure elements and the complete backbone, respectively (Figure 4b). Eleven structures without any violations show high rmsd from the crystal structure. The origin of this can be found in the positioning of the α-helix with respect to the plane defined by the β-sheets, most of the structures with high rmsd having the helix on the opposite side of the β-sheet plane compared to the crystal structure. For the latter, no similar folds could be detected from a DALI search of known protein structures (Holm and Sander, 1993). Two representative structures are shown in Figures 5a and 5b. Both satisfy perfectly the NMR restraints. *A priori* it is thus not possible to discard any of the generated models. This is also illustrated in the plot of restraint energies against the rmsd from the crystal structure in Figure 6a. Although structures with low restraint energies tend to be closer to the crystal structure there are still outliers resulting in a rather poor correlation ($R = 0.47$) between restraint energies and rmsd from crystal.

The structures were analyzed in various ways in order to find a good descriptor of correctly folded structure. The various energy terms used in the structure calculations could not distinguish good from bad models. All models satisfying the NMR data have similar stereochemical quality as assessed by PROCHECK (Laskowski et al., 1993) and similar packing quality as evaluated by PROVE (Pontius et al., 1996). The hydrophobic and hydrophilic solvent accessible surface areas (SASA) of the various structures were calculated with NACCESS (Hubbard and Thornton, 1993). The motivation here was that wrong folds might possibly be identified by an increased exposure of hydrophobic residues to the solvent such as originating, for example, from the positioning of the α-helix in CI2 on the opposite side of the β-sheet plane (Figure 5b). As for the previously discussed parameters this approach failed in identifying the native fold ($R = 0.37$ for hydrophobic SASA versus rmsd from crystal).

Residual dipolar couplings provide another source of experimental information that could be used to identify correct folds. It has previously been shown that their distribution pattern along the residue sequence can be used to search homologues in a database of protein 3D structures (Aitio et al., 1999; Annila et al., 1999; Meiler et al., 2000a). A set of 58 N–H residual dipolar couplings measured in a 4% w/v CPCl/hexanol solution was available for CI2 (Marc Guenneugues, unpublished data). These were used to screen the generated structures. Residual dipolar couplings were back-calculated for each structure by optimizing the parameters of the orientational tensor. The agreement between calculated and measured residual dipolar couplings was expressed in a Q-factor defined as $Q_{DC} = \Sigma(^{DC}J_{exp} - ^{DC}J_{calc})^2/\Sigma^{DC}J^2_{exp}$ (Cornilescu et al., 1998). For this, only residues in secondary structure elements were used as identified from the CSI. The Q-factors range between 0.52 and 0.95 (Figure 6b). For comparison the Q-factor calculated for the crystal structure (2CI2) is 0.25 and the values for the solution structures (3CI2) vary between 0.44

