J Biomol NMR (2008) 42:111–127
DOI 10.1007/s10858-008-9268-9

ARTICLE

Direct methods and residue type specific isotope labeling in NMR structure determination and model-driven sequential assignment

Andreas Schedlbauer · Renate Auer · Karin Ledolter · Martin Tollinger · Karin Kloiber · Roman Lichtenecker · Simon Ruedisser · Ulrich Hommel · Walther Schmid · Robert Konrat · Georg Kontaxis

Received: 1 May 2008 / Accepted: 13 August 2008 / Published online: 2 September 2008

Abstract Direct methods in NMR based structure determination start from an unassigned ensemble of unconnected gaseous hydrogen atoms. Under favorable conditions they can produce low resolution structures of proteins. Usually a prohibitively large number of NOEs is required, to solve a protein structure ab-initio, but even with a much smaller set of distance restraints low resolution models can be obtained which resemble a protein fold. One problem is that at such low resolution and in the absence of a force field it is impossible to distinguish the correct protein fold from its mirror image. In a hybrid approach these ambiguous models have the potential to aid in the process of sequential backbone chemical shift assignment when $^{13}C^{\beta}$ and $^{13}C'$ shifts are not available for sensitivity reasons. Regardless of the overall fold they enhance the information content of the NOE spectra. These, combined with residue specific labeling and minimal triple-resonance data using $^{13}C^{\alpha}$ connectivity can provide almost complete sequential assignment. Strategies for residue type specific labeling with customized isotope labeling patterns are of great advantage in this context. Furthermore, this approach is to some extent error-tolerant with respect to data incompleteness, limited precision of the peak picking, and structural errors caused by misassignment of NOEs.

Keywords Direct NMR methods · Proton density clouds · NMR sequential assignment · Residue type specific isotope labeling

Electronic supplementary material The online version of this article (doi:10.1007/s10858-008-9268-9) contains supplementary material, which is available to authorized users.

A. Schedlbauer · R. Auer · K. Ledolter · M. Tollinger · K. Kloiber · R. Konrat · G. Kontaxis (✉)
Institute of Biomolecular Structural Chemistry, Max F. Perutz Laboratories, University of Vienna, Campus Vienna Biocenter 5/1, 1030 Vienna, Austria
e-mail: georg.kontaxis@univie.ac.at

R. Lichtenecker · W. Schmid
Institute of Organic Chemistry, University of Vienna, Waehringer Strasse 38, 1090 Vienna, Austria

S. Ruedisser · U. Hommel
Novartis Institutes for BioMedical Research, 4057 Basel, Switzerland

Introduction

Traditionally, NMR based protein solution structure determination proceeds in a defined sequence of steps (Cavanagh et al. 1996; Evans 1995): first, sequential assignment of the protein backbone resonances $^{1}H^{N}$, ^{15}N, $^{13}C^{\alpha}$, $^{13}C^{\beta}$, $^{13}C'$ and $^{1}H^{\alpha}$, using pairs of triple-resonance NMR experiments, second, acquisition of ^{1}H proton and ^{13}C carbon relayed TOCSY experiments to correlate the $^{1}H^{N}$ backbone signals with the amino-acid sidechains ^{1}H and ^{13}C signals and third, ^{13}C carbon and ^{15}N nitrogen edited multi 3D/4D dimensional NOE spectra to assign and identify as many distance restraints as possible. This is followed by structure calculation applying restrained simulated annealing molecular dynamics refinement to a covalent template starting from an extended backbone conformation of the protein (Brünger 1993; Brunger et al. 1998). The resulting protein structures can be used to iteratively identify and assign more and more NOE distance restraints to be used in the next round of structure calculation and this procedure is iterated until convergence (Nilges et al. 1997).

While this ‘standard protocol’ is almost guaranteed to work (provided sufficient protein solubility and stability), it

is fairly time consuming and requires a turnaround time of a few months. This has therefore prompted the search for alternative approaches and potentially time-saving shortcuts reducing the time demands in NMR based structural biology.

Usually the sequential $^{13}C^{\alpha}_{i/i-1}$ connectivity information from 3D HNCA (Kay et al. 1990) (in combination with 3D HN(CO)CA) (Grzesiek and Bax 1993) data alone is not sufficient for unambiguous sequential assignment due to the large shift degeneracy of the $^{13}C^{\alpha}$ backbone shifts (Hoffmann et al. 2005). Further information is usually required to make the sequential connectivity unambiguous e.g. $^{13}C^{\beta}_{i/i-1}$ shifts from HNCACB/CBCA(CO)NH (Grzesiek and Bax 1992; Wittekind and Mueller 1993) (or HN(CO)CACB) or $^{13}C'_{i/i-1}$ shifts HNCO/HN(CA)CO (Clubb et al. 1992; Kay et al. 1990), which may prove more difficult to obtain especially in larger molecular weight systems due to fast transverse ^{15}N and ^{13}C T_2 relaxation.

HNCA (and HNCO) and potentially HN(CO)CA spectra can still be acquired with reasonable sensitivity even in large molecular weight systems and in the absence of sidechain deuteration. More complex experiments require full or partial deuteration, especially in conjunction with TROSY methodology (Pervushin et al. 1997; Pervushin et al. 1998), otherwise the quality of the other triple-resonance datasets is seriously compromised by fast relaxation.

Nevertheless, even in such cases NOE spectra of good quality can be obtained without large experimental difficulty (Korzhnev et al. 2004; Tugarinov and Kay 2004; Tugarinov et al. 2005; Tugarinov et al. 2004).

In those cases the NOE connectivities can provide the missing information required for unambiguous sequential linkage (Hoffmann et al. 2005). In the absence of a structural model only sequential to mid-range NOEs are interpretable. The information content of a NOE data set is further enhanced if a three dimensional protein model (e.g. provided by homology modeling) is available, in whose context all $^{1}H^{N}$–$^{1}H^{N}$ NOEs, including long-range and characteristic tertiary contacts e.g. across β-strands, can be used for sequential linkage and placement of spin systems onto a protein structure. This process is greatly improved when a-priori residue typing for a number of spin-systems is possible, which will subsequently serve as pivotal points for sequential assignment and as check points for validation.

One particularly interesting alternative in cases, where no homology model of the protein of interest is available, is to use the NOE information directly without any prior assignment and apply the restrained simulated annealing to an ensemble of unassigned and unconnected and non-interacting gaseous atoms. They are condensed into a low-resolution structural proton distribution a so-called proton cloud, which under favorable circumstances resembles a low-resolution image of a protein (Grishaev and Llinas 2002a, b, c, 2004, 2005; Grishaev et al. 2005; Kraulis 1994). This procedure is repeated a number of times so that a distribution of protein clouds is obtained, from which proton densities can be derived.

In principle, these proton densities carry all the structural information. It has been demonstrated already that if a large enough number of NOE distance restraints is available, tertiary structures of small globular proteins can be obtained 'directly' without resorting to any form of covalent template. The sequential NMR assignment is obtained as a by-product in the course of the structure calculation rather than being its prerequisite.

While such 'direct' approaches sound very appealing per-se, the number of restraints required to obtain meaningful structures ab-initio has been prohibitively high so far.

Different requirements apply in the case of highly deuterated (except Val/Ile/Leu methyl groups) proteins, where due to deuteration consequently the density of NOEs is substantially lower.

In those cases it has been already demonstrated, using the CS-CLOUD protocol that, if the resolution of the resulting cloud is, nevertheless, sufficient to discern individual backbone sites, then the protein chain can be sequentially traced through it using graph theory and consequently the sequential NMR assignment is obtained directly and exclusively from the NOE data (Bermejo and Llinas 2008).

In contrast, we therefore propose a hybrid approach in which we calculate low-resolution proton density clouds using only a minimum number of experimental NOEs, comparable to those used to generate a low-resolution backbone model of a protein (Gardner and Kay 1998; Gardner et al. 1997; Kay and Gardner 1997). In turn, these low resolution proton density clouds can then be used as a 'mock' structural model in combination with a minimal amount of triple-resonance data (HNCA/HN(CO)CA only) to support and drive the sequential assignment of HNCA/HN(CO)CA and the NOE identification by making references to a low-resolution PDB model. Thus we expect more redundancy and stability, since we do not exclusively use NOE connectivities but rely additionally on triple-resonance data to define the chain directionality.

Once the sequential assignment has been established, the structure determination proceeds, in a conventional way, by simulated annealing protocols using standard covalent protein templates. In that case the NOEs previously assigned can be sufficient to obtain the protein backbone fold, and the model obtained such can serve as a starting point for further refinement by iterative assignment of ambiguous NOEs (ARIA) or similar algorithms (Nilges et al. 1997).

Alternatively, more advanced methods of structure generation, other than simulated annealing e.g. based on

the ROSETTA algorithm can be used as engine for fold generation (Bermejo and Llinas 2008; Bowers et al. 2000; Simons et al. 1997).

Here we present a method that is capable of simultaneously obtaining sequential assignment and low- to mid resolution protein backbone folds for small to mid-sized globular proteins requiring only HNCA/HN(CO)CA input data in conjunction with NOE derived unassigned proton density clouds.

We show how this is simplified using different chemical isotope labeling strategies. We also present a synthetic strategy for residue type and position specific isotope labeling, which is capable of producing customized ^{13}C labeling of either specified backbone positions ($^{13}C^{\alpha}/^{13}C'$ of Val and Ile and $^{13}C^{\beta}$ of Leu residues) and/or strategic sidechain methyl positions $^{13}C^{\gamma 1/2}$ of Val, $^{13}C^{\delta 1/2}$ of Leu and $^{13}C^{\delta 1}$ of Ile residues) allowing for spectral simplification and editing of multidimensional spectra at the same time (Goto et al. 1999; Lichtenecker et al. 2004; Rosen et al. 1996).

With this isotope labeling technique both: greatly simplified NOE spectra providing key methyl NOEs, which help define the protein fold, and residue type edited triple-resonance spectra providing pivotal points for sequential assignment can be obtained.

Materials and methods

Position specific isotope labeling

The strategy adopted for the introduction of variable $^{13}C/^{12}C$ and $^{1}H/^{2}H$ isotope labeling patterns into α-ketobutyrate and α-ketoisovalerate has been described in detail previously (Lichtenecker et al. 2004). Generally, α-ketobutyrate and α-ketoisovalerate are efficiently incorporated into proteins yielding high levels (>90%) of selectively labeled Val/Ile/Leu residues (Goto et al. 1999) Specifically isotope labeled precursor compounds are used to direct the isotope labeling into the desired positions (Goto et al. 1999; Lichtenecker et al. 2004; Rosen et al. 1996).

