J Biomol NMR (2012) 52:41–56
DOI 10.1007/s10858-011-9576-3

ARTICLE

VITAL NMR: using chemical shift derived secondary structure information for a limited set of amino acids to assess homology model accuracy

Michael C. Brothers · Anna E. Nesbitt · Michael J. Hallock · Sanjeewa G. Rupasinghe · Ming Tang · Jason Harris · Jerome Baudry · Mary A. Schuler · Chad M. Rienstra

Received: 17 May 2011 / Accepted: 28 September 2011 / Published online: 3 November 2011

Abstract Homology modeling is a powerful tool for predicting protein structures, whose success depends on obtaining a reasonable alignment between a given structural template and the protein sequence being analyzed. In order to leverage greater predictive power for proteins with few structural templates, we have developed a method to rank homology models based upon their compliance to secondary structure derived from experimental solid-state NMR (SSNMR) data. Such data is obtainable in a rapid manner by simple SSNMR experiments (e.g., ^{13}C–^{13}C 2D correlation spectra). To test our homology model scoring procedure for various amino acid labeling schemes, we generated a library of 7,474 homology models for 22 protein targets culled from the TALOS+/SPARTA+ training set of protein structures. Using subsets of amino acids that are plausibly assigned by SSNMR, we discovered that pairs of the residues Val, Ile, Thr, Ala and Leu (VITAL) emulate an ideal dataset where all residues are site specifically assigned. Scoring the models with a predicted VITAL site-specific dataset and calculating secondary structure with the Chemical Shift Index resulted in a Pearson correlation coefficient (−0.75) commensurate to the control (−0.77), where secondary structure was scored site specifically for all amino acids (ALL 20) using STRIDE. This method promises to accelerate structure procurement by SSNMR for proteins with unknown folds through guiding the selection of remotely homologous protein templates and assessing model quality.

Keywords Protein structure prediction · Homology modeling · Solid-state NMR spectroscopy · TALOS database · Chemical shift analysis

Electronic supplementary material The online version of this article (doi:10.1007/s10858-011-9576-3) contains supplementary material, which is available to authorized users.

M. C. Brothers · A. E. Nesbitt · M. J. Hallock · M. Tang · C. M. Rienstra
Department of Chemistry, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA

S. G. Rupasinghe · M. A. Schuler
Department of Cell and Developmental Biology, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA

J. Harris · J. Baudry
Department of Biochemistry, Cellular and Molecular Biology, University of Tennessee, Knoxville, TN 37996, USA

J. Baudry
UT/ORNL Center for Molecular Biophysics, Oak Ridge National Laboratory, Oak Ridge, TN 37830, USA

M. A. Schuler · C. M. Rienstra
Department of Biochemistry, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA

C. M. Rienstra (✉)
Center for Biophysics and Computational Biology, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA
e-mail: rienstra@illinois.edu

Introduction

Structures of membrane proteins and biomolecular assemblies are difficult to obtain due to the inherent limitations of X-ray crystallography and NMR in studying such proteins in their native environments. Although tremendous progress has been made in these fields in recent years (Forrest et al. 2006; Hanson and Stevens 2009; Hiller et al. 2008; Mobarec et al. 2009; Van Horn et al. 2009; Yarnitzky et al. 2010), structural studies of membrane proteins are few in comparison to the number of soluble

proteins and complexes deposited in the Protein Data Bank (White 2009). Hybrid approaches that draw on the strengths of individual methods have recently demonstrated capacity to accelerate structure determination and/or enhance the resolution of protein structures. For example, solution NMR data have been interpreted in conjunction with small angle X-ray scattering, ab initio structure prediction, or homology modeling to determine structures of proteins and protein complexes (Alber et al. 2008). Similar strategies have been successful in elucidating structures of the homodimer Dsy0195 using solution NMR with EPR (Yang et al. 2010), a 82 kDa protein using SAXS and solution NMR data (Grishaev et al. 2008), P450s and other proteins using sequence-function relationships with structure predictions (Baudry et al. 2006, Mercier et al. 2006), zinc-bound proteins using solution NMR and homology modeling (Randazzo et al. 2001), and protein complexes including hemoglobin using X-ray crystallography with structure predictions (Schröder et al. 2010). Methodologies that complement solid-state NMR (SSNMR) data with X-ray crystallography (Tang et al. 2011) or SAXS (Jehle et al. 2010) have also been developed. Such hybrid methodologies enable site-specific restraints to narrow the range of candidate structures. Methods pairing SSNMR with structure prediction have potential as an efficient way to provide rudimentary structures, which can in turn expedite protein assignment by providing initial chemical shift values for auto-assignment algorithms (Monleon et al. 2002; Tycko and Hu 2010), extract greater structural information from poorly diffracting crystals and seed NMR calculations to access higher quality structures as experimental observations accumulate (Schröder et al. 2010). The process of identifying initial structural models is often the most time-consuming step in the NMR structure determination process; once available, a variety of methods are available for structure refinement, both in solution (Bax et al. 2001; Fischer et al. 1999) and in solid state (Franks et al. 2008; Wylie et al. 2009).

SSNMR offers unique capabilities that are well suited for investigating the structures of nanocrystals, macromolecular complexes, fibular and precipitated membrane proteins. Unlike in solution NMR, size is not a fundamental limitation because magic-angle spinning eliminates anisotropic interactions and dipolar couplings; however, the practical data collection and interpretation challenges are significant obstacles to solving de novo structures of high molecular weight proteins. Although higher dimensionality experiments (Ikura et al. 1991; Grzesiek and Bax 1993; Franks et al. 2007) can enhance resolution, site-resolved sensitivity scales in inverse linear proportion to molecular weight, so signal averaging times increase quadratically. Thus, protein structures determined de novo from SSNMR data to date have been limited to proteins with a molecular weight of $\sim$18 kDa or less (Bertini et al. 2010; Castellani et al. 2002; De Angelis et al. 2006; Franks et al. 2008; Loquet et al. 2008; Manolikas et al. 2008; Marassi and Opella 2003; Traaseth et al. 2009; Van Melckebeke et al. 2010). It would therefore be desirable to accelerate the SSNMR structure determination process by improved leveraging of readily obtainable data from NMR.

Substantial structural information is present in the simplest 2D spectra, which can be acquired for proteins as large as 144 kDa (Frericks et al. 2006). For example, ^{13}C chemical shifts provide information on amino acid type and secondary structure (Spera and Bax 1991). As spectral overlap accumulates with protein size, limited isotope labeling strategies are available to assist in simplifying interpretation of type and pairwise assignments. Multiple methods have been developed for this purpose, including cell free expression systems (Baranov et al. 1989; Endo and Sawasaki 2003; Sawasaki et al. 2002; Schwarz et al. 2008) and auxotrophic strains that eliminate scrambling from biosynthetic pathways (Lin et al. 2011; Waugh 1996). With such labeling approaches, it may be possible to obtain partial assignments for much larger membrane proteins. We have recently demonstrated the principle of utilizing chemical shifts to refine both microcrystalline (Wylie et al. 2009) and membrane protein structures (Tang et al. 2011).