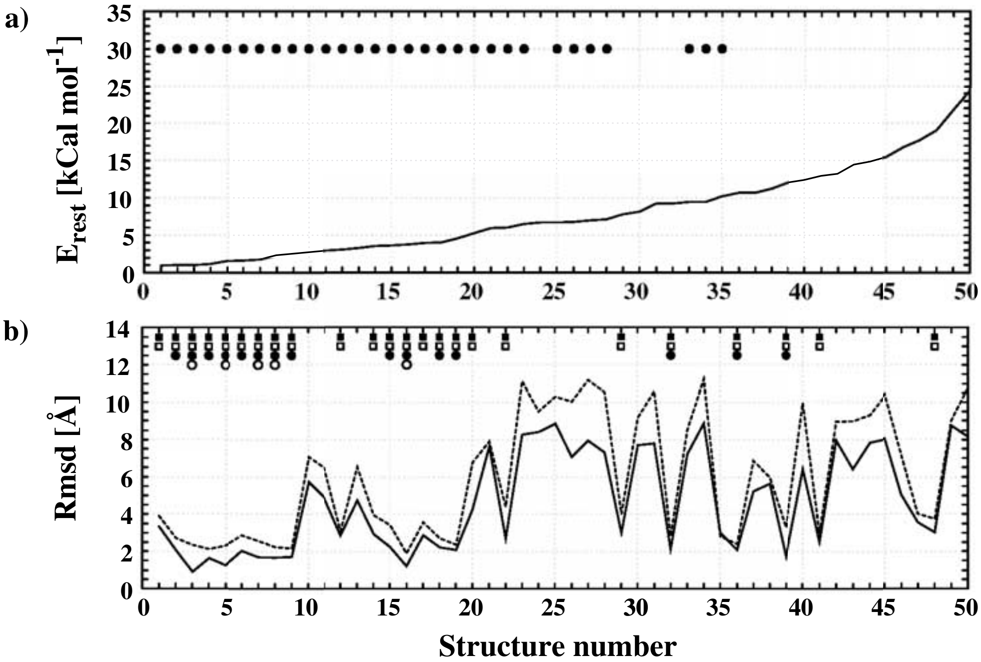

Figure 4. (a) Restraint energy sorted structures and (b) corresponding backbone positional rms deviations from the crystal structure (2CI2) (McPhalen and James, 1987). The restraint energy is the sum of distance and dihedral angle restraint energy terms. The black circles in (a) indicate those structures that do not show any distance violations larger than 0.5 Å and dihedral angle violations larger than 5°. For the rmsd calculation the structures were superimposed on C_α, C, N atoms of residues in secondary structure elements as identified from chemical shift information (see Figure 1). The rms deviations are calculated for the C_α, C, N atoms of residues in secondary structure elements (plain line) and of all residues excluding the first and last residues (dotted line). The symbols in figure (b) indicate structures belonging to the same cluster using rmsd cut-offs of 1.5 Å (open circles), 2.0 Å (filled circles), 2.5 Å (open squares) and 3.0 Å (filled squares).

and 0.74. Although there seems to be a trend that correct folds have a lower Q-factor, the correlation coefficient between rmsd from the crystal and the Q-factor is only 0.48. A few structures with incorrect folds (e.g., around 8 Å rmsd) have reasonably low Q-factors (~0.65) considering the spread in the ensemble of structures ($0.52 \leq Q \leq 0.95$) and could not be discarded based on this criterion.

Next to using residual dipolar couplings to screen the generated structures we also investigated if their inclusion as restraints in the structure calculation protocol could resolve the fold ambiguity (data not shown). The residual dipolar couplings were introduced as interprojection angle restraints (500 restraints for residues in secondary structure elements) following the approach of Meiler et al. (2000b). Although the generated structures have lower Q-factors (0.36–0.60) the fold ambiguity is not resolved and the correlation between rmsd from crystal and Q-factors is completely lost ($R = -0.05$)! The failure of using residual dipolar couplings to discriminate between folds is not surprising since the positioning of the α-helix with respect to the plane of the β-sheets corresponds to a translational component that is not included in a set of residual dipolar couplings.

Since stereochemical, geometric and experimental restraints criteria described above all fail in providing a clear distinction of correct and incorrect folds we finally resorted to empirical potentials for that purpose. These so-called knowledge-based potentials are usually derived from known high resolution three-dimensional structures and typically designed for applications in protein design, structure prediction and folding (for reviews see Sippl, 1995; Moult, 1997; Hao and Scheraga, 1999; Lazaridis and Karplus, 2000). Two such empirical potentials were used to screen the generated structures: Profiler_3D (Bowie et al., 1991) and ProsaII (Sippl, 1993). Profiler_3D assesses protein models based on a three-dimensional profile depending on the environment the residue is in. The environment is described by the buried area of a residue, the fraction of side-chain atoms covered by polar atoms and the local secondary structure. This approach has been proved quite successful in judging the quality of X-ray and NMR structures (Lüthy et al., 1992). The 3D profile score against the rmsd from the crystal is shown in Figure 6c. Although the correlation is somewhat better than for the previous parameters (R = −0.53), there is quite some scattering, even in models very close to the crystal structure. Clearly, the 3D profile would select some bad models and discard some good ones. Interestingly, the solution NMR structures (3CI2) score best here while the crystal structure has a lower score. Recalling that the environment of residues is the key aspect in this empirical potential, crystal contacts might be the reason of this rather low score. The best results are obtained with ProsaII (Figure 6d) with a correlation coefficient of 0.73 between rmsd from crystal and the ProsaII z-score. The ProsaII z-score is built from C_α-C_α and C_β-C_β pair interaction potentials and surface potentials and gives an estimate of the quality of a given structure. The crystal structure (2CI2) scores best, followed by the NMR structures (3CI2). As was the case with the residual dipolar couplings, there is a trend that the lowest scores correspond to the correct fold. Outliers are however also present as illustrated by the model around 7 Å rmsd with a low z-score. Again, it would be difficult to discard the latter based on its ProsaII score. Even various combinations of the above parameters fail in unambiguously identifying the correct folds.

