J Biomol NMR (2013) 57:27–35
DOI 10.1007/s10858-013-9762-6

ARTICLE

Improving 3D structure prediction from chemical shift data

Gijs van der Schot · Zaiyong Zhang · Robert Vernon · Yang Shen · Wim F. Vranken · David Baker · Alexandre M. J. J. Bonvin · Oliver F. Lange

Received: 23 May 2013 / Accepted: 16 July 2013 / Published online: 3 August 2013

Abstract We report advances in the calculation of protein structures from chemical shift nuclear magnetic resonance data alone. Our previously developed method, CS-Rosetta, assembles structures from a library of short protein fragments picked from a large library of protein structures using chemical shifts and sequence information. Here we demonstrate that combination of a new and improved fragment picker and the iterative sampling algorithm RASREC yield significant improvements in convergence and accuracy. Moreover, we introduce improved criteria for assessing the accuracy of the models produced by the method. The method was tested on 39 proteins in the 50–100 residue size range and yields reliable structures in 70 % of the cases. All structures that passed the reliability filter were accurate (<2 Å RMSD from the reference).

Keywords Nuclear magnetic resonance · Protein structure calculation · CS-ROSETTA · Sparse data

Introduction

Knowledge of the three-dimensional (3D) structure of proteins at atomic accuracy is important to understand protein function, protein–ligand interactions and for

Gijs van der Schot and Zaiyong Zhang have contributed equally to this study.

Electronic supplementary material The online version of this article (doi:10.1007/s10858-013-9762-6) contains supplementary material, which is available to authorized users.

G. van der Schot · A. M. J. J. Bonvin (✉)
Computational Structural Biology, Bijvoet Center for Biomolecular Research, Faculty of Science—Chemistry, Utrecht University, Padualaan 8, 3584 CH Utrecht, The Netherlands
e-mail: a.m.j.j.bonvin@uu.nl

Z. Zhang · O. F. Lange (✉)
Biomolecular NMR and Munich Center for Integrated Protein Science, Department Chemie, 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

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

W. F. Vranken
Department of Structural Biology, VIB, Building E, 4th Floor, Pleinlaan 2, 1050 Brussels, Belgium

W. F. Vranken
Structural Biology Brussels, Vrije Universiteit Brussel, Building E, 4th Floor, Pleinlaan 2, 1050 Brussels, Belgium

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

rational drug design. Over the last two decades nuclear magnetic resonance spectroscopy (NMR) has become an established complement to X-ray crystallography for the determination of 3D structures. The most challenging bottleneck in determining NMR structures, the assignment of side-chain chemical shifts and of NOE cross-peaks, can be avoided with methods for computing structures from backbone-only NMR experiments. Backbone chemical shift values reflect a wide array of structural information including backbone and side-chain conformations, secondary structure, hydrogen bond strength, and the position of aromatic rings. This information can be exploited to predict the 3D structure of proteins using software packages such as CS-ROSETTA, CHESHIRE and CS23D (Cavalli et al. 2007; Shen et al. 2008; Wishart et al. 2008).

The convergence and reliability of CS-ROSETTA calculations have been shown to improve by utilizing additional NMR data, such as residual dipolar couplings (RDC) (Raman et al. 2010a), NOE-derived distance restraints (Lange et al. 2012) and pseudo-contact shifts (PCS) (Schmitz et al. 2012). In the context of available RDC and NOE data an iterative sampling scheme, RASREC (Raman et al. 2010a; Lange and Baker 2011; Lange et al. 2012), was shown to greatly extend the applicability towards larger protein structures. Here we introduce a number of algorithmic advances whose cumulative effect significantly improves reliability, convergence and accuracy of final structures for chemical shift-only calculations. Moreover, we describe the WeNMR (Wassenaar et al. 2012) webserver that accesses the European Grid Initiative (EGI, www.egi.eu) computational resources, allowing efficient CS-ROSETTA computations via the simplicity of a web interface to academic users.

The CS-ROSETTA methodology consists of three stages: (1) fragment picking, (2) sampling, and (3) model selection. Originally, backbone chemical shift information was only used in stages (1) and (3). Fragment picking for CS-ROSETTA was originally carried out using the MFR method of the NMRPipe software package (Delaglio et al. 1995; Lange et al. 2012) which combined chemical shift information with peptide sequence matching to score fragment candidates. However, for regions where no experimental data was present the ROSETTA2 method (Rohl et al. 2004; Schmitz et al. 2012) outperformed MFR (Shen et al. 2009b; Lange and Baker 2011). In the present work the chemical shift based fragment picking is incorporated directly into a new ROSETTA3 fragment picker. This new fragment picker (denoted R3FP in the following) combines salient features of both original algorithms (MFR and ROSETTA2) (RV, YS, DB and OFL; J Biomol NMR submitted). The performance of the new method is benchmarked on a set of target proteins that have not been used for development or optimization of the R3FP protocol.

