J Biomol NMR (2013) 57:117–127
DOI 10.1007/s10858-013-9772-4

ARTICLE

Improved chemical shift based fragment selection for CS-Rosetta using Rosetta3 fragment picker

Robert Vernon · Yang Shen · David Baker · Oliver F. Lange

Received: 18 June 2013 / Accepted: 10 August 2013 / Published online: 22 August 2013

Abstract A new fragment picker has been developed for CS-Rosetta that combines beneficial features of the original fragment picker, MFR, used with CS-Rosetta, and the fragment picker, NNMake, that was used for purely sequence based fragment selection in the context of ROSETTA de-novo structure prediction. Additionally, the new fragment picker has reduced sensitivity to outliers and other difficult to match data points rendering the protocol more robust and less likely to introduce bias towards wrong conformations in cases where data is bad, missing or inconclusive. The fragment picker protocol gives significant improvements on 6 of 23 CS-Rosetta targets. An independent benchmark on 39 protein targets, whose NMR data sets were published only after protocol optimization had been finished, also show significantly improved performance for the new fragment picker (van der Schot et al. in J Biomol NMR, 2013).

Keywords Protein structure · NMR · Sparse data · Chemical shifts

Electronic supplementary material The online version of this article (doi:10.1007/s10858-013-9772-4) contains supplementary material, which is available to authorized users.

R. Vernon · O. F. Lange (✉)
Department Chemie, Biomolecular NMR and Munich Center for Integrated Protein Science, Technische Universität München, Lichtenbergstrasse 4, 85747 Garching, Germany
e-mail: oliver.lange@tum.de

R. Vernon · D. Baker
Department of Biochemistry, University of Washington, Seattle, WA 98195, USA

Present Address:
R. Vernon
Program in Molecular Structure and Function, Hospital for Sick Children, Toronto, ON M5G 1X8, Canada

Y. Shen
Laboratory of Chemical Physics, National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Bethesda, MD 20892-0510, USA

D. Baker
Howard Hughes Medical Institute, University of Washington, Seattle, WA 98195, USA

O. F. Lange
Institute of Structural Biology, Helmholtz Zentrum München, 85764 Neuherberg, Germany

Introduction

The CS-Rosetta methodology for protein structure determination from sparse data from nuclear magnetic resonance (NMR) spectroscopy (Raman et al. 2010b; Sgourakis et al. 2011; Lange and Baker 2012; Lange et al. 2012). Remarkably, for small proteins (<15 kDa) chemical shift data alone can be sufficient to yield 3D protein structures with near-atomic accuracy (Cavalli et al. 2007; Shen et al. 2008; Wishart et al. 2008). CS-Rosetta and related methods (Cavalli et al. 2007; Shen et al. 2008) are based on a fragment assembly algorithm (Kraulis and Jones 1987; Simons et al. 1997; Delaglio et al. 2000; Rohl et al. 2004). A fragment denotes a small continuous piece (typically comprising 3–15 residues) of protein backbone with defined 3D structure, which is given by its ϕ, ψ and ω torsion angles. Given libraries of fragments starting at each residue position of the target protein's backbone, the fragment assembly algorithm can efficiently generate a wide variety of compactly folded structural models. A small percentage of accurate fragments usually suffices to

generate a few models close to the native protein structure in a large-scale sampling approach. After all-atom refinement of the fragment-assembled models, CS-Rosetta can often identify the correct models based on its energy function augmented by a comparison of back calculated chemical shifts with the experimental data (Shen et al. 2008).

Fragment picking for CS-ROSETTA was originally carried out using the fragment picker of the multiple fragment replacement (MFR) method of the NMRPipe software package (Delaglio et al. 1995), which combined chemical shift information with peptide sequence matching to score fragment candidates. MFR selects fragments from a database of crystal structures based on three scores, (1) chemical shift similarity between the target's experimentally measured shifts and values predicted for the database structures (CS), (2) sequence identity between target and database (Identity), and (3) the phi/psi probability of the database angles given the target's sequence (Rama). These scores are weighted such that the CS component dominates, which indirectly leads to a strong and largely accurate constraint on secondary structure during fragment selection. However, in contiguous regions with incomplete chemical shifts, the relative contribution of the chemical shift score was decreased, and without its constraint on secondary structure the Rama score term ended up providing a bias towards helices in all cases. This issue is mitigated in the hybrid-CS-Rosetta fragment picker (Shen et al. 2009a), where fragments for regions with insufficient chemical shift data are instead selected using the Rosetta2 fragment picker, R2FP:NNMAKE (Simons et al. 1997; Rohl et al. 2004), which picks fragments based on sequence profile (Altschul et al. 1997) and sequence based secondary structure predictions (Jones 1999; Meiler et al. 2001; Karplus et al. 2003). This substitution removes the helical bias of the MFR Rama score, but introduces sequence based secondary structure assignments, which have the potential to disagree with the experimental data.

In the present work, a new algorithm is introduced which combines salient features of both original algorithms, MFR and R2FP:NNMAKE, while introducing new concepts for the scoring of possible fragment candidates. The new fragment picker, denoted R3FP in the following, was specifically designed with the goal of only providing constraints where they are justified by experimental data. In contrast to its predecessors, R3FP selects a diverse set of fragments in regions where the data is limited, only constrained by what is reasonable given the sequence. We also used this as an opportunity to recode the MFR fragment picker in C++ within the Rosetta3 software framework (Leaver-Fay et al. 2011), producing a versatile platform for future development. This should allow to implement and test alternative approaches to fragment picking (Kalev and Habeck 2011) in conjunction with chemical shift data.

Methods

Development

The CSRosetta3 fragment picker was developed over a series of five steps, where scores were added, reformulated, and reweighted in order to improve the fragment quality as measured by their backbone RMSDs to the target reference structures. Two main quality metrics were used when comparing protocols, the average RMSD of the best 5 % of fragments at each residue position, and the average of the worst 5 % (Fig. 1). The weight ranges tested for each final score are described in Suppl. Tab. 1.

Fragment benchmark

A set of training protein targets was assembled from three previous studies, (1) the original MFR training set (Kontaxis et al. 2005), (2) a previous benchmark of CS-Rosetta together with pseudocontact shift data (PCS; Schmitz et al. 2012), and (3) targets from the CASD experiment (Rosato et al. 2012). The R3FP fragment picker will be evaluated by comparing against MFR fragments. For original MFR targets and PCS targets we use homolog filtering with comprehensive lists of homologs (Schmitz et al. 2012). The CASD targets come from a blind prediction challenge, and for those targets no homolog filtering was used. The benchmark targets are further described in Suppl. Tab. 2.

