J Biomol NMR (2012) 53:149–159
DOI 10.1007/s10858-012-9632-7

ARTICLE

An improved algorithm for MFR fragment assembly

Georg Kontaxis

Received: 16 January 2012 / Accepted: 25 April 2012 / Published online: 13 May 2012

Abstract A method for generating protein backbone models from backbone only NMR data is presented, which is based on molecular fragment replacement (MFR). In a first step, the PDB database is mined for homologous peptide fragments using experimental backbone-only data i.e. backbone chemical shifts (CS) and residual dipolar couplings (RDC). Second, this fragment library is refined against the experimental restraints. Finally, the fragments are assembled into a protein backbone fold using a rigid body docking algorithm using the RDCs as restraints. For improved performance, backbone nuclear Overhauser effects (NOEs) may be included at that stage. Compared to previous implementations of MFR-derived structure determination protocols this model-building algorithm offers improved stability and reliability. Furthermore, relative to CS-ROSETTA based methods, it provides faster performance and straightforward implementation with the option to easily include further types of restraints and additional energy terms.

Keywords Residual dipolar couplings (RDC) · Molecular fragment replacement (MFR) · Rigid body docking · Fragment assembly · Protein backbone fold

Electronic supplementary material The online version of this article (doi:10.1007/s10858-012-9632-7) contains supplementary material, which is available to authorized users.

G. Kontaxis (✉)
Max F. Perutz Laboratories, Department of Structural and Computational Biology, Centre for Molecular Biology, University of Vienna, Campus Vienna Biocenter 5, 1030 Vienna, Austria
e-mail: georg.kontaxis@univie.ac.at

Introduction

Methods that allow the determination of a rough protein structure or a protein fold are of particular interest in the context of structural genomics as they allow classification and subsequently prioritization of a protein structure right from the onset of a project.

Improving the rate at which structures can be determined by solution state NMR also requires faster and more precise structure calculation protocols for proteins; sometimes starting from a minimal amount of input data. Therefore, there is significant interest in protein backbone fold determination protocols exclusively based on backbone data.

Once sequential assignment has been completed and backbone chemical shifts (CS) are obtained, Residual dipolar couplings (RDCs) are of particular interest. Their measurement can be conducted in a reasonable amount of extra experimental time (Rasia et al. 2011). by variants of $^{15}N/^{13}C$ HSQC or variants of backbone triple-resonance experiments (Ottiger et al. 1998a, b; Chou et al. 2000a; Jaroniec et al. 2004; Bax et al. 2001).

In favorable cases the protein backbone shifts alone can be already sufficient to obtain an accurate model of a protein backbone. This has been demonstrated using the CHESHIRE algorithm (Cavalli et al. 2007), which uses secondary structure information and secondary chemical shift derived backbone conformations together with Monte Carlo structure generation or the CS-ROSETTA methodology (Shen et al. 2008, 2009, 2010) which combines mining the PDB database for homologous peptide fragments using the MFR approach with the Monte Carlo type fragment assembly of this MFR fragment library by the ROSETTA software system (Simons et al. 1999; Rohl et al. 2004; Leaver-Fay et al. 2011).

By restricting the conformational space to be searched, much faster convergence can be achieved making routine applications more feasible. CS-ROSETTA is therefore greatly enhanced by the inclusion of RDCs (Raman et al. 2010; Rasia et al. 2011), which are extremely precise probes of backbone geometry (Bowers et al. 2000; Rohl and Baker 2002; Meiler and Baker 2003; Raman et al. 2010).

However, to ensure sufficient sampling of the accessible conformational space and 'convergence' to the lowest energy model (CS-)ROSETTA structure calculations still require substantial computing resources and CPU time for generating a sufficient number of all-atom models. Therefore, more conventional and computationally less challenging assembly algorithms based on the MFR method are still useful to create backbone folds in a more simplified and straight-forward way, especially when highly complete and very precise input data are available.

A simple implementation of the MFR assembly (Kontaxis et al. 2005), orients all fragments such that the coordinate axes of their local alignment tensors become parallel and then translates the fragments on top of each other. It can provide protein backbone structures, even in the complete absence of NOE information and is quite adequate when complete sets of precise RDCs are available, but still suffers from a number of weaknesses or limitations. The sequential chain building process is a purely geometric algorithm, which, due to accumulation of errors, may result in non-physical structures. Additionally, relative fragment orientations may be ill-defined in case of near axially symmetric alignment tensors.

Furthermore it requires highly complete set(s) of RDCs and it is not easily generalizable for optimal simultaneous use of multiple alignment tensors. It can be prone to errors, as accidental incorporation of unsuitable fragments can guide the protein chain into a wrong direction and disrupt the assembly process. As only adjacent fragments are considered, inclusion of long-range NOE distance restraints or H-bond restraints connecting residues far apart in the primary sequence is not easily possible.

The proposed new implementation uses instead a 'rigid body docking' algorithm for chain extension. A new peptide fragment is placed at the end of the growing protein chain by best-fitting the coordinates of overlapping residues and subsequently its orientation with respect to the rest of the protein is refined against RDC restraints by simulated annealing using the IVM dynamics engine in Xplor-NIH (Schwieters et al. 2003). Fragments that fail to converge to an orientation compatible with the rest of the protein are discarded at that stage. The accepted fragments for a particular residue range are averaged 'on the fly' with the previously built protein backbone and the residue range is shifted by one residue at a time until the C-terminus is reached.

The algorithm is in principle capable of operating in the complete absence of NOE restraints. If, however, a limited amount of NOE information or hydrogen bonding geometry (Bax et al. 1999; Cordier and Grzesiek 1999) is available at this stage (e.g. easily obtainable H^N–H^N or sequential H^{α}–H^N NOEs), it can be easily incorporated into the fragment assembly adding extra stability to the model building. This is particularly helpful for mainly β-sheet proteins to better define strand topology and to keep the β-strands 'in register'.

An additional small set of long range H^N–$C^{\gamma/\delta}H_3$ or $C^{\gamma/\delta}H_3$–$C^{\gamma/\delta}H_3$ methyl NOE interactions, which can easily be extracted from a 3-4D NOE data set on a perdeuterated protein sample with protonated amide and/or methyl positions, can result in further great improvements, even though the density of the NOE restraints is not sufficient *per se* to uniquely define the protein fold.

Materials and methods, description of the algorithm

The proposed method is described in the flowchart diagram in Fig. 1 and proceeds in four different stages described below. It is driven by a set of Tcl scrips, running within the nmrWish Tcl interpreter, which is part of the NMRPipe software distribution (Delaglio et al. 1995). For simulated annealing refinement or rigid body docking Xplor-NIH (version 2.19 or newer) (Schwieters et al. 2003) is employed. The required Xplor input scripts and restraint files for Xplor, which have to be customized for each peptide fragment, are generated individually 'on the fly' by Tcl scripts and in turn the Xplor scripts are executed by the Tcl interpreter in a separate processes.

The scripts are available as part of the Supplementary material.

MFR search

In a first step the query protein is broken into small overlapping peptide fragments Empirically, fragment lengths of 7–10 residues were found to be optimal.