RASREC is an iterative sampling algorithm that has been shown to significantly increase sampling efficiency for larger proteins (10–40 kDa), if additional restraint data such as RDCs and NOEs are used (Lange and Baker 2011; Lange et al. 2012). Instead of running 10,000 or more independent structure calculations with increased cycle number as in CS-ABRELAX [the standard CS-ROSETTA algorithm (Shen et al. 2008)], RASREC performs iterative batches of short simulations. Similar to a genetic algorithm, a pool of best performing structures is maintained throughout the iterative procedure and sampling is focused around previously identified conformations. It is crucial for the performance of RASREC that pseudo-energies (e.g., from RDC and NOE restraint data) be available to assist ROSETTA in predominantly selecting structures with native features for this pool. In this study we extend the RASREC algorithm to allow chemical shift rescoring of intermediate structures (CS-RASREC) and test the performance of this extended method.

Chemical shifts are dominated by local backbone conformations, and thus CS-ROSETTA structures based solely on chemical shifts tend to be locally correct but globally unconverged. Here we introduce a post-analysis procedure that identifies locally converged regions of the structure, which have been shown to be generally accurate (Rohl et al. 2004; Shen et al. 2008; Raman et al. 2010a). However, with decreasing convergence there is an increasing probability that also the conformation of the converged part is inaccurate. We address this issue here by testing a number of criteria, including the quality of chemical shift data, the number of converged residues and the significance of the ROSETTA energy gap, to detect inaccurate predictions. These criteria are aggregated to annotate each CS-ROSETTA prediction as *weak* or *strong*, thereby providing users with a reliability metric to assess the results.

Materials and methods

We benchmarked the performance of the new fragment picker (R3FP) and CS-RASREC on a set of 39 proteins in the size range of 50–100 residues that have neither been used for training of the R3FP, nor in CS-ROSETTA, SPARTA+ or TALOS+ development (Suppl. Table 1 + 3). All input files (fragments, reference coordinate and chemical shift files) are available for download at www.csrosetta.org/benchmarks.

Target selection and fragment picking

The benchmark set was selected from a larger set of 206 proteins for which recently released chemical shift

information from the BMRB was linked to coordinate information from the PDB in the CCPN framework (Vranken and Rieping 2009) and re-referenced using the VASCO protocol (Rieping and Vranken 2010). For NMR resolved structures, only proteins of sequence length 50–100 with at least 40 % secondary structure were retained from this set. Homologous proteins using an e-value cutoff of 0.05 (sequence identity >20 %) were excluded from MFR and R3FP fragment picking. The resulting set of 39 proteins covers a wide range of secondary structure content, as determined by DSSP from the PDB deposited structures. Since TALOS+ is used to pre-filter CS-ROSETTA submissions, and because the TALOS+ predicted secondary structure content is similar to what DSSP determines from the coordinates (Suppl. Fig. 6), this set of 39 is expected to be representative of typical CS-ROSETTA input.

Structure generation with CS-ABRELAX (server)

The latest version of the CS-ROSETTA webserver runs ROSETTA 3.3 including the new fragment selection method R3FP. For each target in the benchmark, 50,000 models were automatically generated by the CS-ROSETTA web server, using the standard CS-ABRELAX protocol with the ABRELAX cycle factor (command-line flag *-increase_cycles*) set to 10 as in Ref (Shen et al. 2008). The jobs are automatically distributed to available computational resources part of the worldwide WeNMR grid under the European Grid Initiative (EGI). As input, only a backbone NMR chemical shift list is required, which can be supplied in any of the common NMRPipe (TALOS), NMR-Star 2.1, or NMR-Star 3.1 (BMRB) formats.

The webserver uses SPARTA+ (Shen and Bax 2010) for final model selection in analogy to the original procedure based on SPARTA (Shen et al. 2008). Additionally, the server can combine the chemical shifts score with the DP score (Huang et al. 2005) based on unassigned NOE data for model selection, which has been shown to improve model selection from CS-ROSETTA calculations (Raman et al. 2010b, 2012).

An overview of the CS-ROSETTA web portal workflow can be found in Suppl. Fig. 1.