Additionally, such information could be used to screen initial structures produced by comparative modeling, which can rapidly generate significant numbers of candidate structures for target sequences and can also incorporate data acquired from NMR either to constrain or refine model generation (Sali and Blundell 1993). Identification of the most representative model structure from a pool of candidates is a long-standing problem. Several methods have been developed to assess the quality of comparative models (Bowie et al. 1991; Fasnacht et al. 2007; Hooft et al. 1996; Melo and Feytmans 1997; Monleon et al. 2002; Pieper et al. 2009; Ray et al. 2010; Sippl 1993; Weichenberger and Sippl 2006), and a subset of structure validation methods have been developed for X-ray and NMR derived protein structures (Bowie et al. 1991; Laskowski et al. 1996; Vila et al. 2008). In recent years, new methods have more effectively leveraged data from SAXS, cryo-EM, NMR, and other methodologies in the model generation process (Alber et al. 2008). CS-ROSETTA utilizes NMR chemical shifts in a very early stage of the fragment selection process to accelerate model generation, as well as to score model quality, resulting in atomic resolution structures for an impressive range of proteins up to 20 kDa (Shen et al. 2009). Including nuclear Overhauser effect and/or residual dipolar coupling data set within CS-ROSETTA has further extended the molecular weight range and quality of structures derived in this manner (Raman et al. 2010). Alternative approaches are based

upon molecular dynamics in concert with chemical shift-based potential functions (Robustelli et al. 2010) that can utilize partial assignments of structures to generate a model that agrees with the site-specific chemical shifts determined from solution or SSNMR experiments.

All of the methodologies for rapid chemical shift-based structure determination so far require site-specific assignment information. In contrast, restraint-based structure prediction via homology modeling generates models by aligning known structures to a target template, incorporating known restraints, and then using probability density functions to relax the structure into a final conformation, within which chemical shifts are easily interjected to probe whether the conformations comply with the experimental data. This works well for proteins that are known to be homologous and can be aligned accurately (>30% sequence identity). At lower sequence identity, it becomes increasingly challenging to identify a template and find a proper alignment (Marti-Renom et al. 2000). For poor sequence alignments, NMR restraints that are normally helpful can exacerbate the error by locking the protein into an incorrect conformation, thus generating suboptimal models. This is especially true for proteins well suited for investigations by SSNMR, namely membrane proteins and aggregates, due to the reduced number of potential templates (Oberai et al. 2006). Therefore, a method that marries comparative modeling with partial chemical shift assignments promises to accelerate the pace at which the SSNMR spectroscopist gleans structural information for non-soluble proteins prior to obtaining complete site-specific assignments.

With the intent of producing high quality molecular models from limited experimental NMR data, we present a comparative model scoring function that has been developed and evaluated based on the secondary structural elements reported by chemical shifts. The scoring function was formulated to incorporate different subsets of NMR data ranging from secondary structure by amino acid type to complete site-specific assignments. The inherent flexibility of this model scoring function facilitated the identification of a minimal subset of NMR data where the score achieves maximal correlation with the RMSD of molecular models. To test the scoring function, comparative models were generated for 22 protein targets culled from the TALOS+/SPARTA+ training set of protein structures (Shen and Bax 2007; Shen et al. 2009). For each of the protein targets, chemical shift assignments were simulated with SHIFTX, a measure assuring that the data required for Chemical Shift Index (CSI) predictions was complete. As a control, each model was scored using each target protein's actual secondary structure as assigned by STRIDE (Heinig and Frishman 2004). In this instance, the Pearson correlation coefficient for model score plotted against model-to-structure RMSD was −0.77 (±0.02), compared with −0.73 (±0.02) if the model and target chemical shifts were predicted using SHIFTX and the secondary structure obtained using CSI. We then formulated the scoring function to incorporate truncated datasets, and noted that the amino acid labeling scheme using pairs of the Val, Ile, Thr, Ala, Leu (VITAL) amino acids resulted in a Pearson correlation coefficient commensurate with the control, −0.75 (±0.02). We concluded these investigations with a rigorous test of the model scoring function using actual partial assignments for the membrane protein DsbB. In this case, the absolute difference in RMSD between the best scoring model and best model (ΔRMSD) was only 0.55 Å, demonstrating the power of this methodology to identify the best model from a pool containing 375 models.

Materials and methods

Generation of models

In order to develop and test methods for ranking comparative models based on NMR derived observations, we generated, for twenty-two protein templates, a total of 7,525 homology models based on the sequences of known crystal structures for which publically available, high quality NMR data exists. Twenty-two model targets were culled from the TALOS+/SPARTA+ training set of proteins (Shen et al. 2009; Shen and Bax 2010), proteins that were selected to satisfy several criteria: (1) that high quality NMR chemical shift assignments exist, (2) that high resolution X-ray crystal structures exist (3) that they encompass a diversity of protein folds. The target proteins (Fig. 1) range in size from 137 residues to 502 residues (14.6–59.3 kDa) and represent a diverse mixture of SCOP secondary structure classes (α, β, $\alpha + \beta$, α/β). These measures were taken so that development of the methodology does not favor a particular secondary structure class or protein size (e.g. larger proteins tend to have a disproportionately larger hydrophobic core). Comparative models were produced using MODELLER (version 9.1) (Sali and Blundell 1993; Marti-Renom et al. 2000) along with a published protocol (Eswar et al. 2000) or using MOE (Kelly 1999) with program defaults (Fig. 2) to generate 25 models per template. To identify suitable modeling templates, each of the 22 protein sequences was subjected to a search of the RCSB Protein Data Bank using either the MODELLER build_profile.py function that employed a modified BLAST search (BLOSUM62, Altschul et al. 1990) to determine the best alignment or using the modified FASTA methodology employed by MOE (Pearson 1996).

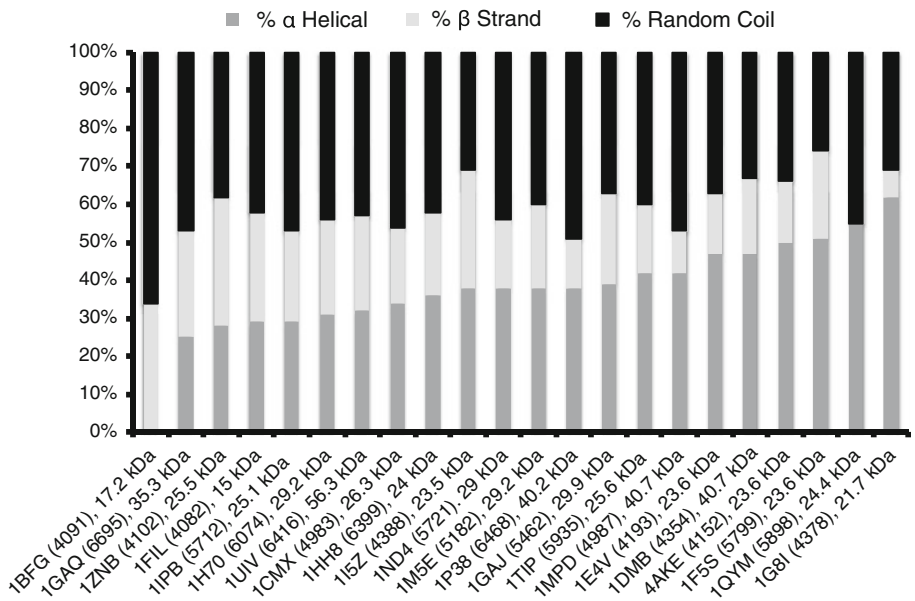

Fig. 1 Structural diversity of the 22 target proteins selected for this investigation. Targets vary in size from 15.0 to 56.3 kDa, and represent all SCOP structural classes. PDB identifiers are indicated along with BMRB codes (in *parentheses*) and molecular masses

The statistical significance of the identified template to sequence alignments varied considerably as exhibited by the E-value calculated by MODELLER (Table S3). If greater than 10 structural templates were available for a given protein target while discarding templates with low *Z*-scores (<5) or high E-values (>1), templates bearing 20–50% sequence identity to the target sequence were selected at random. Twenty-five models were made for each structural template selected and subjected to a conjugate gradient minimization using either the CHARMM22 force field (MODELLER) or AMBER99 force field (MOE). Models with an RMSD greater than 45 Å from the target structure were eliminated from the model pool (51 models, 0.06%) leaving 7,474 total models. The 51 models exhibiting RMSD greater than 45 Å were artifacts of using liberal Z-scores and E-values when selecting modeling templates and possessed residuals greater than 2 standard deviations removed from the best-fit line in the control calculation using STRIDE determined secondary structure. Furthermore, large RMSD values are known to be highly dependent on protein chain length (Carugo and Pongor 2001) and, although no models were generated with RSMD between 30 and 45 Å, models with RMSD at ~50 Å scored as well as models with RSMD at ~25 Å indicating that the linear trend may have plateaued at ~25 Å. Various attributes of the selected structural targets and templates are depicted in Fig. 1 and listed in Table S3.

