J Biomol NMR (2009) 43:63–78
DOI 10.1007/s10858-008-9288-5

ARTICLE

De novo protein structure generation from incomplete chemical shift assignments

Yang Shen · Robert Vernon · David Baker · Ad Bax

Received: 20 August 2008 / Accepted: 28 October 2008 / Published online: 26 November 2008
© US Government 2008

Abstract NMR chemical shifts provide important local structural information for proteins. Consistent structure generation from NMR chemical shift data has recently become feasible for proteins with sizes of up to 130 residues, and such structures are of a quality comparable to those obtained with the standard NMR protocol. This study investigates the influence of the completeness of chemical shift assignments on structures generated from chemical shifts. The Chemical-Shift-Rosetta (CS-Rosetta) protocol was used for de novo protein structure generation with various degrees of completeness of the chemical shift assignment, simulated by omission of entries in the experimental chemical shift data previously used for the initial demonstration of the CS-Rosetta approach. In addition, a new CS-Rosetta protocol is described that improves robustness of the method for proteins with missing or erroneous NMR chemical shift input data. This strategy, which uses traditional Rosetta for pre-filtering of the fragment selection process, is demonstrated for two paramagnetic proteins and also for two proteins with solid-state NMR chemical shift assignments.

Electronic supplementary material The online version of this article (doi:10.1007/s10858-008-9288-5) contains supplementary material, which is available to authorized users.

Y. Shen · A. Bax (✉)
Laboratory of Chemical Physics, National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Bethesda, MD 20892-0520, USA
e-mail: bax@nih.gov

R. Vernon · D. Baker
Department of Biochemistry and Howard Hughes Medical Institute, University of Washington, Seattle, WA 98195, USA

Keywords NMR chemical shift · Protein structure prediction · Solid-state NMR structure determination · Paramagnetic protein · CS-Rosetta

Chemical shifts are key to protein NMR spectroscopy not only because they allow separate observation of each ^{1}H, ^{13}C, and ^{15}N nucleus in the molecule, but also as they carry important information on the local conformation (Saito 1986; Spera and Bax 1991; Williamson and Asakura 1993; Williamson et al. 1995; Asakura et al. 1997; Ando et al. 1998; Cornilescu et al. 1999; Castellani et al. 2003; Neal et al. 2006), including secondary structure (Wishart et al. 1991), hydrogen bonding (Wagner et al. 1983; Shen and Bax 2007) and the position and orientation of aromatic rings (Haigh and Mallion 1979; Case 1995). Protein structural information derived from chemical shifts, such as the backbone φ/ψ torsion angles predicted by the program TALOS (Cornilescu et al. 1999), is widely used in NMR structure determination, but almost invariably as a complement to conventional NOE distance restraints or to internuclear distance restraints obtained by solid-state NMR. Recently, several computational approaches have been developed to use the NMR chemical shifts alone as input for protein structure generation (Cavalli et al. 2007; Gong et al. 2007; Shen et al. 2008; Wishart et al. 2008). These approaches, represented by CHESHIRE (Cavalli et al. 2007), CS-Rosetta (Shen et al. 2008) and CS23D (Wishart et al. 2008), match the experimental chemical shifts of the backbone and ^{13}C$^{\beta}$ atoms, which are commonly available at the early stage of the conventional NMR structure determination procedure, to a structural database to identify protein fragments with similar chemical shifts. Because the structural database of proteins for which actual

NMR assignments are available remains relatively small, empirical relationships (Cornilescu et al. 1999; Neal et al. 2003; Kontaxis et al. 2005; Shen and Bax 2007) are commonly used to "assign" chemical shift values to nuclei in proteins of known structure. Selected protein fragments are then used as input for a fragment assembly procedure, which also aims to optimize empirical energy terms related to hydrogen bonding, hydrophobic packing, etc., to generate an all-atom protein structure. These approaches have been evaluated for over two dozen proteins with sizes of up to 15 kD and a wide variety of folds. For the vast majority, convergence is obtained, which then invariably yields all-atom protein models that compare well with experimental structures, with root-mean-square deviations (rmsd's) from the conventionally determined reference structure in the 0.7–2 Å range for the backbone atoms, and ~1.4–3 Å when considering all atoms. Structures generated by the CS-Rosetta procedure for nine structural genomics target proteins, prior to completion of the conventional NMR structure determination process (Shen et al. 2008), prove the procedure to be a viable alternative for small to medium-size proteins (Gryk and Hoch 2008).

To date, the chemical shift based structure determination methods have been evaluated for proteins with complete or nearly complete NMR chemical shift assignments. In practice, however, resonance assignments are often incomplete, and also may contain a small fraction of erroneous assignments. Often, a completeness of >80–90% of the backbone sequence-specific assignments makes it possible to obtain a sufficient number of side-chain resonance and NOE assignments for deriving a dense network of distance restraints, needed for the conventional NMR structure determination procedure. The present study investigates the impact of incomplete chemical shift assignments on the NMR chemical shift based CS-Rosetta protocol by using chemical shift assignments with various degrees of completeness or correctness, simulated by omission and/or modification of entries in the experimental chemical shift data. For cases where a substantial fraction of the chemical shifts is missing or in error and the standard fragment CS-Rosetta protocol is found to fail, a more robust hybrid fragment selection method is described which largely resolves this limitation.

In recent years, several viable routes to resonance assignment and structure determination of small globular proteins by solid-state NMR (ssNMR) have been demonstrated (Castellani et al. 2002; Igumenova et al. 2004; Siemer et al. 2005; Zech et al. 2005; Nadaud et al. 2007; Loquet et al. 2008; Manolikas et al. 2008), relying mostly on ^{13}C–^{13}C, ^{15}N–^{13}C, and/or indirectly measured ^{1}H–^{1}H distance restraints. Chemical shift assignments of ssNMR spectra typically are obtained by sophisticated two- and three-dimensional ^{13}C-detected analogs of the widely used triple resonance J-connectivity experiments used in solution NMR. However, with few exceptions (Agarwal et al. 2006; Chevelkov et al. 2006), ^{1}H resonance assignments are usually not determined when studying a protein structure by these methods. For a variety of technical reasons, spectral resolution obtained for small proteins by ssNMR is often lower than what can be obtained for such proteins in solution (Tycko 1996), resulting in increased signal overlap and a considerable fraction of missing resonance assignments. For cases where protein structures have been determined both by solution and by solid-state NMR methods, results are generally quite similar (Manolikas et al. 2008), and chemical shifts observed in the solid state generally agree well with those seen in solution (Igumenova et al. 2004; Zech et al. 2005). On the other hand, exceptions are often seen for residues involved in intermolecular contacts, i.e., surface-exposed residues, reflecting the different protein sample conditions. It is therefore interesting to evaluate to what extent the CS-Rosetta approach is applicable to proteins whose chemical shifts have been determined by solid state NMR. Indeed, as we demonstrate for two small proteins, ubiquitin and GB3, CS-Rosetta yields good structural models when using solely the ssNMR chemical shift assignments as input.

A second challenging area, where often a considerable fraction of chemical shift assignments are missing, concerns paramagnetic metalloproteins. About 25% of all proteins in living systems contain metal ions (Andreini et al. 2004) and in many of these cases the metal is paramagnetic ($Fe^{2+/3+}$, Cu^{2+}, Co^{2+}, Ni^{3+}), where the presence of unpaired electrons causes very rapid transverse relaxation for nearby nuclei, interfering with use of the standard ^{1}H-detected triple resonance assignment strategy (Ikura et al. 1990; Montelione and Wagner 1990). Although ^{13}C-detected experiments can yield relief (Bertini et al. 2005; Bermel et al. 2006), collection of ^{1}H–^{1}H NOE restraints remains problematic in the vicinity of paramagnetic centers. The degree of paramagnetic broadening scales with the inverse sixth power of the distance to the metal, resulting in a sphere with radius of ca. 5–15 Å around the metal where assignments are missing. In addition, if protons are observed and assigned, they may contain paramagnetic pseudo-contact contributions to their chemical shifts, which are not easily accounted for in the absence of a known structure, and therefore can impact the molecular fragment search of the CS-Rosetta protocol in a similar manner as assignment errors. We will show, however, that the hybrid CS-Rosetta protocol is quite tolerant to these problems, and demonstrate its application to two small paramagnetic proteins of known structure.

Methods and materials

In this work, the original complete experimental chemical shift assignments, including δ^{15}N, δ^{13}C$'$, δ^{13}C$^{\alpha}$, δ^{13}C$^{\beta}$, δ^{1}H$^{\alpha}$ and δ^{1}H^{N}, for proteins MrR16 (90 residues; PDB code: 1YWX; 514 available chemical shifts from BMRB #6799) and TM1442 (110 residues; PDB code: 1SBO; 647 available chemical shifts from BMRB #5921) are used. The entries of the chemical shift assignments of each protein are regrouped and/or modified to create new datasets that simulate the chemical shift inputs with various degrees of completeness and/or chemical shift errors. The CS-Rosetta protein structure generation protocol is carried out for these differently prepared chemical shift input data sets, but following exactly the same computational procedures. The impact of the incompleteness and/or incorrectness of chemical shift assignments on the CS-Rosetta procedure are evaluated both by monitoring the accuracy of the selected fragments and by the quality and convergence of the generated CS-Rosetta all-atom models.

Preparation of chemical shift datasets

Three groups of incomplete or partially erroneous chemical shift assignments were generated using the original (nearly complete) experimental chemical shift assignments of proteins MrR16 and TM1442. Details regarding the assignments of the two paramagnetic proteins, and the proteins studied by solid-state NMR, are also provided below.

Simulated datasets with missing chemical shifts assignments for certain types of nuclei

Depending on the strategy used for backbone resonance assignment, chemical shift assignments for certain types of backbone may not be available. Table 1 lists the chemical shift datasets generated for MrR16 and TM1442 by omitting the entries of the experimental chemical shift assignments of up to four types of nuclei (represented by datasets *Ia–Ii*). Except for datasets *Ig* (containing δ^{15}N, δ^{13}C$^{\alpha}$, δ^{13}C$^{\beta}$ and δ^{13}C$'$ for all residues) and *Ii* (δ^{13}C$^{\alpha}$ and δ^{13}C$^{\beta}$), these datasets all include δ^{15}N, δ^{13}C$^{\alpha}$ and δ^{1}H^{N}, which constitute the minimum set of protein backbone chemical shifts required for conventional triple resonance assignment. Dataset *Ig*, containing only ^{13}C and ^{15}N chemical shifts, was generated to simulate a typical solid-state NMR chemical shift dataset. Dataset *Ik* contains no chemical shifts and is included to allow comparison of the impact of chemical shifts over standard Rosetta fragment selection.