For each query range (starting at the N-terminus) a representative database of protein structures is mined for homologous fragments using backbone CS and RDCs (Bax et al. 2000; Kontaxis et al. 2005). As database about 900 unique PDB structures, which are part of the NMRPipe software system, are routinely searched. Alternatively, a database of about 5,000 PDB structures, which is part of the CS-ROSETTA distribution, can be used for improved performance.

To prove that this algorithm is capable of determining novel protein folds the reference structure(s) (i.e. the already known structures of the query proteins and its

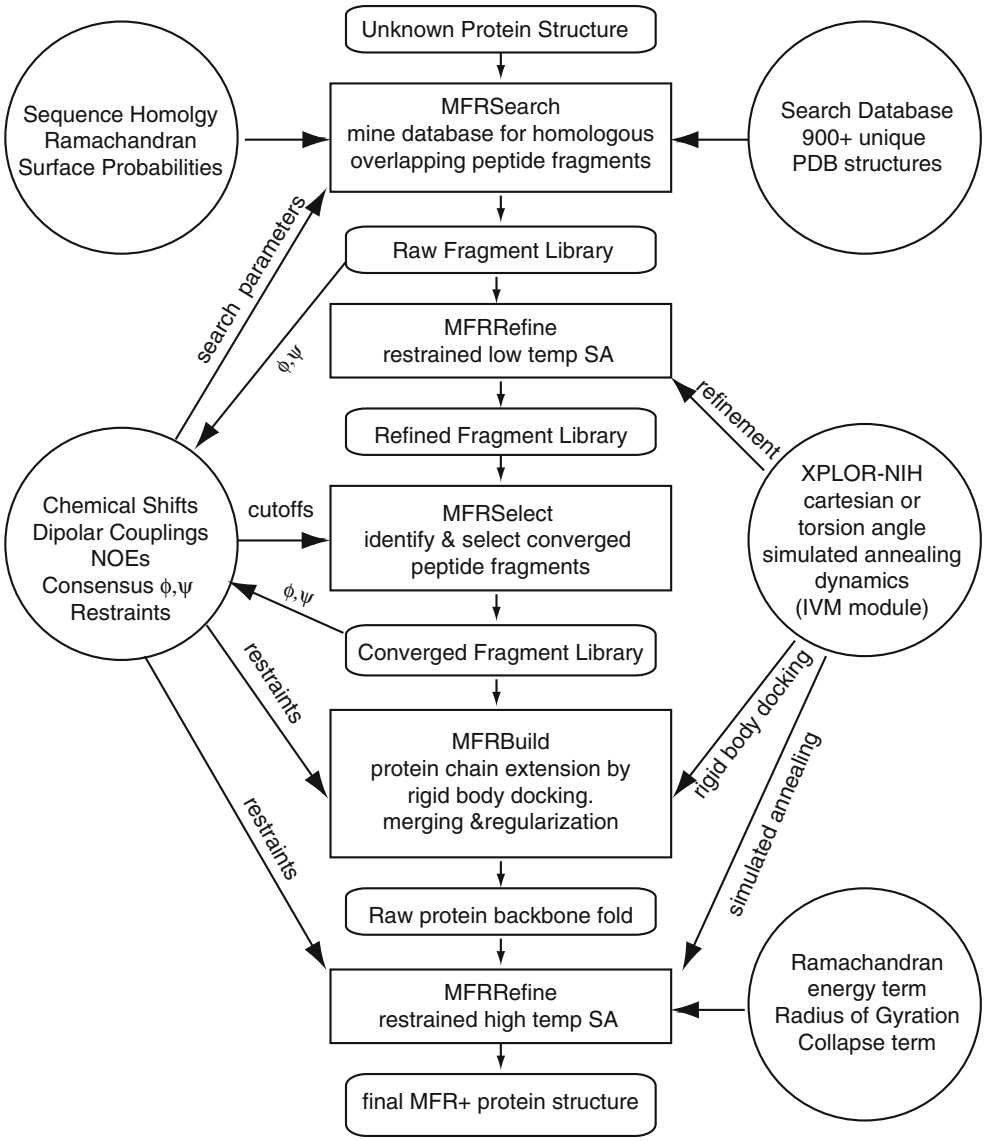

Fig. 1 Flow diagram of the MFR based protein backbone fold determination

homologues) were identified by searching the database for proteins with small backbone coordinate RMSD to the reference proteins and subsequently removed from the search database to avoid structural bias. If this was not the case, the fragments derived from the reference structure(s) (or homologues) were always systematically among the best scoring database hits.

The dipolar term is calculated as RMSD between observed and calculated RDCs by fitting the alignment tensor to the experimental RDCs using the coordinates of the PDB fragment. This is normally performed by an unconstrained SVD algorithm.

The chemical shift term is calculated as RMSD between observed backbone shifts and those predicted from backbone dihedral angles using database derived surfaces surfaces, which describe the (ϕ, ψ) dependence of the secondary chemical shift.

Additional search terms can be included: e.g. a residue type homology and a Ramachandran term. The homology term penalizes 'mutations' between the primary sequence query protein and the homologous trial fragment using a suitable substitution matrix. The Ramachandran surface quality term (Kuszewski et al. 1996, 1997; Kuszewski and Clore 2000), scores the Ramachandran quality of the homologous PDB derived peptide fragment by calculating a (logarithmic) normalized probability of the primary amino acid sequence of the query protein assuming the backbone angles of the homologous trial peptide based on residue type specific (ϕ, ψ) distributions. While not completely excluding rare conformations, it only slightly biases

search results towards the more favoured regions of the Ramachandran plot.

The final score is calculated as the weighted sum of the individual terms. Search results with different parameterizations can be pooled before proceeding. Furthermore a priori knowledge of alignment tensor parameters can be used in the search, by keeping the values of *Da/Dr* fixed in the calculation of the dipolar fit residual, but this comes at a substantial cost in performance and was not used in the present study. Even when applied, this does not completely exclude wrong fragments. Selection the fragments based on *Da/Dr* was therefore deferred until after the following fragment refinement stage where it becomes a more stringent criterion for selection or rejection of fragments.

For each query range the best hits i.e. those showing the best agreement between measured and calculated RDCs as well as between observed and predicted CS are retained. Typically for stretches with highly complete sets of RDCs only 10–20 best hits are kept. In regions with lower density of RDCs they may not be able to define a *unique* solution, but, due to the symmetry of the dipolar interaction, they may be equally compatible with multiple families of conformations. Thus to have a more representative collection of trial fragments it may be preferable to keep a larger number, up to 100.

If, because of exchange-broadening, dynamics or spectral overlap, the density of observable RDCs and CS falls below a certain (adjustable) threshold, and their number is too low to define any structure, then no more fragments are retained for that residue range. Consequently, such regions are later excluded from model building.

The query range is then shifted by one residue and the procedure is repeated until the C-terminus is reached. For a large number of residue positions this results in a number of overlapping trial peptide fragments with very similar backbone torsion angles and for most of the residues well defined consensus backbone angles (ϕ, ψ) can be derived.

MFR refinement

In a next step, the peptide fragment library is refined against the RDCs and other experimental restraints, which may be available at that point (e.g. short-to-mid-range NOE distance constraints, torsion angle restraints, J-couplings) using a very short simulated annealing protocol. This refinement protocol takes no more than a few (5–10) s of CPU time per fragment on a Linux desktop system.