Scoring of models and analysis

Each model was scored based on its adherence to the secondary structure derived from the Chemical Shift Index (CSI) Software Package (Wishart and Sykes 1994) calculated for its target structure (Fig. 2). The program STRIDE (Heinig and Frishman 2004) was used to evaluate the proportion of the α-helical, β-strand and random coil secondary structure elements found for each of the models in addition to those same elements found for the target structures and served as an alternate pathway to measure secondary structure of models versus crystal structures. Coordinates for the 22 target crystal structures were submitted to SHIFTX (Neal et al. 2003), which predicts the backbone chemical shifts (C, CA, CB, N, HN, and HA atoms). The resultant table of predicted resonances was converted into a simplified secondary structure prediction (α-helical, β-sheet, or random coil/unknown) based on three shifts per residue (CA, CB, C). The model secondary structure, represented as A_i for α-helix and B_i for β-sheet, and the target secondary structures, $\hat{A}_i$ and $\hat{B}_i$, were then passed to the scoring function *S* for the equations below.

Equation (1) represents a model score, *S*, based on amino acid type. *S* is summed over residue "type" assignments, where *aatype* represented one of the twenty amino acids used in the scoring function (i.e., Val, Ile, Thr, Ala, Leu; $i = 1, 2, \ldots, 5$) and that N_{aatype} represented the total number of one residue type and *N* corresponded to the total number of assigned residues.

$$Score = 100 \times \frac{\sqrt{\sum_{aatype}\left(\left|N_{aatype} - \frac{\left|A_{aatype}-\hat{A}_{aatype}\right|+\left|B_{aatype}-\hat{B}_{aatype}\right|}{2}\right|\right)^2}}{N} \quad (1)$$

where *N* is the sum of the total number of amino acids used for scoring:

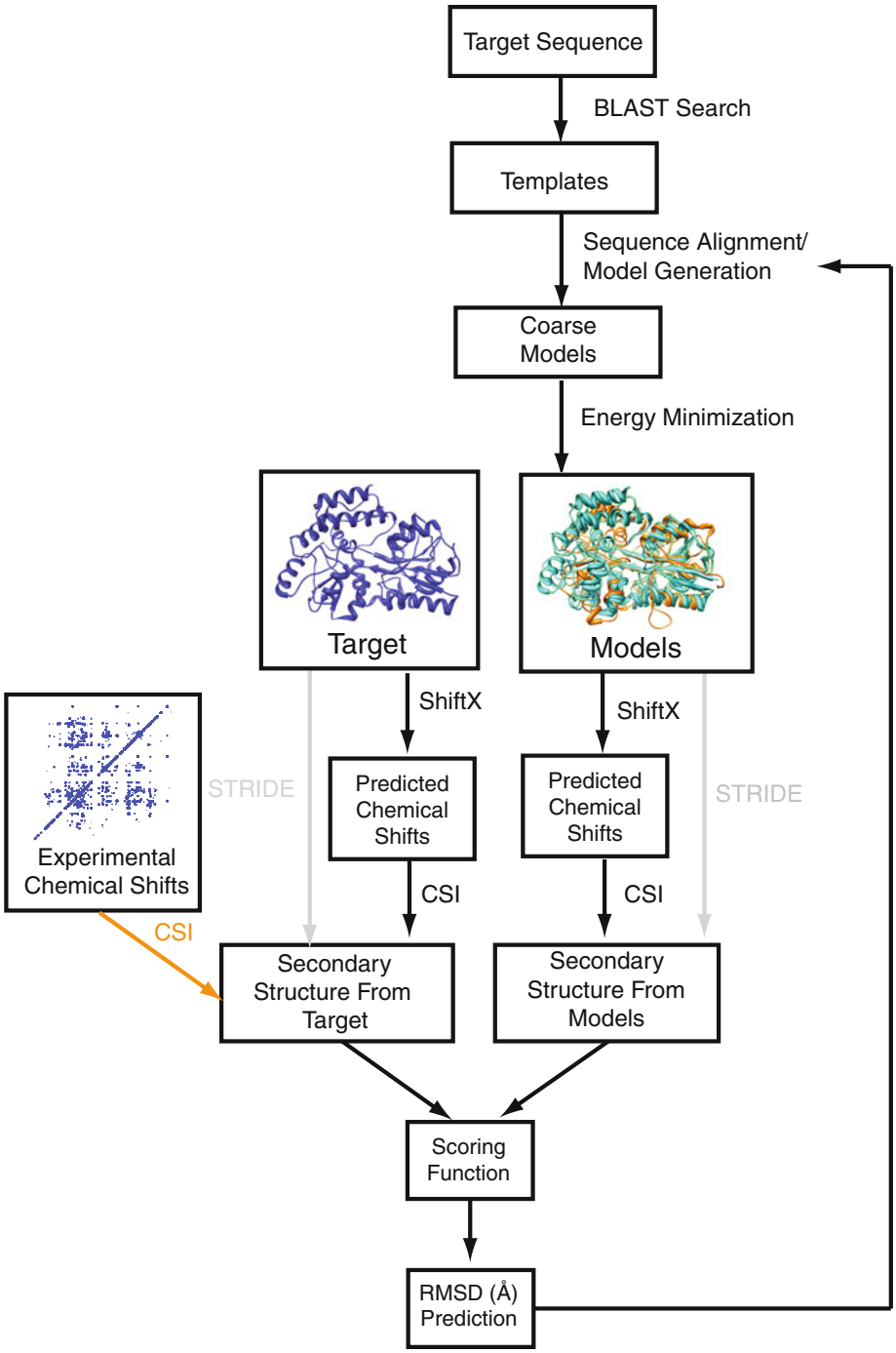

Fig. 2 Flowchart depicting the VITAL NMR modeling and scoring procedures. Briefly, a target sequence from one of the 22 selected proteins from the TALOS+/SPARTA+ training set is submitted to a BLAST search of the RCSB Protein Data Bank leading to the identification of modeling templates with an E-value smaller than 0.1. Coarse models are then produced based upon the suggested BLAST alignment with either MODELLER v9.1 or MOE. Models are then refined with an energy minimization in order to relax side-chains (using the CHARMM22 force field) critical for the determination of secondary structure by SHIFTX/CSI. The secondary structure of the models may be interpreted by the program STRIDE (*grey path*) or ascertained by predicting chemical shifts with SHIFTX and calculating the resultant Chemical Shift Index (CSI). The secondary structure of the model is then compared with the secondary structure determined for the target protein in one of three ways: (1) when the model secondary structure is determined with STRIDE, it is compared to the STRIDE interpreted secondary structure of the target X-ray crystal structure; (2) when the model secondary structure is determined using SHIFTX/CSI it is compared to the secondary structure calculated with SHIFTX/CSI for the target X-ray crystal structure or (3) compared with CSI calculated from chemical shift assignments, a pathway highlighted in *orange*