Considering the rather large scattering of parameters for given rmsd values in Figure 6 we investigated how close to each other the structures are in the ensemble. This was done by applying a simple clustering approach to the pairwise rmsd matrix calculated from the ensemble of 50 structures. Again here the rmsd comparisons were limited to secondary structure elements identified from the CSI to avoid including possibly disordered loop regions. The clustering was performed for increasing rmsd cut-offs from 1.5 to 3.0 Å, respectively. The minimum cluster size was set to 4 (8% of the total number of structures). Results of this clustering is indicated in Figure 4b. Only one cluster is found, even with a 3.0 Å rmsd cut-off, encompassing all the structures close to the crystal structure. This can also clearly be seen in Figure 6 in which the members of the 2.0 Å cluster are indicated by open circles. All structures with high rms deviations from the crystal structure are quite dissimilar and do not fall into a single family. Only when residual dipolar couplings were included in the structure calculations was a second cluster found (data not shown). This latter was, however, less populated than the first cluster and scored on average worse than the first most populated cluster. The 2.0 Å cluster contains 15 structures with average rmsd from the average of 1.5 and 2.1 Å for backbone and all heavy atoms of the secondary structure elements, respectively, and 2.6 and 3.0 Å for backbone and all heavy atoms of the complete sequence. The representative structure of this cluster (the closest to the average) is shown superimposed onto the crystal structure in Figure 5c. The rms deviations from the crystal conformation are 1.3 and 3.0 Å for backbone and all heavy atoms of the secondary structure elements, respectively, and 2.0 and 3.6 Å for backbone and all heavy atoms of the complete sequence. The fit is quite good considering the limited amount of experimental information.

Finally we should remark that the results presented here were obtained following a rather conservative approach using only 100% consistent TALOS predictions and unified, rather loose hydrogen-bond restraints. We performed additional test calculations including the 90% consistent TALOS predictions resulting in ϕ and ψ backbone dihedral angle restraints for 41 residues. Nine of those predictions are only 90% consistent, one of which (D45) having a wrong ψ value (121° instead of −1° in the crystal structure). The additional restraints do however not resolve the fold ambiguity. The resulting ensemble of structures after clustering has a similar resolution as the one obtained with conservative restraints. It is however slightly further away from the crystal structure (2.0 Å rmsd for the backbone of secondary structure elements for the representative structure) than the one obtained with the 100% consistent TALOS predictions (1.3 Å rmsd). Nevertheless, it is encouraging that even with the inclusion of 90% consistent predictions and even a wrong prediction (D45) the correct native fold is still obtained. Similar remarks apply when tighter hydrogen-bond restraints are used, which were obtained using the demonstrated correlation between the cross-hydrogen bond coupling and the N-O distance (Equation 2b in Cornilescu et al., 1999c). Tightening the N-O distance and choosing an H^N-O distance such as to allow a maximum N-H^N-O angle of 120° does not allow to resolve the fold ambiguity and/or generate conformations closer to the crystal structure than when using uniform hydrogen bond restraints. In view of the latter results and considering the experimental errors and the uncertainties in the relationship between the ${}^{3hb}J_{NC'}$ coupling and the N-O distance a quantitative interpretation of cross hydrogen-bond couplings is not required to generate correct folds.