Thus using various combinations of $^{13}C/^{12}C$ and $^{1}H/^{2}H$ introduced into the precursor compounds α-ketobutyrate and α-ketoisovalerate Val/Ile/Leu residues with unique customized $^{13}C,^{12}C,^{2}H,^{1}H$-labeling patterns can be produced in an otherwise uniformly $^{12}C,^{1}H$-labeled (or alternatively $^{13}C,^{2}H$-labeled) protein background.

E.g. 4-^{13}C-α-ketobutyrate and 4-^{13}C-α-ketoisovalerate are used to incorporate selectively $^{13}C^{\gamma 1/2}$ $^{13}C^{\delta 1}$ $^{13}C^{\delta 1/2}$ methyl labeled Val/Ile/Leu residues into a protein (Lichtenecker et al. 2004).

Biosynthetic incorporation of 1,2-$^{13}C_2$-α-ketobutyrate and 1,2-$^{13}C_2$-α-ketoisovalerate into proteins by bacterial growth yields samples that are ^{13}C labeled solely at backbone C^{α} and C' positions of Val and Ile residues (and C^{β} positions in Leu residues).

Briefly, as outlined in Fig. 1, 1,2-$^{13}C_2$-α-ketobutyrate (3) and 1,2-$^{13}C_2$-α-ketoisovalerate (4) were synthesized via a common intermediate, *tert*-butyl-1,2-$^{13}C_2$-2-bromomethyl-2-propenoate (2), using commercially available $^{13}C_2$-bromoacetic acid (1) as labeled starting material. Subsequent incorporation into proteins by bacterial growth in minimal medium supplemented with compounds (3) and (4) yields Val-$^{13}C^{\alpha}/^{13}C'$ and Ile-$^{13}C^{\alpha}/^{13}C'$ labeled protein, while backbone α-positions are ^{1}H or ^{2}H labeled depending on whether the bacterial growth is performed in protonated or deuterated ($^{2}H_2O/^{2}H$-glucose) media, respectively. For Leu, only a single ^{13}C label (originating from the carbonyl ^{13}C of 1,2-$^{13}C_2$-α-ketoisovalerate) is introduced into the β-position of the side-chain.

Furthermore, using suitably, isotope labeled precursor molecules backbone (Val/Ile) and methyl labeling of Val/Ile/Leu residues can be combined and just one sample preparation is required.

Such methyl and/or backbone labeling patterns result in residue type edited (HNCA, HN(CO)CA, HNCO) triple-resonance and greatly simplified NOE experiments (Lichtenecker et al. 2004).

As an example, we have applied this labeling strategy to the C-terminal SH2 domain of the signal transduction protein phospholipase C-γ1, PLCC. PLCC SH2 domain was produced by bacterial growth in ^{12}C-glucose/$^{15}NH_4Cl$ minimal media supplemented with 1,2-$^{13}C_2$-α-ketobutyrate and 1,2-$^{13}C_2$-α-ketoisovalerate, yielding Val-$^{13}C^{\alpha}/^{13}C'$ and Ile-$^{13}C^{\alpha}/^{13}C'$ (and Leu-$^{13}C^{\beta}$) labeling in a $^{12}C/^{15}N$ background. Figure 2 shows two-dimensional H(N)CO and H(N)CA spectra that were obtained for this protein sample. As expected, only correlations belonging to Val and Ile residues are visible in H(N)CO and H(N)CA experiments, while correlations belonging to Leu residues are missing (the PLCC SH2 domain contains 5 Val and 5 Ile residues, as well as 8 Leu residues). V28 is sequentially neighbored to a Pro residue (P29) and can be rapidly identified in the H(N)CA spectrum. Scrambling of isotope labels between residues was not observed.

Input data and systems studied

A set of seven proteins (Table 1 and Supplementary Table S1) with different fold topologies, extent of available chemical shift assignments, varying backbone chemical shift dispersion, and molecular weight served as model cases for testing the procedure. Almost complete chemical shift assignments, restraint lists and high-resolution NMR solution structures of Cyclophilin D (unpublished results), ICln (PDB ID code 1ZYI, BMRB entry 5736) (Furst et al.

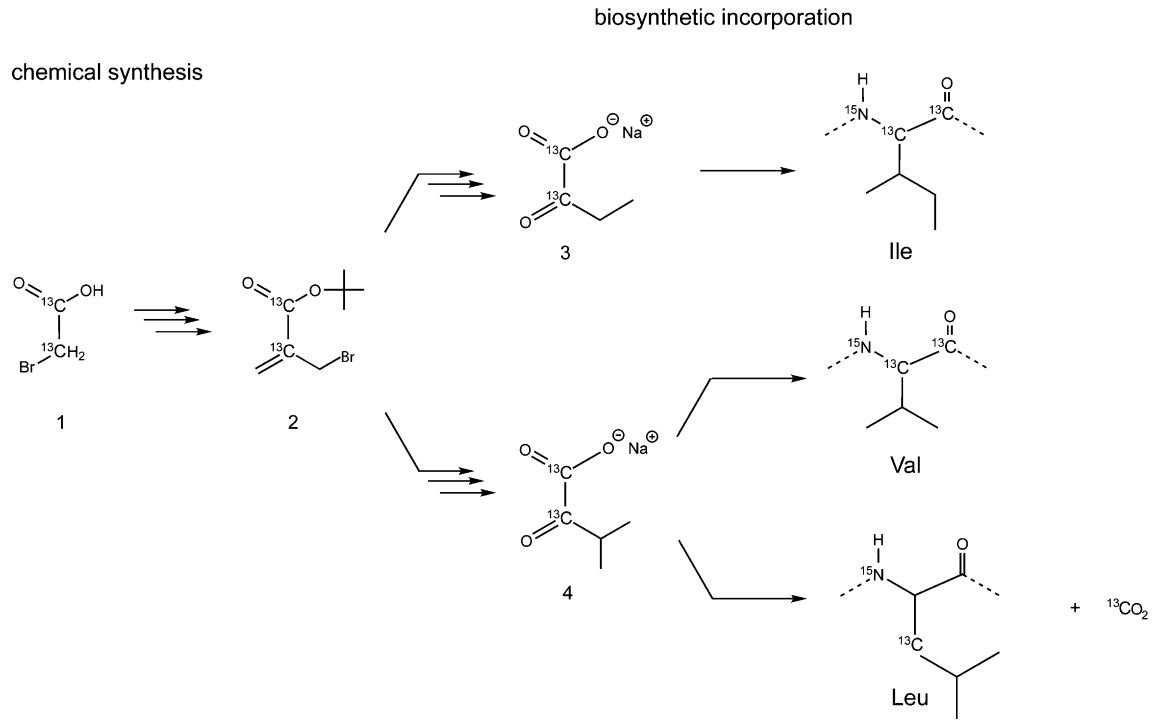

Fig. 1 Outline of the reaction scheme for the production of 1,2-$^{13}C_2$-α-ketobutyrate (**3**) and 1,2-$^{13}C_2$-α-ketoisovalerate (**4**) and biosynthetic incorporation into Ile, Val and Leu residues

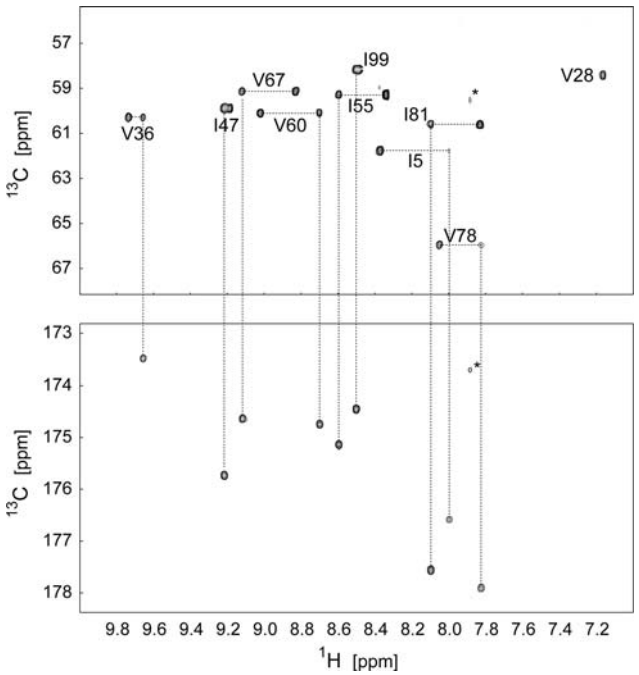

Fig. 2 Example of $^{13}C^{\alpha}$ and $^{13}C'$ selective labeling and residue type edited triple-resonance. Two-dimensional H(N)CO (top) and H(N)CA (bottom) spectra obtained for the PLCC SH2 domain. Resonance assignments for Val and Ile (C' and C^{α}, respectively) are indicated and were taken from the literature. (Pascal et al. 1994) Asterisks denote correlations belonging to the disordered and highly dynamic (Farrow et al. 1994) C-terminus of the proteins (residues S104 and S105). V28 is adjacent to a proline residue (P29)

2005; Schedlbauer et al. 2007) and qCRP2(LIM1) (PDB ID code 1A7I) (Kontaxis et al. 1998) were originally manually determined in our laboratory and could be used for reference (Table 1).

The experimentally derived NOE distance restraints were taken from unambiguously assigned NOESY cross peak tables obtained in the course of conventional structure determination projects taking into account only those NOEs that do not suffer from the problems of shift degeneracy and spectral overlap.

'Experimental' NOEs used in this procedure comprised only those key NOEs, which can be obtained with good sensitivity and resolution in 3D and 4D heteronuclear NOESY experiments using either globally doubly ($^{15}N/^{13}C$) labeled or position specifically labeled (using suitable precursor molecules) proteins.

They comprised backbone–backbone contacts H^N–H^N, H^N–H^{α}, backbone to sidechain contacts: H^N–H^{sc} (where H^{sc} comprises CH_3 methyl groups (Val $C^{\Upsilon 1/2}H_3$, Ile $C^{\delta 1}H_3$, Leu $C^{\delta 1/2}H_3$), aromatic $H^{\delta}/H^{\varepsilon}/H^{\zeta}$ (Phe/Tyr), and sidechain

Table 1 Proteins, and completeness of chemical shift data used

Protein	# Residues	BMRB ID	PDB ID	$C^{\alpha}_i/C^{\alpha}_{i-1}$	$C^{\beta}_i/C^{\beta}_{i-1}$
Cyclophilin D	165	7310		95.6/94.9	97.1/95.6
ICln	168	5736	1ZYI	69.0/68.3	67.1/68.3
qCRP2(LIM1)	81		1A7I	81.1/84.6	83.3/84.8

amide $N^{\delta/\varepsilon}H_2$ (Asn/Gln) atoms, respectively), and sidechain to sidechain NOEs especially methyl to methyl contacts between (Val, Ile and Leu residues) and contacts between methyl groups (Val/Ile/Leu) and amide side chains of (Asn/Gln) (Table 2A).