RMSD quality metrics

Fragment quality was quantified by the coordinate RMSD of all backbone heavy-atoms when compared to the reference structure at the respective backbone position. The average fragment quality is given by the average RMSD over the entire population of fragments, while *best* and *worst* fragment metrics are given by the average RMSD over either the best or worst 5 % of fragments by RMSD at a given position.

Secondary structure quality metric

The secondary structure of each residue in a given fragment is taken from the DSSP assignment (Kabsch and Sander 1983) in the crystal structure the fragment derives from. These assignments were compared against the DSSP assignments in the target reference structure. The percentage of fragment-residues whose assignment matches those of the corresponding target-residues yields the secondary structure quality metric of a fragment.

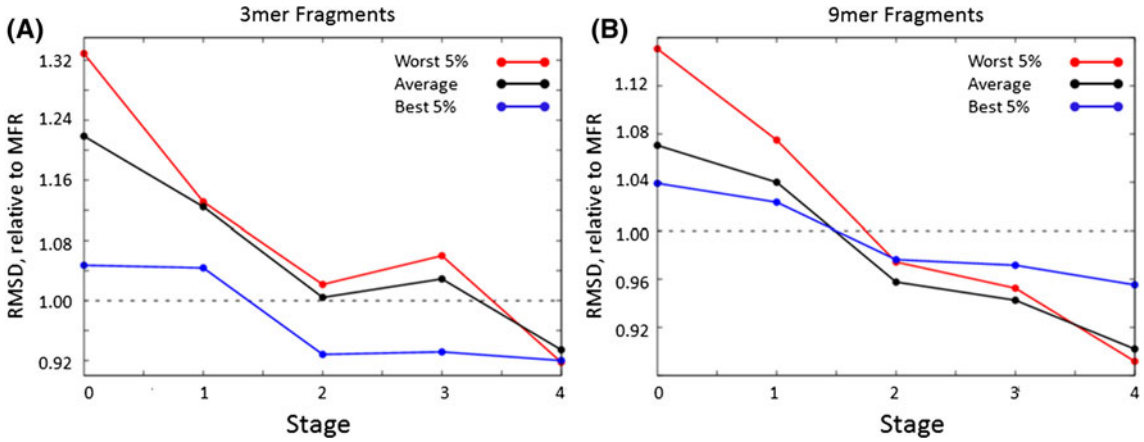

Fig. 1 Fragment RMSDs during the development process. The RMSD-fragment metrics ("Methods") were computed at each stage during the development process, and their relative change compared to the MFR reference. The stages reflect discrete steps during the development phase and correspond to (0) MFR's original CS score (Eq. 1), no sequence or RAMA components (1) R3FP's new sigmoidal CS Score (Eq. 3) (2) Added preliminary versions of sequence (Eq. 4), Ramachandran (Eq. 6), and TALOS+ secondary structure (Eq. 7) scores. (3) Optimized A and B constants in sigmoidal CS score (Eq. 1) (4) Added TALOS+ phi/psi score (Eq. 8) and reformulated secondary structure score (Eq. 9). The individual stages are discussed in detail in the "Results" section

Structural modeling benchmark

CS-Rosetta was run using standard command lines (Shen et al. 2008) using either the original MFR or the new R3FP fragments. For the fragment picker benchmark a diverse assortment, based on length, secondary structure, and CS-Rosetta's performance, was selected from targets used in previous studies involving the MFR fragment picker (SI Table).

Ca. 4000 CPU-hours were devoted to each run, which produces different numbers of models, depending on the target's size and Rosetta's ability to produce models that satisfy its basic quality filters, ranging from roughly six thousand to forty thousand models. C_α-RMSDs to the reference structures were calculated to assess model quality. A bootstrap analysis was used to determine the standard deviation of the best 0.1 % by C_α-RMSD.

Fragment picking (final protocol)

The fragment picker uses the default *bounded protocol* as in the previously reported Rosetta3 generalized fragment picker (Gront et al. 2011), which prepares 9-mer and 3-mer fragment files by selecting the lowest scoring 200 fragments for each residue position. A database of ~2.3 million fragment candidates has been generated from 9,523 proteins, comprising the same set as used previously (Shen et al. 2009a). Derived data have been attached to each database residue as follows: Sequence profiles from PSI-BLAST (Altschul et al. 1997), predicted secondary chemical shifts from SPARTA+ (Shen and Bax 2010) and secondary structure assignments from DSSP (Kabsch and Sander 1983). For each target, every residue in the database is scored against every residue in the target sequence using the five independent scoring functions detailed below. The calculations use the same input data as the previous MFR (Kontaxis et al. 2005) method, but scoring is performed differently as detailed in the following. Instructions for running the final protocol are provided in Suppl. Text 1.

Results

Development process

Stage 1: rebuilding the CS score on its own

MFR CS-score: M_δ We began by seeking to optimize the selection of fragments using chemical shifts alone. To compute the MFR CS-Score, the experimental backbone ($C,C_\alpha,C_\beta,N,H_N,H_\alpha$) chemical shifts for each target are converted into secondary shifts (Wishart et al. 1995) and scored for similarity versus the database's predicted secondary shifts (Kontaxis et al. 2005).

$$M_\delta = \sum_{\text{shifts}}^{N_{\text{DB}}} \left(\frac{\delta_T - \delta_{\text{DB}}}{\Delta\delta_{\text{DB}}} \right)^2 \quad (1)$$

with δ_T, δ_{DB} and $\Delta\delta_{\text{DB}}$ as the target shift, database shift, and prediction error, respectively.

Analysis of fragment candidate ranking using the MFR chemical shift score M_δ revealed that precise matches of all the data dominate in the better half of fragments; differences in a few non-matching data points dominate the ranking within the worse half of fragments. This results

from the use of a harmonic potential for the error between predicted and calculated chemical shifts, which results in high penalties for a few or even individual badly matching data points.

CS-score: I_δ To avoid distorting the overall score by a few non matching positions, we experimented with a sigmoid potential,

$$S(x,a,b) = \frac{1}{1+e^{-ax+b}} \tag{2}$$

that better reflects the error distribution of deviations than the original harmonic penalty function (Eq. 1; Kontaxis et al. 2005) and final scores range from ~ 0 for a perfect match, to ~ 1 for a failure to match. When comparing fragments against each other, this serves to rank fragments based on how many database shifts match the target, ignoring the extent of differences once they fail to match.