$$N = \sum_{aatype} N_{aatype} \tag{2}$$

Another approach is to score based on pairwise assignments. Equation (3) represents a model score, *S*, based on pairwise assignments where amino acids are grouped by pairs (i.e., all Ala-Gly pairs or all Gly-Pro pairs). *S* is calculated from a summation of the score of each individual pair of amino acids selected, where *pairstype* is the numerical value for one type of amino acid pair from a limited subset of amino acids (i.e., one of 25 pairs derived from Val, Ile, Thr, Ala, Leu) that occurred in the amino acid sequence of the model structure. In this case, the secondary structure of both amino acids is treated as a single unit in the following equation,

$$Score = 100 \times \frac{\sqrt{\sum_{pairstype}\left(\left|N_{pairstype} - \frac{\left|A_{pairstype}-\hat{A}_{pairstype}\right|+\left|B_{pairstype}-\hat{B}_{pairstype}\right|}{2}\right|\right)^2}}{N} \quad (3)$$

where N is the sum of the total number of pairs observed in the sequence:

$$N = \sum_{pairstype} N_{pairstype} \quad (4)$$

Equation (5) represents the model score, S, derived from site specifically assigned data. S is calculated from a summation of the score of each site-specific assignment used, where *#aa* represents each of the residues in the target's amino acid sequence with carbon backbone chemical shift assignments (CA, CB, C). N corresponds to the total number of assigned residues in the target sequence,

$$Score = 100\frac{\sqrt{\sum_{\#aa}\left(\left|N_{\#aa} - \frac{\left|A_{\#aa}-\hat{A}_{\#aa}\right|+\left|B_{\#aa}-\hat{B}_{\#aa}\right|}{2}\right|\right)^2}}{N} \quad (5)$$

where N is the total number of amino acids in the structure.

$$N = \sum_{aa\#} N_{aa\#} \quad (6)$$

Linear regression analyses were used to test the ability of the model ranking score, S, to predict the model/target RMSD values. For all Pearson correlation coefficients calculated (R), the 99% confidence intervals are reported in parentheses or as error bars. The VMD RMSD plugin was used to first align each model to its corresponding target and then calculate the Cα coordinate root-mean-square difference between the model and the structure (RMSD) (Humphrey et al. 1996). ΔRMSD was calculated as the absolute difference in RMSD for the score-selected best model and the actual best model as shown in (7).

$$\Delta\text{RMSD} = |\text{RMSD}(\text{Best scoring Model, Target}) - \text{RMSD}(\text{Best Model, Target})| \quad (7)$$

In instances where several models achieved the same score, the RMSDs of these models were averaged before calculating the ΔRMSD. Tables 1 and S1 give the Pearson correlation coefficients, standard error of estimate and ΔRMSD determined for each implementation of S described in the "Results" and "Discussion".

The CS-ROSETTA score reported in Shen et al. (2008), E' in (8), was calculated for all models generated in this study.

Table 1 Pearson correlations coefficients (R), standard error of estimates (SEE), average ΔRMSDs, and the standard deviation (SD) of ΔRMSD for different subsets of the amino acids scored with either the type (T), pairwise (PW), or site-specific (SS) scoring functions [(1), (3) and (5), respectively] tested in order to produce a score predictive of the RMSD between a given model and its true structure

Amino acids scored	Score type	R (99% conf.)	SEE	ΔRMSD (Å)	SD of ΔRMSD (Å)
ALL 20	Control	−0.77 (±0.01)	14.0	1.6	3.0
VITLHRMKGS	PW	−0.75 (±0.02)	13.1	1.9	2.9
VITLHRMKGS	SS	−0.75 (±0.01)	14.3	1.9	2.6
VILHYGS	SS	−0.75 (±0.01)	14.6	2.1	2.6
VITALS	SS	−0.75 (±0.01)	14.9	2.8	4.6
VITALGS	SS	−0.74 (±0.01)	13.9	2.8	4.2
ALL 20	SS	−0.73 (±0.02)	14.5	2.9	4.6
VITALGS	PW	−0.71 (±0.01)	12.5	1.9	2.7
VITALH	PW	−0.71 (±0.02)	16.1	2.2	3.1
TAGS	SS	−0.69 (±0.02)	14.5	2.8	4.5
VILHYGS	PW	−0.68 (±0.02)	16.1	3.2	4.2
ALHRMKYGS	SS	−0.58 (±0.02)	22.8	2.2	2.2
VITAL	T	−0.57 (±0.02)	14.5	3.4	4.8
VITALGS	T	−0.54 (±0.02)	13.7	2.9	4.6
VITLHRMKGSY	T	−0.52 (±0.02)	13.8	2.8	4.9
VILHYGS	T	−0.51 (±0.02)	15.2	3.3	3.5
TAGS	T	−0.49 (±0.02)	14.0	4.0	5.1
TAGS	PW	−0.48 (±0.02)	17.0	4.7	5.5
ALL 20	T	−0.49 (±0.02)	14.0	5.8	3.5

A comprehensive table may be found in the "Supplementary Material"

Table 2 A comparison of Pearson correlation coefficients derived from models scored with the CS-ROSETTA full atom energy, χ^2_{CS}, VITAL pairs and all amino acids site-specifically for 5 different windows of RMSD

RMSD window (Å)	# of models in RMSD window	CS-ROSETTA	χ^2_{CS} BMRB	χ^2_{CS} SPARTA+	VITAL pairs	All amino acids SS
0–2.5	415	0.35	0.27	0.52	−0.14	−0.06
0–5	1,693	0.43	0.21	0.22	−0.31	−0.22
0–10	3,448	0.45	0.44	0.37	−0.33	−0.39
0–15	4,651	0.32	0.22	0.31	−0.52	−0.52
0–20	6,511	0.28	0.30	0.31	−0.66	−0.70

$$E' = E + c \times \chi^2_{CS} \quad (8)$$

The ROSETTA full atom energy, represented as E in (8), was first calculated using the *score_jd2* function in ROSETTA 3.1. E was then adjusted by χ^2_{CS}, (9), multiplied by a weighting factor c set to 0.25. χ^2_{CS} was calculated as a summation over backbone spins for a given residue, i, and residues, j as follows:

$$\chi^2_{CS} = \sum_i \sum_j \left(\delta^{\exp}_{i,j} - \delta^{pred}_{i,j}\right)^2 \Big/ \sigma^2_{i,j} \quad (9)$$

where $\delta^{\exp}$ represents either the experimental chemical shift (BMRB) or the SPARTA+ predicted chemical shift from the crystal structure (SPARTA+), δ^{pred} the SPARTA+ predicted chemical shift for the model and σ, the uncertainty in the chemical shift. The SPARTA+ csout option was used to facilitate the calculation of χ^2_{CS}. Linear regressions were then conducted for either the CS-ROSETTA score versus RMSD or χ^2_{CS} as calculated with experimental or predicted chemical shifts versus RMSD. Five RMSD windows were selected for this comparison, and the Pearson correlation coefficients are reported in Table 2.