Table 1 Chemical shift datasets with partial assignment of backbone nuclei

Dataset name	δ^{15}N	δ^{1}H^{N}	δ^{13}C$^{\alpha}$	δ^{13}C$^{\beta}$	δ^{13}C$'$	δ^{1}H$^{\alpha}$
Ia	●	●	●	●	×	●
Ib	●	●	●	×	●	●
Ic	●	●	●	●	●	×
Id	●	●	●	×	×	●
Ie	●	●	●	●	×	×
If	●	●	●	×	●	×
Ig	●	×	●	●	●	×
Ih	●	●	●	×	×	×
Ii	×	×	●	●	×	×
Ij	●	●	●	●	●	●
Ik	×	×	×	×	×	×

●, Present chemical shifts; ×, absent chemical shifts

Datasets with unassigned residues

To simulate the situation of proteins with unassigned residues, for both MrR16 and TM1442 two sets of 'incomplete' chemical shift assignments were generated by omitting all chemical shifts (δ^{15}N, δ^{13}C$'$, δ^{13}C$^{\alpha}$, δ^{13}C$^{\beta}$, δ^{1}H$^{\alpha}$ and δ^{1}H^{N}) for ~10% or ~20% of the residues from their original complete chemical shift datasets. Two different sets of partial chemical shift assignments were generated in this manner. First, a favorable but perhaps unrealistic set was generated where the unassigned residues are evenly distributed along the protein sequence by deleting chemical shifts of residue numbers $N \times 10$ (data set *IIa*) or $N \times 5$ (data set *IIb*), where $N = 1, 2, 3,\ldots$. Second, two more realistic sets of partial assignments were generated where the unassigned residues are consecutive along the protein sequence, exemplifying the situation where residues of one or two segments in the protein are not assigned. Considering that such unassigned stretches of residues are often located in loop or turn regions, we arbitrarily selected such regions with length of ca. 8–10% of the entire sequence, and removed their chemical shift assignments from the datasets. For MrR16, the deleted regions comprise two loops, residues 24–32 (between the second β-strand and the first α-helix, referred to as loop I) and 43–50 (which connects the first α-helix and the third β-strand, referred to as loop II); for TM1442, the two loops, comprise residues 21–29 (between the third β-strand and the first α-helix, loop I) and 52–59 (between the fourth β-strand and the second α-helix, loop II). For each protein, three chemical shift assignment datasets were generated and named as follows: dataset *IIc*, for which all assignments of loop I are omitted, simulating the situation that the residues in loop I are "unassigned"; dataset *IId*, for which the residues in loop II are "unassigned"; and dataset *IIe*, for which the residues in both loops I and II, comprising ~16-19% of the total number of residues in the protein, are "unassigned".

Simulated datasets with artificial errors

In practice, various kinds of chemical shift assignment errors can occur during the protein resonance assignment process, either resulting from mistakes during automated resonance assignment, or from human errors. In order to evaluate the impact of such "random" errors on the CS-Rosetta structure generation, several chemical shift assignment datasets were generated by swapping the chemical shift assignments for two dipeptides with identical amino acid types along the protein sequence: dataset *IIIa*, for which the chemical shift assignments of the two dipeptides have the same secondary structures (for MrR16, Val^{42}-Leu^{43} and Val^{85}-Leu^{86}, both in α-helices; for TM1442, Ile^{16}-Val^{17} and Ile^{47}-Val^{48}, both in β-strands) were swapped; dataset *IIIb*, for which the chemical shift assignments of two dipeptides in different secondary structures were swapped (for MrR16, Leu^{39}-Val^{40} in the first α-helix, Leu^{50}-Val^{51} in the third β-strand; for TM1442, Ser^{52}-Ser^{53} in the loop between the fourth β-strand and second α-helix, and Ser^{82}-Ser^{83} in the last β-strand).

Chemical shift referencing errors also are common, and the resulting "artificial" chemical shift offsets are easily simulated by systematically altering chemical shifts of certain types of nuclei. Here, we evaluate two such datasets: *IIIc*, for which 1.0 ppm was added to all ^{13}C$^{\alpha}$ and ^{13}C$^{\beta}$ chemical shifts as the artificial chemical shift referencing error; and dataset *IIId*, for which 1.7 ppm was added to all ^{13}C$^{\alpha}$ and ^{13}C$^{\beta}$ chemical shifts.

Experimental chemical shifts from solid state NMR

The ^{15}N, ^{13}C$^{\alpha}$, ^{13}C$^{\beta}$ and ^{13}C$'$ chemical shifts of GB3 and ubiquitin, as determined by ssNMR spectroscopy, were taken from the BMRB (accession codes 15283 (Nadaud et al. 2007) and 7111 (Manolikas et al. 2008)). For both proteins, the high-resolution solution NMR structures (PDB entries 2OED (Ulmer et al. 2003) and 1D3Z (Cornilescu et al. 1998), respectively), were used as the experimental reference structures to evaluate the CS-Rosetta all-atom models.

Experimental chemical shifts from paramagnetic proteins

The ^{15}N, ^{13}C$^{\alpha}$, ^{13}C$^{\beta}$, ^{13}C$'$, ^{1}H$^{\alpha}$ and ^{1}H^{N} chemical shift assignments of two paramagnetic proteins, calbindin (75 residues; with a paramagnetic Yb^{3+} ion in the C-terminal metal binding site and Ca^{2+} in the N-terminal site) and ferredoxin (98 residues; with a [2Fe–2S] cofactor), were taken from BMRB entries 15594 (Barnwal et al. 2008) and 5148 (Muller et al. 2002), respectively. The experimental structure of calbindin is taken from a 1.6 Å X-ray structure (PDB entry 4ICB) of diamagnetic Ca^{2+}-calbindin (Svensson et al. 1992); the NMR structure (PDB entry 1JQ4) of the [2Fe–2S] ferredoxin (Muller et al. 2002), for which the above NMR chemical shift assignments were obtained, is used as the experimental reference structure for this protein.

Protein fragment selection and structure generation protocols

The newly extended Rosetta protein structural database, comprising a total of 9,523 proteins, was supplemented with predicted ^{13}C$^{\alpha}$, ^{13}C$^{\beta}$, ^{13}C$'$, ^{15}N, ^{1}H$^{\alpha}$ and ^{1}H^{N} chemical shifts by the program SPARTA (Shen and Bax 2007). Then, for each 3-residue and 9-residue fragment in the query protein, selection of database fragment candidates was performed in two different ways:

(1) *Standard MFR fragment selection*: 200 fragment candidates with best matched backbone NMR chemical shifts and amino acid sequence patterns were selected by using a standard MFR search of the protein structural database (Kontaxis et al. 2005; Shen et al. 2008).

(2) *Hybrid fragment selection*: As indicated in Fig. 1, an exhaustive search was first conducted throughout the protein structural database by using the standard Rosetta method (Rohl et al. 2004) to find the 2,000 database fragments with the best matched amino acid sequence and sequence-derived secondary structure patterns. A second search was then performed on these 2,000 fragment candidates to select the 200 fragments with the best matched chemical shifts pattern according to a chemical shift score of

$$E_{\mathrm{CS}} = \sum_{i,j} c_i \times \left[\frac{\delta_{i,j}^{\mathrm{exp}} - \delta_{i,j}^{\mathrm{calc}}}{\sigma_{i,j}^{\mathrm{calc}}}\right]^2 \Big/ N \qquad (1)$$

as defined in Eq. 3 of Shen et al. (2008), where $\delta_{i,j}$ stands for the chemical shifts of atom i (i = ^{13}C$^{\alpha}$, ^{13}C$^{\beta}$, ^{13}C$'$, ^{15}N, ^{1}H$^{\alpha}$ and ^{1}H^{N}) of residue j in the fragment; $\delta_{i,j}^{\mathrm{exp}}$is the experimental chemical shift in the target segment; $\delta_{i,j}^{\mathrm{calc}}$ and $\sigma_{i,j}^{\mathrm{calc}}$ denote the SPARTA-derived chemical shifts and uncertainties, respectively, for the fragments in the protein structural database; N is the total number of chemical shifts in the fragment; c_i is the weighting factor for each atom type (1.0 for ^{13}C$^{\alpha}$, ^{13}C$^{\beta}$, ^{13}C$'$, ^{1}H$^{\alpha}$; 0.9 for ^{1}H^{N} and ^{15}N). For all tests, proteins with significant sequence homology, as judged by a PSI-BLAST (Altschul et al. 1997) e-score <0.05 to the target protein were excluded from the protein structural database before fragment searching. Note that this removal is only needed for the tests carried out

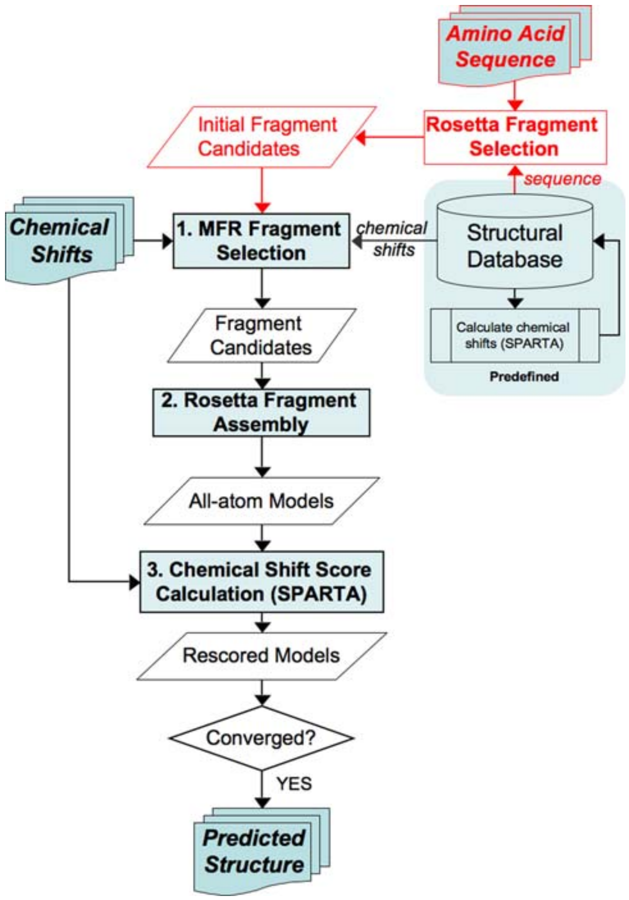
Amino Acid Sequence
Rosetta Fragment Selection
Initial Fragment Candidates
sequence
Chemical Shifts
1. MFR Fragment Selection
chemical shifts
Structural Database
Calculate chemical shifts (SPARTA)
Predefined
Fragment Candidates
2. Rosetta Fragment Assembly
All-atom Models
3. Chemical Shift Score Calculation (SPARTA)
Rescored Models
Converged?
YES
Predicted Structure

Fig. 1 Flow chart of CS-Rosetta structure generation protocol. In the hybrid fragment selection procedure, shown in *red*, step 1 selects 200 fragments from an initial cohort of 2,000 fragments which has been extracted from the structural database by standard Rosetta methods. In the standard CS-Rosetta method, step 1 takes its fragments directly from the 2,200,000 fragments present in the structural database

in this study; in real applications the presence of homologous proteins will increase the quality of the resulting structures.