Usually direct refinement against RDCs can pose some computational problems. At early stages of refinement the dipolar energy term can create large, erratic forces. To avoid structural divergence in the absence of NOEs, harmonic restraints are usually applied to C^{α} backbone atoms (Kontaxis et al. 2005) or backbone (ϕ, ψ) torsions (Chou et al. 2000b). To not introduce any wrong structural bias in this application the backbone torsions were restrained to the consensus values obtained in the MFR search. At residue positions where no such consensus value could be extracted no restraint was applied.

Fragments that are already very close to the 'true' query structure undergo only very minor structural rearrangements in the order of a few tenths of Å, which nevertheless greatly improve the agreement between the experimental restraints and those predicted by the PDB fragment. Such 'well-behaved' fragments readily converge to a global minimum energy conformation. Only 'false-positives', fragments accidentally picked up in the MFR search stage, will fail to do so and get trapped in energetically unfavorable conformations and can thus be identified as such.

MFR select

Following the refinement, re-scoring of the structural quality of the refined fragments and elimination of the 'false positives' is performed using a number of criteria (Kontaxis et al. 2005). The agreement between measured and predicted backbone CS or the agreement between the measured and calculated RDCs is the main criterion for the fragment selection. The Ramachandran surface quality is a further important quantity for fragment selection.

The database fragments have to converge to a certain fraction of the lowest value found within each query residue range and not to exceed a maximum value for each of these indicators to be accepted. For the dipolar residual this cut-off is typically $0.3 \times \mathrm{sqrt}\ [(0.8\ Da)^2 + (0.6Dr)^2]$, for the chemical shift residual it is 4–5 standard deviations between observed and predicted backbone shifts and for the (logarithmic) Ramachandran surface quality score it is to 4–5. Furthermore, the value(s) for magnitude and rhombicity (*Da* and *Dr*) of the best-fit alignment tensor(s) must reflect the target value(s) that can be inferred from their distribution in the MFR search results or from a powder pattern analysis of the distribution of experimental RDCs.

If RDCs from more than one alignment medium are available, relative orientation of the individual alignment tensors as well as their 'scalar product' $\mathrm{Tr}(A \cdot B)$ (Sass et al. 1999) can be a criterion for convergence. These values should not exceed 1–2 standard deviations from the median values of the whole fragment library. Further details of the selection procedure can be found in (Kontaxis et al. 2005).

When the distribution of backbone angles of the refined and converged fragments only is analyzed this distribution has become much tighter that those of the original raw peptide trial fragments and extremely precisely defined backbone angles can be derived for a large majority of the residues. Remaining ambiguities can be resolved during fragment assembly.

MFR build

The algorithm for fragment assembly is illustrated in Fig. 2.

This alogrithm differs from previous implementations. Previously, all the fragments were rotated such, that the coordinate axes of their local alignment tensors become parallel to the principal axis frame(s) of the alignment tensor(s) and are then translated on top of each other for best coordinate overlap (with their relative orientation kept fixed as determined before). The weakness of this algorithm is that, each peptide fragment is added to the chain by minimizing its coordinate RMSD with the previously built fragment. If sections of the backbone built previously could not be taken into account, this then results in non-physical protein geometries, e.g. the protein backbone folding back onto itself resulting in severe steric problems or knotted structures.

The proposed new implementation is conceptually similar to previous methods, proposed by (Hus et al. 2001; Giesen et al. 2003; Walsh et al. 2005; Walsh and Wang 2005; Wang et al. 2007; Bouvignies et al. 2006a, b, 2007) which sequentially place individual peptide planes at the end of a growing peptide chain and orient them relative to the protein chain built so far using its experimental RDCs and sometimes NOEs where available.

However, rather than using individual single peptide units at a time, in this new proposed application chain extension is performed by placing the PDB derived 7-10-residue peptide fragments—as obtained by the MFR homology search and refinement—at the end of the growing protein chain similar to (Berardi et al. 2011). This avoids numerical instabilities or singularities, which may arise in previous implementation when attempting to place a peptide plane with an incomplete set of experimental RDCs such as Proline residues or exchange-broadened resonances.

Instead of simple geometric considerations, e.g. searching for best coordinate overlap with the previously

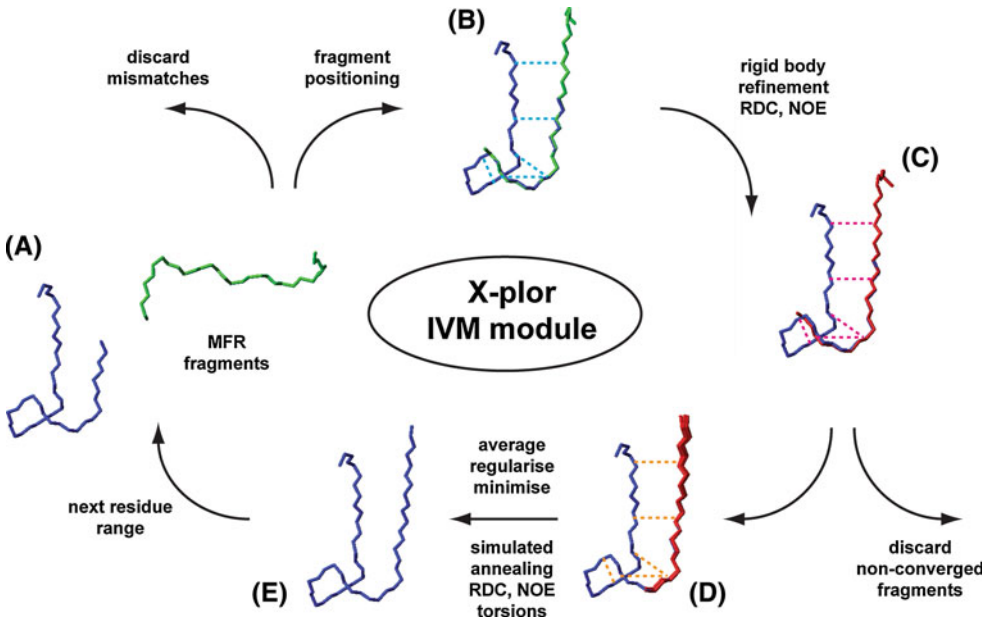

Fig. 2 Flow diagram illustrating the chain extension algorithm based on rigid body docking using the N-terminal domain of murine γS-crystallin as example. PDB fragments (in *green* before, in *red* after rigid body docking) covering residue range Phe 15 to Cys 24 are added to the previously built protein backbone ranging from residue Gly 5 to Asp 21 (in *blue*). Experimental H^N–H^N distance restraints from Ile 7 to Cys 22, Phe 9 to Tyr 20, Glu 11 to Arg 18, Asp12 to Gln 16, Asp 12 to Arg 18, Arg 13 to Phe 15, Arg 13 to Gln 16 and Asn 14 to Gln 16 are shown as *dashed lines*. The figure was created using the program MOLMOL (Koradi et al. 1996). **a** New peptide fragments are added to the already built fraction of the protein backbone by best fitting the coordinates of the overlapping residue range. Mismatches i.e. peptide fragments whose backbone coordinate RMSD is larger than a (predefined) cutoff RMSD are considered incompatible and are discarded at that stage. **b** The remaining fragments, whose coordinates are compatible with the protein backbone, built so far, are placed at the end of the protein backbone and NMR restraints are applied. **c** The peptide fragments are subject to a brief rigid body annealing refinement protocol, which optimizes their orientation with respect to the protein chain using RDCs from one or more alignment tensors and backbone NOEs (when available) as restraints while maintaining coordinate overlap with the protein chain. Only minor repositioning is expected of suitable and converged fragments. Peptide fragments which diverge to an orientation, incompatible with the previously built protein backbone are again discarded. **d** The remaining accepted fragments are averaged and merged with the previously built fraction of the protein chain after discarding their N- and C-termini. **e** The resulting potentially slightly non-physical geometry is regularized by cartesian simulated annealing refinement using RDCs NOEs and backbone torsion angles (derived as consensus backbone dihedrals from the fragment refinement stage) applied as restraints. Then the chain building algorithm moves to the next residue range and is repeated until the C-terminus is reached