To count the number of database shifts δ_{DB} that failed to match the target chemical shifts δ_T within their prediction error $\Delta\delta_{\mathrm{DB}}$, we define

$$I_\delta = \frac{N_T}{N_{\mathrm{DB}}} \sum_{\mathrm{shifts}}^{N_{\mathrm{DB}}} S\left(\frac{|\delta_T - \delta_{\mathrm{DB}}|}{\Delta\delta_{\mathrm{DB}}}, 2, 4\right) \tag{3}$$

with S the sigmoid function defined above (Eq. 2), and N_{DB} and N_T the number of comparable shifts available in the database or target, respectively. Systematically lower scores for residues where C_β shifts are compared against database glycines are avoided by normalizing with N_T/N_{DB}.

As shown in Fig. 2, the deleterious effect of individual bad matches to the input data that dominated the harmonic CS-score in MFR leads to a large number of bad matching (by RMSD) outlier fragment candidates that score significantly lower than the correct fragment candidates. This problem is significantly reduced using the new CS-Score. With the improved score, high-RMSD fragment candidates no longer score significantly lower than low-RMSD fragment candidates, and only a handful of high-RMSD candidates score similar to low-RMSD fragment candidates. This improvement is reflected in the overall performance of the new CS-score as indicated in Stage 1 in Fig. 1: worst fragment RMSD is significantly improved, from 133 to 113 % of the full MFR reference for 3mers and from 115 to 107 % for 9mers.

Stage 2: addition of sequence profile, Ramachandran, and secondary structure scores

We next incorporated score terms for sequence identity (Profile), Ramachandran compatibility (Rama) and secondary structure (SS) as present in the original fragment pickers MFR and R2FP:NNMAKE.

ProfileScore: I_p For sequence identity, the Profile-Score from R2FP:NNMAKE (Rohl et al. 2004; Shen et al. 2009a) was used, which compares the sequence profiles of two residues according to their L1 block distance. It is given by the Manhattan distance of the target residue's sequence profile P_T, computed using PSIBLAST, to the candidate residues profile P_{DB}.

$$I_p = \sum_{aa} |P_T - P_{\mathrm{DB}}| \tag{4}$$

This is superior to MFR's matrix based sequence identity score (Kontaxis et al. 2005), as it compares residues based on conserved sequence profiles rather than just residue similarity.

RamaScore: M_R and I_R MFR's Ramachandran score computes the log-likelihood of a candidate residues ϕ,ψ positions given the residue type (aa) at the respective sequence position. Accordingly,

$$M_R = \sum_{k\in\{h,e,l\}} -\log\left[\frac{R(\phi,\psi,k,aa)}{R_{\max}(aa)}\right], \tag{5}$$

where $R(\phi,\psi,k,aa)$ is the Ramachandran density for the given residue type aa, secondary structure type k and torsion angles ϕ,ψ. $R_{\max}$ (aa) is the maximum observed density for that residue type.

For the new fragment picker, this score was reformulated as the sigmoid of the density in order to flatten out the extremes and reduce its helical bias. Additionally, we weighted the score-contributions from different secondary structure types using the TALOS+ predicted (Shen et al. 2009b) secondary structure propensities P_{TSS} for helix, sheet and loop (h,e,l). This yields,

$$I_R = S\left(\sum_{k\in\{h,e,l\}} \log[R(\phi,\psi,k,aa)P_{\mathrm{TSS}}(k)], 1, 0\right) \tag{6}$$

evaluating to near-zero for allowed angles, and near-one for angles that are unlikely to occur.

SS-similarity-score Finally, the secondary structure score from R2FP:NNMAKE was added but with the chemical shift based secondary structure profiles P_{TSS} as its input, instead of the purely sequence based prediction used by R2FP:NNMAKE. We compute

$$I_{SS} = 1 - P_{TSS}(k_{DSSP}) \tag{7}$$

where k_{DSSP} denotes the DSSP (Kabsch and Sander 1983)-assigned secondary structure (helix, sheet or loop).

Optimization stage 2 Weights were chosen by first matching each scores dynamic range to the equivalent ranges observed for MFR, and then varying individual