Structure generation with RASREC-ROSETTA

RASREC structure calculations (Lange and Baker 2011) with a pool-size of 500 conformers (command-line flag *-iterative:pool_size 500*) were started from the same fragment libraries as CS-ABRELAX calculations. As in the standard protocol (Lange and Baker 2011), *Recombination-stages* were terminated when the acceptance ratio into the pool dropped below 10 % (*-iterative:accept_ratio* 0.1) and the cycle factor was set to 2.0 (*-increase_cycles 2*). The protocol was modified to add chemical shift pseudo-energies with a weight of 5.0 to the ROSETTA energy to bias the RASREC pool of low-energy structures towards conformations in agreement with the experimental chemical shifts. Chemical shifts were computed from conformations using SPARTA+ (Shen and Bax 2010) and compared to the experimental chemical shifts to yield a pseudo-energy as described previously (Shen et al. 2008). To improve the prediction of chemical shifts from intermediate low-resolution structures a shortened refinement procedure was applied that uses only 1 of the usual 5 relax cycles (Raman et al. 2010a). SPARTA+ was implemented as a module of ROSETTA to allow computation of chemical shift pseudo-energies during RASREC iterations.

Calculation of converged regions

To determine the converged region of a protein structure predicted with CS-ROSETTA an adaption of the Gaussian-weighted RMSD method (Damm and Carlson 2006) was implemented in ROSETTA. The 30 lowest energy structures were superimposed using a scaling factor of 2 $Å^2$ (Damm and Carlson 2006). This procedure iteratively determines a set of rigid residues on which the structures can be superimposed; residues with a root-mean square fluctuation (RMSF) below 2 Å are considered to be converged. Gaps of <3 residues between regions of low RMSF (<2 Å) are ignored.

RMSD calculations

All reported RMSDs are C_α-RMSD to the PDB deposited reference structure or its first model. If the reference structure stems from an NMR solution ensemble only residues that superimpose within 1 Å in the deposited ensemble were used. Where indicated, C_α-RMSD computations are further *restricted* to regions converged in the ROSETTA calculations (see "Materials and methods").

Criteria used for annotations

The criteria of *strong/weak* prediction annotation are slightly different between CS-RASREC and CS-ABRELAX. *cs-consensus, convergence* and *energy gap* are used to annotate the prediction from CS-RASREC, and for CS-ABRELAX the criteria are *cs-class, convergence* and *energy gap*. *cs-consensus* is the fraction of residues for which TALOS+ finds more than 7 consensus matches in the database. *cs-class* is the fraction of residues annotated by TALOS+ with 'GOOD'. *convergence* is the fraction of residues which are

converged with an RMSF cutoff of 2 Å (see "Materials and methods"). The *energy gap* is the difference in ROSETTA all-atom energy between the lowest energy decoys and the lowest energies obtained for decoys far away (>4 Å) from the lowest energy decoys. Specifically, it is calculated as follows: the median energy of the 10 lowest energy structures is subtracted from the median energy of the 10 lowest energy structures within the subset of structures that are more than >4 Å (converged region; see "Materials and methods") from the lowest energy structure. In CS-RASREC annotations, the raw energy gap is divided by the number of residues and mapped to an interval 0.0.1 using a sigmoidal function with its inflection point at 0.05 Rosetta energy units (REU) per residue (see Suppl. Methods). Differently, the raw energy gap is directly mapped to [0,1] using sigmoidal function in CS-ABRELAX annotations.

Results

Performance of new fragment picker (R3FP)

Figure 1a shows the mean C_α-RMSD of the best 10 generated models with respect to the reference structure for MFR and R3FP. As can be seen, more targets appear above the diagonal, i.e., ABRELAX samples closer to the native structure, if R3FP fragments are used. Necessary for the success of a CS-ROSETTA structure calculation is that sufficient conformations below a C_α-RMSD of 2.0 Å to the reference structure are generated (Shen et al. 2008). This is the case for significantly more targets, if R3FP fragments are used (Fig. 1a; Table 1).

We also compared the performance in sampling near-native conformations of ABRELAX between software versions Rosetta 2.6 [used here (Shen et al. 2008; Wassenaar et al. 2012)] and Rosetta 3.x [used here (Raman et al. 2010b; Schmitz et al. 2012)]. As shown in Table 1, a performance gain is observed for Rosetta 3.x.

RASREC with chemical shift rescoring

As shown previously (Rohl et al. 2004; Shen et al. 2008), chemical shift rescoring improves precision and accuracy of final target selection for the CS-ABRELAX method. We have now implemented SPARTA+ rescoring directly into ROSETTA which allows us to apply the chemical shift score as a filter between iterations of the RASREC method (Shen et al. 2009b; Lange and Baker 2011). However, chemical shift rescoring is usually applied to fully refined structures, whereas intermediate structures in RASREC are without atomic detail (i.e., they use only a single *centroid* to represent the side-chain). To allow chemical shift rescoring nevertheless, a short refinement to atomic detail models that requires only ca. 20 % of the usual computer time is applied (see "Materials and methods").