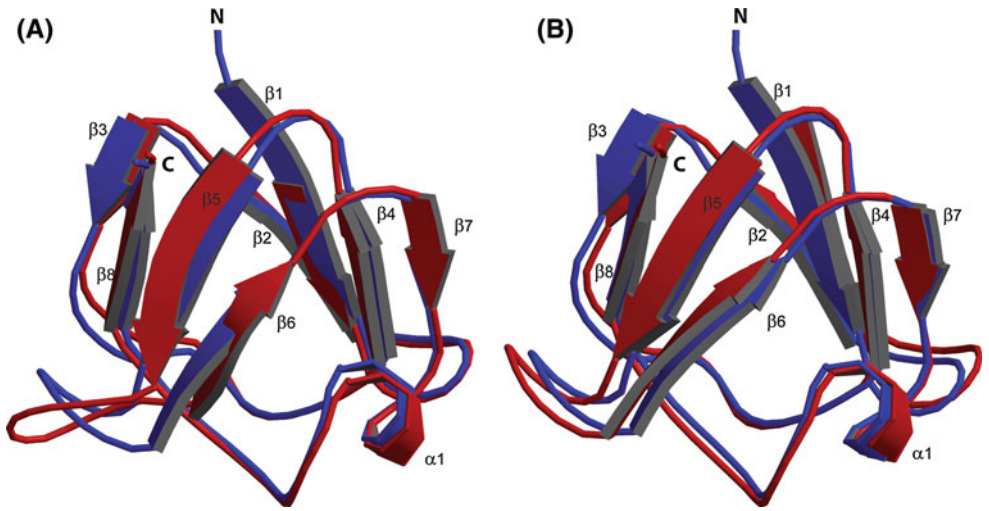

Fig. 3 Protein backbone folds generated for the N-terminal domain of murine γS-crystallin by the modified MFR algorithm compared to reference high-resolution structures. The MFR backbone models are shown in red and the reference structure PDB ID code 1ZWO is shown in blue. Elements of secondary structure are indicated. The renderings were created using MOLSCRIPT (Kraulis 1991) and Raster 3D (Merritt and Murphy 1994; Merritt and Bacon 1997). **a** MFR model generated by searching a database of 900 unique PDB structures and explicitly *excluding* crystallins from the MFR homology searach. Due to consistently wrong torsion angle predictions at residue position Glu 50 the reverse turn connecting strands β_5 and β_6 is distorded. **b** Improved MFR model generated by searching a database of ~5,000 PDB structures *including* crystallins in the MFR homology search otherwise applying exactly the same parameters as in **a**

built protein chain, (Andrec et al. 2001; Kontaxis et al. 2005) or agreement of overlapping (ϕ, ψ) values (Berardi et al. 2011) a 'rigid body docking' algorithm (Clore 2000; Clore and Bewley 2002; Clore and Schwieters 2003) is used for chain extension. A new 7-10-mer peptide fragment is initially placed at the end of the growing protein chain by first best fitting its coordinates to the protein backbone and then its orientation with respect to the rest of the protein is further refined against the RDCs by rigid body simulated annealing using the IVM dynamics engine of Xplor-NIH (Schwieters et al. 2003). Again this process does not last more than a few seconds per PDB fragment. This is repeated for all the fragments of a specific residue range. For suitable fragments a very minor re-positioning in the order of a few tenths of an Å is required and this results in a substantial decrease in dipolar energy.

Fragments that fail to converge to an orientation compatible with the rest of the protein are discarded at that stage. Reasons for that may be steric clashes with the rest of the protein backbone or an unacceptably large resulting coordinate RMSD between the added fragment and the protein backbone. All the remaining accepted fragments for a particular residue range, which now form a tight bundle at the residue position where they are added to the protein chain, are averaged 'on the fly' with the previously built protein backbone, which is subsequently regularized by cartesian simulated annealing using RDCs (and NOEs, where available) as restraints. To avoid divergence, tight dihedral restraints are applied again, which have been derived previously as the consensus angles from the refined PDB fragment library.

Advantages of this approach are that any kind of other restraints, e.g. NOE contacts between the peptide fragment and the protein built so far, can be used at that stage, too. This algorithm can easily accommodate RDC data from several (i.e. more than two) alignment media and restrain the relative orientations of the individual alignment tensors. Any further type of energy term implemented in the energy function of Xplor e.g. anisotropic CS or paramagnetic restraints (Banci et al. 2004), (such as pseudo contact shifts (PCS) or paramagnetic relaxation enhancements (PRE)) can be used in the model building.

The chain building algorithm is then shifted by one residue and the whole procedure is repeated until a terminus is reached. Chain building normally proceeds from the N-terminus towards the C-terminus, but the direction can be also reversed, if desired.

If over a longer stretch of residues no suitable fragments are found, which fit onto the growing protein chain, then the chain building is aborted and restarted at the residue position immediately following the last accepted peptide fragment. Thus a number of larger pieces with or without overlapping residue ranges are generated which, if necessary, can be docked together in a second pass, using the same algorithm (see below) if the pieces overlap or, if this is not the case, joined by gap filling as described in (Berardi et al. 2011). The final model is then subject to one further round of refinement using a protocol similar to the