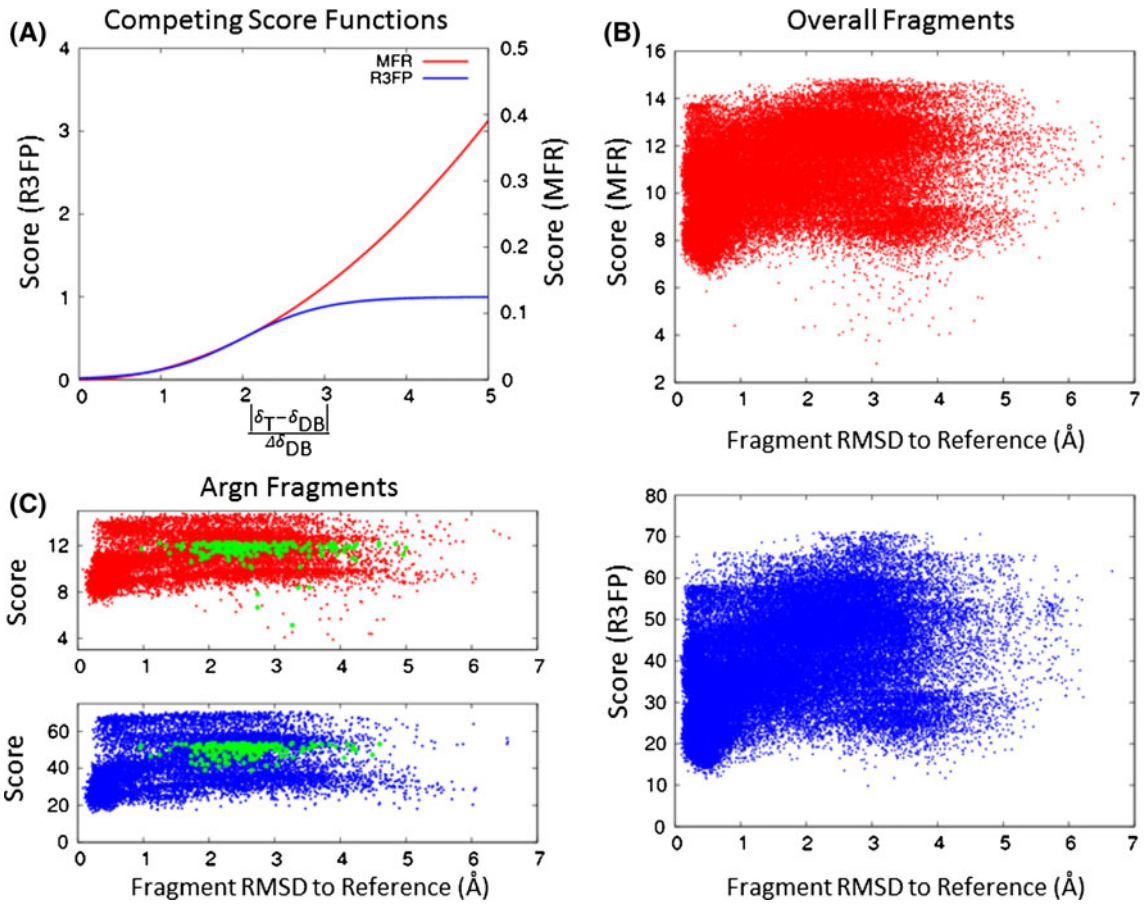

Fig. 2 Comparison between the harmonic MFR CS-score and the sigmoidal R3FP version. **a** The harmonic (Eq. 1) versus sigmoidal CS score (Eq. 3) functions are plotted against normalized shift deviation to demonstrate that the curves are similar when distances between observed and database shifts are low while the sigmoidal score soon plateaus as the distances grow larger. **b** Fragment CS-scores versus fragment RMSD to reference structure are shown for all fragments in the MFR (*red*) and R3FP (*blue*) sets, for fragments containing at least four shifts per residue. Overall the scores are similar, though the R3FP score (Eq. 3) has significantly less low scoring fragments above 2 Å. **c** The fragment scores versus RMSD, as before, but from the single target ARGN, with fragments from residue 38 being highlighted in green to show how the subtle differences between scores appear in different contexts. Low scoring, high RMSD outliers appear to be a common problem with MFR

scores in order to manually reduce best fragment RMSDs. The additional information provided by these scores improved best, average and worst fragment RMSD metrics (Stage 2 of Fig. 1). The best fragments were 2.4 and 7.2 % lower than the MFR reference for the 9mers and 3mers, respectively, while the average fragment RMSD was 4.3 % lower for the 9mers and equal to the reference for the 3mers.

Stage 3: CS score optimization

During the initial development of the CS-Score we assigned preliminary constants for the sigmoid potential by roughly fitting the lowest scoring region (highest similarity) to the harmonic used by MFR. In order to empirically determine the best slope and inflection point for discriminating between shifts in the context of the scores added in stage 2 we then optimized the weight constants of the sigmoidal CS-Score. We performed a grid search of the two constants which define the sigmoid curve $S(x,a,b)$ of the CS-Score I_δ defined above (Stage 1). We varied a and b in the ranges 1–5 and 3–7, respectively, with steps of 1 (Suppl. Fig. 1). The constants control which shifts are considered as matched, unmatched, or partially matched, and set the slope for discriminating between different partial matches. The optimized constants resulted in an overall improvement to the 9mers but made the 3mers worse by a similar margin (Stage 3 of Fig. 1). Despite the bad performance on 3mers the optimized constants were used in the final protocol since 9mers make a larger contribution towards final structure quality in Rosetta fragment assembly (Handl et al. 2011).

Stage 4: addition of phi/psi score and secondary structure score reformulation

Phi/psi-squarewell: $I_{\phi\psi}$ A fifth score was added to incorporate the chemical shift based phi/psi predictions that TALOS+ provides alongside the secondary structure predictions. The angular-distance d between TALOS+ (Shen et al. 2009b) predictions ϕ_T, ψ_T and candidate angles ϕ_{DB}, ψ_{DB} is calculated as the number of TALOS+ prediction errors, $\Delta\phi_T, \Delta\psi_T$ past the first error bar.

$$I_{\phi\psi} = \sqrt{S\left(\frac{\max(0, \mathrm{d}(\phi_\mathrm{T}, \phi_\mathrm{DB}) - \Delta\phi_\mathrm{T})}{\Delta\phi_T}, 2.5, 5\right) + S\left(\frac{\max(0, \mathrm{d}(\psi_\mathrm{T}, \phi_\mathrm{DB}) - \Delta\psi_\mathrm{T})}{\Delta\psi_T}, 2.5, 5\right)} \quad (8)$$

Preliminary results using a harmonic penalty resulted in improved fragment quality in regions where already accurate fragments where identified, but reduced it significantly where accurate fragment picking was already difficult. As previously for the CS-Score, we found that individual bad predictions would dominate the scores, causing partially matching fragments to be discarded. Thus, we reformulated the score using a sigmoid to flatten the extremes.

Revised SS-similarity: I_{TSS} In the refined and final version of the new TALOS+ secondary structure score, replacing I_{SS} (Eq. 7), a sigmoid function is used, and the probability of the occurrence of k_DSSP is weighted with the TALOS+ (Shen et al. 2009b) secondary structure prediction's confidence C_TSS to yield

$$I_\mathrm{TSS} = \sqrt{C_\mathrm{TSS}} S(P_\mathrm{TSS}(k_\mathrm{DSSP}), 7, 5) \quad (9)$$

Rank weighted fragment scores The Phi/Psi-SquareWell (Eq. 8) and TALOS-SS-Similarity (Eq. 9) scores can yield large penalties for individual residues which disagree with the predictions, which is not necessarily desirable when the overall match is otherwise still good. To down-weight these individual outliers and improve the odds of picking fragments that disagree with the predictions at only one or two positions the individual residue scores are ranked and a weight is assigned for each rank $r \in \{1 \ldots M\}$. With M being the fragment length, we compute the rank-dependent weight as

$$w(r) \equiv M^{-1} S(r, -10, -7). \quad (10)$$

This step has the effect of consistently down-weighting the contribution made by the worst residue scores when calculating the final fragment score. Finally, the fragment scores are normalized by fragment size M to decrease their relative contribution as fragment size increases.