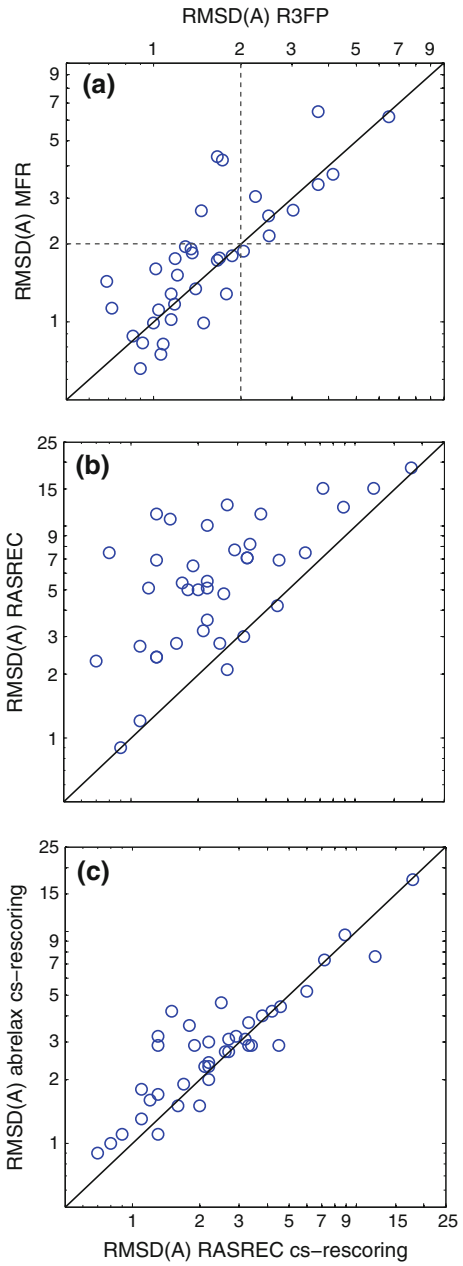

Fig. 1 Performance comparison. **a** Comparison of MFR and R3FP fragment picking methods using the ABRELAX sampling protocol. Shown are the mean C_α-RMSD of the lowest 10-RMSD structures (Table 1). *Dashed lines* indicate the 2 Å RMSD threshold, which is often predictive whether CS-Rosetta yields converged ensembles after energy-based selection. **b** Comparison of CS-RASREC (*x-axis*) and RASREC (*y-axis*). Shown are the median RMSDs of the ten lowest energy models selected by Rosetta energy and chemical shift score. **c** Comparison of CS-RASREC (*x-axis*) with CS-ABRELAX (*y-axis*). Shown are RMSDs of 10 lowest energy models selected by Rosetta energy and chemical shift score as in **b**

Table 1 Success of structure generation for MFR and R3FP fragment picker

Version[a]	Fragments[b]	Native sampling rate[c] (%)	RMSD (Å)[d]
Rosetta2-ABRELAX[e]	MFR	62	1.12 ± 0.42
Rosetta3-ABRELAX[e]	MFR	72	1.23 ± 0.48
Rosetta3-ABRELAX[e]	R3FP	77	1.18 ± 0.39
Rosetta3-RASREC	R3FP	64[f]	1.31 ± 0.41[f]
Rosetta3-CS-RASREC	R3FP	74[f]	1.27 ± 0.43[f]

[a] Major version number of Rosetta

[b] Fragment picking protocols

[c] Success rate of the structural sampling step defined as the percentage of targets for which the mean C_α-RMSD of the 10 lowest RMSD structures is lower than 2.0 Å; this reflects if the method samples the native structure, not how well it predicts it. C_α-RMSDs are calculated over all residues that are converged within 1 Å in the reference NMR structural ensemble (Suppl. Table 1)

[d] Average and standard deviation of the distribution of mean C_α-RMSDs, when restricted to those targets where the mean C_α-RMSD of 10 lowest RMSD structures is lower than 2.0 Å

[e] In CS-ABRELAX the chemical shifts are only used for final model selection. The native sampling rate is independent of final model selection, and thus CS-ABRELAX and ABRELAX are equivalent in this analysis. Note, however, that chemical shifts are used for fragment picking for all protocols analysed in this table

[f] For RASREC protocols, the native sampling rate is systematically lower than for ABRELAX, since instead of 50,000 full-atom models in ABRELAX, only ca. 1,500 full-atom models are generated in RASREC

We investigated whether chemical shift rescoring of intermediate structures improves the performance of the new CS-RASREC protocol on the benchmark set of 39 proteins. Indeed, a significant improvement in the RMSDs of the final energy selected models (Fig. 1b) is seen for CS-RASREC (points left of diagonal). Thus, CS-RASREC (but not RASREC) can further improve the accuracy of final models in comparison to CS-ABRELAX with R3FP fragments (Fig. 1c).