The selected fragments, represented by their idealized backbone torsion angles and the secondary structure classification for each residue, were used in the standard Rosetta manner as inputs for a Monte Carlo assembly and relaxation process to generate ca. 10,000 Rosetta all-atom models for each protein. These all-atom models were further evaluated in terms of fitness with respect to the input chemical shift data, following the same procedure used in the standard CS-Rosetta protocol (Shen et al. 2008), contributing to the empirical energy term that is used for the selection of final all-atom models.

All CS-Rosetta structure generations were performed using Rosetta@home (http://boinc.bakerlab.org/rosetta/) supported by the BOINC server or the Biowulf PC/Linux cluster at the NIH (http://biowulf.nih.gov).

Evaluation of CS-Rosetta structure generation

To evaluate the influence of the completeness of chemical shift assignments on the CS-Rosetta protein structure generation process, the following parameters are monitored and analyzed:

1. Average value for the coordinate rmsd between 200 selected fragments and the experimental coordinates of the corresponding target segment, representing the average accuracy or "quality" of the selected fragments.
2. Lowest coordinate rmsd of any of the 200 selected fragments relative to the experimental coordinates of the corresponding target segment, representing the accuracy or quality of the best fragment.
3. Raw Rosetta all-atom empirical energy of the assembled full-atom models.
4. Re-scored Rosetta all-atom empirical energy, which includes the agreement with the input chemical shift data (Shen et al. 2008).
5. C^{α} coordinate rmsd of Rosetta all-atom models relative to the experimental protein structure, representing the "accuracy" of the generated all-atom models.
6. C^{α} coordinate rmsd of the ten models with the lowest re-scored empirical Rosetta all-atom models relative to the model with the lowest energy, representing the convergence of the generated all-atom models. Clustering of these lowest energy models within ~2.0 Å from the model with the lowest energy, or within ~2.0 Å from the experimental structure, is taken as the criterion for a successful prediction (Shen et al. 2008).

Results and discussion

During the CS-Rosetta structure generation, the input chemical shifts serve two major functions: fragment selection and re-scoring of the Rosetta models (Fig. 1). Use of the chemical shift information during the fragment search process significantly increases the accuracy of selected fragments over the use of sequence information alone (Shen et al. 2008), and dramatically improves convergence of the structure generation process. Evaluation of the agreement between the final Rosetta-generated models and the input experimental chemical shifts also provides an important selection criterion for eliminating structures whose backbone angles have diverged from those of the original input fragments during the Rosetta optimization procedure. In practice, frequently not all chemical shifts (δ^{15}N, δ^{13}C′, $\delta^{13}C^{\alpha}$, $\delta^{13}C^{\beta}$, $\delta^{1}H^{\alpha}$ and $\delta^{1}H^{N}$) of all residues are available, depending on the resonance assignment strategy chosen and/or missing connectivities in the assignment pathway, most often resulting from conformational exchange on an intermediate time scale. The completeness of the chemical shift assignment will impact both the fragment selection and the re-scoring steps, and thereby the entire CS-Rosetta structure generation procedure. The impact of missing chemical shifts on each of these steps will be discussed below.

Absence of assignments for certain types of nuclei

It is well recognized that secondary chemical shifts of different nuclei in any given residue are correlated (Supplementary Fig. S1), and this correlation can be used effectively to identify potential errors in chemical shift referencing (Wang et al. 2005). The structural information contained in the chemical shifts of the different types of backbone nuclei therefore may be partly redundant. The standard CS-Rosetta protocol utilizes chemical shifts of all backbone and $^{13}C^{\beta}$ atoms to select the best matched 3-residue and 9-residue fragments. This redundancy suggests that the absence of assignments for some of this set of six nuclei (N, H^{N}, C^{α}, H^{α}, C^{β}, C') may not significantly decrease the accuracy of the selected fragments. This issue will be evaluated below for the chemical shift combinations listed in Table 1.

Omission of a single type of chemical shift ($\delta^{13}C'$, $\delta^{13}C^{\beta}$ or $\delta^{1}H^{\alpha}$) is found to have very little adverse impact on the quality of selected fragments (Fig. S2A–C), either when using the regular MFR selection protocol or the hybrid method. There also appears little systematic difference in the accuracy of fragments selected with the regular MFR protocol or the hybrid method when using these sets of chemical shifts, although the individual sets of fragments selected by the two methods can differ substantially. This holds true both when considering the average backbone rmsd relative to the reference structure, and for the rmsd of the fragment most closely matching the reference structure (Fig. S2). In passing, we note that the moderate differences in the quality of the fragments are not that easy to evaluate from Figures such as S2, but these differences propagate during the Monte Carlo Rosetta structure generation process, dramatically impacting the yield of converged structures.

Although the accuracy of the fragments selected when omitting two types of chemical shifts (either $\delta^{13}C'/\delta^{13}C^{\beta}$, $\delta^{13}C'/\delta^{1}H^{\alpha}$, $\delta^{1}H^{\alpha}/\delta^{13}C^{\beta}$, or $\delta^{1}H^{N}/\delta^{1}H^{\alpha}$) decreases somewhat (Fig. S2D–G), this decrease is small compared to the variation in accuracy seen for different fragments along the sequence of the two proteins.

For MrR16, the quality of fragments obtained by using the chemical shift assignments of only $^{1}H^{N}$, ^{15}N, and $^{13}C^{\alpha}$, or sets containing only $\delta^{13}C^{\alpha}$ and $\delta^{13}C^{\beta}$ is not much lower than for sets derived using more complete assignments (Fig. S2). As a result, the convergence of the CS-Rosetta structure generation process remains adequate and permits assembly of reasonable Rosetta models, albeit with raw Rosetta all-atom energies that are not as low as for structures obtained from using all six types of chemical shifts for fragment searching (Fig. S3). Similar results are obtained for TM1442 (Fig. 2).

Remarkably, even though the accuracy of the resulting structures decreases when just using $^{1}H^{N}$, ^{15}N, and $^{13}C^{\alpha}$ chemical shifts, or just $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ chemical shifts, lowest energy structures remain close to the reference structure, in particular when the hybrid fragment selection method is used. A survey of the energies of the Rosetta-assembled structures and their accuracies (Fig. S3) indicates that the original MFR fragment selection results in higher yields during structure generation than the hybrid fragment selection method when assignments are relatively complete. However, for MrR16, the hybrid method outperforms the regular MFR method for datasets *Id*, *If* and *Ih* (Fig. S3); for TM1442 the hybrid method outperforms the regular MFR approach for datasets *Ib*, *If*, and *Ih* (Fig. S4). For the case where no chemical shifts are available, only the standard Rosetta approach can be used. No convergence is then reached for MrR16, whereas for TM1442 the lowest energy models fall within 4 Å from the reference structure and relaxed convergence criteria are met (Fig. S5).

The calculations discussed above, and summarized in Fig. 2 and S2–S5 indicate that the resonance assignments of not all six types of nuclei are required for success of the CS-Rosetta structure generation process. The order of importance of each type of chemical shift can be ranked as $\delta^{13}C^{\alpha} \sim \delta^{13}C^{\beta} > \delta^{1}H^{\alpha} \sim \delta^{13}C' > \delta^{15}N \sim \delta^{1}H^{N}$. For proteins where all or the vast majority of these chemical shifts are available, the standard MFR fragment selection protocol tends to yield better accuracy of the selected MFR fragments and higher convergence, as well as lower energy when generating the all-atom Rosetta models. The calculations also suggest that the chemical shift assignment dataset needed for the CS-Rosetta protocol at a minimum comprises $\delta^{15}N$, $\delta^{1}H^{N}$ and $\delta^{13}C^{\alpha}$, which also are the cornerstone nuclei during triple resonance backbone assignment, complemented by either $\delta^{13}C'$, $\delta^{13}C^{\beta}$ or $\delta^{1}H^{\alpha}$.

Absence of assignments for subsets of residues

The standard MFR fragment selection procedure, implemented in the previously described CS-Rosetta protocol, relies primarily on the match between the experimental $^{13}C^{\alpha}$, $^{13}C^{\beta}$, $^{13}C'$, ^{15}N, $^{1}H^{N}$ and $^{1}H^{\alpha}$ secondary chemical shift values of each residue in any given 3- or 9-residue query fragment, and the SPARTA-generated secondary shift values for the corresponding residues in any fragment present in the structural database. The similarity in amino