For cases where the availability of such a complete NOE set may be overly optimistic, the effect of a less complete restraint set involving only $^1H^N$ and CH_3 groups was studied using Cyclophilin D as an example.

In this case: first, NOE distances originating from H^α atoms were omitted, and second, the distance restraints involving aromatic $H^\delta/H^\varepsilon/H^\zeta$ (Phe/Tyr) or sidechain amide $N^{\delta/\varepsilon}H_2$ (Asn/Gln) atoms were removed.

To investigate the influence of further different protein fold topologies idealized NOE data sets were simulated for a number of different proteins of varying secondary and tertiary structures.

Backbone chemical shifts and structural data for Calmodulin, Carbonic Anhydrase I, Plastocyanin, and Thioredoxin were taken from BioMagResBank and Protein Data Bank (Supplementary Tables S1 and S2 summarize the number of restraints available. For those test sets theoretical NOE distances comprising H^N and CH_3 (Val/Ile/Leu,) atoms were inferred directly from the PDB structures using an interatomic distance cutoff of 6.0 Å. NOE cross peaks were assumed to occur between protons separated by less than 6 Å.

The generation of structures is described in the flow chart of Fig. 3 and its individual stages are briefly described below.

Restraint classification

Backbone NOEs, encompassing H^N and optional H^α protons (H^N–H^N, H^N–H^α) were roughly classified according to their NOESY crosspeak volumes as strong, medium and weak, corresponding to distance limits of 1.8–2.8, 1.8–3.4 and 1.8–5.0 Å, respectively. Due to the influence of spin diffusion and consequently problems with the quantification of NOEs originating from sidechain atoms, distances related to backbone to sidechain and inter-sidechain NOEs (H^N–CH_3, H^α–CH_3, CH_3–CH_3) were uniformly restrained between 1.8 and 6.0 Å.

This tight classification scheme for restraints, which sometimes tends to underestimate distances, was empirically chosen to counterbalance the fact that distance restraints identified in later rounds of restraint identification (see below) tend to be overestimates of the 'true distance'.

For weak NOEs small upper limit violations of up to 2.0 Å i.e. <7.0 Å for backbone NOEs and <8.0 Å for sidechain NOEs are only penalized with a very small force constant. This applies only a weak bias towards the distance range <5 Å (for backbone–backbone)/6 Å (NOEs of

Table 2 Proteins and available experimental NOE and derived restraints

Protein (atoms included)	Experimental NOEs			
	# bb-bb	# bb-sc	# sc-sc	Total
(A)[a]				
Cyclophilin D H^N, CH_3 (I,L,V)	247	165	50	462
Cyclophilin D H^N, CH_3 (I,L,V), H^{aro} (F,Y), NH_2 (N,Q)	247	285	102	634
Cyclophilin D H^N, H^α, CH_3 (I,L,V), H^{aro} (F,Y), NH_2 (N,Q)	846	361	218	1425
ICln H^N, H^α, CH_3 (I,L,V), H^{aro} (F,Y), NH_2 (N,Q)	336	235	82	653
qCRP2(LIM1) H^N, H^α, CH_3 (I,L,V), H^{aro} (F,Y), NH_2 (N,Q)	245	127	92	464
Protein (atoms included)	Number of derived restraints			
	NOEs	Round #1	Round #2	Total
(B)[b]				
Cyclophilin D H^N, CH_3 (I,L,V)	462	281	1251	1994
Cyclophilin D H^N, CH_3 (I,L,V), H^{aro} (F,Y), NH_2 (N,Q)	634	1686	3181	5501
Cyclophilin D H^N, H^α, CH_3 (I,L,V), H^{aro} (F,Y), NH_2 (N,Q)	1425	6401	7977	15803
ICln H^N, H^α, CH_3 (I,L,V), H^{aro} (F,Y), NH_2 (N,Q)	653	3607	3605	7865
qCRP2(LIM1) H^N, H^α, CH_3 (I,L,V), H^{aro} (F,Y), NH_2 (N,Q)	464	352	4941	5757

[a] The number of available restraints depends on the number of atoms included in the atom density cloud calculation and the protein fold topology, mainly β-stranded proteins have a larger number of sc-sc restraints (long-range contacts) than mainly α-helical proteins

[b] The number of derived restraints that are extracted as being well-conserved distances depends on the initial number of atoms included in the atom density cloud calculation and the protein fold topology, mainly β-stranded proteins have a larger number of restraints (long-range contacts) than mainly α-helical proteins

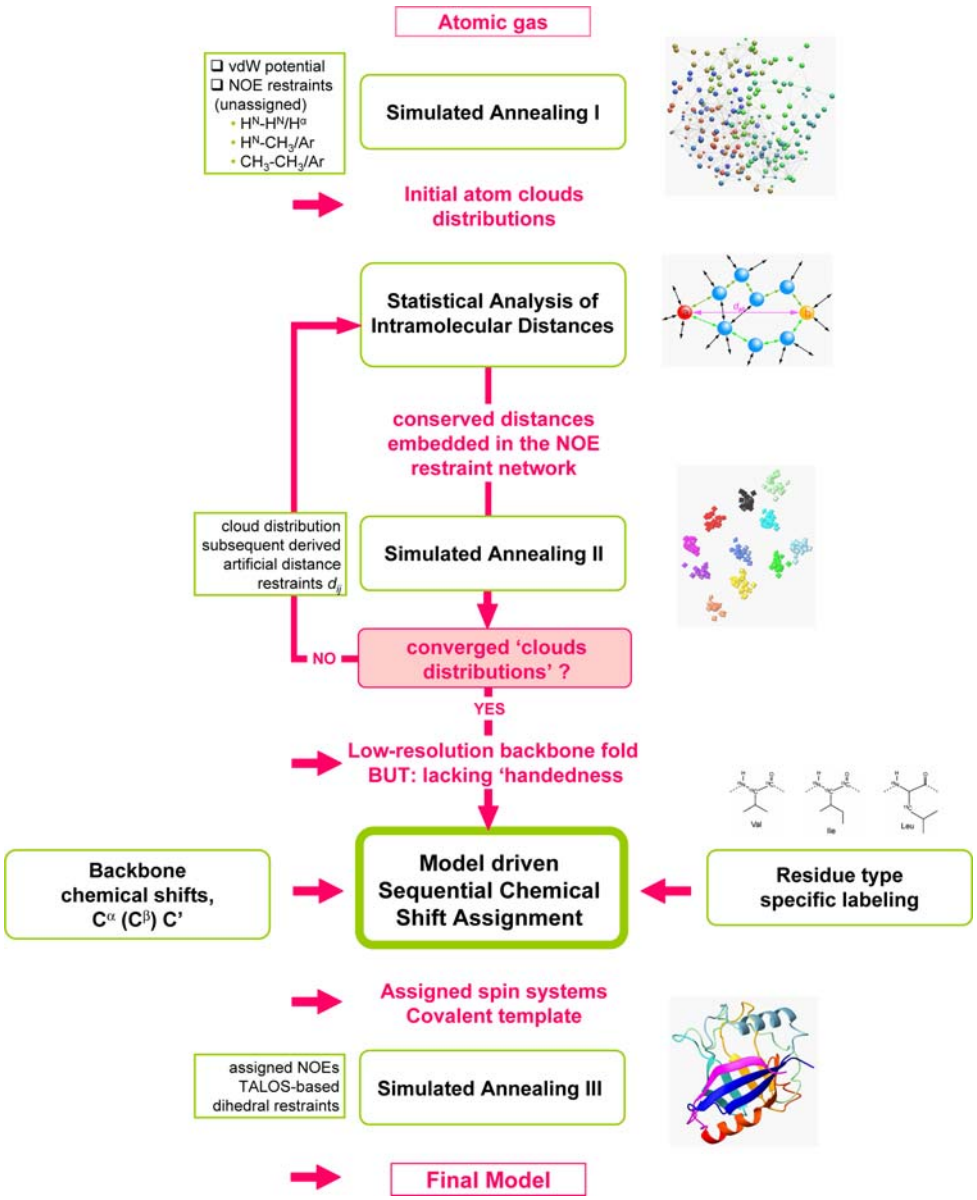

Fig. 3 Flow diagram and overview of the structure generation procedure and backbone assignment procedure. The individual stages are described in more detail in the main body of the text

sidechains) while practically allowing distances up to 7–8 Å accounting for possible effects of spin-diffusion.

In practice this is implemented by using a sum of two distance restraints (1.8–5.0/6.0 Å and 1.8–7.0/8.0 Å) with two different force constants.

While, strictly speaking, not absolutely required, this was empirically found to improve convergence of structure calculations.

Intra-residue contacts (e.g. between Val $C^{\gamma 1/2}H_3$ and Leu $C^{\delta 1/2}H_3$) can be identified through strong NOEs or from the analysis of HCCH-COSY/TOCSY experiments, if available. For intra residue $C^{\Upsilon\#/\delta\#}H_3$–$C^{\Upsilon\#/\delta\#}H_3$ (Leu, Val) NOE distances an upper bound of 5.0 Å was employed. The protons of the methyl groups –CH_3, Gly $H^{\alpha 1}$/$H^{\alpha 2}$, aromatic $H^{\delta 1}$/$H^{\delta 2}$and $H^{\varepsilon 1}$/$H^{\varepsilon 2}$ (Phe, Tyr), as well as sidechain amides –$N^{\delta/\varepsilon}H_2$ (Asn, Gln) were represented as grouped pseudoatoms $H^{\delta\#/\varepsilon\#}$ in the relevant NOE databases and molecular structure templates, respectively. Analogously, aromatic residues can be identified by their characteristic intraresidue NOE contacts or by HCCH-COSY/TOCSY experiments, if available. Corresponding distance restraints were defined between aromatic intra-residue Phe ($H^{\delta\#}$ to $H^{\varepsilon\#}$, $H^{\varepsilon\#}$ to H^{ζ} and $H^{\delta\#}$ to H^{ζ}) and Tyr ($H^{\delta\#}$ to $H^{\varepsilon\#}$) pseudoatoms in order to achieve consistent spatial vicinity and a roughly linear arrangement of these entities.