Restriction to converged regions

Figure 2a shows the C_α-RMSD to the reference structure of the lowest 10 scoring models from CS-ABRELAX calculations. As can be seen, only a small fraction of targets ($\sim$25 %) yields accurate (<2 Å) solutions. The reason for this apparent bad accuracy of CS-ROSETTA predictions is that RMSDs were computed on regions that are not converged in the CS-ROSETTA ensemble. To address this issue we added an auxiliary application called *ensemble_analysis* to the CS-ROSETTA toolbox (www.csrosetta.org), which detects residues whose RMSD fluctuations are <2 Å (see "Materials and methods"). Restriction of the structural prediction to these converged residues drastically changes the appearance of the results and shows that the converged regions are actually quite accurate for the majority of targets, with only five targets where the accuracy is worse than 2.5 Å (Fig. 2b). However, Fig. 2b also reveals that for many targets significant portions of the structures remain unconverged in the CS-ABRELAX calculations. As can be seen in Fig. 2c, the convergence is significantly improved in CS-RASREC calculations.

Reliability measure: annotation of weak/strong predictions

Originally, CS-Rosetta calculations were discarded if they did not converge on all residues (Shen et al. 2008; Schmitz et al. 2012). However, as shown above, some of the calculations that contain converged segments yield quite acceptable models. Thus, we looked for additional criteria to detect accurate predictions. We speculated that, in addition to (a) the overall convergence of the calculation, also the significance of (b) the chemical shift consensus or class (Shen et al. 2009a) and (c) the ROSETTA energy gap (Raman et al. 2010a; Fleishman and Baker 2012) should be informative on the likelihood of obtaining accurate structures.

To this purpose we define a predictor model that yields the signal *strong* if the weighted sum of the criteria c_i

$$P_{\text{sum}} = \sum_{i=1}^{3} w_i c_i$$

exceeds a threshold of 0.82 or 0.69 for CS-RASREC and CS-ABRELAX calculations, respectively (see Suppl. Methods). Optimizing the predictor model against manual classifications of the benchmark results, we obtained for CS-RASREC the weights 0.58, 0.29 and 0.13 for the criteria *cs-consensus, convergence,* and *energy-gap*, respectively. For CS-ABRELAX the weights 0.08, 0.54 and 0.38 for criteria *cs-class, convergence,* and *energy-gap*. The criteria are defined in "Materials and methods" section. In 100 rounds of cross-validated training using a different random selection of 25 % of the data as test-set for each

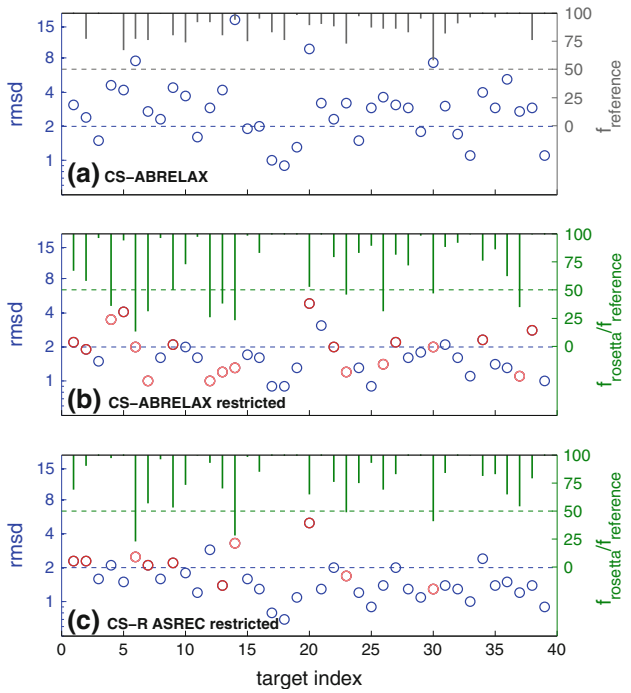

Fig. 2 Overview of accuracy of 10 lowest scoring structures from the 39 protein benchmark. C_α-RMSD to the reference structure (*circles*) are calculated over a subset of residues (*bars*). Predictions annotated as *weak* are shown in *red* (convergence is more than 50 %) or *pink* (convergence is <50 %) (Suppl. Table 2 and 4). **a** The RMSDs are calculated over all residues that are converged within 1 Å in the reference NMR structural ensemble (Suppl. Table 1). The number of residues used for RMSD calculation are shown as fraction of total length $f_{\text{reference}}$ (*gray*). **b** The RMSD calculation is restricted to residues converged within 2 Å in the CS-ROSETTA structural ensemble (and within 1 Å in the references) (Suppl. Table 1). The additional restriction in RMSD calculation is given as ratio $f_{\text{rosetta}}/f_{\text{reference}}$ (*green*). **c** RMSD restriction as in **b** but using the CS-RASREC method