Results

To predict the RMSD (Å) of comparative models using NMR-derived secondary structure, we developed (1), (3) and (5) and used STRIDE to define the model secondary structure. As a positive control, we first tested the extent to which the known secondary structures of various target structures could be used to select the best molecular models according to (5). Thus, all 7,474 models were scored based on their adherence to the secondary structure observed for each individual amino acid in the target structure. The scores were then plotted (Fig. 3a) against the model-to-target RMSDs resulting in a Pearson correlation coefficient (R) of −0.77 (±0.01), and an average ΔRMSD of 1.60 Å, where ΔRMSD is calculated according to (7) (see "Materials and methods"). This result captures the essential relationship between secondary structure and RMSD and

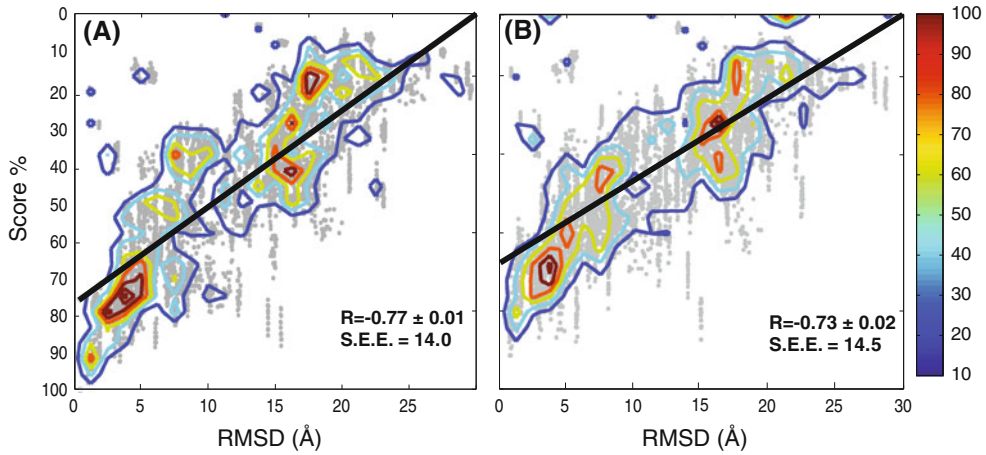

Fig. 3 Site-specific scoring of all amino acids using **a** STRIDE versus **b** SHIFTX/CSI methods for determining secondary structure reveals a similar correlation between model score and RMSD (Å). In both instances, the model scores are determined by (5) and (6), and are plotted against RMSD (*grey scatter plots*). To better visualize densely populated regions in the scatter plot, each point is binned in a 25 × 25 matrix using nearest neighbors interpolation over the same range and domain of the scatter plot. Contours are drawn according to the number of points within a bin ranging from 100 (*red*) to 10 (*dark blue*). Finally, linear regression analyses are performed on each of the scatter plots, and the resultant linear relationships are plotted

represents an upper threshold for agreement between models and actual structures as measured by secondary structure within the context of known protein structures.

This analysis was then repeated in order to evaluate the effect of filtering raw secondary structural information through an intermediate of predicted NMR chemical shift information using SHIFTX. We did so by replacing the actual secondary structure of both the target structures and the models with the secondary structure determined by predicting the resonances of the target structures using SHIFTX, and then deriving the secondary structure from the Chemical Shift Index (CSI). This serves as an intermediate step to our final goal, to take experimental chemical shifts directly and compare the secondary structures derived from these to the model in order to generate the score. Provided that the SHIFTX and CSI algorithms are robust, we would expect a high similarity with the positive control noted above. In this case, R was reduced minimally to −0.73 (±0.02), while the standard error of estimate was increased minimally to 14.5 and ΔRMSD increased to 2.91 Å (Fig. 3b). Aside from the change in ΔRMSD, only minute differences exist between the STRIDE and SHIFTX/CSI formulations of secondary structure. Therefore, in all subsequent evaluations of *S* presented, unless otherwise stated, the target and model secondary structure were derived from the SHIFTX/CSI pathway illustrated in Fig. 2.

We next sought to identify the minimal dataset that optimized correlation according to (1), (3), or (5). All comparative models were scored according to (1), the amino acid type formulation. This facilitated assessment of each amino acid's relative value to the overall correlation. Model scores based on a single amino acid type demonstrated modest-to-negligible correlation with RMSD where *R* ranged between −0.52 (±0.02) (Ala) and −0.10 (±0.03) (Cys). The inherent assumption motivating our selection of a subset of amino acid types was that scores based on amino acid types garnering higher correlations would when paired have roughly additive correlations. Thus, we also scored all combinations of two amino acids to test the hypothesized additivity of model scores. The heat map shown in Fig. 4 illustrates a symmetric matrix of R for the 190 combinations of two amino acids with the diagonal representing values for the lone amino acid. By analyzing the individual 20 amino acid types and all combinations, patterns of primarily score additivity but also destructive non-additivity were observed, as visualized in the heat map. Only a few significant destructive patterns were observed, for instance, when Ser was paired with Ile the correlation (−0.40± 0.02) was diminished with respect to Ile in isolation (−0.47 ± 0.02). From these calculations, pairings that included five particular amino acids displayed a general additive trend in terms of their correlations: Val, Ile, Thr, Ala and Leu (VITAL) as visualized in Fig. 4.

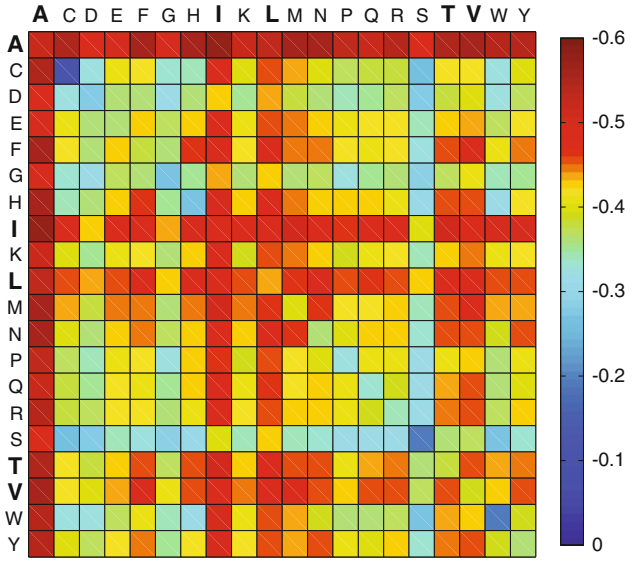

Fig. 4 Heat map of Pearson correlation coefficients for structure-to-model RMSD and model scores derived for the 190 combinations of amino acid type (*off-diagonal* elements) and the 20 individual amino acids (*diagonal*) using (1) and (2). The aliphatic amino acids A, I, M, L and V along with T and N have the greatest individual correlation to RMSD, and as indicated by the *bolded font*, pairings including the amino acids A, I, L, T and V (VITAL) have larger correlation coefficients overall

The aliphatic residues VIAL along with T thus became the focus of our analyses in addition to His, Pro, Gly, Met, Arg, Lys, Tyr and Ser. The amino acids His, Arg, Lys and Tyr amino acids were selected due to their inclusion in several auxotrophic *E. coli* strains. Ser and Gly were selected for their ease in assignment and because it was unclear whether their slight deleterious trend would be propagated to the site-specific and pairwise scoring mechanisms. Table 1 lists R, the Standard Estimate Error (SEE) and ΔRMSD for the site-specific (SS), type-specific (T), and pairwise (PW) applications of a select group of the scoring schemes; others are included in Table S1. To investigate the impact of subsets of information on these scoring functions, TAGS, VITAL, VITALGS, and VITLHRMK were compared to ALL 20 site-specific or ALL 20 type-specific (Fig. 5; Table 1); the trend shows that looking at only well scoring subsets of the amino acids increases the correlation for type-specific scoring (primarily VITAL), while both the number of amino acids and the correlations of those amino acids are important for SS or pairs scoring, with a modest increase in correlation available upon appropriate selection of amino acids.