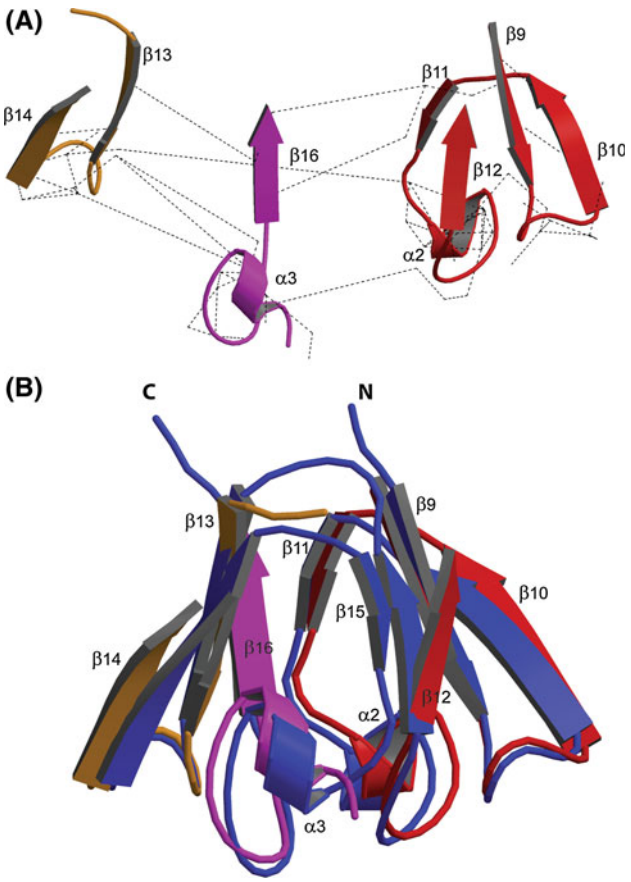

Fig. 4 **a**. Due to exchange broadening a number of backbone N–H^N resonances of the C-terminal domain γS-crystallin of cannot be detected (Lys 130, Val 131, Thr 135, Trp 136 and Asp 152 to Tyr 156) therefore the protein backbone cannot be built in one pass. With the chain building method described above three large pieces of the C-terminal of γS-crystallin can be obtained: residue ranges: Lys 94 to Val 131 (*red*), Glu 133 to Leu 151 (*orange*) and Arg 157 to Arg 174 (*magenta*). The fragments can be docked to form a protein backbone fold using experimental backbone H^N–H^N NOEs (shown as *dashed lines*) as restraints to position them relative to each other. **b** C-terminal domain of murine γS-crystallin. The three parts of the MFR backbone model are shown in red/orange/magenta respectively and the reference structure PDB ID code 1ZWO is shown in *blue*. The rendering was created using MOLSCRIPT (Kraulis 1991) and Raster 3D (Merritt and Murphy 1994; Merritt and Bacon 1997)

one used for regularization, but additionally including a collapse energy term, which restrains the radius of gyration (Clore et al. 1999) as well as a Ramachandran database derived potential of mean force (Kuszewski et al. 1996, 1997; Kuszewski and Clore 2000), which biases sidechain orientations to those most commonly found in the PDB. The backbone RMSD of the ensemble of structures obtained by the final round of simulated annealing allows an estimate of the coordinate precision of the MFR model.

The algorithm described here is, in principle, capable of operating in the complete absence of NOE restraints although stability is greatly improved by just a modest number of backbone (H^N–H^N and H^α–H^N) NOE contacts, which can be obtained easily experimentally from ^{15}N edited NOE experiments, and do not require time consuming sidechain experiments. Of course sidechain contacts, when available, can be incorporated, too, when available, e.g. in specifically methyl group labeled (Val, Leu, Ile $^{13}C^{\gamma/\delta}H_3$) but otherwise per-deuterated proteins.

Results and discussion

The algorithm described above was applied to a number of small globular proteins for which multiple almost complete published sets of RDC data exist in the literature. Backbone structures were calculated using the algorithm described above for each protein using different sets of input data and compared to already known X-ray crystal structures or high-resolution NMR solution structures. These proteins were: Ubiquitin (PDB ID code 1UBQ, 76 residues), DinI (PDB ID code 1GHH, 81 residues) and murine γS-crystallin (PDB ID code 1ZWO, 177 residues). To demonstrate that this algorithm is capable of generating novel protein folds not present in the search database the reference structure(s) and homologues were excluded from the MFR search. Homologous structures were identified by searching the MFR database using the backbone RMSDs to the reference structure(s) as the sole criterion. The proteins found there were excluded from further consideration. (For multi-domain proteins such as γS-crystallin this was done separately for each domain to account for potential insertions between domains.)

Various input factors were studied in silico. The number of different alignment tensors used in the structure calculations was varied to see how much of an improvement one can expect from using two independent sets of RDCs over just one. Then the number of dipolar coupling restraints per residue was deliberately reduced to investigate if the RDCs that can be obtained on a fully deuterated (except for the amide protons) sample are sufficient for definition of a unique backbone fold. Whereas up to six RDCs per residue ($^1D_{NHN}$, $^1D_{NC'}$, $^2D_{C'HN}$, $^1D_{C'C\alpha}$, $^1D_{C\alpha C\beta}$, $^1D_{C\alpha H\alpha}$) can be measured with a protonated protein sample, sometimes only as little as three couplings per residue ($^1D_{NHN}$, $^1D_{NC'}$, $^2D_{C'HN}$) can be obtained on a fast relaxing per-deuterated protein. Inclusion or exclusion of backbone NOE distance restraints (H^N–H^N NOEs for perdeuterated proteins and additionally H^α–H^N NOEs for protonated preparations) can make a substantial difference in such cases. The effect of reversing the direction of the chain building (from the C- to the N-terminus) was found to be insignificant. The resulting models were found to differ by only a few tenths of Å in backbone RMSD. All different combinations of input parameters with their results are summarized in Table 1.

Table 1 Results of MFR based structure calculations with different sets of input data

Protein	No. of tensors	Types of RDCs	NOEs	bb RMSD Å to ref./to mean
Ubq	A, B	$N–H^N$, $C^\alpha–H^\alpha$, $C'–N$, C'-H^N, C'-C^α	$H^N–H^N$	0.60/0.18 (0.70)[a]
Ubq	A, B	$N–H^N$, $C^\alpha–H^\alpha$, $C'–N$, $C'–H^N$, $C'–C^\alpha$		0.70/0.35 (0.70)[a]
Ubq	A, B	$N–H^N$, $C'–N$, $C'–H^N$	$H^N–H^N$	0.64/0.31
Ubq	A, B	$N–H^N$, $C'–N$, $C'–H^N$		0.76/1.05
Ubq	A	$N–H^N$, $C^\alpha–H^\alpha$, $C'–N$, $C'–H^N$, $C'–C^\alpha$	$H^N–H^N$	0.62/0.20 (0.81)[a]
Ubq	A	$N–H^N$, $C^\alpha–H^\alpha$, $C'–N$, $C'–H^N$, $C'–C^\alpha$		0.82/0.43 (0.81)[a]
Ubq	A	$N–H^N$, $C'–N$, $C'–H^N$	$H^N–H^N$	0.79/0.30
Ubq	B	$N–H^N$, $C^\alpha–H^\alpha$, $C'–N$, $C'–H^N$, $C'–C^\alpha$	$H^N–H^N$	0.81/0.28
Ubq	B	$N–H^N$, $C^\alpha–H^\alpha$, $C'–N$, $C'–H^N$, $C'–C^\alpha$		0.80/1.24
Ubq	B	$N–H^N$, $C'–N$, $C'–H^N$	$H^N–H^N$	0.79/0.63
DinI	A, B	$N–H^N$, $C^\alpha–H^\alpha$, $C'–N$, ($C'–H^N$), $C'–C^\alpha$	$H^N–H^N$, $H^\alpha–H^N$	1.42/0.36 (1.51)[a]
DinI	A, B	$N–H^N$, $C^\alpha–H^\alpha$, $C'–N$, ($C'–H^N$), $C'–C^\alpha$		1.58/0.86 (1.51)[a]
DinI	A, B	$N–H^N$, $C'–N$, ($C'–H^N$), $C'–C^\alpha$	$H^N–H^N$	2.07/0.38
DinI	A, B	$N–H^N$, $C'–N$, ($C'–H^N$), $C'–C^\alpha$		2.61/0.87
DinI	A	$N–H^N$, $C^\alpha–H^\alpha$, $C'–N$, ($C'–H^N$), $C'–C^\alpha$	$H^N–H^N$, $H^\alpha–H^N$	1.72/0.55
DinI	A	$N–H^N$, $C'–N$, ($C'–H^N$), $C'–C^\alpha$	$H^N–H^N$	1.86/0.80
γS-Crys N-term.	A, B	$N–H^N$, $C'–N$, $C'–C^\alpha$, $C^\alpha–C^\beta$	$H^N–H^N$	2.10/0.48
γS-Crys C-term.	A, B	$N–H^N$, $C'–N$, $C'–C^\alpha$, $C^\alpha–C^\beta$	$H^N–H^N$	3.10/1.88
γS-Crys[b] N-term.	A, B	$N–H^N$, $C'–N$, $C'–C^\alpha$, $C^\alpha–C^\beta$	$H^N–H^N$	1.26/0.32
γS-Crys[b] C-term.	A, B	$N–H^N$, $C'–N$, $C'–C^\alpha$, $C^\alpha–C^\beta$	$H^N–H^N$	1.74/0.87