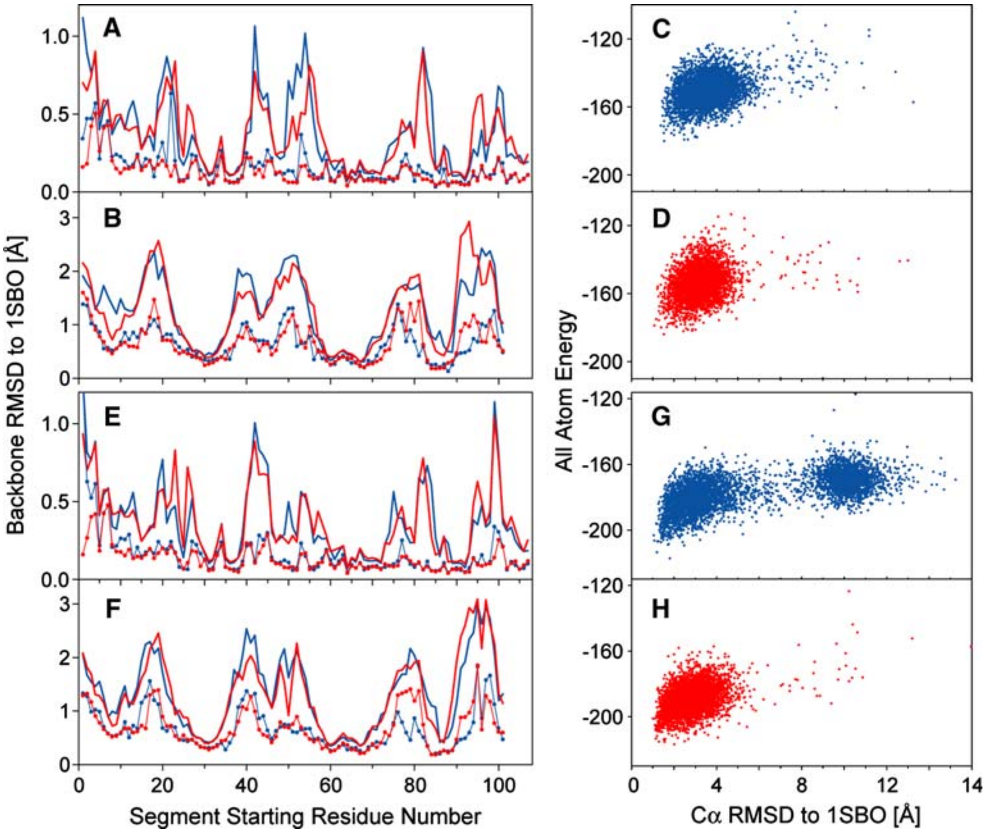

Fig. 2 CS-Rosetta structure generation for TM1442 with missing chemical shift assignments for certain types of nuclei. **A**, **B** Plots of accuracy of fragments selected using the MFR (*blue*) and hybrid (*red*) methods with the chemical shift inputs for δ^{15}N, δ^{1}H^{N} and δ^{13}C$^{\alpha}$ (as contained in dataset *Ih*). Quality of 3-residue (**A**) and 9-residue (**B**) fragments is represented by the average (*bold lines*) and lowest (*lines with dots*) rmsd of 200 selected fragments relative to the experimental coordinates of the corresponding TM1442 segment. **C**, **D** plots of Rosetta all-atom energy, rescored by using the input chemical shifts (as contained in dataset *Ih*), versus C$^{\alpha}$ rmsd relative to the experimental TM1442 structure, for CS-Rosetta models obtained using MFR (**C**) and hybrid (**D**) fragment selection methods. **E**–**H** CS-Rosetta fragment selections and structure generations for TM1442 using only δ^{13}C$^{\alpha}$ and δ^{13}C$^{\beta}$ (as contained in dataset *Ii*)

acid sequence is also used in this scoring process but carries a much weaker weighting. However, when most or all chemical shifts are missing for any given residue or group of residues, the relative importance of similarity in residue type increases and eventually becomes the only criterion when no chemical shifts are available at all. Clearly, the absence of chemical shift information leads to a decrease in accuracy of the fragments that can optimally be selected from any structural database (Shen et al. 2008).

For the relatively favorable situation, where residues with missing chemical shifts are distributed evenly throughout the protein sequence, the chemical shift patterns encoded in the 9-residue target fragments only sustain a small fractional loss in information content when a single residue in such a fragment is missing. Indeed the quality of the MFR-selected fragment candidates for chemical shift assignments *IIa* and *IIb* (see Preparation of chemical shift datasets section) remains quite good (Fig. S6). For the 3-residue fragments, where the loss of assignments for one residue represents 33% loss in information contents, results are less favorable. In particular, when the backbone angles within the 3-residue fragment strongly differ from one another, i.e., when the fragment is not embedded in an α-helix or β-strand, results from the MFR search can be poor. For example, for the 3-residue TM1442 fragments containing residue Lys^{85} (an N-terminal helix capping residue), omission of its chemical shifts (dataset *IIb*), causes a large spike in the coordinate rmsd when using the regular MFR fragment search (Fig. S6B″). Nevertheless, because the adverse impact of lacking chemical shift assignments on the quality of the 9-residue fragments remains small, the Rosetta fragment assembly process remains capable of generating high quality models. This result applies for both MrR16 and TM1442 (Figs. S7 and S8), but for both proteins convergence to the correct structure is lower compared to using a complete set of chemical shift assignments.

A more realistic but also more challenging situation occurs when the unassigned residues cluster along the protein sequence. The MFR fragment selection then becomes dominated by residue type similarity between the query fragment and fragments present in the structural database. The accuracy of fragments that include such unassigned segments, selected by the standard MFR method, is severely affected (Fig. S6C, D), in particular when the missing assignments are located outside regions of secondary structure (datasets *IIc* and *IId*). Interestingly, the quality of these fragments tends to be much lower than what is achieved with the standard Rosetta fragment selection method (Fig. S5), highlighting that the simple residue similarity scoring used by the MFR method performs much worse than the far more elaborate Rosetta fragment selection protocol (Rohl et al. 2004). Unsurprisingly, the subsequent Rosetta structure assembly protocol, using standard MFR fragments as input, can fail to obtain a converged low-energy fold (Figs. S7, S8). On the other hand, for MrR16 the CS-Rosetta structure generation for dataset *Ic*, lacking assignments for residues 24–32, remains successful and finds a converged low-energy fold, where the backbone of the lowest energy model deviates by 1.8 Å from the experimental reference structure (Fig. S7). Even while the quality of 9-residue fragments encompassing this region with missing assignments is poor, the accuracy of the best 3-residue fragments selected remains quite good for this region, and it is the powerful combinatorial engine of Rosetta which can exploit the presence of a relatively small subset of accurate fragments for this single region during the assembly process. For the case where two regions with missing assignments are present in the protein (dataset *IIe*), CS-Rosetta with standard MFR selection no longer is able to obtain converged low energy structures (Fig. 3, S7 and S8).

One way to improve the selection of suitable fragments, and thereby the CS-Rosetta structure generation process, for proteins with extended segments of missing chemical shift assignments is to take advantage of the standard Rosetta fragment selection procedure (Rohl et al. 2004), which searches for matched database fragments based on a relatively sophisticated procedure that simultaneously exploits residue type similarity and predicted secondary structure. Amino acid sequence similarity alone provides less structural information than the backbone chemical shifts, and therefore results in a wider distribution of selected peptide conformations. The average quality of Rosetta-selected fragments therefore is significantly lower than for MFR selection based on chemical shifts, but the quality of the best fragments (out of 200 selected) remains quite good, in particular for the 3-residue fragments (Shen et al. 2008). A preferred way to score the fragments therefore would directly combine, with suitable weight factors, the amino acid sequence based Rosetta fragment score with the chemical shift component of the MFR score. For technical reasons, however, this is not easily accomplished and we therefore resort to a simpler protocol which equally takes advantage of the strengths of both approaches. This hybrid fragment selection procedure first uses standard Rosetta to select the 2000 database fragments (out of over 2,200,000) that are most compatible in terms of amino acid sequence, and then uses MFR chemical shift scoring to narrow down this set to fragments that are most compatible with the

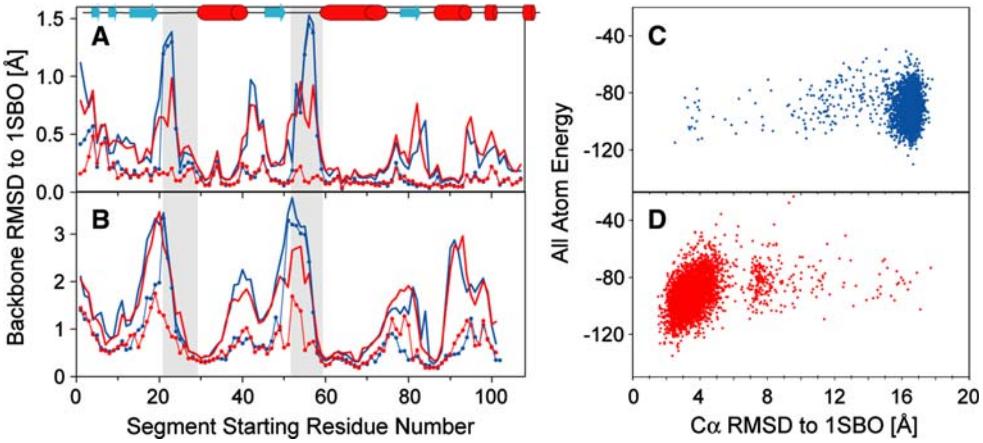

Fig. 3 CS-Rosetta structure for TM1442 with missing chemical shifts. **A**, **B** Plots of accuracy of fragment candidates selected using the MFR (*blue*) and hybrid (*red*) methods using chemical shift values δ^{15}N, δ^{1}H^{N}, δ^{13}C$^{\alpha}$, δ^{13}C$^{\beta}$, δ^{13}C$'$ and δ^{1}H$^{\alpha}$ for residues 1–20, 30–51 and 60–120 (as contained in the dataset *IIe*). For each 3-residue (**A**) and 9-residue (**B**) segment of TM1442, 200 fragments were selected. Average (*bold lines*) and lowest (*lines with dots*) rmsd of these fragments relative to the experimental coordinates of the corresponding TM1442 segment are plotted with respect to the position of the first segment residue in the TM1442 sequence. The regions corresponding to the "unassigned" residues are shaded; the secondary structure elements are displayed at the top. **C**, **D** Plots of Rosetta all atom energy, rescored by using the input chemical shifts (as contained in dataset *IIe*), versus C$^{\alpha}$ rmsd relative to the experimental TM1442 structure, for CS-Rosetta models obtained using MFR (**C**) and hybrid (**D**) fragment selection methods

experimental shifts. When complete chemical shifts are available, this hybrid method performs slightly worse than the regular MFR procedure (Figs. S2–S4). However, when significant segments in the protein lack assignments, the hybrid method remains successful at generating low energy, converged results. For example, when using the 'hybrid' fragments selected with chemical shift datasets *IIc–IIe*, lacking chemical shifts for two extended loop regions, the Rosetta fragment assembly and relaxation protocol results in near-convergence for TM1442, yielding lowest energy models that are within 2.5 Å C$^{\alpha}$ rmsd relative to the reference structure (Fig. 3).

Impact of chemical shift assignment errors

A potential error during conventional and/or automated backbone resonance assignments is exemplified by the case where chemical shift assignments of two di- or tripeptide sequences of similar amino acid sequence, embedded between residues with similar chemical shifts, are accidentally interchanged. Below, we consider the case where assignments for two dipeptides with identical amino acid types are interchanged.