Evaluating improvements of stage 4 The addition of phi/psi constraints and reformulation of the secondary structure constraints improved all three score metrics for both 3mers and 9mers (Stage 4 of Fig. 1). The improvement was larger for the average and worst fragments, showing that this formulation of the constraints primarily serves to reject bad fragments. For 9mers, this scoring scheme resulted in a 10–11 % improvement over MFR for the *worst-* and *average fragment* metric, and an 8 % improvement for the *best fragment* metric. For 3mers, the improvements were 6 % for *average-*, and 8 % for the *worst-* and *best fragment* metric.

As a final check, because the phi/psi and secondary structure predictions used in this stage are derived from the chemical shift data we tested their independence from each other and from the chemical shift score (Eq. 3), showing that the scores contribute to fragment quality in different ways and that the addition of the individual score provides a cumulative benefit (Suppl. Fig. 2).

Final fragment quality

In its final form, the score of a fragment at position k comprising of M residues is obtained by

$$M^{-1} \sum_{i=k}^{M+k-1} I_\delta^{(i)} + 1.5 I_P^{(i)} + I_R^{(i)} + 0.25 I_\mathrm{TSS}^{(i)} w\left(r_\mathrm{TSS}^{(i)}\right) + 5.0 I_{\phi\psi}^{(i)} w\left(r_{\phi\psi}^{(i)}\right) \quad (11)$$

The weights were optimized for picking fragments with low C_α-RMSD against reference structures as well as for sampling low C_α-RMSD conformations in CS-Rosetta structure calculations (data not shown). The targets used for development of the fragment picker were not used in the final benchmark set.

For both 9mer and 3mer fragments, less fragments with RMSDs of 0.2–0.3 Å per residue and more fragments with top-RMSDs (<0.15 Å per residue) are generated compared to MFR (Fig. 3). The difference is less pronounced in the 3mers, probably because the proportion of 3mers in the 0.2–0.3 Å range was already low in MFR.

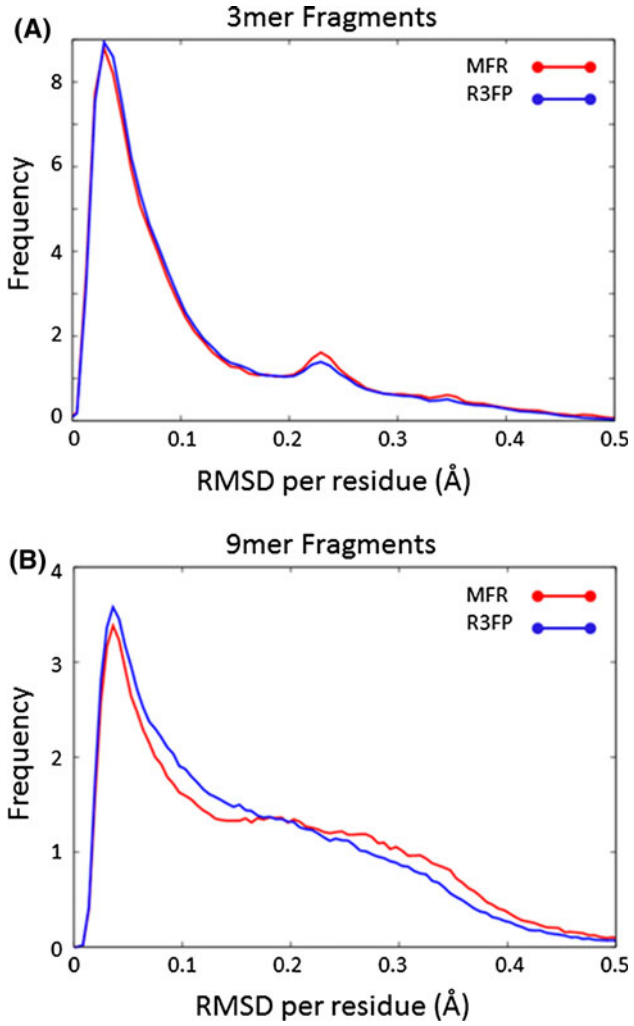

Fig. 3 Fragment RMSD, MFR versus R3FP. Full backbone fragment RMSDs are shown for the entire set of **a** 3mers and **b** 9mers, with fragments selected by MFR in *red* and by the final version of R3FP in *blue*. Both fragment sizes show a shift to lower RMSDs when using R3FP, but the effect is larger for 9mers

Because fragment RMSD alone is not sufficient to predict the impact of the fragments on 3D structure generation in fragment assembly, we also quantified the secondary structure accuracy of the fragments (Fig. 4). As expected, using the chemical-shift based secondary structure predictions of TALOS+ for fragment selection (Eq. 9) has a significant and positive impact on this metric. For both, 3mers and 9mers, the frequency of fragments with 80–100 % correct secondary structure assignments is increased from ~60 % for MFR to ~80 % in R3FP (Fig. 4a). More importantly, focusing on loop-residues one finds in the 80–100 % category the dramatic improvement in frequency from ~20 % for MFR to ~70 % for R3FP (Fig. 4b). Thus, we were able to remove a strong anti-loop bias that existed in MFR. A drawback of coordinate RMSD as a quality metric is that compact structures such as helices have an advantage relative to extended structures such as loops, thus it is possible that some of the observed helical bias in MFR stems from training exclusively towards a reduction in RMSDs.

It is illuminating to examine the fragment quality at individual positions rather than as an aggregated histogram. Some representative examples for targets VpR247, argn and Ncalmodulin, are shown in Fig. 5. The improvements do not result in a general shift towards slightly more accurate fragments, but rather result in dramatic improvements of fragment accuracy in some of those regions, often where no useful fragments were found by MFR. The opposite case, that regions with reasonably accurate MFR fragments are lost when switching to R3FP, is rarely seen.

The improvements in the 9mers tend to cover a wide range of positions, while the differences between MFR and R3FP in 3mers change rapidly with residue position (Fig. 5). This behavior corrobates our previous analysis, which suggested that MFR is overly sensitive to either bad data or bad predictions on specific residues. For instance, the narrow spike in 3mer quality from residue 31–33 in NCalmodulin (Fig. 5c) is broadened in the 9mer quality to residues 25–33 whose 9mer fragments all overlap with residue 33, too. The scoring of both, 3mers and 9mers, in MFR is affected by problems at the single position 33 of NCalmodulin.