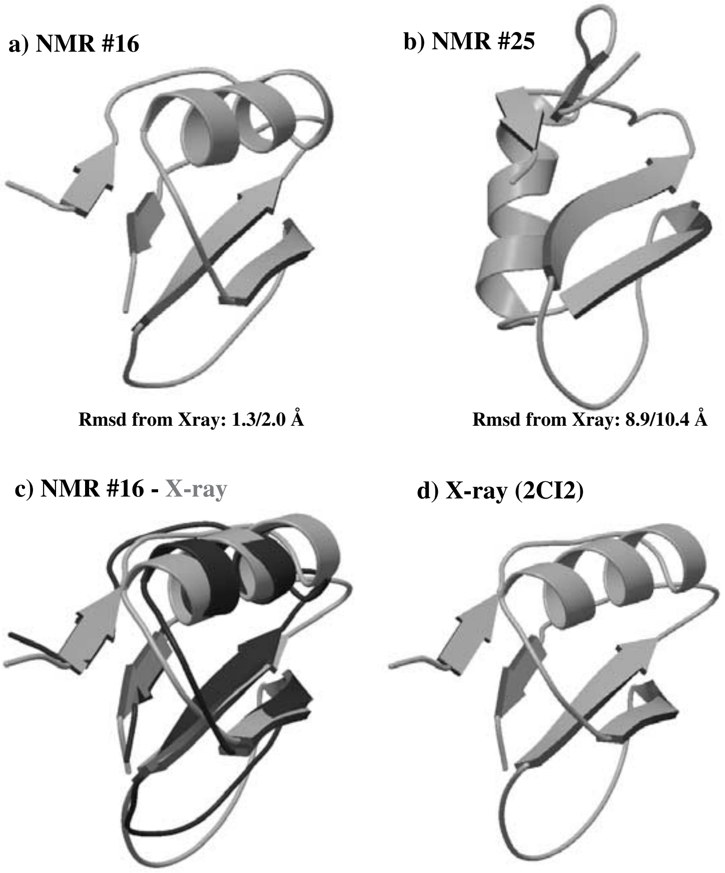

Figure 5. Comparison of the crystal structure of CI2 (2CI2) (McPhalen and James, 1987) with two representative NMR structures calculated from hydrogen bond and chemical shift derived restraints only. (a) NMR model #16 (see Figure 3), (b) NMR model #25, (c) superposition of NMR model #16 onto the crystal structure and (d) crystal structure. Both NMR models (#16 and #25) satisfy the NMR restraints. Model 16 corresponds to the structure that is closest to the average structure from the main cluster (see text). Backbone rms deviations from the crystal structure are given for residues in secondary structure element (see Figure 1) and the complete backbone. These figures were generated with the programs Molscript (Kraulis, 1991) and Raster3D (Merritt and Murphy, 1994).

Conclusions

The solution structure of CI2 was calculated using sparse experimental information available at the stage of backbone assignment. The experimental data consisted of backbone ϕ/ψ dihedral angles predictions for 32 residues obtained from secondary chemical shifts analysis with TALOS and 18 hydrogen bond restraints identified from cross-hydrogen bond $^{3hn}J_{NC'}$ couplings. This information was sufficient to generate models as close as 1.3/2.0 Å backbone rms deviations from the crystal structure for secondary structure elements and complete backbone, respectively. The fold was, however, not uniquely defined. Correct folds could be identified from a combination of clustering and knowledge-based potentials, while geometric and stereochemical criteria failed in distinguishing between native and non-native folds. The discrimination ability of knowledge-based potentials was greatly improved after refining the structures in explicit water using full van der Waals and electrostatic energy terms.

Although there is an increasing literature on cross-hydrogen bond couplings, their measurement is not a trivial task and particular care should be used in setting up such measurements. Considering the long evolution periods in these experiments, this approach is likely to be more successful for small to medium size proteins than larger proteins although cross-hydrogen bond scalar couplings have been reported for a 30 kDa protein (Wang et al., 1999). The amount of structurally important information will also be limited by the fold