For the favorable situation where the two dipeptides are located in segments with the same secondary structure, as exemplified in dataset *IIIa*, the chemical shift patterns in the 3-residue and 9-residue fragments are virtually unchanged and there is essentially no adverse impact on the fragment selection, neither for the standard MFR nor the hybrid approach (Fig. S9). Clearly, generation of Rosetta structures also remains unaffected (Figs. S10 and S11).

For the case where the two miss-assigned dipeptides are engaged in different types of secondary structure, the incorrect chemical shift values are likely to favor selection of fragments with backbone torsion angles that deviate substantially from the true values, resulting in a significant decrease in the quality of MFR-selected fragments. This is particularly true for the 3-residue fragments (Fig. 4A, B; Fig. S9), where the fraction of erroneous assignments equals two thirds. Not surprisingly, the subsequent Rosetta fragment assembly and relaxation protocol has trouble generating well converged models. Although the lowest (re-scored) energy models exhibit folds that are essentially correct, these differ by ~2.56 and ~3.47 Å (C$^{\alpha}$-rmsd) from the experimental structures of MrR16 and TM1442, respectively (Fig. 4C; Figs. S10 and S11). The re-scored all-atom energies of these models are also systematically higher than obtained when using the correct chemical shift assignments.

When using the hybrid fragment selection method, the impact of erroneous assignments is reduced considerably, and acceptable convergence is achieved (Fig. 4; Figs. S9–S11).

As pointed out by Wang et al. (2005), nearly 30% of the deposited chemical shift data in the BMRB have chemical shift referencing problems. Such referencing errors are most prevalent for $^{13}C^{\alpha}/^{13}C^{\beta}$, but also are common for $^{13}C'$ and ^{15}N. Below, we evaluate the impact of $^{13}C^{\alpha}/^{13}C^{\beta}$ referencing errors. As will be shown, the fragment search procedure is relatively insensitive to moderate errors in $^{13}C^{\alpha}/^{13}C^{\beta}$ chemical shift referencing, in part because $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ secondary shifts are anti-correlated. For example,

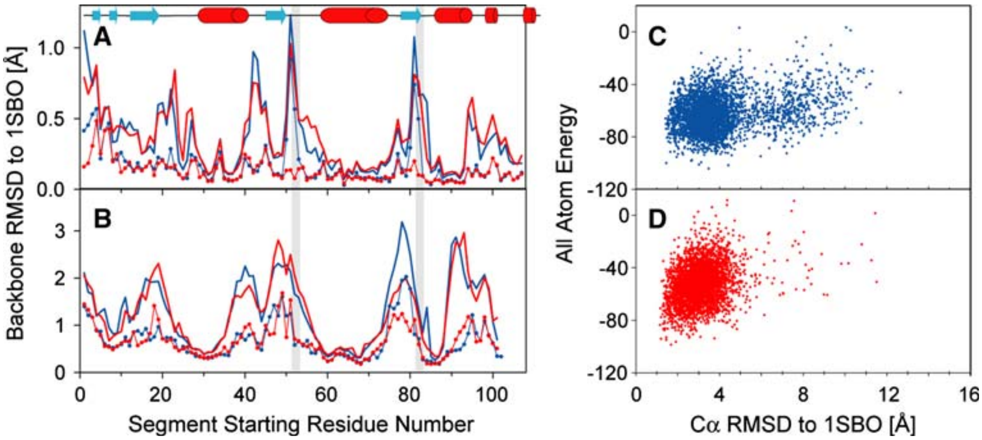

Fig. 4 CS-Rosetta structure generation of TM1442 with chemical shift errors. **A**, **B** Plots of accuracy of fragments selected using the MFR (*blue*) and hybrid (*red*) methods, with the inputs swapped for the $\delta^{15}N$, $\delta^{1}H^{N}$, $\delta^{13}C^{\alpha}$, $\delta^{13}C^{\beta}$, $\delta^{13}C'$ and $\delta^{1}H^{\alpha}$ assignments of dipeptides Ser^{52}-Ser^{53} and Ser^{82}-Ser^{83} (as contained in the dataset *IIIb*). For each 3-residue (**A**) and 9-residue (**B**) segment of TM1442, 200 fragments were selected. Average (*bold lines*) and lowest (*lines with dots*) rmsd of these fragments relative to the experimental coordinates of the corresponding TM1442 segment are plotted with respect to the position of the first segment residue in the TM1442 sequence. The regions corresponding to the "miss-assigned" residues are shaded; secondary structure elements are displayed at the top. **C**, **D** Plots of Rosetta all atom energy, rescored by using the input chemical shifts (as contained in the dataset *IIIb*), versus C$^{\alpha}$ rmsd relative to the experimental TM1442 structure, for CS-Rosetta models obtained using MFR (**C**) and hybrid (**D**) fragment selection methods

a 4 ppm reference error could change a typical β-sheet secondary $^{13}C^{\alpha}$ shift of -1 ppm to an α-helical 3 ppm value. However, the +2 ppm β-sheet secondary $^{13}C^{\beta}$ shift would become +6 ppm, completely incompatible with a helical conformation, preventing the residue from being misidentified as helical. To first order, the impact of $^{13}C^{\alpha}/^{13}C^{\beta}$ referencing errors is small when both $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ shift data are available, and manifests itself mainly as a steeper $^{13}C^{\alpha}/^{13}C^{\beta}$ chemical shift gradient when selecting fragments, and increased total energies when rescoring the energies of the Rosetta models.

The impact of $^{13}C^{\alpha}/^{13}C^{\beta}$ chemical shift referencing errors on CS-Rosetta structure generation was evaluated using the chemical shift assignment datasets *IIIc* and *IIId*. When 1.0 ppm offset was added to $\delta^{13}C^{\alpha/\beta}$ (dataset *IIIc*), comparable to the average $\delta^{13}C^{\alpha/\beta}$ prediction errors (σ in Eq. 1) (Gong et al. 2007; Shen and Bax 2007), the accuracy of the selected fragments slightly decreases (Fig. S9), with a very small adverse impact on subsequent Rosetta structure generation (Figs. S10 and S11). The impact of chemical shift referencing errors appears to be insensitive to the type of fragment selection method used: For MrR16, standard MFR yields slightly better results (Fig. S10); for TM1442, the hybrid method is slightly favorable (Fig. S11).

When the $\delta^{13}C^{\alpha/\beta}$ offset error is increased to 1.7 ppm (dataset *IIIc*), convergence and accuracy of the resulting structures decreases noticeably (Figs. S10 and S11), but the folds remain essentially correct. However, when the offset error is increased to 2.7 ppm, which corresponds to the approximate difference between $\delta^{13}C^{\alpha/\beta}$ values referenced to TMS and DSS (Wishart et al. 1995; Markley et al. 1998), fragment selection results are poor and no acceptable structures are obtained with the CS-Rosetta protocol (data not shown).

When the chemical shift referencing error affects only a single type of nucleus, e.g. $^{13}C^{\alpha}$ or $^{13}C'$, an erroneous bias towards selection of helical or extended fragments can occur, resulting in poorer fragment quality and decreased performance of the CS-Rosetta protocol (results not shown). Even in these cases, the impact of ^{15}N or ^{13}C chemical shift referencing errors of up to 1 ppm have very little adverse effect on CS-Rosetta performance.

Chemical shift referencing errors readily can be detected by automated methods (Moseley et al. 2004; Wang et al. 2005). For this purpose, a script has been added to the CS-Rosetta package which applies reference error corrections when the referencing error exceeds the average uncertainty in the database chemical shifts (1.0 ppm for $\delta^{13}C^{\alpha/\beta}$ and $\delta^{13}C'$; 0.3 ppm for $\delta^{1}H^{\alpha}$). These referencing corrections are based on the method described by Markley and coworkers (Wang et al. 2005), and correlations between ($\Delta\delta^{13}C^{\alpha}$– $\Delta\delta^{13}C^{\beta}$) and $\Delta\delta^{13}C^{\alpha/\beta}/\Delta\delta^{13}C'/\Delta\delta^{1}H$ are shown in Fig. S1.

A situation similar to the chemical shift referencing problem discussed above can arise when chemical shifts are measured from TROSY spectra (Pervushin et al. 1998), when the displacement between the observed resonance frequency and the true chemical shift ($^{1}J_{NH}/2$ for $\delta^{15}N$ and $\delta^{1}H^{N}$) is not taken into account. However, considering that this error is much smaller than the standard error in the predicted database chemical shifts, no adjustment of the chemical shift values is required.

A larger apparent referencing error can result from deuteration effects (Venters et al. 1996; Gardner et al. 1997) on $\delta^{13}C^{\alpha}$ (with deuterium isotope shifts of -0.5 to -0.9 ppm) and $\delta^{13}C^{\beta}$ (-0.7 to -1.3 ppm). These isotope effects on the backbone chemical shifts are relatively uniform and mostly smaller than the 1 ppm referencing error, discussed above. Although it is beneficial to apply uniform isotope shift corrections of +0.7 and +0.9 ppm to $\delta^{13}C^{\alpha}$ and $\delta^{13}C^{\beta}$ values, respectively, the absence of such corrections shows little adverse impact on the performance of CS-Rosetta (data not shown). Nevertheless, a script has been added to the CS-Rosetta package which adjusts the $\delta^{13}C^{\alpha}$ and $\delta^{13}C^{\beta}$ chemical shifts by the residue-type-specific values reported by Cavanagh et al. (2007).

Structures from solid-state NMR chemical shifts

The backbone chemical shifts $\delta^{15}N$, $\delta^{13}C'$, $\delta^{13}C^{\alpha}$ and $\delta^{13}C^{\beta}$ obtained by ssNMR for the proteins GB3 and ubiquitin were used as inputs for the CS-Rosetta structure generation protocols. For GB3, nearly complete ssNMR backbone chemical shift assignments, including 55 $\delta^{15}N$, 56 $\delta^{13}C'$, 56 $\delta^{13}C^{\alpha}$ and 52 $\delta^{13}C^{\beta}$, are taken from (Nadaud et al. 2007). For the most part, these chemical shifts closely agree with values observed by solution NMR (Fig. S12). For ubiquitin, the ssNMR backbone chemical shift assignments taken from (Igumenova et al. 2004) are about ~90% complete, and include 65 $\delta^{15}N$, 65 $\delta^{13}C'$, 67 $\delta^{13}C^{\alpha}$ and 63 $\delta^{13}C^{\beta}$ values, with no chemical shift assignments for residues 8–11. With the exception of several residues involved in intermolecular contacts, these chemical shifts also agree well with values observed in solution (Igumenova et al. 2004) (Fig. S12).