Ubiquitin: 37 long-range $H^N–H^N$ NOEs, alignment medium A: positively doped bicelles (DMPC:DHPC:CTAB = 15:5:1; 5 % w/v), 67 $^1D_{NHN}$, 66 $^1D_{C\alpha H\alpha}$, 54 $^1D_{C'C\alpha}$, 69 $^1D_{NC'}$, 69 $^2D_{HNC'}$; alignment medium B: undoped bicelles (DMPC:DHPC:CTAB = 3:1:0; 5 % w/v), 69 $^1D_{NHN}$, 72 $^1D_{C\alpha H\alpha}$, 58 $^1D_{C'C\alpha}$, 67 $^1D_{NC'}$, 67 $^2D_{HNC'}$

DinI: 70 (58 sequential, 12 mid- to long-range) $H^N–H^N$ NOEs, 311 (148 intra-residue, 122 sequential, 27 mid-range, 14 long-range) $H^\alpha–H^N$ NOEs, alignment Medium A: positively doped bicelles (DTPC:DHPC:CTAB = 30:10:1; 5 % w/v), 69 $^1D_{NHN}$, 70 $^1D_{C\alpha H\alpha}$, 69 $^1D_{C'C\alpha}$, 61 $^1D_{NC'}$, 64 $^2D_{HNC'}$; alignment Tensor B: 8 mg/ml Pf_1 phage 58 $^1D_{NHN}$, 65 $^1D_{C\alpha H\alpha}$, 69 $^1D_{C'C\alpha}$, 33 $^1D_{NC'}$

γS-crystallin: 177 (95 sequential, 29 mid-range, 53 long-range) $H^N–H^N$ NOEs; alignment medium A: stretched polyamide gel 6 %, 147 $^1D_{NHN}$, 153 $^1D_{C'C\alpha}$, 135 $^1D_{C\alpha C\beta}$, 139 $^1D_{NC'}$; alignment medium B: gelled Pf_1 (3 mg/ml) 144 $^1D_{NHN}$, 150 $^1D_{C'C\alpha}$, 111 $^1D_{C\alpha C\beta}$, 134 $^1D_{NC'}$

[a] (in brackets): values obtained using the original MFR assembly procedure (Kontaxis et al. 2005)

[b] using a larger search database of ~5,000 PDB entries *including* crystallins

In the case of Ubiquitin very precise input data are available. Therefore, when using two complete sets of RDCs as input for the structure generation algorithm described above, non-surprisingly, a very accurate structure was generated with a quality comparable, or slightly better than previously published methods as judged by the backbone rmsd to the reference structure (PDB ID code 1UBQ). There was also only very little room for improvement when a limited number of $H^N–H^N$ NOE restraints (Giesen et al. 2003) were included in the model building. With only three RDCs per residue ($^1D_{NHN}$, $^1D_{NC'}$, $^2D_{C'HN}$) for each of the two alignment conditions the impact of the additional NOEs became noticeable.

When RDC data from only one alignment medium were used as input still fairly accurate backbone folds were achieved as long as all five types of RDCs were used as restraints. Slightly better models were produced using RDC restraints from 'tensor A' (positively charged bicelles), which can be rationalized by the fact that this alignment tensor is more rhombic than 'tensor B' (neutral bicelles). Again the limited set of NOEs proved beneficial but not essential.

The advantages of the improved fragment assembly algorithm became apparent, when no more than three RDCs per residue from one alignment medium were available. With the now essential support of the $H^N–H^N$ NOEs a backbone fold could be determined for each of the two alignment conditions. Previous implementations of MFR have not been able to produce a backbone fold with this set of input data.

The protein DinI (Ramirez et al. 2000) was already a more challenging test case. Due to a slight tendency for aggregation, NMR data collection was conducted at lower concentration resulting in less precise sets of RDCs. Additionally, RDCs acquired in alignment medium 'B' (filamentous phage) suffer from broadening in

the proton dimension due to strong alignment and are therefore less complete. Again, using the full set of restraints available as input, one could obtain correct backbone folds for DinI, which were substantially closer to the reference structure (PDB ID code 1GHH) than those of previous implementations of MFR model building. A problematic region was the loop from residue Lys 9 to Gly 16 connecting strand β_1 and helix α_1, for which, due to the lower density of RDCs (it contains two Proline residues: Pro 12 and 14), the MFR search did not converge towards a unique structure and did not find any structurally close matches among the best 20 hits. Therefore the best 100 fragments were retained for that residue range. The inclusion of backbone to backbone NOEs (H^N–H^N, H^α–H^N) better brought β-strands β_{1-3} into register and thus corrected structural errors in the loop region mentioned above. When the number of RDCs was reduced by excluding the $^1D_{C\alpha H\alpha}$ for both alignment conditions to mimic protein deuteration a correct backbone fold could still be obtained and the H^N–H^N NOEs further improved this fold considerably.

Using only data from one alignment medium as restraints, in the complete absence of NOEs neither set of RDCs was able to produce a correct protein fold. Indeed, problems arose reproducing the precise conformation of the loop connecting strand β_1 and helix α_1 (see above) and the loop connecting strand β_3 and helix α_2, which contains a Glycine (Gly 54), with a H^N signal broadened beyond detection, The more complete set of RDCs acquired under alignment conditions A (doped bicelles) yielded a correct protein fold in both a 'protonated' and a 'deuterated' scenario, when supplemented with backbone NOEs. RDCs from the other alignment medium B (phage) were found not to be sufficient to define the protein fold, because the relative orientation of the two helices α_1 and α_2 could not uniquely be established with respect to the sheet β_{1-3}. In that case backbone only NOEs could not provide any improvement, either, since the position of the helices are not restrained by backbone NOEs only.