Structure calculations: simulated annealing I

Proton density cloud calculations (round 0) were performed using nih-xplor software version 2.9.6 or newer

(Schwieters et al. 2003). A simulated annealing protocol was applied to a gaseous randomized distribution of atoms, initially randomly spread out over a cube of 50–100 Å, which were connected only by distance restraints. Gross errors in the NOE list can be detected at this point by analyzing the distributions of NOEs connectivities/pathways between all the atoms of the distance restraint database. Only a 'soft square' NOE potential and a hard sphere vdW energy terms but no covalent (bond, angle, torsion, improper) energy terms were applied. In brief: The annealing protocol starts with a high temperature phase of 10 ps molecular dynamics at 4000 K using a time step of 1 fs. The initial force constants applied were $f_{NOE} = 2$ kcal/mol Å^2 and $f_{VdW} = 4$ cal/mol Å^4. The force constant f_{upper} for the 'transition region' of the distance restraints between 5/6 and 7/8 Å (i.e. for upper limit violations <2.0 Å) was scaled relative to the experimental NOE restraints by a factor of 100 and set to 0.02 kcal/mol Å^2.

During a cooling period of 100 ps the temperature was stepwise lowered to 10 K while the force constants were ramped up to their final values $f_{NOE} = 300$ kcal/mol Å^2 $f_{upper} = 3$ kcal/mol Å^2 and $f_{VdW} = 4$ kcal/mol Å^4. Finally 1000 steps of Powell energy minimization were applied. Only atomic clouds with zero restraint violations were selected for further analysis.

Statistical analysis of conserved distances

Analysis or the resulting atomic density clouds was performed using home written perl and .tcl scripts. Structural superposition and visualization was done using MOLMOL version 2k.2 (Koradi et al. 1996) or vmd-xplor version 1.4 (Schwieters and Clore 2001).

When analyzing the distribution of internuclear distances of the distribution of atom clouds generated above, internuclear distances between two atoms, that are not connected by a direct NOE restraint, are nevertheless well-conserved. This happens when both atoms concerned are well embedded in the network of NOEs, and connected by a short unbroken chain of individual NOEs. The number of intervening NOEs is typically referred to as the NOE path length. Such consistently conserved distances can potentially serve as restraints for further rounds of structure calculations. But being conserved is necessary but not sufficient for its restraining power. Figure 4a, c (black) correlates conserved internuclear distances extracted from a family of clouds with their target values measured in a reference structure. In general lack of NOEs leads to more expanded structures. From Fig. 4a, c it is obvious that there is a large number of such distances, which show a clear tendency to be systematically overestimated and there is therefore a need to filter them as illustrated in Fig. 4b, d).

The slope of the correlation between the distances is somewhat dependent on the classification of the original NOE distance restraints. With our empirically optimized NOE classification we find a slope smaller than one at this stage.

Conserved distances were only converted into distance restraints D_{ij}, when the two involved atoms i and j were connected by a NOE path length l of $2 \leq l \leq 6$ individual NOEs and when they were conserved within relatively narrow margins. Only interatomic distances r_{ij} with a standard deviation $\sigma_{ij} \leq 20\%$ of their mean values d_{ij} (calculated over the whole family of atomic clouds) or $\sigma_{ij} \leq 5$ Å (in absolute numbers) were considered. The corresponding upper bound d_{upper} and lower bound d_{lower} were set to $d_{ij} \pm 2\sigma_{ij}$ unless σ_{ij} became smaller than 10% of the mean d_{ij}. In that case d_{upper} and d_{lower} were redefined as $d_{ij} \pm 0.2d_{ij}$, in order not to artificially 'overrestrain' the system.

Furthermore, for the rarely occurring cases when the shortest individual distance in the family of calculated structures $\min(r_{ij})$ was smaller than d_{lower} or the largest individual distance of the family of clouds and $\max(r_{ij})$ was longer than d_{upper} the corresponding distance restraint D_{ij} was redefined to range from $[\min(r_{ij}) - \sigma_{ij}]$ to $[\max(r_{ij}) + \sigma_{ij}]$.

Simulated annealing II and statistical analysis II

The artificial distance restraints $D_{ij}^{(1)}$ derived from the initial family of atom clouds, as described above, were merged with the original NOE derived distance restraints and used for a subsequent round of structure calculation (iteration 1) using essentially the same protocol as described above. However, the force constant $f_{artificial}$ of those artificial statistical D_{ij} was scaled down by a factor of five relative to f_{NOE} to ensure that all original experimental NOE distance restraints remain fulfilled. The resulting ensemble of atom cloud distributions showed no violation >0.5 Å.

From the best structures (with respect to lowest NOE energy values) new statistically conserved distances r^{ij} were extracted and selected to yield another new set of artificial distance restraints $D_{ij}^{(2)}$. Because of the presence of the $D_{ij}^{(1)}$ during the previous, first round of structure calculation more stringent criteria had to be applied in their selection:

First, that the two atoms i and j involved should be connected by a NOE path length l of $2 \leq l \leq 4$ individual NOEs. Second, that at least two independent NOE pathways should be present (i.e. pathways that do not share any atom along their individual NOE route between the atoms i and j). Third, that all atoms along the NOE pathway had at least two additional NOE contacts with surrounding atoms.

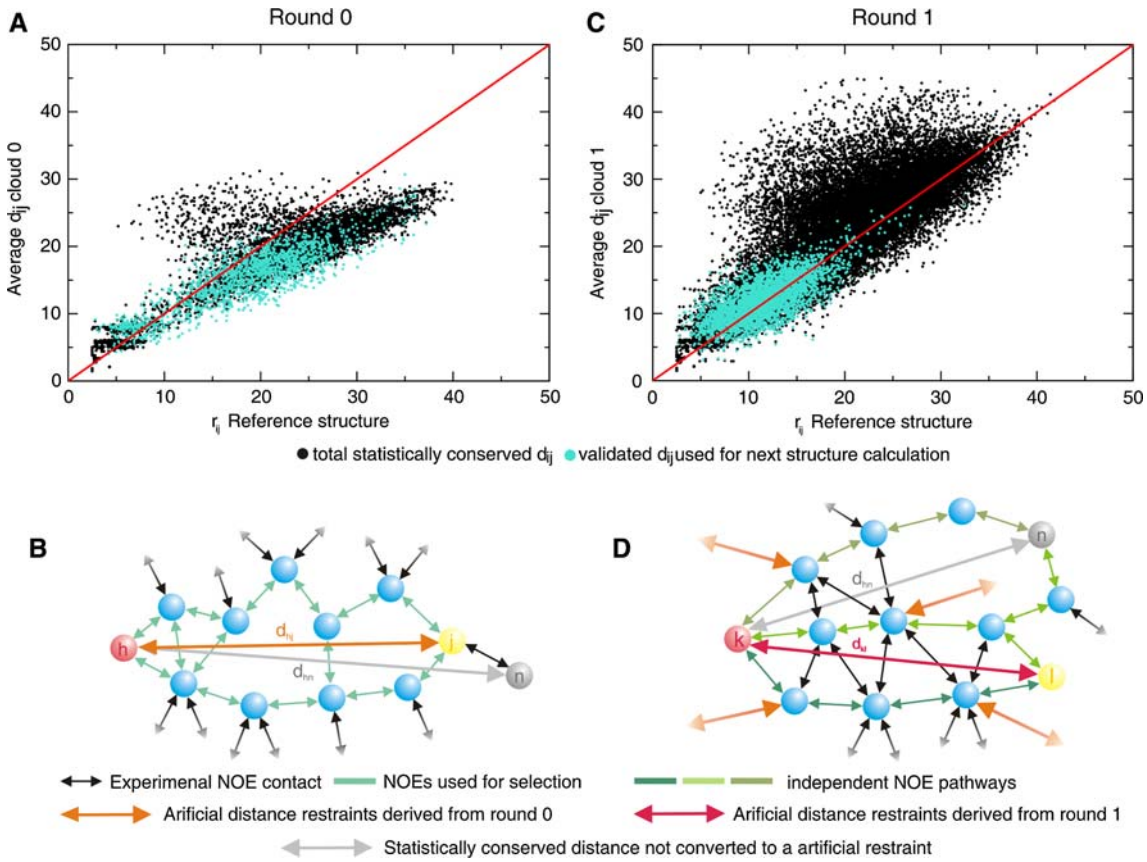

Fig. 4 Distance correlation plot for the protein Cyclophilin D. (**a**) Atomic cloud distances r_{ij} averaged over an ensemble of 50 structures. The clouds were generated (Simulated Annealing I, round 0) from a random gaseous proton distribution using the experimental NOE distance restraints only. Distances found to be statistically 'conserved' are plotted against those of the reference structure (black). Many distances have the tendency to be overestimated. Application of selection criteria based also on the experimental input NOE density (Statistical Analysis I) lead to a filtered set of distance restraints where the (mostly) overestimated erroneous d_{ij} are efficiently removed (cyan). (**b**) Illustration of the selection criteria used in the Statistical Analysis I (**c**) Atomic cloud distances r_{ij} averaged over an ensemble of 50 structures. The clouds were generated (Simulated Annealing II, round 1) from a random gaseous proton distribution using the experimental NOE distance restraints and those extracted as conserved in the previous round. Distances found to be statistically 'conserved' are plotted against those of the reference NMR structure (black). Again, application of selection criteria based also on the experimental input NOE density (Statistical Analysis II) leads to a filtered set of distance restraints where the erroneous d_{ij} are efficiently removed (cyan). (**d**) Illustration of the selection criteria used in the Statistical Analysis II

And finally, that all distances between atoms i and j to any third atom of the NOE pathway were larger than the bounds of the experimental NOEs.

Nevertheless, the distance restraints extracted in this second round involve atom pairs not yet directly restrained and show a slight tendency to be overestimated. This general trend partly compensates to some extent for the fact that original experimental NOE restraints were set to fairly tight values and puts the slope of the correlation back to unity.

Applying these newly-derived artificial distances restraints in a new round of structure calculations a new set of atom density clouds was calculated (iteration 2).

This procedure of structure calculations and identification of conserved distances can be repeated and iterated until convergence (i.e. until no more indirect or remote distance restraints could be identified.) Typically, convergence occurred after two iterations.

Applying the structure generation protocol outlined above, convergent proton densities were obtained, which resembled low-resolution images of a protein or parts thereof. Converged parts of a protein structure, which were superimposable, could be identified by analyzing the matrix of mutual distances between atoms r_{ij}. Distances r_{ij} with small variations σ_{ij} are indicative of regions with defined structures.

Similar to Distance Geometry (Clore et al. 1987a; Clore et al. 1986; Clore et al. 1987b; Havel et al. 1983a; Havel et al. 1983b), the resulting atomic clouds clustered in two degenerate families of structures which are related to each other as mirror images of each other, because the NOE term is invariant to inversion or reflection. Given the

typical resolution of the atomic clouds the correct protein fold could not be distinguished from its inverse one and the two families of structures remained degenerate.

The three dimensional proton density clouds obtained were of great value in the process of sequential signal assignment as they were capable of improving the information content of triple-resonance data, as only distance information is used, both the families of structures can be used equally well.