CS-Rosetta benchmark comparison

To test the performance of the fragment libraries in actual 3D structure generation we ran CS-Rosetta fragment assembly on 23 targets for both, MFR and R3FP fragments. The most important criterion for the success of CS-Rosetta is how close to the native structure the fragment assembly stage can sample. We selected the lowest 0.5 % by Ca-RMSD and compute the average RMSD over this set (Fig. 6), which we call *spearhead-RMSD* in the following. A bootstrap analysis was used to calculate the standard deviation of the spearhead-RMSD, and targets were labeled *better* (green) or *worse* (red) if it decreased or increased by more than two standard deviations, respectively. The new fragment picker improves model quality for 6 of the 23 targets tested, and decreases it for 3. The remaining 14 targets maintain a similar RMSD. The percentage of targets sampled within 2 Å to the reference is increased from 29 to 42 %, and the percentage of targets where sampling is stuck over 4 Å is decreased from 38 to 29 %. Improved spearhead-RMSDs coincide with an overall shift towards lower RMSDS (Fig. 7a–f) as well as a shift towards lower rosetta energies (Suppl. Fig. 3), indicating an overall improvement in the models generated by fragment assembly using the R3FP fragments. The R3FP fragment picker protocol has also been tested on an independent set

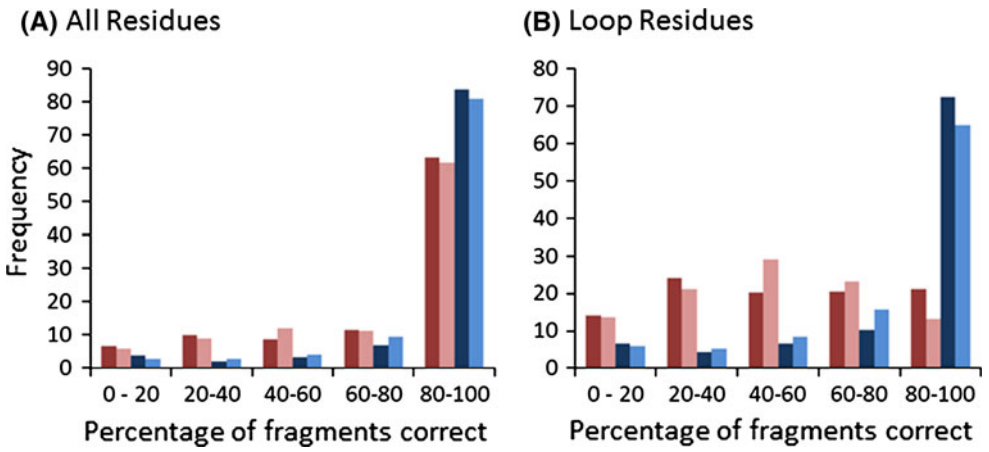

Fig. 4 Secondary Structure Accuracy, MFR versus R3FP. For each residue we determine the percentage of the selected fragment candidates that have the correct DSSP assignments. A histogram for 3mers (*Dark tone*) and 9mers (*Light tone*) produced by both MFR (*Red*) and R3FP (*Blue*) shows how often the fragment's secondary structure assignments are accurate. **a** R3FP shows a marked improvement in accuracy on the overall benchmark set, where the number of residues with a >80 % match to the reference increased from 60 to 80 % in both the 9 and 3mers. **b** This improvement is more dramatic when restricted to target residues which DSSP assigns as loop or as unstructured. MFR appears to assign a large amount of wrong secondary structure to these residues, and when using R3FP the number of residues obtaining a >80 % match to the reference increased from 12 and 21 % to 65 and 85 %, for 3- and 9mers, respectively

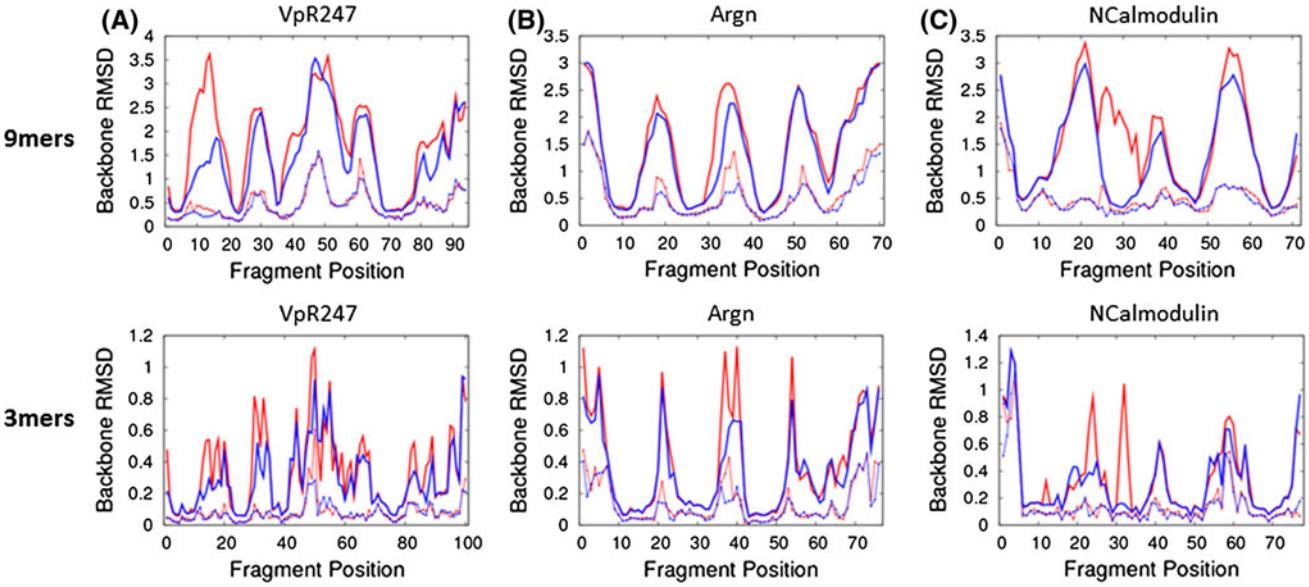

Fig. 5 Fragment RMSDs of representative targets. Fragment RMSDs of MFR (*red*) and R3FP (*blue*) fragments of targets **a** VPR247, **b** Argn, and **c** NCalmodulin, respectively. *Solid lines* show the average full backbone RMSD at each fragment position, while the *dashed lines* show the lowest RMSD values observed

of targets, not used during training, with similar results (van der Schot et al. 2013).