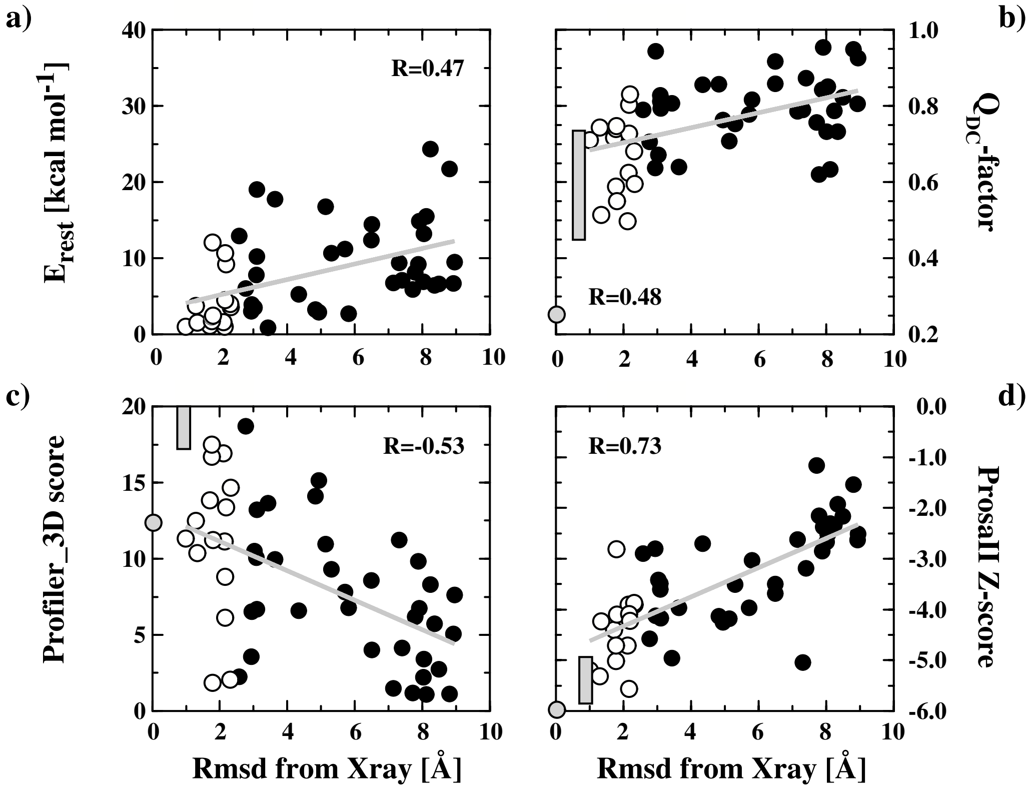

Figure 6. Correlation plots of the backbone positional rms deviations from the crystal structure (2CI2) (McPhalen and James, 1987) versus (a) sum of distance and dihedral angle restraint energies, (b) ^{15}N-H^{N} residual dipolar couplings Q-factor ($Q_{DC} = \Sigma[^{DC}J_{exp} - ^{DC}J_{calc}]^2 / \Sigma^{DC}J^2_{exp}$), (c) 3D profile score (Bowie et al.,1991; Lüthy et al., 1992) and (d) Prosa-II z-score (Sippl, 1993). The rms deviations were calculated for C_α, C, N atoms of the secondary structure elements identified from the CSI (see Figure 1). The linear regression coefficients R are given. Open symbols correspond to structures belonging to the main cluster using a 2.0 Å rmsd cut-off. The corresponding values for the crystal (2CI2) and solution (3CI2) structures are indicated by gray circles and bars, respectively.

of the protein under investigation. We demonstrated here its suitability for an α/β fold. Similar results should be obtained for all β folds while all α folds will be problematic due to the lack of long range structural information. When moving toward larger systems for which perdeuteration might be used, similar structural information can be obtained from amide-amide NOE data. With increasing protein size, however, the fold complexity and therefore the fold ambiguity when generating models from sparse NMR data is expected to increase. It will thus become crucial to have proper scoring functions to identify native folds. For that we will clearly benefit from the ongoing developments in the protein structure prediction field.

Finally, one should remark that generating low resolution folds should not be a final goal *in se*. If we consider all the efforts that go into cloning, expressing and labeling the protein of interest, it should be clear that in most cases a low resolution structure will not be satisfactory as an end point. Generating such low resolution structures early on in the structure determination process will have, however, several advantages. In terms of a structural genomic approach, these initial models can help prioritizing targets that should go through the classical and rather lengthy high resolution full structure determination process. Further, they will provide a very good starting point for automated NOE assignment and structure calculation methods. Having reliable starting structures should make these automated methods more robust and trustful, thereby reducing the overall time needed for protein structure determination.

Acknowledgements

Dr A. Fehrst and Mr N.W. Foster (MRC Cambridge) are acknowledged for providing the expression system of CI2 and A.V.E. George (Utrecht University) for expressing and purifying doubled-labelled CI2. Eugene Tischenko (Utrecht University) is acknowledged for his software and helpful discussion concerning residual dipolar couplings. Financial support from the EC 'Biotechnology' programme BIO4-CT97-2085 to R.B. and from the Netherlands Foundation for Chemical Research (NWO/CW) are gratefully acknowledged.

References

Aitio, H., Annila, A., Heikkinen, S., Thulin, E., Drakenberg, T. and Kilpelainen, I. (1999) *Protein Sci.*, **8**, 2580–2588.

Annila, A., Aito, H., Thulin, E. and Drakenberg, T. (1999) Recognition of protein folds via dipolar couplings *J. Biomol. NMR*, **14**, 223–230.

Berman, H.M.,Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. and Bourne P.E (2000) *Nucl. Acids Res.*, **28**, 235–242.