Model driven sequential chemical shift assignment

To demonstrate the usefulness of low-resolution atom density clouds and assess their impact in the context of sequential assignment, they were included together with backbone chemical shift data into sequential chemical shift assignment. For proof of principle this was evaluated using Monte Carlo simulated annealing (MCSA) using 'Full Monte' version 2.02 assignment software (Hitchens et al. 2003) though alternative assignment software would have probably performed equally e.g. (Jung and Zweckstetter 2004; Leutner et al. 1998; Moseley et al. 2004; Wang et al. 2005; Zimmerman et al. 1997).

Input data comprised ^{15}N and ^{13}C edited 3D (and 4D) NOE information and backbone $^{13}C^{\alpha}_{i}$ and $^{13}C^{\alpha}_{i-1}$ shifts (and for comparison $^{13}C^{\beta}_{i}$ $^{13}C^{\beta}_{i-1}$ $^{13}C'_{i}$ and $^{13}C'_{i-1}$ when available), which were obtained from additional, uniformly $^{15}N/^{13}C$ labeled protein, preparations, classified by $^{1}H^{N}$–^{15}N spin system. Furthermore, either the primary sequence with or without a three-dimensional PDB model or a proton density cloud was supplied. Residue type specific labeling as described above (when used) was included as additional boundary condition.

Using the labeling scheme based on α-ketobutyrate and α-ketoisovalerate facilitates a-priori identification of Val and Ile and sequentially neighboring residues through simplified HNCO/HNCA/HN(CO)CA spectra.

Furthermore, through careful analysis of their $^{13}CH_3$ chemical shifts and NOE patterns, which must be complementary to the corresponding $^{1}H^{N}$ NOE patterns they can be uniquely linked up with their backbone $^{1}H^{N}$ (unpublished results). Again, simplified NOE spectra based on the α-ketobutyrate/α-ketoisovalerate labeling scheme are best suited for that purpose due to their simplicity.

With the information of $^{13}C^{\alpha}$, $^{13}C'$ backbone and $^{13}C^{\gamma/\delta}$ methyl shift combined Val, Ile and Leu residues could be almost uniquely identified.

A number of different input scenarios were tested for comparison.

Different extents of available ^{13}C backbone chemical shifts ($^{13}C^{\alpha/\beta}{}_{i/i-1}$ or $^{13}C^{\alpha}{}_{i/i-1}$ alone), different extents of precision and completeness of ^{13}C backbone chemical shifts were used. The impact of a three dimensional model was evaluated by inclusion or omission of atomic density clouds of different resolution. To create a suitable structural model as input for assignment by Monte Carlo simulated annealing (using the Full Monte software system). The atom density clouds generated by structure calculations were superimposed and clustered. And if equivalent families of solutions existed (see below), a number of the lowest energy density clouds were merged for each cluster. The resulting equivalent averaged models which were considered as representative and used as input in the sequential assignment algorithm.

In a similar way different residue type specific labeling strategies were evaluated, too.

The goal was to reproduce as many of the (a-priori known) assignments as possible. For each input set of data, ten cycles of Monte Carlo simulated annealing were performed and ten assignment configurations were calculated and evaluated with respect to correctness and uniqueness. A sequential assignment was considered correct, if it was reproduced in the majority of the cases, and it was considered unique, if it was consistently reproduced in ten out of ten cases.

Simulated annealing III

With the sequential protein backbone and NOE assignment available, all requirements for classical structure calculations were fulfilled and the structure generation process could proceed to its final stage. In our hands conventional methods using standard simulating annealing protocols starting from extended covalent protein templates and using assigned NOE distance restraints were completely adequate. Only NOEs originating from residues from the parts of the protein, that were deemed to be uniquely and reliably assigned, were included. Optionally distance restraints extracted from cloud models can be included improving convergence of structure calculations. The remainder of the protein structure was left unrestrained (except for the restrictions imposed by the covalent bond geometry). Similar parameters as in Simulated annealing I/II were used.

A detailed flow diagram of the complete structure generation process is summarized in Fig. 3.

Results and discussion

Residue type and position specific isotope labeling

The incorporation of isotope labels at backbone positions of selected residues can significantly reduce the complexity of crowded NMR spectra and allows rapid identification of Val and Ile (and potentially Leu) residues in such proteins. The frequency of occurrence of Ile and Val residues in

proteins is 5.3 and 6.6%, respectively (Creighton 1992). This is illustrated in Fig. 2 showing greatly simplified residue type edited clean triple-resonance spectra with a minimum of undesired 'cross labeling' between residue types.

Extraction of artificial intramolecular distance restraints

Starting from initial atomic cloud distributions generated using experimental or structure derived, theoretical NOE distances only (*Simulated Annealing I*, round 0), the degree of freedom of the atomic system is successively decreased by adding further artificial D_{ij} to the distance restraints in subsequent round of structure calculation. The $D_{ij}^{(1)}$ and $D_{ij}^{(2)}$ etc. for the following rounds of structure calculation were extracted through analysis of the distances of the atomic cloud distributions and filtering by the consistency criteria defined in "Materials and Methods".

Low density of restraints, as it is the case in the absence of a covalent template, systematically resulted in overly extended structures and distances were systematically overestimated. Therefore small standard deviations σ_{ij} of distances d_{ij} alone are not a sufficient criterion. For that reason more stringent criteria taking input NOE density around the concerned atoms into account have to be applied.

Figure 4 shows the (filtering) effect of the selection criteria. Initially, a large number of distances appeared to be conserved (black dots). When compared to the corresponding distances in the reference structure, many of them were systematically overestimated due to an insufficient restraint density. The green dots represent those distances that were left after application of the selection criteria defined above. Their number is substantially reduced but their correlation with the 'true' distance is greatly improved.

The number of newly obtained artificial distance restraints in a structured part of a protein is of course proportional to the original density of experimental NOE. Hence, more artificial D_{ij} were defined in regions of extended β-sheets and β-barrels, than for predominantly α-helical regions.

Table 2 gives a statistics of the number of additional restraints, which can be derived from the statistical analysis of the atom clouds. It varies with the original number of NOE distance restraints and to some extent with the fold topology of the proteins studied (Results for further simulated datasets are compiled in the supplementary material, Table S2).

Structural results and quality of obtained atomic density models

Applying the structure calculation protocol outlined above, convergent atom densities can be obtained, which resemble a low-resolution image of a protein (Table 3; Fig. 5). Superimposable atoms in the converged cloud ensemble (after round 2) could be identified by the analysis of their distance matrix. The occurrence of a high density of conserved r_{ij} indicates the presence of structurally well-defined regions. Depending on the extent and correctness of the initial experimental input NOE restraints, the obtained atomic densities resemble more or less a low resolution image of a protein.

One obvious, fundamental problem encountered at this stage was that the resulting handedness or chirality of the resulting protein backbone fold was undefined, because the NOE term is invariant to inversion or reflection. Therefore, in the absence of a covalent protein template and improper energy terms, which define the stereochemistry around chiral centers, the correct protein fold could not be distinguished from the inverse, wrong one. On the basis of NOEs alone, one is left with at least two families of solutions with degenerate energy.

Only when the resolution is sufficient that the stereochemistry around chiral centers (backbone $^{13}C^{\alpha}$, Ile $^{13}C^{\beta}$, Thr $^{13}C^{\beta}$, which were not included in the calculation of the atomic cloud distribution) or other chiral features e.g. the handedness of α-helices or twist of β-sheets is resolved, the correct model can be identified. However, this was not the case in the present application, where typical coordinate precision is limited. The results of the simulated annealing stage II are summarized in Table 3A for simulated annealing stage III in Table 3B (Results for further simulated datasets are compiled in the supplementary material, Table S3).

Nevertheless, the three dimensional proton clouds obtained thereby are of great value. E.g. when used to aid in the process of sequential signal assignment and NOE identification. Having even limited structural information at this stage is still useful in guiding the sequential assignment process by making references to a low-resolution PDB model. In this context the problem distinguishing the correct mirror image from the wrong one ceases to be a limitation, at least at the assignment stage. The apparent disadvantage turns out in fact to be an advantage as either possible stereoisomer can be used in the model driven assignment process and the selection of the correct stereoisomer can be delayed until later.

To assess the value of a structural model in the process of sequential shift assignment sequential assignment was performed using Monte Carlo simulated annealing methods with different input scenarios.

First, an idealized, i.e. complete and error-free set of $^{13}C^{\alpha/\beta}$ backbone shifts was assumed. The assignment was performed using the NOEs database with and without a structural model in the form of an atomic density cloud. The results are compiled in Table 4 (Results for further

Table 3 Results of structure calculation experimental data

Protein (PDB reference)	Atoms/NOEs used	bb RMSD [Å] to mean	bb RMSD [Å] to reference
(A) Structural statistics of converged atom clouds[A]			
Cyclophilin D	H^N, CH3	5.61[a] (1.74[b])	5.84[a] (1.92[b])
	H^N, CH_3, NH_2, H^{aro}	2.62[a] (0.57[b])	2.76[a] (0.70[b])
	H^N, CH_3, NH_2, H^{aro}, H^α	1.61[a] (0.41[b])	1.71[a] (0.51[b])
ICln (1ZYI)	H^N, CH_3, NH_2, H^{aro}, H^α	0.84[a] (0.14[b])	1.06[a] (0.30[b])
qCRP2(LIM1) (1A7I)	H^N, CH_3, NH_2, H^{aro}, H^α	1.89[a] (1.86[n], 1.72[c])	2.08[a] (1.99[n], 1.87[c])
(B) Structural statistics of covalent all atom models[B]			
Cyclophilin D	H^N, CH_3	2.01[a] (0.82[b])	2.13[a] (0.87[b])
ICln (1ZYI)	H^N, CH_3, NH_2, H^{aro}, H^α	1.65[a] (0.63[b])	1.84[a] (0.72[b])
qCRP2(LIM1) (1A7I)	H^N, CH_3, NH_2, H^{aro}, H^α	2.35[a] (2.38[n], 0.85[c])	3.30[a] (3.07[n], 1.70[c])

[A] Produced by Simulated Annealing stage II. RMSDs are quoted for the cluster of atom density clouds with the correct stereochemistry

[B] Produced by Simulated Annealing stage III

Remarks: CH_3: stands for Ile $C^{\delta 1}H_3$, Leu $C^{\delta 1/2}H_3$, Val $C^{\Upsilon 1/2}H_3$, Ala $C^\beta H_3$, Thr $C^{\Upsilon 2}H_3$

NH_2: Asn $H^{\delta 1/2}$, Gln $H^{\varepsilon 1/2}$, H^{aro}: Phe $H^{\delta 1/2 \varepsilon 1/2 \zeta}$, Tyr $H^{\delta 1/2 \varepsilon 1/2}$; When H^α atoms were included, only the sequential H^N–H^α NOEs were used

[a] All converged residues

[b] β barrel (structural core)

[n] N-terminal domain

[c] C-terminal domain

simulated datasets are compiled in the supplementary material, Table S4).