The extent to which four of these scoring schemes sample the secondary structure of Maltose-binding Protein (MBP) is illustrated for VITAL, VITAL pairs, and VITLHRMK pairs in Fig. 6, a comparison that highlights how the reduced site-specific/type-specific and pairwise datasets

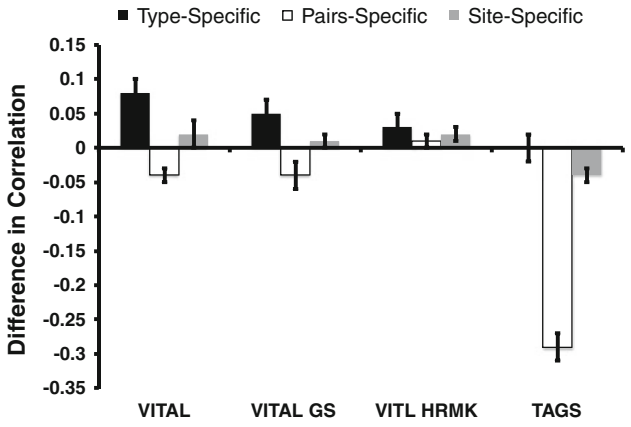

Fig. 5 Histograms indicating the change in correlation with respect to either type-specific (*black*), pairwise (*white*), or site-specific (*grey*) correlations calculated when presuming 100% complete assignments. Differences are presented for four of the amino acid subsets tested: VITAL, VITAL GS, VITL HRMK and TAGS. *Error bars* denote the square of the sum of the 99% confidence intervals for the two correlation coefficients subtracted

uniformly sample the entire structure of MBP. Figure 6 also captures the compromise between fewer sites with relatively easier to obtain pairwise assignments (Fig. 6c) versus more sites with more challenging to obtain site-specific assignments (Fig. 6b) and the subsequent loss of conformational sampling that results. Using a VITLHRMK auxotroph in order to limit scattering provides coverage commensurate to our VITAL scheme (Fig. 6d).

With the above information in mind, the following scoring schemes were tested, analyzed and listed in Table 1: (1) Amino acid type scores (1) that relied on the following combinations of amino acids: ALL 20, TAGS, VITAGSP, VITAL, VITALGS, VILHY, and VITLHRMK; (2) pairwise and site-specific scores (3) relying on the following combinations of amino acids: ALL 20, TAGS, VITAGSP, VITAL, VITALGS, VILHY, and VITLHRMK; and (3) site-specific scores (5) relying on the following combinations of amino acids: ALL 20, TAGS, VITAGSP, VITAL, VITALGS, VILHY, and VITLHRMK. Additional combinations included various pairings of the residues VITALEFGHMQW and are presented in Table S1. Scoring schemes reliant on amino acid type held only modest-to-negligible correlation with RMSD that ranged from −0.48 (±0.02) for all amino acid types (Fig. 1b) to −0.57 (±0.02) for VITAL types. Strikingly, using the pairing of Ile and Ala achieved a correlation equivalent to VITAL types, −0.58 (±0.02), and using Ala alone fared as well as using all amino acid types ($R = -0.52 \pm 0.02$). Upon evaluating (5) using site-specific information for the same schemes enumerated above, correlations were universally

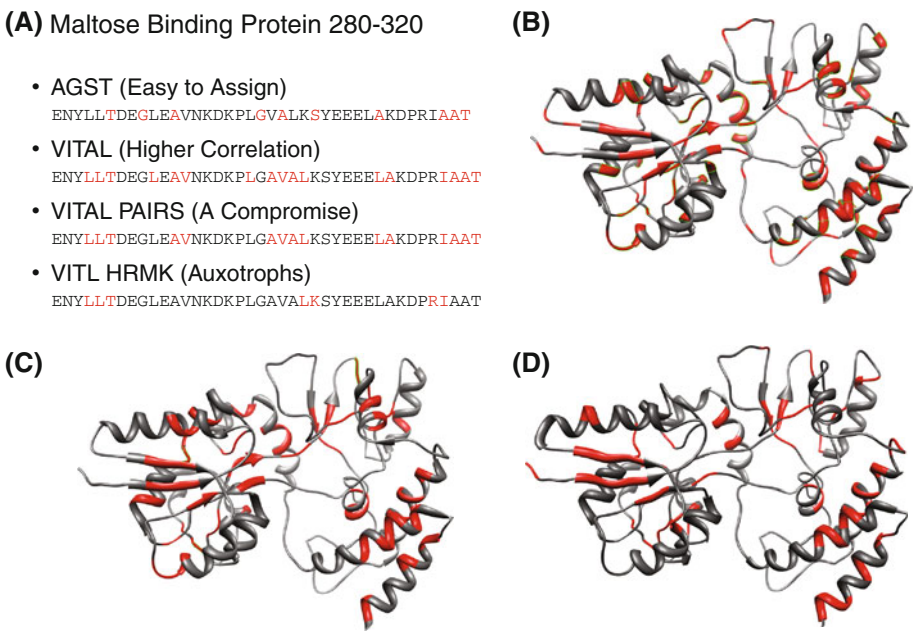

Fig. 6 Pictorial representation of how sampling by various scoring schemes differs for maltose-binding protein. Subsets of amino acids may be selected to achieve different aims: (1) AGST residues are in many cases unambiguously typed and consequently easier to assign relative to other amino acids, but the Pearson correlation coefficient observed for the site-specific AGST labeling scheme indicates that it is less predictive of RMSD than site-specific VITAL assignments. (2) VITAL site-specific assignments (illustrated in **b**) are more difficult to obtain than AGST, but are as predictive of RMSD as complete site-specific assignments. (3) VITAL pairwise assignments (illustrated in **c**) are easier to obtain, but have correlations on par with AGST site-specific assignments. (4) VITLHRMK pairwise assignments (illustrated in **d**) may be easier to obtain than VITAL pairwise assignments since an auxotropic *E. coli* strain exists whereby VITLHRMK residues may be exclusively isotopically labeled