Bowers, P.M., Strauss, C.E.M. and Baker, D. (2000) *J. Biomol. NMR,* **18**, 311–318.

Bowie, J.U., Lüthy, R. and Eisenberg, D. (1991) *Science*, **253**, 164–170.

Brünger, A.T., Adams, P.D., Clore, G.M., DeLano, W.L., Gros, P., Grosse-Kunstleve, R.W., Jiang, J.-S., Kuszewski, J., Nilges, N., Pannu, N.S., Read, R.J., Rice, L.M., Simonson, T., and Warren, G.L. (1998) *Acta Cryst.,* **D54**, 905–921.

Cavanagh, J. and Rance, M. (1993) *Annu. Rev. NMR Spectrosc.*, **27**, 1–58.

Clore, G.M., Robien, M.A. and Gronenborn, A.M. (1993) *J. Mol. Biol.* **231**, 82–102.

Cordier, F. and Grzesiek, S. (1999) *J. Am. Chem. Soc.*, **121**, 1601–1602.

Cornilescu, G., Marquard, J.L., Ottiger, M. and Bax, A. (1998) *J. Am. Chem. Soc.*, **120**, 6836–6837.

Cornilescu, G., Delaglio, F. and Bax, A. (1999a) *J. Biomol. NMR*, **13**, 289–302.

Cornilescu, G., Hu, J.-S. and Bax, A. (1999b) *J. Am. Chem. Soc.*, **121**, 2949–2950.

Cornilescu, G., Ramirez, B.E., Frank, M.K., Clore, G.M., Gronenborn, A.M. and Bax, A. (1999c) *J. Am. Chem. Soc.*, **121**, 6275–6279.

Delaglio, F., Grzesiek, S., Vuister, G. W., Zhu, G., Pfeifer, J. and Bax A. (1995) *J. Biomol. NMR,* **6**, 277–293.

Delagio, F., Kontaxis, G. and Bax, A. (2000) *J. Am. Chem. Soc.*, **122**, 2142–2143.

Dingley, A. and Grzesiek, S. (1998) *J. Am. Chem. Soc.*, **120**, 8293–8297.

Duggan, B.M., Legge, G.B, Dyson, J. and Wright, P.E. (2001) *J. Biomol NMR,* **19**, 321–329.

Engh, R. and Huber, R. (1991) *Acta Crystallogr.*, **A47**, 392–400.

Fowler, C.A., Tian, F., Al-Hashimi, H.M. and Prestegard, J. (2000) *J. Mol. Biol.*, **304**, 447–460.

Gardner, K.H. and Kay, L.E. (1998) *Annu. Rev. Biophys. Biomol. Struct.*, **27**, 357–406.

Grzesiek, S. and Bax, A. (1992) *J. Magn. Reson.*, **96**, 432–440.

Grzesiek, S. and Bax, A. (1993) *J. Biomol. NMR*, **3**, 185–204.

Grzesiek, S., Anglister, J. and Bax, A. (1993) *J. Magn. Reson.*, **B101**, 114–119.

Güntert, P., Mumenthaler, C. and Wüthrich, K. (1997) *J. Mol. Biol.*, **273**, 283–298.

Hao, M.-H. and Scheraga, H. A. (1999) *Curr. Opin. Struct. Biol.*, **9**,184–188.

Holm, L. and Sander, C. (1993) *J. Mol. Biol.*, **233**, 123–138.

Hubbard, S.J. and Thornton, J.M. (1993) 'NACCESS', Computer Program, Department of Biochemistry and Molecular Biology, University College London.

Hus, J.-C., Marion, D. and Blackledge, M. (2001) *J. Am. Chem. Soc.*, **123**, 1541–1542.

Jorgensen, W. and Tirado-Rives, J. (1988) *J. Am. Chem. Soc.*, **110**, 1657–1666.

Jorgensen, W.L., Chandrasekhar, J., Madura, J.D., Impey, R.W. and Klein, M.L. (1992) *J. Chem. Phys.*, **79**, 926–935.

Karimi-Nejad, Y., Warren, G.L., Schipper, D., Brünger, A.T. and Boelens, R. (1998) *Mol. Phys.*, **95**, 1099–1112.

Kay, L.E., Xu, G.Y, Singer, A.U., Muhandiram, D.R. and Forman-Kay, J.D. (1993) *J. Magn. Reson.*, **B101**, 333–337.