Except for the ubiquitin target fragments that involve the missing residues 8–11, the quality of fragments selected on the basis of ssNMR shift values is good, with little difference apparent between results from the standard MFR and the hybrid fragment selection method (Fig. 5). As expected based on the evaluations carried out above for regions with missing assignments, the regular MFR method fares poorly when selecting fragments that include residues

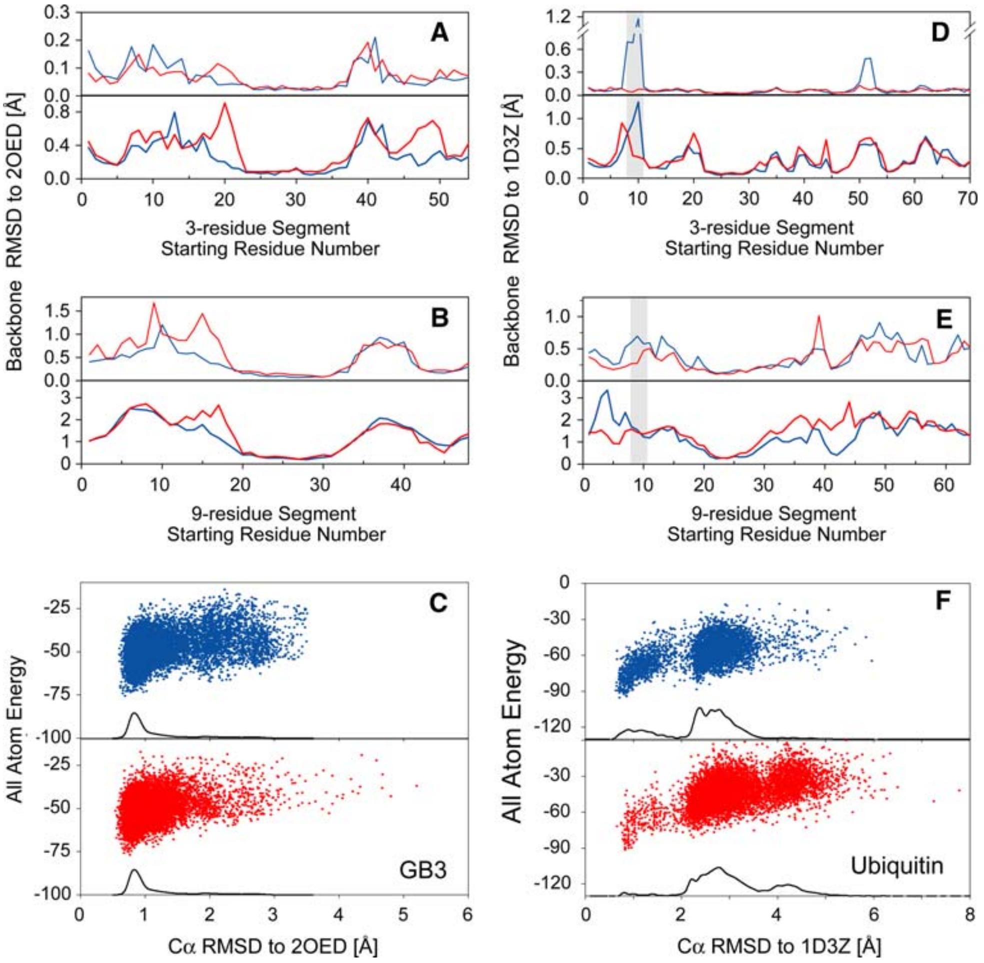

Fig. 5 CS-Rosetta fragment selection and structure generation for GB3 (**A**–**C**) and ubiquitin (**D**–**F**), using chemical shift assignments from solid-state NMR. **A**, **D** Plots of the lowest (upper panel) and average (lower panel) backbone coordinate rmsds (N, C^{α} and C′) between query segment and two hundred 3-residue fragments, selected using the MFR (*blue*) and hybrid methods (*red*), as a function of starting position in the sequence. **B**, **E** Same as (**A**, **D**), but for 9-residue fragments. **C**, **F** Plots of Rosetta all atom energy, rescored by using the experimental ssNMR chemical shifts, versus C^{α} rmsd relative to the experimental NMR structures of GB3 and ubiquitin for the CS-Rosetta all-atom models obtain using MFR-selected (upper panel, *blue dots*) and the hybrid method (lower panel, *red dots*) fragments. The *solid black lines* in (**C**, **F**) represent the normalized number of structures found at a given C^{α}-rmsd

8–11, whereas the hybrid method shows no decrease in structural quality for this region.

Importantly, either selection method yields fragments from the ssNMR chemical shifts that suffice for generating converged, high quality all-atom models for both proteins (Fig. 5C, F). When the MFR method is used to select the fragments, the coordinate rms deviations for GB3 between the lowest energy model and the experimental solution NMR structure are 0.71 Å for the backbone atoms (N, C^{α} and C′) and 1.28 Å for all non-hydrogen atoms. For ubiquitin these numbers are 0.69 and 1.22 Å. When the fragments are selected by the hybrid procedure, the coordinate rmsd's are slightly higher: 0.73 and 1.70 Å for backbone and all non-hydrogen GB3 atoms, respectively, and 0.86 and 1.49 Å for ubiquitin.

Considering the generally somewhat lower spectral resolution attainable by ssNMR compared to solution NMR, detailed structural studies of globular proteins by ssNMR mostly have remained restricted to relatively small systems, typically less than ~80 residues. Clearly, CS-Rosetta provides a powerful new complementary tool for generating structural models of such proteins once chemical shift assignments have been completed, without requiring the extensive internuclear distance information which sometimes can be difficult to obtain.

Paramagnetic protein structure from chemical shifts

Two small paramagnetic proteins for which chemical shifts are available in the BMRB (Doreleijers et al. 2005) have been used to evaluate the applicability of CS-Rosetta to such systems: calbindin and ferredoxin. The backbone chemical shift assignments of calbindin, chelating a paramagnetic Yb^{3+} ion in its C-terminal metal binding site and Ca^{2+} in the N-terminal site, include 52 $\delta^{15}N/\delta^{1}H^{N}$, 43 $\delta^{13}C'$, 37 $\delta^{13}C^{\alpha}/\delta^{1}H^{\alpha}$ and 33 $\delta^{13}C^{\beta}$ shifts, but no chemical shift assignments for residues 18 to 24 and 47 to 66; the completeness of the backbone chemical shift assignments is ~60% (Barnwal et al. 2008). The backbone chemical shift assignments of ferredoxin include 78 $\delta^{15}N/\delta^{1}H^{N}$, 83 $\delta^{13}C'$, 86 $\delta^{13}C^{\alpha}/\delta^{1}H^{\alpha}$ and 78 $\delta^{13}C^{\beta}$ values, and lack assignments for residues 41–50 and 80–82; the completeness of the backbone chemical shift assignments is ~80% (Muller et al. 2002).

With the absence of chemical shift assignments for long segments in each of these two proteins, the standard CS-Rosetta protocol, using MFR fragment selection, fails to converge for both proteins (Fig. S13). However, the hybrid fragment selection procedure performs much better, in particular for those target fragments involving the unassigned residues (Fig. 6A, B, D, E), permitting the structure assembly phase to be successful (Fig. 7). Interestingly, this improved performance is not dominated by recognition of the relatively common EF-hand and Fe–S metal-binding sites as, for testing purposes, proteins with a PSI-BLAST e-score <0.05 had been removed from the database. Subsequent manual evaluation of the 9-residue fragments covering the regions lacking chemical shifts showed the presence of six 9-residue fragments for calbindin segment 54–62, which were taken from EF-hand containing proteins that had escaped detection by the PSI-BLAST filter.

For both proteins, the Rosetta fragment assembly and relaxation procedure generates a number of good all-atom models, with the lowest energy models having backbone coordinates that differ by less than 2 Å from their respective reference structures when only including residues involved in secondary structure (Fig. 6C, F). Although, the standard convergence criterion (10 lowest energy structures cluster with 2 Å from the lowest energy structure) is not met for either protein (Fig. S13), when relaxing this limit to 3.3 Å both structures are converged.

For calbindin, the coordinate rmsd's between the lowest energy all-atom model and the 1.6-Å X-ray structure of calbindin D9K (Svensson et al. 1992) are 1.5 and 2.1 Å for the backbone atoms (N, C^{α} and C′) and for all heavy atoms involved in secondary structure, respectively. The Ca^{2+} binding loops of both metal binding sites are remarkably well formed in the CS-Rosetta structures (Fig. 7A), even