Murine γS-crystallin (Wu et al. 2005) comprises two (almost) entirely β-stranded domains consisting of two Greek key motifs each, but this a priori information was neither required nor used in the chain building process. As the linker connecting them is partly flexible and disordered from residue Ser 88 to Ala 93 resulting in a systematic reduction of the observed RDCs due to dynamics, it could not entirely be built by the MFR algorithm. With the structure of the linker undetermined and no backbone NOE contacts between them the two domains were obtained separately. Due to its relatively complicated topology, previous implementations of MFR fragment assembly have consistently failed to reproduce the correct protein topology in the complete absence of NOE constraints. Including just the H^N–H^N NOEs was necessary and sufficient to keep the β-strands in register and maintain the correct topology. Thus high resolution models for the two individual domains could be obtained with much less complete sets of RDCs than the proteins previously solved. In the case of the N-terminal domain the application of the new improved MFR build algorithm was straightforward, and the whole backbone could be built in an automated way from Lys 6 to Ser 88 (Fig. 3). In contrast, for the C-terminal domain the situation was complicated due to two stretches of 'missing' residues, Lys 130, Val 131, Thr 135, Trp 136 and Asp 152 to Tyr 156, whose amide resonances were broadened beyond detection due to conformational exchange (Wu et al. 2005). Therefore, the backbone could only be built in three separate and non-overlapping pieces (Lys 94 to Val 131, Glu 133 to Leu 151 and Arg 157 to Arg 174) and some form of user intervention was necessary to correctly assemble them (Fig. 4). First, the orientational degrees of freedom relative to each other were fixed by aligning them with the principal axis system of one alignment tensor according to (Losonczi et al. 1999). Remaining ambiguities were resolved by cross-validation with the other alignment tensor (Al-Hashimi et al. 2000). Second, the translational degrees of freedom were resolved by docking these fragments using the available set of H^N–H^N NOEs and applying a brief simulated annealing protocol. Nevertheless, as a consequence the backbone RMSD for the C-terminal domain remained somewhat worse than for the N-terminal domain.

In the case of γS-crystallin the whole structure determination algorithm was performed twice: In one case the reference structure(s) was deliberately removed from the PDB database used in the MFR search to prove that the algorithm was capable of creating the correct fold de novo. Under this condition, an unusual kink at residue Glu 50 in the N-terminal domain, which seems to be structural feature specific to crystallins (Wu et al. 2005), could not be correctly reproduced, resulting in a somewhat distorted geometry of this loop. This problem was also evident at the equivalent position Glu 140 of the C-terminal domain.

When homologous protein structures of other crystallines were deliberately included in the MFR search they were systematically found among the best ranking fragments. Then the algorithm had no problems reproducing the unusual stretches mentioned above. This improved the backbone RMSD relative to the reference structure, whose structure was already known (PDB ID code 1ZWO) for both domain by roughly one Å, proving that the precision of this method is essentially limited by the quality and completeness of the database used in the search phase.

Conclusion

The presented MFR-based method for protein backbone fold determination requires only backbone CS and RDCs and no or only a minimal number of backbone NOE restraints. More NOEs between sidechains could in principle be integrated in the structure generation algorithm, too. The new method is more robust than similar previously published algorithms with respect to the required completeness of RDC data sets. It is better able to cope with fewer RDCs per residue and to cope with stretches of residues with missing RDC data. To some extent, it is also able to correct structural errors and mistakes. Compared to ROSETTA, computational requirements are less demanding and implementation of any other type of energy term, including simultaneous use of multiple sets of RDCs, is straight forward. A potential future application is a hybrid approach, in which MFR is used to create an initial model, which can be further optimized using ROSETTA without the need to search large fragment libraries.

Software availability

Two large scripts are included as supplementary material:

(1) mfrBuild.tcl which is a.tcl script for building a protein backbone from MFR-derived PDB fragments, using rigid body docking for fragment assembly
(2) docking.inp which is an X-plor input script for docking a PDB fragment with the protein chain using RDCs and backbone NOEs as restraints.

The complete distribution of scripts (including sample input data) is available from the author upon request.

References

Al-Hashimi HM, Valafar H, Terrell M, Zartler ER, Eidsness MK, Prestegard JH (2000) Variation of molecular alignment as a means of resolving orientational ambiguities in protein structures from dipolar couplings. J Magn Reson 143(2):402–406

Andrec M, Du P, Levy RM (2001) Protein backbone structure determination using only residual dipolar couplings from one ordering medium. J Biomol NMR 21(4):335–347

Banci L, Bertini I, Cavallaro G, Giachetti A, Luchinat C, Parigi G (2004) Paramagnetism-based restraints for Xplor-NIH. J Biomol NMR 28(3):249–261

Bax A, Cornilescu G, Hu JS (1999) Identification of the hydrogen bonding network in a protein by scalar couplings. J Am Chem Soc 121(12):2949–2950

Bax A, Delaglio F, Kontaxis G (2000) Protein structure determination using molecular fragment replacement and NMR dipolar couplings. J Am Chem Soc 122(9):2142–2143

Bax A, Kontaxis G, Tjandra N (2001) Dipolar couplings in macromolecular structure determination. Methods Enzymol 339:127–174

Berardi MJ, Shih WM, Harrison SC, Chou JJ (2011) Mitochondrial uncoupling protein 2 structure determined by NMR molecular fragment searching. Nature 476(7358):109–113

Bouvignies G, Markwick P, Bruschweiler R, Blackledge M (2006a) Simultaneous determination of protein backbone structure and dynamics from residual dipolar couplings. J Am Chem Soc 128(47):15100–15101

Bouvignies G, Meier S, Grzesiek S, Blackledge M (2006b) Ultrahigh-resolution backbone structure of perdeuterated protein GB1 using residual dipolar couplings from two alignment media. Angew Chem Int Ed Engl 45(48):8166–8169

Bouvignies G, Markwick PR, Blackledge M (2007) Simultaneous definition of high resolution protein structure and backbone conformational dynamics using NMR residual dipolar couplings. ChemPhysChem 8(13):1901–1909

Bowers PM, Strauss CE, Baker D (2000) De novo protein structure determination using sparse NMR data. J Biomol NMR 18(4):311–318

Cavalli A, Salvatella X, Dobson CM, Vendruscolo M (2007) Protein structure determination from NMR chemical shifts. Proc Natl Acad Sci USA 104(23):9615–9620

Chou JJ, Delaglio F, Bax A (2000a) Measurement of one-bond 15 N–13C' dipolar couplings in medium sized proteins. J Biomol NMR 18(2):101–105

Chou JJ, Li S, Bax A (2000b) Study of conformational rearrangement and refinement of structural homology models by the use of heteronuclear dipolar couplings. J Biomol NMR 18(3):217–227

Clore GM (2000) Accurate and rapid docking of protein-protein complexes on the basis of intermolecular nuclear overhauser enhancement data and dipolar couplings by rigid body minimization. Proc Natl Acad Sci USA 97(16):9021–9025

Clore GM, Bewley CA (2002) Using conjoined rigid body/torsion angle simulated annealing to determine the relative orientation of covalently linked protein domains from dipolar couplings. J Magn Reson 154(2):329–335

Clore GM, Schwieters CD (2003) Docking of protein-protein complexes on the basis of highly ambiguous intermolecular distance restraints derived from 1H/15 N chemical shift mapping and backbone 15 N-1H residual dipolar couplings using conjoined rigid body/torsion angle dynamics. J Am Chem Soc 125(10):2902–2912