Kleywegt, G.J., Vuister, G.W., Padilla, A., Knegtel, R.M.A., Boelens, R. and Kaptein, R. (1993) *J. Magn. Res.*, **102**, 166–176.

Kraulis, P.J. (1991) *J. Appl. Cryst.*, **24**, 946–950.

Laskowski, R.A., MacArthur, M.W. and Thornton, J.M. (1993) *J. Appl. Cryst.,* **26**, 283–291.

Lazaridis, T. and Karplus, M. (2000) *Curr. Opin. Struct. Biol.*, **10**, 139–145.

Levitt, M. (1983) *J. Mol. Biol.*, **170**, 723–764.

Linge, J.P. and Nilges, M. (1999) *J. Biomol. NMR*, **13**, 51–59.

Liu, A., Hu, W., Majumdar, A., Rosen, M.K. and Patel, D. (2000a) *J. Biomol. NMR*, **17**, 79–82.

Liu, A., Hu, W., Majumdar, A., Rosen, M.K. and Patel, D. (2000b) *J. Biomol. NMR*, **17**, 305–310.

Ludvigsen, S., Shen, H., Kjaer, M., Madsen, J.C. and Poulsen, F.M. (1991) *J. Mol. Biol.*, **222**, 621–635.

Lüthy, R., Bowie, J.U. and Eisenberg, D. (1992), *Nature*, **356**, 83–85.

McPhalen, C.A. and James, M.N.G. (1987) *Biochemistry*, **26**, 261–269.

Medek, A., Olejniczak, E.T., Meadows, R.P. and Fesik, S.W. (2000) *J. Biomol. NMR*, **18**, 229–238.

Meiler, J., Peti, W. and Griesinger, C. (2000a) *J. Biomol NMR*, **17**, 283–294.

Meiler, J., Blomberg, N. Nilges, M. and Griesinger, C. (2000b) *J. Biomol NMR*, **16**, 245–252.

Meissner, A. and Sørensen, O.W. (2000a), *J. Magn. Reson.*, **143**, 387–390.

Meissner, A. and Sorensen, O.W. (2000b), *J. Magn. Reson.*, **143**, 431–434.

Melacini, G., Boelens, R. and Kaptein, R. (1999) *J. Biomol. NMR*, **15**, 189–201.

Merritt, E.A. and Murphy, M.E.P. (1994) *Acta Cryst.*, **D50**, 869–873.

Mogens, K., Ludvigsen, S., Sorensen, O.W. Denys, L., Kindtler, J. and Poulsen. F.M. (1987) *Carlsberg Res. Commun.*, **52**, 327–354.

Moult, J. (1997) *Curr. Opin. Struct. Biol.,* **7**, 194–199.

Mueller, G.A., Choy, W.Y., Yang, D., Forman-Kay, J.D., Venters, R.A. and Kay, L.E. (2000) *J. Mol. Biol.*, **300**, 197–212.

Muhandiram, D.R. and Kay, L.E. (1994) *J. Magn. Reson.*, **B103**, 203–216.

Nilges, M. and O' Donoghue, S. (1998) *Prog. NMR Spectrosc.*, **32**, 107–139.

Pervushin, K., Ono, A.,Fern‡ndez, C., Szyperski, T., Kainosho, M. and Wüthrich, K. (1998) *Proc. Natl. Acad. Sci. USA*, **95**, 14147–14151.

Pontius, J., Richelle, J. and Wodak, S. (1996) *J. Mol. Biol.*, **264**, 121–136.

Sippl, M.J. (1993) *Prot. Struct. Funct. Genet.*, **17**, 355–362.

Sippl, M.J. (1995) *Curr. Opin. Struct. Biol.,* **5**, 229–235.

Skolnick, J., Kolinski, A. and Ortiz, A.R. (1997) *J. Mol. Biol.*, **265**, 217–241.

Stein, E.G., Rice, L.M. and Brünger, A.T. (1997) *J. Magn. Reson.*, **B124**, 154–164.

Wang, Y.-X., Jacob, J., Cordier, F., Wingfield, P., Stahl, S., Lee-Huang, S., Torchia, D., Grzesiek, S. and Bax, A. (1999) *J. Biomol. NMR*, **14**, 181–184.

Wishart, D.S. and Sykes, B.D. (1994) *J. Biomol. NMR*, **4**, 171–180.

Mumenthaler, Ch. and Braun, W. (1995) *J. Mol. Biol.*, **254**, 465–480.