round, for CS-RASREC the cutoff was selected by fixing the false positive rate (FPR) to 3 % and 6 % for CS-RASREC and CS-ABRELAX, respectively. The resulting thresholds of 0.82 ± 0.06 and 0.69 ± 0.05 yielded true positive rates (TPR) of (89 ± 20) %and (80 ± 33) %, respectively. The standard deviation of the weights trained on the 100 different selections of training data during cross-validation were 0.08, 0.10 and 0.08, for CS-RASREC and 0.25, 0.13, 0.14 for CS-ABRELAX. The higher variation of weights for CS-ABRELAX reflects the less pronounced energy gap and the lower rate of convergence observed in CS-ABRELAX simulations (Suppl. Fig. 4). Alternative predictor models were discarded based on inferior receiver operating characteristic (ROC) in the cross-validation and the compound predictor model outperforms the individual criteria (Suppl. Fig. 2). The final set of weights was obtained by optimizing the most successful predictor model against all data points.

Indeed, the classification scheme successfully annotates those predictions as *weak* that yield bad accuracy (red in Fig. 2b, c). 20 of 39 targets (51 %) listed in Suppl. Table 2 computed with CS-ABRELAX are considered as *strong* structure calculations. For 18 of these the accuracies range from 0.9 to 2.0 Å, and for the remaining two, accuracies are 3.1 and 2.1 Å for targets #21 (2jvf) and #31 (2k3d), respectively. From the targets computed with CS-RASREC, 29 of 39 (74 %) results are considered *strong*. For 26 of the *strong* targets, accuracies range from 0.7 to 2.0 Å and for the remaining three, targets #4 (2dm2), #12 (2jov) and #34 (2k5c), accuracies are 2.1, 2.9 and 2.4 Å, respectively (Suppl. Table 4).

For these three targets (#4, #12, and #34), CS-RASREC predicted structures have the same fold as the reference structure, but show better packing with less and smaller solvent inaccessible cavities in the protein core (Suppl. Fig. 3) (Sheffler and Baker 2008). Given the clear packing deficiencies in the deposited NMR ensembles, we believe that the 2.1–2.9 Å RMSDs do not actually reflect the accuracy of the CS-RASREC structures, and that these targets can be ignored for the overall assessment of CS-RASREC accuracy of *strong* predictions. Representative examples of the remaining *strong* predictions are shown in Fig. 3.

The WeNMR CS-ROSETTA web server

The most time consuming part of a typical CS-ROSETTA run consists of a large number (500–2,500) of independent Monte Carlo calculations to calculate in the order of 10,000–50,000 structures. The WeNMR (www.wenmr.eu) CS-Rosetta web server (Wassenaar et al. 2012) conveniently distributes those calculations over the grid resources made available through the European Grid Infrastructure (EGI, www.egi.eu). The original server has now been extended to allow DP scoring and include the reliability measure described above. Table 2 shows the results of the DP rescoring option (using the CS-ABRELAX setup), using a different benchmark of 6 CASD-NMR targets (Rosato et al. 2009, 2012). Consistent with previous observations (Raman et al. 2010b; Rosato et al. 2012) the combination of DP rescoring (Huang et al. 2005; Raman et al. 2010b) and CS rescoring outperforms the other rescoring option, including CS rescoring, both in successful predictions and reliability (100 %).

Discussion

We have considerably improved the scope, convergence and reliability of CS-ROSETTA calculations from chemical shifts only. On a representative benchmark of 39 small proteins in the size range of 50–100 residue size range, we