For obvious reasons, the ideal scenario is to have a complete set of protein backbone shifts ($^{13}C'$, $^{13}C^\alpha$, $^{13}C^\beta$, ^{15}N, $^1H^N$) available and consequently this was used as a 'gold standard' for further comparisons. It turned out that once a complete set of triple-resonance data and a complete set of $^{13}C^\alpha$ and $^{13}C^\beta$ shifts is available 90% or more of the sequential assignments can be correctly and unambiguously made right away and the improvement in the accuracy and extent of the sequential assignment by including $^1H^N$–$^1H^N$ NOEs together with a low resolution proton cloud is almost negligible.

From there we could conclude that, whenever possible, under favorable circumstances it is still preferable to follow the standard protocol for sequential backbone assignment using a full set of triple-resonance experiments.

The situation is different, when, for reasons of fast relaxation, the sensitivity of triple-resonance experiments is severely compromised, and at best only HNCO, HNCA and HN(CO)CA experiments can be conducted with sufficient quality. Typically, sequential backbone assignment, using $^{13}C^\alpha_{i/i-1}$ shifts only, while possible in theory, is usually precluded in practice due to the large $^{13}C^\alpha$ shift degeneracy. This is reflected in the extent of achievable sequential backbone assignment, which, while still tentatively correct (50–80% were correct), becomes very instable (only 0–25% of the sequential backbone assignment can be unambiguously made with confidence).

In conjunction with NOE spectra and a low-resolution proton density cloud distribution the task becomes feasible. NOEs, especially sequential $^1H^N$–$^1H^N$ contacts help resolving $^{13}C^\alpha$ chemical shift overlap and alleviate the problem of degeneracy in making sequential connections between spin-systems. If, furthermore, a structural model, e.g. a homology model or, in our case, a low resolution proton density cloud, mid-range $^1H^N$–$^1H^N$ NOEs and other tertiary contacts can be interpreted in the context of the model to improve placement of the spin-systems onto the *tertiary* structure. In these cases the achievable assignment is back to 70–97% (Table 4).

One problem is that for most residue types $^{13}C^\alpha$ shifts, in contrast to $^{13}C^\beta$, do not carry a lot of residue type information. This means that, while it may still possible to sequentially link spin-systems, using their $^{13}C^\alpha_{i/i-1}$ shifts, it is impossible to correctly place them in the primary sequence of a protein. This was also reflected in our test cases where long stretches of correctly linked spin systems failed to be unambiguously mapped onto the amino-acid sequence resulting in the low reliability of the assignments (see Table 4).

This could be dramatically improved by the application of residue type selective isotope labeling, of $^{13}C^\alpha$ and $^{13}C'$ of Val and Ile and $^{13}C^{\gamma/\delta}$ methyl of Val, Leu and Ile residues which provides pivotal points in the sequential assignment process and helps to correctly map the stretches of linked spin systems onto the *primary* structure. In cases where an atomic cloud model is not completely sufficient

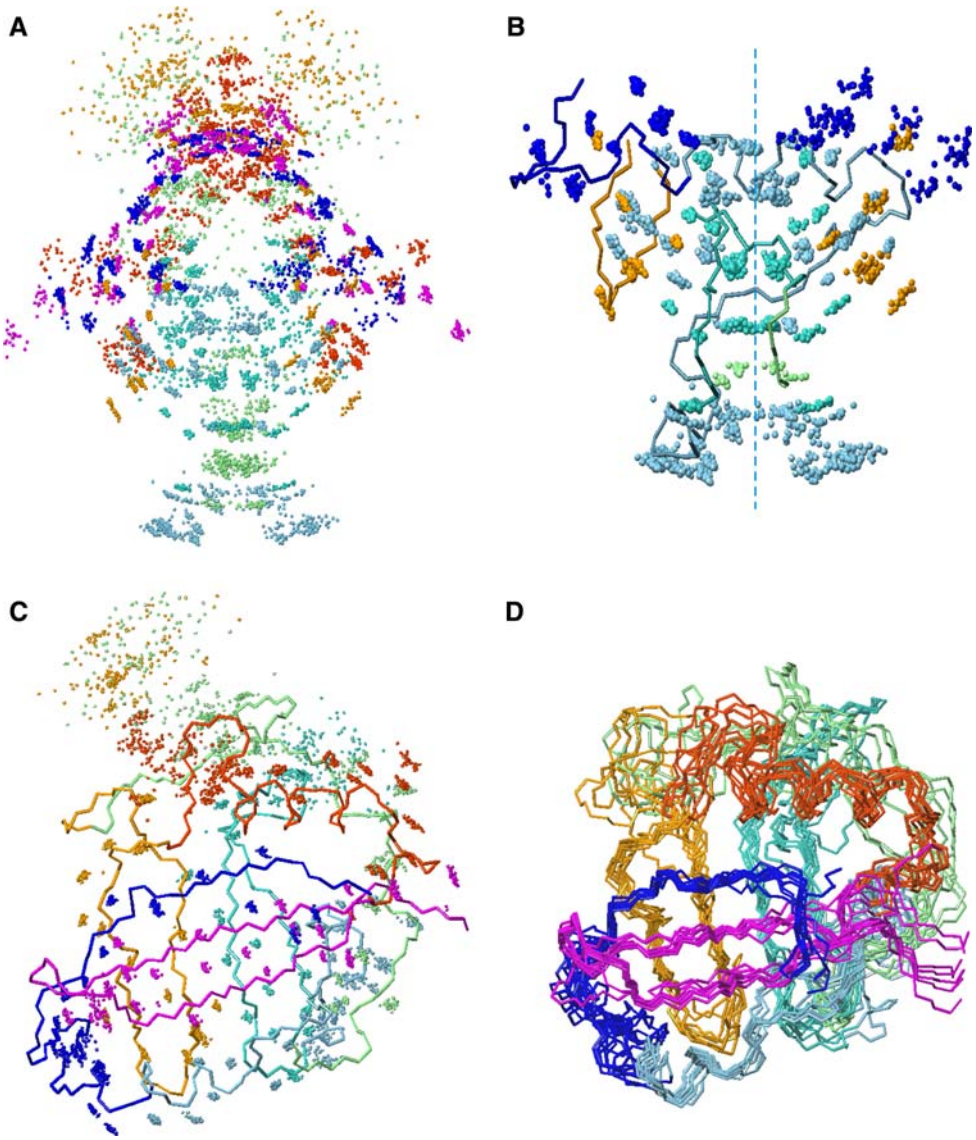

Fig. 5 A complication arises from the fact, that in the absence of any force-field terms, other than distance restraints the 'handedness' or 'chirality' of structures is undefined and two equivalent symmetry related clusters of solutions with identical low energy exist, which exhibit an inverted overall fold (**a**, **b**). The two degenerate results for the protein Cyclophilin D are shown. They cannot be a-priori distinguished at this stage. Nevertheless, they can be used to drive the sequential assignments using minimal triple-resonance data. Once the sequential shift assignment becomes available, this ambiguity can be automatically resolved at the final structure calculation stage of the protocol, when a full all-atom force field is employed (including improper potentials). Side chain NOEs are particularly important in the definition of the correct fold (**c**, **d**)

to assign all $^{13}C^{\alpha}$ (as in the case of ICln) the last remaining ambiguities can be resolved through residue type specific labeling.

Interestingly, such residue type specific labeling schemes can provide the sequential assignment of (sometimes substantial) parts of a protein even in complete absence of triple-resonance data, given a sufficient density of experimental NOEs (data not shown).

Interestingly, although only $^{1}H^{N}$–$^{1}H^{N}$ NOEs are used directly in the assignment algorithm, and only proton cloud distributions with zero violations are used, the assignment score improves slightly in the course of the structure generation process from iteration 0 to iteration 2, as the accidental occurrence of false-tertiary contacts, which can misguide the assignment process, is gradually eliminated.

A number of further factors were investigated using the example of cyclophilinD (Table 5).

We investigated the influence of peak picking accuracy. Typically, the precision at which a peak position can be determined depends on both the signal to noise ratio (S/N) and the linewidth at half height (LW1/2). It can be expressed as $^{LW1/2}/_{S/N}$. Using typical numbers of S/N=20 and a linewidth, determined by an acquisition time in the $^{13}C^{\alpha}$ dimension of 7 ms, random noise of an RMS of $\pm$0.2 ppm was added to each intra $^{13}C^{\alpha}_{i}$ and inter $^{13}C^{\alpha}_{i-1}$ chemical shift value.

As a result of this the extent of achievable assignment using $^{13}C^{\alpha}_{i,i-1}$ only dropped by a factor of two to 45% (only 4.2% could be made with confidence) and 50% were tentatively wrong.

Under such circumstances including a structural model in the assignment resulted in a big improvement. Correctness was back up to 97% and reliability to 64%. Virtually complete and uniquely correct assignment could again be

Table 4 Results of assignment calculations: extent of sequential assignment achieved (%)

Protein	PDB model cloud[a]	Residue specific labeling[b]	Shifts used	Correctly (uniquely) assigned [%]	Incorrectly assigned [%]
CyclophilinD			C^{α}, C^{β}	99.4 (99.4)	0.6
			C^{α}	80.5 (14.9)	19.4
	Yes		C^{α}	99.4 (97.4)	0
ICln			C^{α}, C^{β}	87.3 (84.5)	13.0
		Yes	C^{α}, C^{β}	99.1 (98.2)	0.9
			C^{α}	52.8 (0)	49.0
	Yes		C^{α}	72.7 (70.9)	27.8
		Yes	C^{α}	97.3 (90.0)	2.7
	Yes	Yes	C^{α}	97.3 (95.5)	2.7
qCRP2(LIM1)			C^{α}, C^{β}	94.3 (91.4)	5.7
			C^{α}	85.7 (25.7)	14.3
	Yes		C^{α}	94.3 (85.7)	5.7
		Yes	C^{α}	91.4 (90.0)	8.6
	Yes	Yes	C^{α}	100 (91.4)	0

Remarks: [a] An atomic density cloud including H^N, CH_3 (I,L,V), H^{aro} (F,Y), NH_2 (N,Q) atoms was used as a PDB model to support the assignment

[b] Specific ^{13}C labeling of Val, Ile residues in C^{α}, C' and Val, Ile and Leu in $C^{\gamma 1/2/\delta 1/\delta 1/2}$ positions as described in Materials and Methods

Table 5 Effects of limited precision of chemical shifts and completeness of chemical shift database