Clore GM, Kuszewski J, Gronenborn AM (1999) Improving the packing and accuracy of NMR structures with a pseudopotential for the radius of gyration. J Am Chem Soc 121(10):2337–2338

Cordier F, Grzesiek S (1999) Direct observation of hydrogen bonds in proteins by interresidue (3 h) J (NC ') scalar couplings. J Am Chem Soc 121(7):1601–1602

Delaglio F, Grzesiek S, Vuister GW, Zhu G, Pfeifer J, Bax A (1995) NMRPipe: a multidimensional spectral processing system based on UNIX pipes. J Biomol NMR 6(3):277–293

Giesen AW, Homans SW, Brown JM (2003) Determination of protein global folds using backbone residual dipolar coupling and long-range NOE restraints. J Biomol NMR 25(1):63–71

Hus JC, Marion D, Blackledge M (2001) Determination of protein backbone structure using only residual dipolar couplings. J Am Chem Soc 123(7):1541–1542

Jaroniec CP, Ulmer TS, Bax A (2004) Quantitative J correlation methods for the accurate measurement of 13C'-13Calpha dipolar couplings in proteins. J Biomol NMR 30(2):181–194

Kontaxis G, Delaglio F, Bax A (2005) Molecular fragment replacement approach to protein structure determination by chemical shift and dipolar homology database mining. Methods Enzymol 394:42–78

Koradi R, Billeter M, Wuthrich K (1996) MOLMOL: a program for display and analysis of macromolecular structures. J Mol Graph 14(1):51–55 29–32

Kraulis PJ (1991) Molscript: a program to produce both detailed and schematic plots of protein structures. J Appl Crystallogr 24:946–950

Kuszewski J, Clore GM (2000) Sources of and solutions to problems in the refinement of protein NMR structures against torsion angle potentials of mean force. J Magn Reson 146(2):249–254

Kuszewski J, Gronenborn AM, Clore GM (1996) Improving the quality of NMR and crystallographic protein structures by means of a conformational database potential derived from structure databases. Protein Sci 5(6):1067–1080

Kuszewski J, Gronenborn AM, Clore GM (1997) Improvements and extensions in the conformational database potential for the refinement of NMR and X-ray structures of proteins and nucleic acids. J Magn Reson 125(1):171–177

Leaver-Fay A, Tyka M, Lewis SM, Lange OF, Thompson J, Jacak R, Kaufman K, Renfrew PD, Smith CA, Sheffler W, Davis IW, Cooper S, Treuille A, Mandell DJ, Richter F, Ban YE, Fleishman SJ, Corn JE, Kim DE, Lyskov S, Berrondo M, Mentzer S, Popovic Z, Havranek JJ, Karanicolas J, Das R, Meiler J, Kortemme T, Gray JJ, Kuhlman B, Baker D, Bradley P (2011) ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. Methods Enzymol 487:545–574

Losonczi JA, Andrec M, Fischer MW, Prestegard JH (1999) Order matrix analysis of residual dipolar couplings using singular value decomposition. J Magn Reson 138(2):334–342

Meiler J, Baker D (2003) Rapid protein fold determination using unassigned NMR data. Proc Natl Acad Sci USA 100(26): 15404–15409

Merritt EA, Bacon DJ (1997) Raster3D: photorealistic molecular graphics. Methods Enzymol 277:505–524

Merritt EA, Murphy ME (1994) Raster3D Version 2. 0. A program for photorealistic molecular graphics. Acta Crystallogr D Biol Crystallogr 50(Pt 6):869–873

Ottiger M, Delaglio F, Bax A (1998a) Measurement of J and dipolar couplings from simplified two-dimensional NMR spectra. J Magn Reson 131(2):373–378

Ottiger M, Delaglio F, Marquardt JL, Tjandra N, Bax A (1998b) Measurement of dipolar couplings for methylene and methyl sites in weakly oriented macromolecules and their use in structure determination. J Magn Reson 134(2):365–369

Raman S, Lange OF, Rossi P, Tyka M, Wang X, Aramini J, Liu G, Ramelot TA, Eletsky A, Szyperski T, Kennedy MA, Prestegard J, Montelione GT, Baker D (2010) NMR structure determination for larger proteins using backbone-only data. Science 327(5968):1014–1018

Ramirez BE, Voloshin ON, Camerini-Otero RD, Bax A (2000) Solution structure of DinI provides insight into its mode of RecA inactivation. Protein Sci 9(11):2161–2169

Rasia RM, Lescop E, Palatnik JF, Boisbouvier J, Brutscher B (2011) Rapid measurement of residual dipolar couplings for fast fold elucidation of proteins. J Biomol NMR 51(3):369–378

Rohl CA, Baker D (2002) De novo determination of protein backbone structure from residual dipolar couplings using Rosetta. J Am Chem Soc 124(11):2723–2729

Rohl CA, Strauss CE, Misura KM, Baker D (2004) Protein structure prediction using Rosetta. Methods Enzymol 383:66–93

Sass J, Cordier F, Hoffmann A, Cousin A, Omichinski JG, Lowen H, Grzesiek S (1999) Purple membrane induced alignment of biological macromolecules in the magnetic field. J Am Chem Soc 121(10):2047–2055

Schwieters CD, Kuszewski JJ, Tjandra N, Clore GM (2003) The Xplor-NIH NMR molecular structure determination package. J Magn Reson 160(1):65–73

Shen Y, Lange O, Delaglio F, Rossi P, Aramini JM, Liu G, Eletsky A, Wu Y, Singarapu KK, Lemak A, Ignatchenko A, Arrowsmith CH, Szyperski T, Montelione GT, Baker D, Bax A (2008) Consistent blind protein structure generation from NMR chemical shift data. Proc Natl Acad Sci USA 105(12):4685–4690

Shen Y, Vernon R, Baker D, Bax A (2009) De novo protein structure generation from incomplete chemical shift assignments. J Biomol NMR 43(2):63–78

Shen Y, Bryan PN, He Y, Orban J, Baker D, Bax A (2010) De novo structure generation using chemical shifts for proteins with high-sequence identity but different folds. Protein Sci 19(2):349–356

Simons KT, Bonneau R, Ruczinski I, Baker D (1999) Ab initio protein structure prediction of CASP III targets using ROSETTA. Proteins Suppl 3:171–176

Walsh JD, Wang YX (2005) Periodicity, planarity, residual dipolar coupling, and structures. J Magn Reson 174(1):152–162

Walsh JD, Kuszweski J, Wang YX (2005) Determining a helical protein structure using peptide pixels. J Magn Reson 177(1):155–159

Wang J, Walsh JD, Kuszewski J, Wang YX (2007) Periodicity, planarity, and pixel (3P): a program using the intrinsic residual dipolar coupling periodicity-to-peptide plane correlation and phi/psi angles to derive protein backbone structures. J Magn Reson 189(1):90–103

Wu Z, Delaglio F, Wyatt K, Wistow G, Bax A (2005) Solution structure of (gamma)S-crystallin by molecular fragment replacement NMR. Protein Sci 14(12):3101–3114