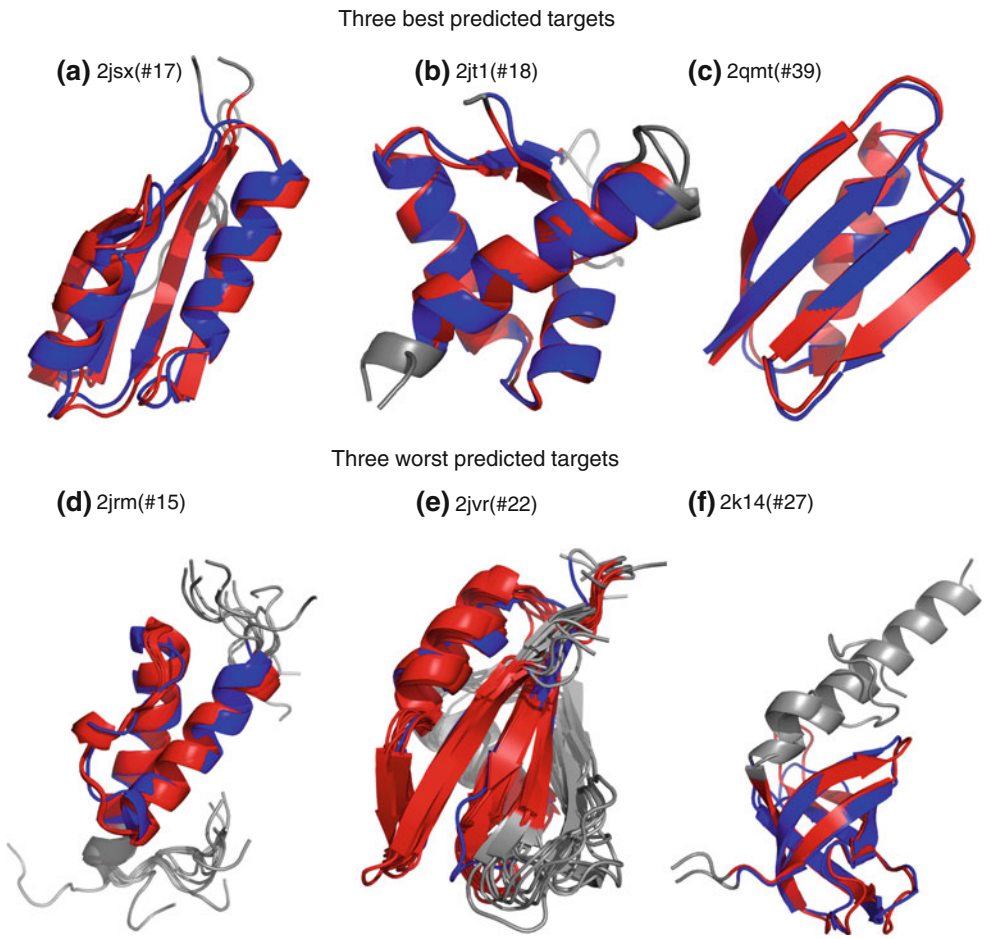

Fig. 3 Overview of structures obtained with RASREC structure calculations that passed the filter (i.e., annotated as *strong prediction*). Shown are the three best, 2jsx (0.8 Å), 2jt1 (0.7 Å) and 2qmt (0.9 Å), respectively, and the three worst, 2jrm (1.6 Å), 2jvr (2.0 Å) and 2k14 (2.0 Å), respectively. For each target, the reference structure is in *blue* and the predicted structures are in *red* with unconverged regions (see "Materials and methods") shown in *gray*

Table 2 Reliability of different structure selection methods

Selection[a]	Converged[b]		Not converged[c]	Reliability[d] (%)
	TP	FP		
Raw	2	2	2	50
cs	3	1	2	75
dp	5	1	0	83
dpcs	5	0	1	100

[a] Final structure selection methods, raw: rosetta score; cs: cs-rescoring; dp: dp-rescoring; dpcs: cs-rescoring + dp-rescoring

[b] Number of targets for which the average RMSD of selected models is below the threshold of 2.0 Å and are counted as true/false positive

[c] N umber of targets for which the average RMSD of selected models is above the threshold of 2.0 Å

[d] Reliability of different structure selection methods

demonstrated that CS-ROSETTA calculations yield successful and accurate 3D structure predictions in 74 % of the cases when using the new CS-RASREC method. CS-ABRELAX is still successful in 51 % of the cases but generally yields less converged residues per target. Most importantly, we introduced a classification scheme that can be used to detect whether a successful prediction has been made, which increases the reliability to >89 and >80 % for CS-RASREC and CS-ABRELAX calculations, respectively. Reliable predictions have accuracies of 2 Å and better on the converged residues. This renders the presented CS-ROSETTA structure calculation protocols a reliable tool for rapid and accurate structure determination at atomic resolution.

CS-ROSETTA calculations entail a considerably computational effort; a reliable structure prediction requires 10,000 or more models to be generated with an overall cost of several thousand CPU-hours. We implemented a webserver that utilizes the WeNMR grid infrastructure to farm out the time-consuming model generation part of CS-ROSETTA calculations. The service is available for the whole scientific community and is free of charge to academic users. It only

requires a backbone chemical shift list as input and offers several options to re-evaluate the generated models, including NOE based rescoring with the DP-score (Huang et al. 2005; Raman et al. 2010b).