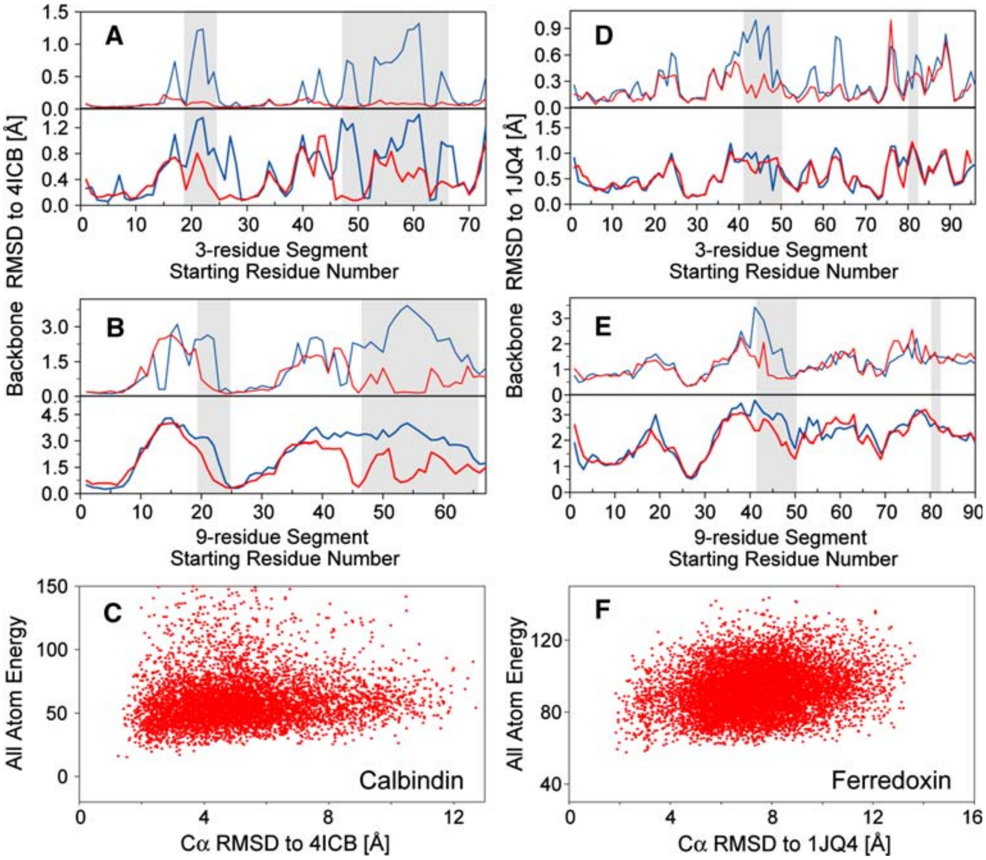

Fig. 6 CS-Rosetta structure generation for paramagnetic calbindin (**A**–**C**) and ferredoxin (**D**–**F**). **A**, **D** Plots of the lowest (upper panel) and average (lower panel) backbone coordinate rmsds (N, C^{α} and C′) between query segment and two hundred 3-residue fragment candidates, selected using the MFR (*blue*) and hybrid methods (*red*), as a function of starting position in the sequence. The regions lacking chemical shift assignments are shaded. **B**, **E** Same as (**A**, **D**), but for 9-residue fragments. **C**, **F** Plots of Rosetta all-atom energy, rescored by the experimental chemical shifts, versus C^{α} rmsd of final al-atom models (including only residues located in elements of secondary structure) relative to the corresponding X-ray (calbindin) and NMR (ferredoxin) structure. Only results from CS-Rosetta all-atom models obtained by the hybrid fragment selection procedure are shown; when using fragments from the standard MFR method, Rosetta fails to converge. Residues included in the backbone rmsd calculation include 3–14, 25–40, 46–53 and 63–74 for calbindin, and 4–11, 15–22, 27–34, 54–56, 71–75 and 91–93 for ferredoxin

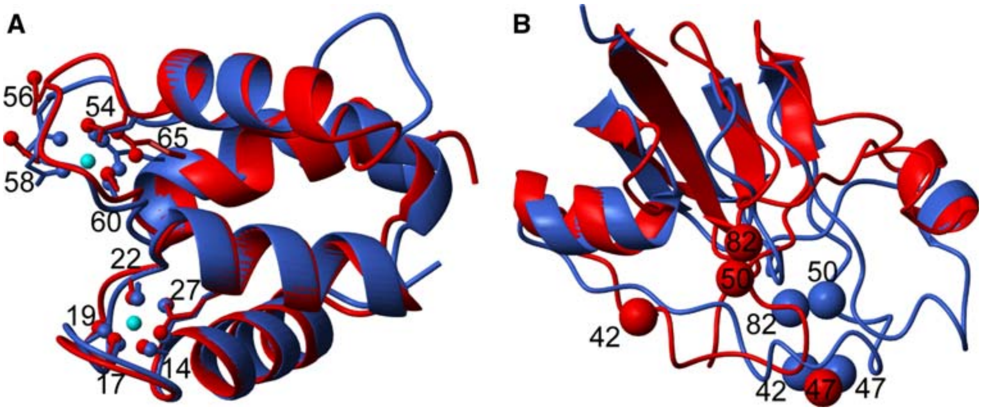

Fig. 7 Comparison of experimental (*blue*) and lowest energy CS-Rosetta (*red*) structure for paramagnetic calbindin (**A**) and ferredoxin (**B**). Superposition is optimized for residues in secondary structure, defined in the caption to Fig. 6. The sidechains of residues involved in metal binding including their metal-ligating oxygen atoms, as well as the X-ray positions of the Ca^{2+} ions (*cyan*) are shown. Metal-ligating residues (atoms) include Ala^{14} (O), Glu^{17} (O), Asp^{19} (O), Gln^{22} (O), Glu^{27} (O$^{\varepsilon1}$/O$^{\varepsilon2}$), Asp^{54} (O$^{\delta1}$), Asn^{56} (O), Asp^{58} (O$^{\delta1}$), Glu^{60} (O) and Glu^{65} (O$^{\varepsilon1}$/O$^{\varepsilon2}$). (**B**) Backbone ribbon representation of the lowest-energy CS-Rosetta structure (*red*) superimposed on the experimental X-ray structure (*blue*) for ferredoxin, with superposition optimized for the residues in secondary structure (See caption to Fig. 6). The sidechain S atoms of Cys^{42}, Cys^{47}, Cys^{50} and Cys^{82}, which coordinate the [2Fe–2S] cluster are marked as *solid spheres*. Figures made using Molmol (Koradi et al. 1996)

with the second metal binding site lacking all of its chemical shift assignments and the absence of any restraints on metal chelation for both metal binding sites. For the first Ca^{2+} binding loop, a pseudo-EF-hand, the four backbone carbonyl groups are properly positioned and point towards the location where Ca^{2+} is found in the X-ray structure. Even the bidentate sidechain ligating group of Glu^{27} adopts a conformation suitable for metal chelation. For the second site, a regular EF-hand, the backbone carbonyl of Glu^{60} and the sidechains of Asp^{54} and Glu^{65} are well positioned for metal binding, but the sidechains of Asn^{56} and Asp^{58} point away from the position where the metal ion is observed in the X-ray structure.

For the secondary structure elements of ferredoxin, the lowest energy Rosetta model deviates from the experimental NMR structure obtained for the same protein by 2.06 Å for the backbone and by 3.54 Å for all non-hydrogen atoms. Two of the four Cys sidechains that ligate the [2Fe–2S] cluster are in close proximity, even though the loop conformations differ substantially from the experimentally determined structure (Fig. 7B).

Concluding remarks

Although previous reports have clearly demonstrated the potential of using chemical shifts to determine good quality all-atom structures for small proteins (Cavalli et al. 2007; Shen et al. 2008), these studies were based on relatively ideal cases where complete or nearly complete backbone assignments were available, in the absence of assignment errors. Our present study demonstrates that the CS-Rosetta procedure and its new variant, which uses a hybrid fragment selection procedure, are remarkably tolerant to such incompleteness and errors. Clearly, a study such as the present one, which evaluates the impact of missing or erroneous assignments, is never complete. We simply have evaluated the impact for two proteins, and have made an attempt to evaluate representative cases of missing assignments. Both proteins chosen for the current study, MrR16 and TM1442, yielded good (albeit not exceptional) results when originally studied with complete data sets, and these systems therefore are likely to be more robust to incompleteness or assignment errors than proteins which only yield borderline convergence to begin with.

The CS-Rosetta protocol uses the chemical shift information at two stages: first for fragment selection, and then again when evaluating the final full-atom models. There are two primary reasons for the improved performance of the CS-Rosetta protocol over a conceptually similar, earlier attempt to integrate chemical shift information into Rosetta (Bowers et al. 2000). First, the quality of fragments selected has improved considerably by the use of SPARTA to “assign” better chemical shifts to a structural database. SPARTA uses both a more advanced algorithm to assign these chemical shifts, but also benefits from a considerable expansion of entries in the BMRB for which complete chemical shift and high resolution structural information is available (Doreleijers et al. 2005). Second, a number of improvements in the Rosetta Monte Carlo assembly process have been made in recent years, most notably the incorporation of explicit all atom refinement with a physically realistic force field (Das and Baker 2008).

The adverse impact of errors and incompleteness on the CS-Rosetta protocol results primarily from decreased quality of the fragment library, and has relatively little impact on the rescoring of the final full-atom models. The hybrid CS-Rosetta protocol first limits the selection of fragments to a ~0.1% fraction of the total structural database on the basis of the standard Rosetta selection mechanism. In the next step, it uses MFR to select the 200 fragments from this ensemble that agree best with experimental chemical shifts. This reduces the impact of chemical shift errors because only fragments compatible with standard Rosetta criteria are available for selection. Moreover, in the absence of any chemical shift information, the Rosetta pre-selection of the top 0.1% fragments yields better results than the less sophisticated MFR procedure, which had been designed primarily to find fragments with similar chemical shifts and/or RDCs (Delaglio et al. 2000; Kontaxis et al. 2005). In the absence of assignment errors or missing assignments, the initial Rosetta pre-selection used in the hybrid procedure is not beneficial and actually results in a small decrease in performance. On the other hand, for cases where significant fractions of assignments are missing or ambiguous, the hybrid procedure is considerably more robust.

For all evaluations, including those of the two paramagnetic proteins, homologous proteins were first eliminated from the structural database. In practice, this is clearly disadvantageous as Rosetta no longer can take advantage of standard structural elements, such as Ca^{2+}-ligating EF-hand sequences, present in the database. Indeed 30 proteins containing a total of 64 EF-hands were removed prior to fragment searching. Similarly, proteins containing the relatively common Fe_2S_2 cluster were removed prior to searching for fragments for ferredoxin assembly. While for calbindin the CS-Rosetta protocol resulted in remarkably good backbone structures for its metal binding sites, even in the absence of chemical shift information, loop conformations in ferredoxin were poor. Nevertheless, using the hybrid protocol, CS-Rosetta was able to generate the remainder of the ferredoxin structure quite well, suggesting that even for these challenging systems the method will be quite useful.

For the two proteins for which a structure was generated from solid-state NMR chemical shifts, lacking ^{1}H chemical shifts, the standard MFR-based protocol and the hybrid CS-Rosetta method performed comparably well. For both proteins, the final structures obtained from these smaller input data sets approach the quality of structures obtained from solution NMR chemical shifts, indicating that CS-Rosetta may be a particularly useful complement when working with samples in the solid state.

Although CS-Rosetta considerably reduces the amount of spectral data collection time required for structure generation compared to conventional procedures, the amount of computational time required typically is very high. Although for simple systems such as GB3, generation of less than one hundred structures may suffice to reach convergence (Shen et al. 2008), for many other proteins as many as 10,000 models may be required. Rosetta assembly and minimization of each model takes 5–10 min on a single CPU, and in practice use of a large cluster or a central server such as BOINC is required to take advantage of this technology.