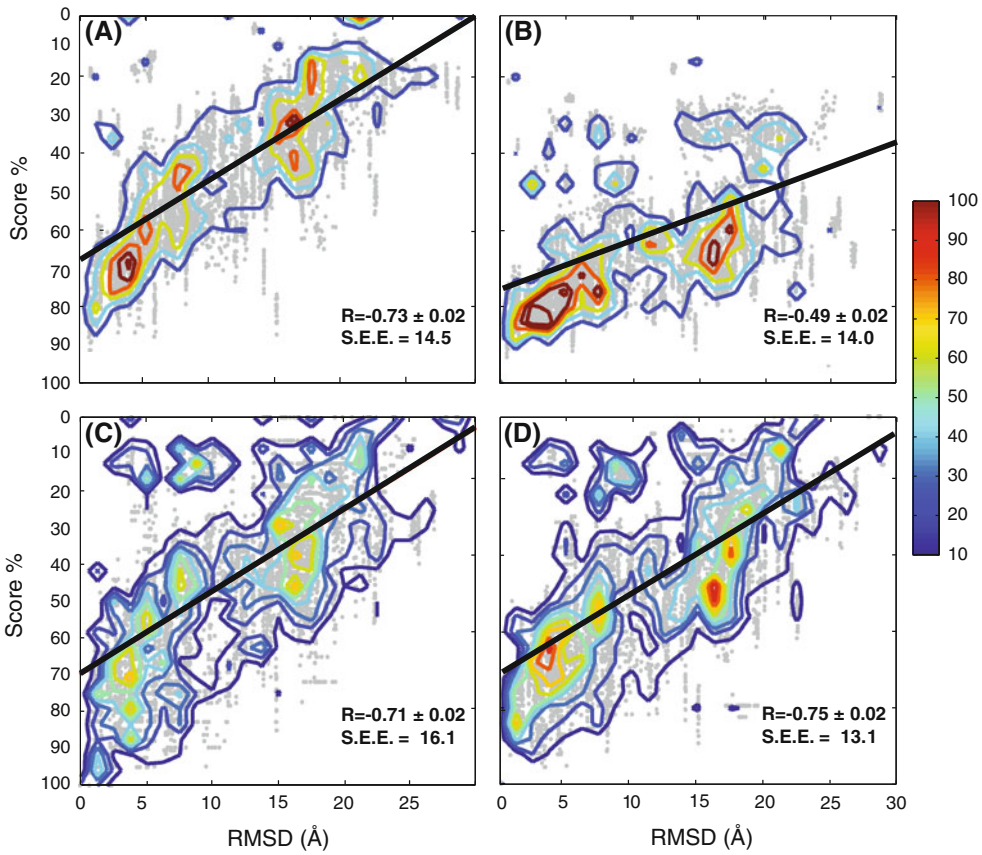

Fig. 7 Linear regression analyses of four scoring schemes: **a** ALL 20 site-specifically scored with (3) and (4); **b** all 20 type-specifically scored with (1) and (2); **c** VITAL pairs scored with (5) and (6); and **d** VITLHRMK pairs scored with (5) and (6). In all instances, the model scores are plotted against RMSD (*grey scatter plots*). To better visualize densely populated regions in the scatter plot, each point is binned in a 25 × 25 matrix using nearest neighbors interpolation over the same range and domain of the scatter plot. Contours are drawn according to the number of points within a bin ranging from 100 (*red*) to 10 (*dark blue*). Finally, linear regression analyses are performed on each of the scatter plots, and the resultant linear relationships are indicated

improved by an increment of at least −0.18 over their amino acid type counterpart, with the sole exception of site specifically labeled Ala, which had a correlation of −0.58 (±0.02). The significantly improved correlations ranged between −0.70 (±0.02) and −0.75 (±0.01), where VITAL and VITLHRMK site-specific assignments achieved a slightly higher correlation than 100% site-specific assignments (Fig. 7a). Although this latter difference is not outside of the 99% confidence interval, the site-specific score correlations are in general well outside of the 99% confidence interval of the correlations for the type-specific scores.

When using the pairwise scoring function (3), there was also a general improvement in correlation compared to type-specific scoring. Relative to site-specific scoring (5), pairwise scoring suffered a significant decrease in the number of comparisons between model and structure, which lead to fewer data points used in the scoring function and thus poor discrimination of model quality when using four or fewer amino acids. However, by moving to pairwise combinations of five or more amino acids (Fig. 5; Table 1), enough comparisons were generated to discriminate between models as indicated by a general trend of lower ΔRMSDs. This is not surprising since the percentage of amino acid pairs evaluated for a given 200 residue model increases from eight for pairs generated from four amino acid types to twelve for pairs generated from five amino acid types, assuming a random distribution of amino acids and that a given amino acid type only has a 1 in 20 probability of appearing in a protein. Weighting for the likelihood of specific amino acids appearing in a sequence modifies the result, but not the overall conclusion; each additional amino acid used in pairwise scoring confers an exponentially increasing number of sites. Astonishingly, we find that when using a subset of site specifically assigned amino acids, each observed to have a high correlation with RMSD (e.g., VITAL), a higher correlation is observed than for all of the amino acids site specifically assigned. This observation is supported by previous studies where motifs derived from pairs of amino acids have been used in discrimination of extracellular versus intracellular proteins, illustrating how pairs of amino acids can be used as a substitute for complete site-specific information (Nakashima and Nishikawa 1994). This implies that selection of the amino acids is important to maximize correlation and that we can obtain somewhat greater correlations by eliminating amino acids that have poor secondary structure tendencies (and/or complex relationships between chemical shift and secondary structure, such as cysteine), and thus poor correlation with RMSD.

In practice, the VITAL pairwise scoring scheme had a significantly better correlation (−0.71 ± 0.02) than the

TAGS scheme (−0.49 ± 0.02). By increasing the number of amino acids used in the pairwise scoring function to seven and eight residues (VITALGS and VITLHRMK pairs) the scatter as measured by the SEE was significantly reduced while the correlations increased in magnitude. Accordingly, VITLHRMK pairwise assignments outscored all other methods in terms of the correlation (−0.75), scatter (13.1) and ΔRMSD (1.9), all values essentially equivalent to the control calculation. While VITAL pairwise assignments (Fig. 7c) had a slightly worse correlation (−0.71 ± 0.2) and scatter (16.1), the ease of assigning fewer sites generally outweighs the better correlation and decreased scatter of VITLHRMK pairwise assignments (Fig. 7d). Furthermore, for membrane proteins, the VITLHRMK may be of little additional benefit due to the comparatively smaller number of charged residues found in the transmembrane segments of this class of proteins and the general scarcity of His. However, VITLHRMK has an auxotroph that could allow for easy labeling without scrambling, as is prevalent with alanine, and could be used to great effect by simplifying the labeling process. Therefore, the general robustness of site-specific scores including three or more residues and pairwise scores including more than four amino acids, suggests that the score may be tailored to the sequence of the protein of interest, the methods available, and/or the majority of amino acids assigned. Fig. 8 illustrates the VITAL pairwise scores plotted against RMSD for MBP (40.7 kDa, Tables S2, S3). Not surprisingly, the scores appear in clusters around a tight range of RMSD, and each cluster is associated with a different template (Fig. 8a). This is due to the comparative models adopting a structure closely akin to the template. Selecting the best and worst scoring models in this manner (compared to the structure determined via X-ray crystallography) is an additional control. The result supports the idea that a good template results in not only good alignment, but also an overall good model compared to a poor scoring model (Fig. 8b), which in some instances has no overlap with the structure at all (Fig. 8c). The template for this particular best model possessed 30% sequence identity to the target sequence, and thus supports our hypothesis that VITAL NMR can screen not only for the sequence identity matches, but more critically the secondary structure matches, to determine the best templates for future modeling and structure refinements.

Next, we examined additional amino acids (His, Met, Phe, Trp, Glu, and Gln) to determine their relative utility. Glu is a strong α-helical forming residue; combining it with VITAL provides an optimal blend of keeping the number of amino acids to a minimum, minimizing ΔRMSD (2.1 Å for site-specific, 1.5 Å for pairs, Table S1), and providing the best correlation of any sixth amino acid used for testing secondary structure (−0.72); however the minimal increase