Discussion

For fragment assembly to succeed, the fragment picker needs to exclude non-productive fragments to narrow down the search (Handl et al. 2011), while not making the ultimate mistake of excluding the correct structure itself. These can be assessed using two fragment quality measures, *fragment coverage*, which assesses whether the fragment space includes the local structure of the target, *and fragment accuracy*, which becomes lower as more and more fragments deviate from the reference structure. Given any possible fragment metric, its best values inform on the fragment coverage, while its worst inform on accuracy.

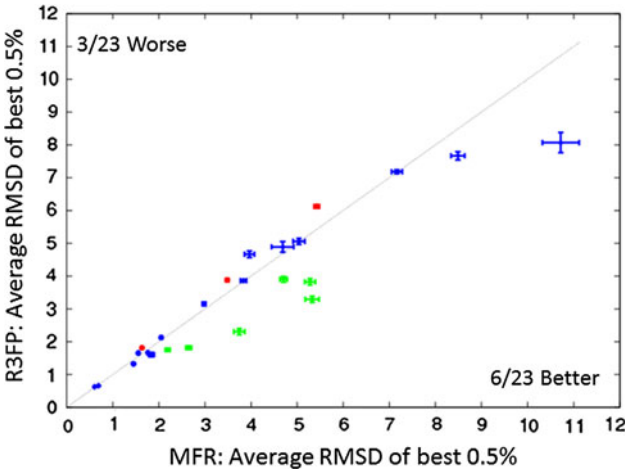

Fig. 6 Final fragment assembly benchmark. Compared are the average Cα-RMSDs of the lowest 0.5 % by RMSD models (spearhead-RMSD) for each target when using MFR (*x*-axis) or R3FP (*y*-axis) fragments to generate models. *Errors* were calculated by the standard deviation observed in a bootstrap analysis, and individual targets were assigned as better or worse when the difference in spearhead-RMSD was >4 standard deviations. From this analysis, 6 of 23 targets show significant improvement, and 3 targets get significantly worse when using R3FP fragments

Generally, sampling space increases exponentially with sequence length; so insufficient accuracy is the primary reason why the naive fragment assembly algorithm of CS-Rosetta fails for larger proteins. This problem, however, can be overcome by using more sophisticated sampling algorithms (Lange and Baker 2012) and additional sparse restraints (Raman et al. 2010a; Lange et al. 2012). Inaccurate constraints within the fragments, however, are hard to overcome even with more thorough sampling methods. Moreover, we observed that some rarely picked fragments can provide a linchpin (Kim et al. 2009) and thus loss of coverage in those (few) positions does not compensate for gain of accuracy in the majority of positions.

The trade-off between accuracy (excluding less likely fragments) and coverage (allowing all admissible fragments) thus has to be chosen carefully. Unfortunately, probing this trade-off requires carrying out the actual fragment assembly to observe the functional utility of the fragments during structure generation itself. A better understanding of how to rank the functional utility of fragments a priori, would greatly facilitate the training of a fragment picker. Lacking such a hypothetical metric,

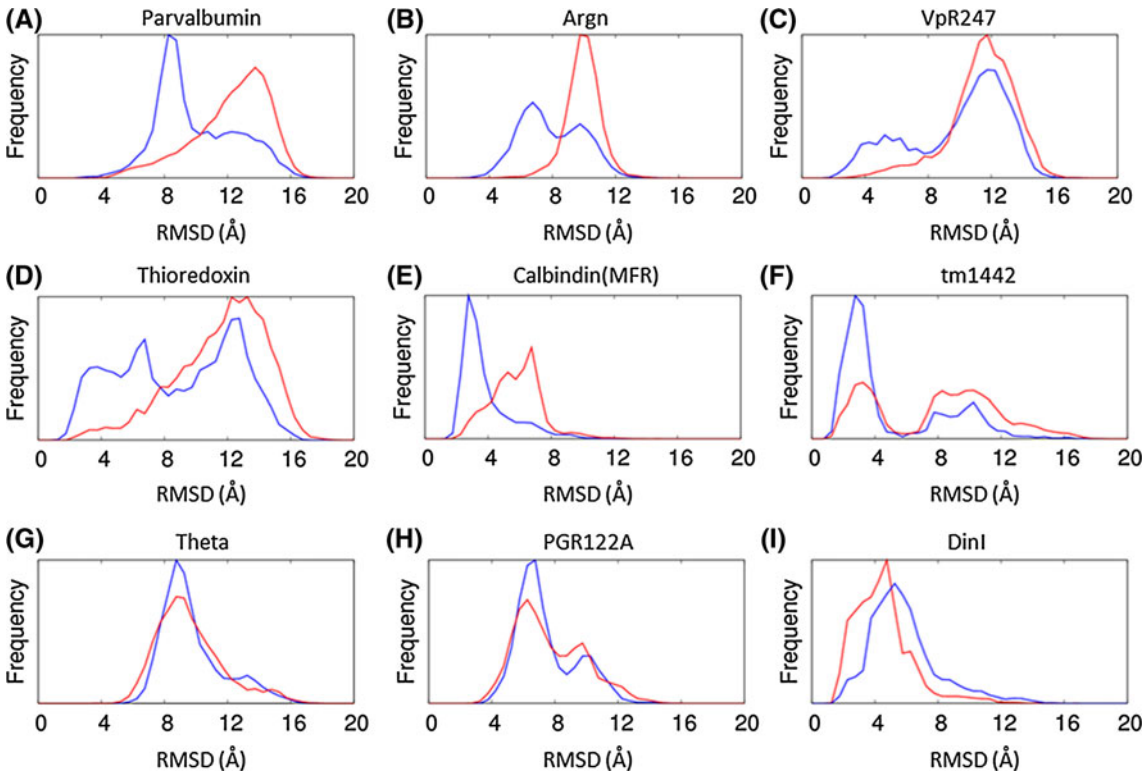

Fig. 7 RMSD distribution of CS-Rosetta generated models. Histograms show the distribution of RMSD to reference structure for models generated with CS-Rosetta from R3FP (*blue*) and MFR (*red*) fragments, respectively. **a**–**f** Six significant improvements for R3FP. **g**–**i** Three targets that got slightly worse using R3FP fragments

however, we primarily optimized against the best fragment RMSD while ruling out changes that cause large disturbances in secondary structure metrics. Additionally, we frequently carried out full structure calculations to check on the behavioral changes of fragment assembly caused by changes to the fragment selection.