Currently, the WeNMR grid cannot support CS-RASREC calculations due to the requirement of communication between RASREC processes that is not supported by the grid-infrastructure. However, RASREC calculations are considerably more time-efficient than CS-ABRELAX; for targets in the size range addressed here, they require on the order of 200–1,000 CPU hours, which is available on medium sized in-house clusters or at adjunct computer centers of universities. We made considerable advances to simplify running these calculations by providing a Python-based toolbox for pre- and post-processing of CS-ROSETTA related data files and fragment picking. This allows easy setup of CS-ABRELAX and RASREC CS-ROSETTA structure generation runs including integrated support for queuing systems such as SLURM and MOAB. The computational infrastructure has to support jobs that utilize the common Message Passing Interface (MPI) protocol (e.g., openMPI, LAM, MPICH, MPICH2) for inter-process communication. Additionally, a website providing documentation and tutorials (www.csrosetta.org) has been launched in support of the growing user community.

The main advantage of CS-RASREC calculations over the CS-ABRELAX is that a larger fraction of residues converges and that the energy gap becomes more pronounced. This in turn generates a higher chance of a *strong* prediction. On the 39 benchmark cases, the average fraction of converged residues (as shown in Fig. 2b, c (green bars) is 72 % for CS-ABRELAX and 80 % for CS-RASREC. From the 9 targets that are classified *strong* in CS-RASREC but *weak* in CS-ABRELAX, 4 have improved classification due to a drastic increase in convergence (from $\sim$30 to $>$70 %), whereas the remaining 5 have similar convergence but improved energy-gaps (Suppl. Table 2 + 4). Finally, the mean accuracy (RMSD) for *strong* predictions is 1.76 Å for CS-ABRELAX and 1.44 Å for CS-RASREC. Thus, if local computer resources can be obtained it is advisable to run CS-RASREC rather than CS-ABRELAX, if such resources cannot be secured, running just the webservice-based CS-ABRELAX remains a reasonable and valuable alternative. Adaption of the RASREC protocol to a grid or cloud computing platform is in principle possible as only very low-bandwidth communication is required, but technically involved as the entire communication layer of the protocol has to be adapted.

A program to apply the reported classification scheme into *strong* and *weak* 3D structure predictions is provided with the CS-ROSETTA toolbox versions 2.x and higher at www.csrosetta.org and is implemented in the CS-ROSETTA web server.

Acknowledgments This work was supported by the German Science Foundation (DFG) Grant LA 1817/3-1 (to Z.Z. and O.F.L.), the Brussels Institute for Research and Innovation (Innoviris) grant BB2B 2010-1-12 (to W.F.V.), and the Intramural Research Program of the NIDDK (to Y.S.). The WeNMR project (European FP7 e-Infrastructure Grant, Contract No. 261572, www.wenmr.eu), supported by the European Grid Initiative (EGI) through the national GRID Initiatives of Belgium, France, Italy, Germany, the Netherlands (via the Dutch BiG Grid project), Portugal, Spain, UK, South Africa, Malaysia, Taiwan and the Latin America GRID infrastructure via the Gisela project is acknowledged for the use of web portals, computing and storage facilities.

References

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

Damm KL, Carlson HA (2006) Gaussian-weighted RMSD superposition of proteins: a structural comparison for flexible proteins and predicted protein structures. Biophys J 90:4558–4573

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

Fleishman SJ, Baker D (2012) Role of the biomolecular energy gap in protein design, structure, and evolution. Cell 149:262–273

Huang YJ, Powers R, Montelione GT (2005) Protein NMR recall, precision, and F-measure scores (RPF scores): structure quality assessment measures based on information retrieval statistics. J Am Chem Soc 127:1665–1674

Lange OF, Baker D (2011) 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

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

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

Rieping W, Vranken WF (2010) Validation of archived chemical shifts through atomic coordinates. Proteins 78:2482–2489

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

Rosato A, Bagaria A, Baker D, Bardiaux B, Cavalli A, Doreleijers JF, Giachetti A, Guerry P, Güntert P, Herrmann T et al (2009) CASD-NMR: critical assessment of automated structure determination by NMR. Nat Methods 6:625–626

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(5):668–677

Sheffler W, Baker D (2008) RosettaHoles: rapid assessment of protein core packing for structure prediction, refinement, design and validation. Protein Sci 18(1):229–239

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

Shen Y, Zhang Z, 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

Vranken WF, Rieping W (2009) Relationship between chemical shift value and accessible surface area for all amino acid atoms. BMC Struct Biol 9:20

Wassenaar TA, van Dijk M, Loureiro-Ferreira N (2012) WeNMR: structural biology on the grid. J Grid 10:743–767

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