PDB model cloud[a]	Residue specific labeling[b]	Completeness C^{α}_i [%]	Completeness C^{α}_{i-1} [%]	Correctly (uniquely) assigned [%]	Incorrectly assigned [%]
		100	100	45.8 (4.2)	50.3
Yes		100	100	97.3 (64.2)	2.6
		100	38.1	29.9 (3.2)	66.4
Yes		100	38.1	72.2 (16.1)	27.1
		83.9	38.1	16.9 (5.2)	79.3
Yes		83.9	38.1	86.8 (11.6)	12.9
	Yes	100	100	83.8 (64.9)	15.5
Yes	Yes	100	100	96.7 (96.7)	3.2
	Yes	100	38.1	85.5 (57.9)	14.2
Yes	Yes	100	38.1	98.0 (90.3)	1.9
	Yes	83.9	38.1	86.2 (62.5)	13.5
Yes	Yes	83.9	38.1	91.5 (88.9)	8.4

H^N-H^N NOEs and $^{13}C^{\alpha}_i$ and $^{13}C^{\alpha}_{i-1}$ were used for Monte Carlo assignment by simulated annealing. The precision of the $^{13}C^{\alpha}$ chemical shifts was assumed to be 0.2 ppm. Various extents of C^{α}_i and C^{α}_{i-1} shifts were removed to assess the effect of gaps in the assignment. First 25 of 155 (16.1%) of the C_i^{α} shifts (those residues located in bend regions) and then 96 of 155 (61.9%) of the C^{α}_{i-1} shifts (those residues located outside of regular secondary structure) were removed. CyclophillinD (165 residues)

Remarks: [a]An atomic density cloud including H^N, CH_3 (I,L,V), atoms was used as a PDB model to support the assignment

[b] Specific ^{13}C labeling of Val/Ile/Leu residues in C^{α}, C' and $C^{\gamma/\delta 1/\delta 1/2}$ positions as described in Materials and Methods

obtained when residue type selective labeling was added as further constraint.

Next, the influence of incompleteness of the $^{13}C^{\alpha}$ chemical shift database was studied. First, by deleting 96 sequential $^{13}C^{\alpha}_{i-1}$ chemical shifts i.e. those $^{13}C^{\alpha}_{i-1}$ associated with ($^1H^N_i$ $^{15}N_i$) pairs of residues located outside of regular secondary structure. And second, by additionally deleting a selection of 25 intra $^{13}C^{\alpha}_i$ chemical shifts associated with residues found in bend regions of the protein.

Using only the remaining $^{13}C^{\alpha}$ chemical shifts only 30 and 17% rsp. were correctly assigned and the percentage of incorrect assignments went up to 66 or 80%. Resorting to an atom density cloud structural model brought the percentage of correct assignments back up to more than 70% but only with a modest percentage of unambiguous

assignments. To further improve upon this residue type selective labeling is required.

Finally, the influence of substantial structural errors, which could arise from mis-assignment of NOEs, was studied. In our case four 'wrong' NOEs were deliberately introduced by adding extra 'fake' distance restraints between pairs of atoms far apart in the tertiary structure. These restraints were chosen to simulate gross errors in the assignment with severe structural consequences. They included a restraint between the N- and the C-cap of an α-helix, between the N- and the C-terminus of a β-strand, a restraint between the top and bottom of a β-barrel and a contact between β-strands not adjacent in a β-sheet (Indicated in Fig. 6a on the solution structure of Cyclophilin D). It should be mentioned that in practice, in this particular case, those wrong restraints could have been spotted in advance by an unusual distribution of NOE contacts in the analysis of the density of restraints. All these errors, while detrimental for the resulting proton density cloud distributions per-se, (which were severely distorted and did no longer resemble the protein tertiary structure), had surprisingly little influence on the results of the model-driven assignment process as they left the overwhelming majority of native tertiary contacts intact (Fig. 6b; Table 6). The extent of the sequential assignment obtained is only marginally lower that that achieved with a 'correct' structural model applied to the same input data.

Conclusion

To sum up, we have presented a method capable of generating low-resolution representations of protein tertiary structures e.g. for fold validation in the complete absence of any sequential assignment by iteratively applying simulated annealing to an ensemble of gaseous unconnected atoms which are progressively condensed into so-called atom density clouds, which resemble a protein structure. A potential weakness of this method is that, with NOE distance restraints as the only source of restraints, the stereochemistry around chiral centers is undefined, because the correct protein fold cannot be distinguished from its mirror image on the basis of distance restraint violation only. Typically, obtaining a resolution high enough to be able to identify the correct mirror image on the basis of a-priori stereochemical knowledge (e.g. the configuration around chiral centers, or the handedness of helices) requires a prohibitively large number of restraints. Therefore this decision can be deferred until later. Regardless of

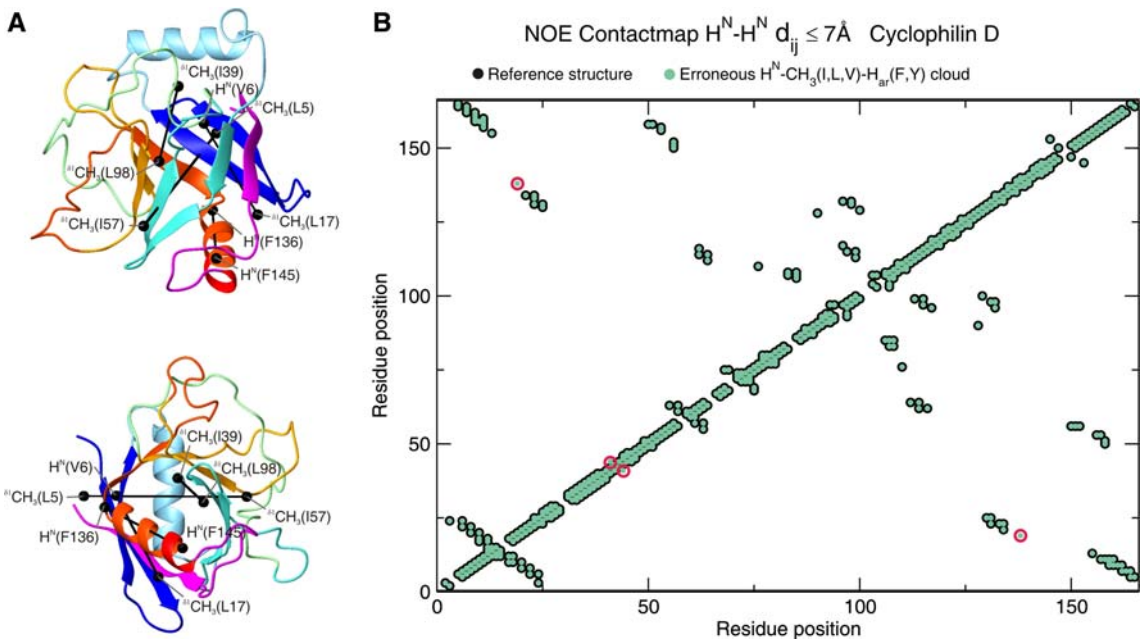

Fig. 6 Structural errors, generated by misassignment of NOEs may be detrimental for the outcome of the atomic density calculations, which may no longer resemble a protein structure. Four such erroneous distance restraints as indicated in (**a**) were introduced into the calculation of an atomic density cloud for cyclophilinD. Interestingly, even introduction of serious structural 'mistakes' does not significantly affect the distribution of tertiary contacts. (**b**) compares the tertiary contacts of the erroneous structural model generated by the application of the additional erroneous distance restraints (shown in **a**) with those of the reference solution structure (determined in a conventional way). The additional erroneous tertiary backbone contacts are highlighted by red circles in the contact map. Their only limited effect on the assignment process is summarized in Table 6

Table 6 Effect of structural errors of the atom density cloud for the assignment process

Residue specific labeling[a]	Completeness C^{α}_{i} %	Completeness C^{α}_{i-1} %	Correctly (uniquely) assigned %	Incorrectly assigned %
	100	100	99.3 (91.0)	0.7
Yes	100	100	99.3 (99.3)	0.7
	83.9	38.1	13.1 (0)	81.3
Yes	83.9	38.1	87.6 (79.7)	12.3

A distorted atomic density cloud PDB file including H^N, CH_3 (I,L,V), H^{aro} (F,Y), NH_2 (N,Q) atoms was generated by adding four incorrect long range NOE to the experimental input NOE database (1) L5-$^{\delta 1}CH_3$ (facing outwards at the top of β-barrel) to I57-$^{\delta 1}CH_3$ (exterior at bottom opposite site of β-barrel), (2) backbone V6-H^N (begin strand β2) to L17-$^{\delta 1}CH_3$ (end of strand β2), (3) L39-$^{\delta 1}CH_3$ (core of β-barrel) to L98-$^{\delta 1}CH_3$ (core of β-barrel), and (4) backbone M136-H^N (N-cap α2 helix) to F145- H^N (C-cap α2 helix)

To assess the effect of structural errors in the atom cloud (e.g. brought about by misassignment of NOEs) this model was used to support the assignment process

Remarks: [a] Specific ^{13}C labeling of Val/Ile/Leu, residues in C^{α}, C′ and $C^{\gamma/\delta 1/\delta 1/2}$ positions as described in Materials and Methods

their absolute stereochemistry, proton density clouds can be used as models to drive or guide the sequential backbone chemical shift assignment using only minimal triple-resonance data as input. We have shown that their inclusion into the assignment procedure substantially enhances the information content of HNCA/HN(CO)CA spectral information, which by itself is usually insufficient for unique backbone assignment.

Further improvement is obtained in combination with residue type selective isotope labeling.

We have additionally presented a novel isotope-labeling pattern for the biosynthetic precursor compounds α-ketobutyrate and α-ketoisovalerate, which allows selective incorporation of isotope labels at backbone C^{α} and C′ positions of Val and Ile residues (C^{β} of Leu residues) by standard laboratory procedures. Such a labeling pattern facilitates identification of Val, Ile and Leu and sequentially neighbored residues in proteins.

Using position specifically isotope labeled precursor compounds allows spectral editing of triple-resonance spectra and spectral simplification of multi-dimensional NOE spectra and provides important pivotal points for the sequential assignment process.

Once the assignment has become available for the majority of the residues, and the key methyl groups have been linked up to the correct backbone $^1H^N$ ^{15}N pair, the question of the absolute stereochemistry can be resolved by repeating the structure calculations starting from a covalent template including all improper and chiral restraints of a full force-field.

Alternatively, methods of bioinformatics can be used to evaluate the structural alternatives and identify the best solution by calculating a structural quality score.

Finally, we have shown that this algorithm is to some extent error-tolerant with respect to data incompleteness, limited precision of the peak picking and structural errors caused by misassignment of NOEs.