A general problem we detected in the original MFR and R2FP:NNMAKE methods is the overuse of probability-based empirical scoring terms such as Eq. 5, which yield the penalty score as $-\log(p)$, where p denotes the probability of the scored feature in a database of known structures. We found two problems associated to such score terms. First, rare fragment conformations are highly penalized, which often leads to removal of important linchpin fragments. Second, double-counting occurs, since naturally the database contains more fragment candidates with typical conformations than with rare conformations, which already gives them a competitive advantage based on their probability.

The overuse of $-\log(p)$-scores is a pitfall that was hard to detect at first, since predictions improve overall, if the predictor is biased towards the more likely features. Indeed, the bias towards helical fragments in the MFR picker has a positive effect on average fragment RMSDs, since choosing helices is often a good guess if nothing better is known. Moreover, helices are compact and thus yield lower RMSD than extended conformations that are equally wrong. Our experience with fragment assembly showed, however, that guessing with a bias towards common fragments is not as effective as keeping unclear stretches of the target sequence unconstrained. This strategy demands more from the sampling algorithm, but entails less danger of catastrophic failure.

Consequently, while trying to improve both coverage and accuracy, we ended up primarily flattening scores (Fig. 2a), increasing tolerance for bad matching input (resonances, phi/psi-predictions, and secondary structure predictions), and overall reducing the reliance on $-\log(p)$-scores. The final score now just sums up how many comparisons are in violation of the data, while relying for the prior information about fragment conformations more on their frequency in the candidate database than on additional $-\log(p)$-rewards.

Conclusions

We have improved the overall quality of the CS-Rosetta fragment picker by increasing its robustness against missing or difficult input data, and by adding the highly accurate secondary structure and backbone torsion angle predictions from TALOS+. This new protocol shows an overall improvement in the quality of the generated fragments and in the quality of models produced by CS-Rosetta. The new code base also simplifies the task of adding new score terms, constraints, or experimental data to the existing protocol, setting the groundwork for future development.

The new fragment picker algorithm is integrated into the CS-Rosetta software available at www.csrosetta.org and is activated by the *pick_fragments* command.

Acknowledgments Approximately half a million CPU hours donated from the public to the Rosetta@Home project on BOINC made this project possible.

References

Altschul SF, Madden TL, Schäffer AA, Zhang J, 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

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

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 6:277–293

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

Gront D, Kulp DW, Vernon RM, Strauss CEM, Baker D (2011) Generalized fragment picking in Rosetta: design, protocols and applications. PLoS ONE 6:e23294

Handl J, Knowles J, Vernon R, Baker D, Lovell SC (2011) The dual role of fragments in fragment-assembly methods for de novo protein structure prediction. Proteins 80:490–504

Jones DT (1999) Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 292:195–202

Kabsch W, Sander C (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22:2577–2637

Kalev I, Habeck M (2011) HHfrag: HMM-based fragment detection using HHpred. Bioinformatics 27:3110–3116

Karplus K, Karchin R, Draper J, Casper J, Mandel-Gutfreund Y, Diekhans M, Hughey R (2003) Combining local-structure, fold-recognition, and new fold methods for protein structure prediction. Proteins 53(Suppl 6):491–496

Kim DE, Blum B, Bradley P, Baker D (2009) Sampling bottlenecks in de novo protein structure prediction. J Mol Biol 393:249–260

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

Kraulis PJ, Jones TA (1987) Determination of three-dimensional protein structures from nuclear magnetic resonance data using fragments of known structures. Proteins 2:188–201

Lange OF, Baker D (2012) Resolution-adapted recombination of structural features significantly improves sampling in restraint-guided structure calculation. Proteins 80:884–895

Lange OF, Rossi P, Sgourakis NG, Song Y, Lee H-W, Aramini JM, Ertekin A, Xiao R, Acton TB, Montelione GT et al (2012) Determination of solution structures of proteins up to 40 kDa using CS-Rosetta with sparse NMR data from deuterated samples. Proc Natl Acad Sci USA 109:10873–10878

Leaver-Fay A, Tyka M, Lewis SM, Lange OF, Thompson J, Jacak R, Kaufman K, Renfrew PD, Smith CA, Sheffler W et al (2011) ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. Meth Enzymol 487:545–574

Meiler J, Müller M, Zeidler A, Schmäschke F (2001) Generation and evaluation of dimension-reduced amino acid parameter representations by artificial neural networks. J Mol 7(9):360–369

Raman S, Huang YJ, Mao B, Rossi P, Aramini JM, Liu G, Montelione GT, Baker D (2010a) Accurate automated protein nmr structure determination using unassigned NOESY data. J Am Chem Soc 132:202–207

Raman S, Lange OF, Rossi P, Tyka M, Wang X, Aramini JM, Liu G, Ramelot TA, Eletsky A, Szyperski T et al (2010b) NMR structure determination for larger proteins using backbone-only data. Science 327:1014–1018

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

Rosato A, Aramini JM, Arrowsmith C, Bagaria A, Baker D, Cavalli A, Doreleijers JF, Eletsky A, Giachetti A, Guerry P et al (2012) Blind testing of routine, fully automated determination of protein structures from NMR data. Structure 20:227–236

Schmitz C, Vernon R, Otting G, Baker D, Huber T (2012) Protein structure determination from pseudocontact shifts using ROSETTA. J Mol Biol 416:668–677

Sgourakis NG, Lange OF, DiMaio F, Andre I, Fitzkee NC, Rossi P, Montelione GT, Bax A, Baker D (2011) Determination of the structures of symmetric protein oligomers from NMR chemical shifts and residual dipolar couplings. J Am Chem Soc 133:6288–6298

Shen Y, Bax A (2010) SPARTA plus: a modest improvement in empirical NMR chemical shift prediction by means of an artificial neural network. J Biomol NMR 48:13–22

Shen Y, Lange OF, Delaglio F, Rossi P, Aramini JM, Liu G, Eletsky A, Wu Y, Singarapu KK, Lemak A et al (2008) Consistent blind protein structure generation from NMR chemical shift data. Proc Natl Acad Sci USA 105:4685–4690

Shen Y, Delaglio F, Cornilescu G, Bax A (2009a) TALOS+: a hybrid method for predicting protein backbone torsion angles from NMR chemical shifts. J Biomol NMR 44:213–223

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

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

van der Schot G, Zhang Z, Vernon R, Shen Y, Vranken WF, Baker D, Bonvin AMJJ, Lange OF (2013) Improving 3D structure prediction from chemical shift data. J Biomol NMR. doi:10.1007/s10858-013-9762-6

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:W496–W502