We also note that the CS23D program (Wishart et al. 2008) performs very well for the test datasets used in our study (Supplementary material). The major strength of CS23D is that it takes optimal advantage of sequence homologues present in the database during fragment selection. Such homologues were present in the structural database for all six proteins evaluated in our work (see Supplementary material Table S2), but were excluded from the database for CS-Rosetta testing. On the other hand, based on a limited number of tests, techniques such as CS-Rosetta and Cheshire are believed to be superior for proteins that lack significant homology to previously solved structures.

Software availability

The CS-Rosetta software package with its newly implemented hybrid fragment selection module can be downloaded from http://spin.niddk.nih.gov/bax/.

Acknowledgments This work was funded by the Intramural Research Program of the NIDDK, NIH, and by the Intramural AIDS-Targeted Antiviral Program of the Office of the Director, NIH; the NIGMS, NIH, and the Howard Hughes Medical Institutes (to D. B.). We also thank Rosetta@home participants and the BOINC project for contributing computing power.

References

Agarwal V, Diehl A, Skrynnikov N, Reif B (2006) High resolution H-1 detected H-1, C-13 correlation spectra in MAS solid-state NMR using deuterated proteins with selective H-1, H-2 isotopic labeling of methyl groups. J Am Chem Soc 128:12620–12621

Altschul SF, Madden TL, Schaffer AA, Zhang JH, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402

Ando I, Kameda T, Asakawa N, Kuroki S, Kurosu H (1998) Structure of peptides and polypeptides in the solid state as elucidated by NMR chemical shift. J Mol Struct 441:213–230

Andreini C, Bertini I, Rosato A (2004) A hint to search for metalloproteins in gene banks. Bioinformatics 20:1373–1380

Asakura T, Demura M, Date T, Miyashita N, Ogawa K, Williamson MP (1997) NMR study of silk I structure of *Bombyx mori* silk fibroin with N-15- and C-13-NMR chemical shift contour plots. Biopolymers 41:193–203

Barnwal RP, Rout AK, Chary KVR, Atreya HS (2008) Rapid measurement of pseudocontact shifts in paramagnetic proteins by GFT NMR spectroscopy. Open Magn Reson J 1:16–28

Bermel W, Bertini I, Felli IC, Piccioli M, Pierattelli R (2006) C-13-detected protonless NMR spectroscopy of proteins in solution. Prog Nucl Magn Reson Spectrosc 48:25–45

Bertini I, Luchinat C, Parigi G, Pierattelli R (2005) NMR spectroscopy of paramagnetic metalloproteins. Chembiochem 6: 1536–1549

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

Case DA (1995) Calibration of ring-current effects in proteins and nucleic acids. J Biomol NMR 6:341–346

Castellani F, van Rossum B, Diehl A, Schubert M, Rehbein K, Oschkinat H (2002) Structure of a protein determined by solid-state magic-angle-spinning NMR spectroscopy. Nature 420:98–102

Castellani F, van Rossum BJ, Diehl A, Rehbein K, Oschkinat H (2003) Determination of solid-state NMR structures of proteins by means of three-dimensional ^{15}N–^{13}C–^{13}C dipolar correlation spectroscopy and chemical shift analysis. Biochemistry 42:11476–11483

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

Cavanagh J, Fairbrother WJ, Palmer AG, Rance M, Skelton NJ (2007) Protein NMR spectroscopy: principles and practice, 2nd edn. Academic Press, San Diego, CA

Chevelkov V, Rehbein K, Diehl A, Reif B (2006) Ultrahigh resolution in proton solid-state NMR spectroscopy at high levels of deuteration. Angew Chem Int Ed 45:3878–3881

Cornilescu G, Marquardt JL, Ottiger M, Bax A (1998) Validation of protein structure from anisotropic carbonyl chemical shifts in a dilute liquid crystalline phase. J Am Chem Soc 120:6836–6837

Cornilescu G, Delaglio F, Bax A (1999) Protein backbone angle restraints from searching a database for chemical shift and sequence homology. J Biomol NMR 13:289–302

Das R, Baker D (2008) Macromolecular modeling with Rosetta. Annu Rev Biochem 77:363–382

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

Doreleijers JF, Nederveen AJ, Vranken W, Lin JD, Bonvin A, Kaptein R, Markley JL, Ulrich EL (2005) BioMagResBank databases DOCR and FRED containing converted and filtered sets of experimental NMR restraints and coordinates from over 500 protein PDB structures. J Biomol NMR 32:1–12

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

Gong HP, Shen Y, Rose GD (2007) Building native protein conformation from NMR backbone chemical shifts using Monte Carlo fragment assembly. Protein Sci 16:1515–1521

Gryk MR, Hoch JC (2008) Local knowledge helps determine protein structures. Proc Natl Acad Sci USA 105:4533–4534

Haigh CW, Mallion RB (1979) Ring current theories in nuclear magnetic resonance. Prog Nucl Magn Reson Spectrosc 13: 303–344

Igumenova TI, McDermott AE, Zilm KW, Martin RW, Paulson EK, Wand AJ (2004) Assignments of carbon NMR resonances for microcrystalline ubiquitin. J Am Chem Soc 126:6720–6727

Ikura M, Kay LE, Bax A (1990) A novel approach for sequential assignment of ^{1}H, ^{13}C, and ^{15}N spectra of larger proteins: heteronuclear triple-resonance three-dimensional NMR spectroscopy. Application to calmodulin. Biochemistry 29:4659–4667

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

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

Loquet A, Bardiaux B, Gardiennet C, Blanchet C, Baldus M, Nilges M, Malliavin T, Boeckmann A (2008) 3D structure determination of the Crh protein from highly ambiguous solid-state NMR restraints. J Am Chem Soc 130:3579–3589

Manolikas T, Herrmann T, Meier BH (2008) Protein structure determination from C-13 spin-diffusion solid-state NMR spectroscopy. J Am Chem Soc 130:3959–3966

Markley JL, Bax A, Arata Y, Hilbers CW, Kaptein R, Sykes BD, Wright PE, Wuthrich K (1998) IUPAC-IUBMB-IUPAB inter-union task group on the standardization of data bases of protein and nucleic acid structures determined by NMR spectroscopy. Pure Appl Chem 70:117–142

Montelione GT, Wagner G (1990) Conformation-independent sequential NMR connections in isotope-enriched polypeptides by ^{1}H–^{13}C–^{15}N triple-resonance experiments. J Magn Reson 87:183–188

Moseley HNB, 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

Muller J, Lugovskoy AA, Wagner G, Lippard SJ (2002) NMR structure of the [2Fe–2S] ferredoxin domain from soluble methane monooxygenase reductase and interaction with its hydroxylase. Biochemistry 41:42–51

Nadaud PS, Helmus JJ, Jaroniec CP (2007) 13C and 15N chemical shift assignments and secondary structure of the B3 immunoglobulin-binding domain of streptococcal protein G by magic-angle spinning solid-state NMR spectroscopy. Biomol NMR Assign 1:117–120

Neal S, Nip AM, Zhang HY, Wishart DS (2003) Rapid and accurate calculation of protein H-1, C-13 and N-15 chemical shifts. J Biomol NMR 26:215–240

Neal S, Berjanskii M, Zhang HY, Wishart DS (2006) Accurate prediction of protein torsion angles using chemical shifts and sequence homology. Magn Reson Chem 44:S158–S167

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

Rohl CA, Strauss CEM, Misura KMS, Baker D (2004) Protein structure prediction using Rosetta. Meth Enzymol 383:66–93

Saito H (1986) Conformation-dependent C13 chemical shifts—a new means of conformational characterization as obtained by high resolution solid state C13 NMR. Magn Reson Chem 24: 835–852

Shen Y, Bax A (2007) Protein backbone chemical shifts predicted from searching a database for torsion angle and sequence homology. J Biomol NMR 38:289–302

Shen Y, Lange O, Delaglio F, Rossi P, Aramini JM, Liu GH, Eletsky A, Wu YB, 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 105:4685–4690

Siemer AB, Ritter C, Ernst M, Riek R, Meier BH (2005) High-resolution solid-state NMR spectroscopy of the prion protein HET-s in its amyloid conformation. Angew Chem Int Ed 44:2441–2444

Spera S, Bax A (1991) Empirical correlation between protein backbone conformation and Cα and Cβ ^{13}C nuclear magnetic resonance chemical shifts. J Am Chem Soc 113:5490–5492

Svensson LA, Thulin E, Forsen S (1992) Proline cis-trans isomers in calbindin D9K observed by X-ray crystallography. J Mol Biol 223:601–606

Tycko R (1996) Prospects for resonance assignments in multidimensional solid-state NMR spectra of uniformly labeled proteins. J Biomol NMR 8:239–251

Ulmer TS, Ramirez BE, Delaglio F, Bax A (2003) Evaluation of backbone proton positions and dynamics in a small protein by liquid crystal NMR spectroscopy. J Am Chem Soc 125: 9179–9191

Venters RA, Farmer BT, Fierke CA, Spicer LD (1996) Characterizing the use of perdeuteration in NMR studies of large proteins C-13, N-15 and H-1 assignments of human carbonic anhydrase II. J Mol Biol 264:1101–1116

Wagner G, Pardi A, Wuthrich K (1983) Hydrogen-bond length and H-1-NMR chemical-shifts in proteins. J Am Chem Soc 105:5948–5949

Wang LY, Eghbalnia HR, Bahrami A, Markley JL (2005) Linear analysis of carbon-13 chemical shift differences and its application to the detection and correction of errors in referencing and spin system identifications. J Biomol NMR 32:13–22

Williamson MP, Asakura T (1993) Empirical comparisons of models for chemical-shift calculation in proteins. J Magn Reson B 101:63–71

Williamson MP, Kikuchi J, Asakura T (1995) Application of H1 NMR chemical shifts to measure the quality of protein structures. J Mol Biol 247:541–546

Wishart DS, Sykes BD, Richards FM (1991) Relationship between nuclear magnetic resonance chemical shift and protein secondary structure. J Mol Biol 222:311–333

Wishart DS, Bigam CG, Holm A, Hodges RS, Sykes BD (1995) 1H, 13C and 15N random coil NMR chemical shifts of the common amino acids. I. Investigations of nearest-neighbor effects. J Biomol NMR 5:67–81

Wishart DS, Arndt D, Berjanskii M, Tang P, Zhou J, Lin G (2008) CS23D: a web server for rapid protein structure generation using NMR chemical shifts and sequence data. Nucleic Acids Res 36:496–502

Zech SG, Wand AJ, McDermott AE (2005) Protein structure determination by high-resolution solid-state NMR spectroscopy: application to microcrystalline ubiquitin. J Am Chem Soc 127:8618–8626