References

Bermejo GA, Llinas M (2008) Deuterated protein folds obtained directly from unassigned nuclear overhauser effect data. J Am Chem Soc 130:3797–3805

Bowers PM, Strauss CE, Baker D (2000) De novo protein structure determination using sparse NMR data. J Biomol NMR 18: 311–318

Brünger AT (1993) X-PLOR (Version 3.1) a system for X-ray crystallography and NMR. Yale University Press, New Haven

Brunger AT, Adams PD, Clore GM, DeLano WL, Gros P, Grosse-Kunstleve RW, Jiang JS, Kuszewski J, Nilges M, Pannu NS, Read RJ, Rice LM, Simonson T, Warren GL (1998) Crystallography & NMR system: a new software suite for macromolecular structure determination. Acta Crystallogr D Biol Crystallogr 54:905–921

Cavanagh J, Fairbrother AG, Palmer AG, Skelton NJ (1996) Protein NMR Spectroscopy. Academic Press, New York

Clore GM, Nilges M, Sukumaran DK, Brunger AT, Karplus M, Gronenborn AM (1986) The three-dimensional structure of alpha1-purothionin in solution: combined use of nuclear magnetic resonance, distance geometry and restrained molecular dynamics. EMBO J 5:2729–2735

Clore GM, Nilges M, Brunger AT, Karplus M, Gronenborn AM (1987a) A comparison of the restrained molecular dynamics and distance geometry methods for determining three-dimensional structures of proteins on the basis of interproton distances. FEBS Lett 213:269–277

Clore GM, Sukumaran DK, Nilges M, Zarbock J, Gronenborn AM (1987b) The conformations of hirudin in solution: a study using nuclear magnetic resonance, distance geometry and restrained molecular dynamics. EMBO J 6:529–537

Clubb RT, Thanabal V, Wagner G (1992) A constant-time 3-dimensional triple-resonance pulse scheme to correlate intra-residue H-1(N), N-15, and C-13(‘) chemical-shifts in N-15-C-13-labeled proteins. J Magn Reson 97:213–217

Creighton TE (1992) Proteins: structures and molecular properties. Freeman, W. H, San Francisco, USA

Evans JNS (1995) Biomolecular NMR spectroscopy. Oxford University Press, New York

Farrow NA, Muhandiram R, Singer AU, Pascal SM, Kay CM, Gish G, Shoelson SE, Pawson T, Forman-Kay JD, Kay LE (1994) Backbone dynamics of a free and phosphopeptide-complexed Src homology 2 domain studied by 15N NMR relaxation. Biochemistry 33:5984–6003

Furst J, Schedlbauer A, Gandini R, Garavaglia ML, Saino S, Gschwentner M, Sarg B, Lindner H, Jakab M, Ritter M, Bazzini C, Botta G, Meyer G, Kontaxis G, Tilly BC, Konrat R, Paulmichl M (2005) ICln159 folds into a pleckstrin homology domain-like structure. Interaction with kinases and the splicing factor LSm4. J Biol Chem 280:31276–31282

Gardner KH, Kay LE (1998) The use of 2H, 13C, 15N multidimensional NMR to study the structure and dynamics of proteins. Annu Rev Biophys Biomol Struct 27:357–406

Gardner KH, Rosen MK, Kay LE (1997) Global folds of highly deuterated, methyl-protonated proteins by multidimensional NMR. Biochemistry 36:1389–1401

Goto NK, Gardner KH, Mueller GA, Willis RC, Kay LE (1999) A robust and cost-effective method for the production of Val, Leu, Ile (delta 1) methyl-protonated 15N-, 13C-, 2H-labeled proteins. J Biomol NMR 13:369–374

Grishaev A, Llinas M (2002a) CLOUDS, a protocol for deriving a molecular proton density via NMR. Proc Natl Acad Sci USA 99:6707–6712

Grishaev A, Llinas M (2002b) Protein structure elucidation from NMR proton densities. Proc Natl Acad Sci USA 99:6713–6718

Grishaev A, Llinas M (2002c) Sorting signals from protein NMR spectra: SPI, a Bayesian protocol for uncovering spin systems. J Biomol NMR 24:203–213

Grishaev A, Llinas M (2004) BACUS: a Bayesian protocol for the identification of protein NOESY spectra via unassigned spin systems. J Biomol NMR 28:1–10

Grishaev A, Llinas M (2005) Protein structure elucidation from minimal NMR data: the CLOUDS approach. Methods Enzymol 394:261–295

Grishaev A, Steren CA, Wu B, Pineda-Lucena A, Arrowsmith C, Llinas M (2005) ABACUS, a direct method for protein NMR structure computation via assembly of fragments. Proteins 61:36–43

Grzesiek S, Bax A (1992) Correlating backbone amide and side-chain resonances in larger proteins by multiple relayed triple resonance Nmr. J Am Chem Soc 114:6291–6293

Grzesiek S, Bax A (1993) Amino acid type determination in the sequential assignment procedure of uniformly 13C/15N-enriched proteins. J Biomol NMR 3:185–204

Havel TF, Crippen GM, Kuntz ID, Blaney JM (1983a) The combinatorial distance geometry method for the calculation of molecular conformation. II. Sample problems and computational statistics. J Theor Biol 104:383–400

Havel TF, Kuntz ID, Crippen GM (1983b) The combinatorial distance geometry method for the calculation of molecular conformation. I. A new approach to an old problem. J Theor Biol 104:359–381

Hitchens TK, Lukin JA, Zhan Y, McCallum SA, Rule GS (2003) MONTE: an automated Monte Carlo based approach to nuclear magnetic resonance assignment of proteins. J Biomol NMR 25:1–9

Hoffmann B, Eichmuller C, Steinhauser O, Konrat R (2005) Rapid assessment of protein structural stability and fold validation via NMR. Methods Enzymol 394:142–175

Jung YS, Zweckstetter M (2004) Mars–robust automatic backbone assignment of proteins. J Biomol NMR 30:11–23

Kay LE, Gardner KH (1997) Solution NMR spectroscopy beyond 25 kDa. Curr Opin Struct Biol 7:722–731

Kay LE, Ikura M, Tschudin R, Bax A (1990) 3-dimensional triple-resonance Nmr-spectroscopy of isotopically enriched proteins. J Magn Reson 89:496–514

Kontaxis G, Konrat R, Krautler B, Weiskirchen R, Bister K (1998) Structure and intramodular dynamics of the amino-terminal LIM domain from quail cysteine- and glycine-rich protein CRP2. Biochemistry 37:7127–7134

Koradi R, Billeter M, Wuthrich K (1996) MOLMOL: a program for display and analysis of macromolecular structures. J Mol Graph 14(51–55):29–32

Korzhnev DM, Kloiber K, Kanelis V, Tugarinov V, Kay LE (2004) Probing slow dynamics in high molecular weight proteins by methyl-TROSY NMR spectroscopy: application to a 723-residue enzyme. J Am Chem Soc 126:3964–3973

Kraulis PJ (1994) Protein three-dimensional structure determination and sequence-specific assignment of 13C and 15N-separated NOE data. A novel real-space ab initio approach. J Mol Biol 243:696–718

Leutner M, Gschwind RM, Liermann J, Schwarz C, Gemmecker G, Kessler H (1998) Automated backbone assignment of labeled proteins using the threshold accepting algorithm. J Biomol NMR 11:31–43

Lichtenecker R, Ludwiczek ML, Schmid W, Konrat R (2004) Simplification of protein NOESY spectra using bioorganic precursor synthesis and NMR spectral editing. J Am Chem Soc 126:5348–5349

Moseley HN, Sahota G, Montelione GT (2004) Assignment validation software suite for the evaluation and presentation of protein resonance assignment data. J Biomol NMR 28:341–355

Nilges M, Macias MJ, O'Donoghue SI, Oschkinat H (1997) Automated NOESY interpretation with ambiguous distance restraints: the refined NMR solution structure of the pleckstrin homology domain from beta-spectrin. J Mol Biol 269:408–422

Pascal SM, Singer AU, Gish G, Yamazaki T, Shoelson SE, Pawson T, Kay LE, Forman-Kay JD (1994) Nuclear magnetic resonance structure of an SH2 domain of phospholipase C-gamma 1 complexed with a high affinity binding peptide. Cell 77:461–472

Pervushin K, Riek R, Wider G, Wuthrich K (1997) Attenuated T2 relaxation by mutual cancellation of dipole-dipole coupling and chemical shift anisotropy indicates an avenue to NMR structures of very large biological macromolecules in solution. Proc Natl Acad Sci USA 94:12366–12371

Pervushin K, Riek R, Wider G, Wuthrich K (1998) Transverse relaxation-optimized spectroscopy (TROSY) for NMR studies of aromatic spin systems in C-13-labeled proteins. J Am Chem Soc 120:6394–6400

Rosen MK, Gardner KH, Willis RC, Parris WE, Pawson T, Kay LE (1996) Selective methyl group protonation of perdeuterated proteins. J Mol Biol 263:627–636

Schedlbauer A, Hoffmann B, Kontaxis G, Rudisser S, Hommel U, Konrat R (2007) Automated backbone and side-chain assignment of mitochondrial matrix cyclophilin D. J Biomol NMR 38:267

Schwieters CD, Clore GM (2001) The VMD-XPLOR visualization package for NMR structure refinement. J Magn Reson 149: 239–244

Schwieters CD, Kuszewski JJ, Tjandra N, Clore GM (2003) The Xplor-NIH NMR molecular structure determination package. J Magn Reson 160:65–73

Simons KT, Kooperberg C, Huang E, Baker D (1997) Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. J Mol Biol 268:209–225

Tugarinov V, Kay LE (2004) An isotope labeling strategy for methyl TROSY spectroscopy. J Biomol NMR 28:165–172

Tugarinov V, Sprangers R, Kay LE (2004) Line narrowing in methyl-TROSY using zero-quantum 1H-13C NMR spectroscopy. J Am Chem Soc 126:4921–4925

Tugarinov V, Kay LE, Ibraghimov I, Orekhov VY (2005) High-resolution four-dimensional 1H-13C NOE spectroscopy using methyl-TROSY, sparse data acquisition, and multidimensional decomposition. J Am Chem Soc 127:2767–2775

Wang J, Wang T, Zuiderweg ER, Crippen GM (2005) CASA: an efficient automated assignment of protein mainchain NMR data using an ordered tree search algorithm. J Biomol NMR 33: 261–279

Wittekind M, Mueller L (1993) Hncacb, a high-sensitivity 3d Nmr experiment to correlate amide-proton and nitrogen resonances with the alpha-carbon and beta-carbon resonances in proteins. J Magn Reson B 101:201–205

Zimmerman DE, Kulikowski CA, Huang Y, Feng W, Tashiro M, Shimotakahara S, Chien C, Powers R, Montelione GT (1997) Automated analysis of protein NMR assignments using methods from artificial intelligence. J Mol Biol 269:592–610