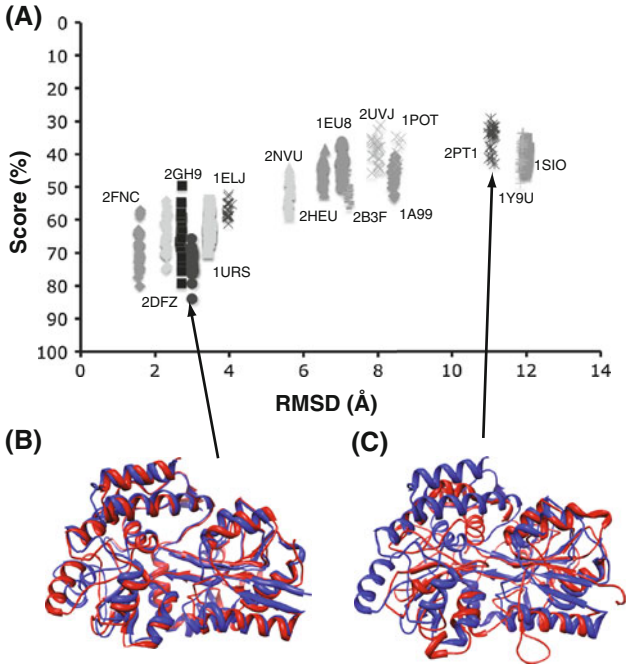

Fig. 8 VITAL pairwise analysis of maltose-binding protein (MBP: PDB ID: 1DMB; BMRB: 4354. **a** A plot of model scores calculated using VITAL pairwise assignments versus model-to-structure RMSD. **b** A superposition of the true structure (*blue*) with the best scoring model (*red*), which had a score of 83.9% and an RMSD to the crystal structure of 3.0 Å. **c** A superposition of the true structure (*blue*) with a poor scoring model (*red*), which had a score of 43.0% and an RMSD of 11.2 Å. *Arrows* point to the template used to generate the particular model. The models and structure of MBP were aligned with the VMD RMSD plugin and were rendered using Chimera. The best overall model had a score of 77.5% and an RMSD of 1.6 Å

in correlation as well as the difficulty in labeling and assigning Glu renders this an interesting finding and less practical than other labeling schemes suggested. Addition of further amino acids minimally increases correlation and has a negligible effect on ΔRMSD (Table S1), thus indicating that a small subset is comparable to a larger subset and that there is little need to further enhance the complexity of analysis and labeling schemes to increase the scoring function above 7–8 residues.

In order to understand how these global correlations and ΔRMSD values correspond to individual targets, the targets and their templates were analyzed to determine the conditions under which our scoring algorithm works best (Tables S2, S3). While the optimal correlation had little to do with number of templates or the sequence identity of the templates, the ΔRMSD correlated solely to the number of templates, implying that scaling up the number of templates should result in better and more accurate starting points to finding the true structure. This helps to explain why in some instances we observed large ΔRMSD values for some targets, and thus higher than desired standard deviations. This also demonstrated that the best use of this

methodology for protein fold identification is sampling the entire database of available templates to generate a full library for each target structure and then using experimental data to determine the most appropriate template for model generation and subsequent analysis.

We next sought to test whether the CS-ROSETTA full atom energy selects the best available protein templates and is correlated to RMSD over a range typically observed for comparative modeling. Previously, Pearson correlation coefficients were calculated for the 22 protein targets in one linear regression. Doing the same for CS-ROSETTA results in an artificially low correlation of 0.33. For a more fair comparison to the CS-ROSETTA full atom energy, the targets were considered separately and averaged, since the ROSETTA energy is highly target dependent. When averaging the Pearson correlation for each individual template CS-ROSETTA held a correlation to RMSD of 0.52 ± 0.28, as opposed to -0.68 ± 0.20 for VITAL pairwise scores and -0.73 ± 0.21 when assuming 100% site-specific assignments. Unfortunately, due to the relative paucity of comparative models between 0 and 5 Å for several of the 22 targets, we were unable to calculate an average Pearson correlation coefficient for various windows of RMSD for each individual target. The total Pearson correlation coefficient for all models within specified RMSD windows are reported in Table 2, and are consistent with the knowledge that the ROSETTA full atom energy better distinguishes models within 5 Å of the correct structure. The score based on chemical shift differences alone using either BMRB data or SPARTA+ predicted data, χ^2_{CS} from (9), performs slightly worse than the CS-ROSETTA in all RMSD windows between 2.5 and 15 Å. The difference between the CS-ROSETTA and χ^2_{CS} in the 0 to 20 Å window (0.02) is statistically insignificant relative to the performance of VITAL pairs and all amino acid site-specific scoring. Taken together, the results in Table 2 suggest that the CS-ROSETTA score complements VITAL PAIRS by its superior ability to select the best model once models are known to be within 2.5 Å of the actual structure.

Finally, to establish the utility of the VITAL pairwise score on a moderately sized membrane protein, we applied our method to the *E. coli* disulfide bond-generating protein DsbB, whose 21 kDa molecular weight is slightly beyond the current range for which SSNMR can produce de novo structures. From partial resonance assignments of DsbB made in previous studies (Li et al. 2007, 2008) and from recently acquired data (Tang et al. 2011), we calculated the likely secondary structure using CSI and tested whether our method could discriminate models based on incomplete assignments from experimental SSNMR data. With the VITAL pairwise score, we obtained a ΔRMSD value of 0.55 Å. The two best scoring models shared the same score

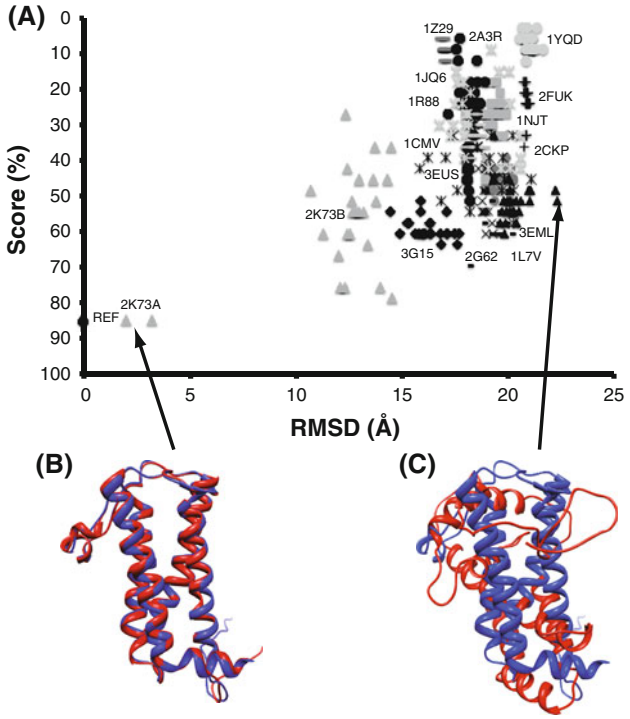

Fig. 9 VITAL pairwise analysis of the *E. coli* membrane protein DsbB. **a** A plot of model scores calculated using VITAL pairwise assignments versus model-to-structure RMSD for the *E. coli* disulfide bond generating membrane protein DsbB (PDB ID: 2ZUQ). **b** A superposition of the true structure (*blue*) with the best scoring model (*red*). **c** A superposition of the true structure (*blue*) with a poor scoring model (*red*). *Arrows* point to the template used to generate the particular model. The models and structure of DsbB were aligned with the VMD RMSD plugin and were rendered using Chimera. The two best scoring models and also the two best models had scores of 84.9% and RMSDs of 3.2 and 2.0 Å to the crystal structure from 2ZUQ

and had an average RMSD of 2.65 Å, while the minimum RMSD observed was 2.1 Å. These were the only models that had a score of greater than 80% out of 375 models generated from 15 templates (Fig. 9). Using VITAGSP pairs, AGST pairs, VITAL or AGST, we were unable to discriminate between the best models, supporting that selection of the VITAL residues or the analogous auxotroph VITLHRMK and the pairwise scoring function for experimental data with incomplete assignments. This provides a proof-of-concept that our methodology can be applied to proteins of unknown structure and with incomplete backbone assignments.

Discussion

Numerical measures of the agreement between the secondary structure of a comparative model and that predicted from its amino acid sequence have been used previously to

rank molecular models (Eramian et al. 2008; Wallner and Elofsson 2003). Indeed, secondary structure was identified as the primary component of the composite model score SVMod, with a relative contribution of ~63%, where an accessible surface score comprised the second largest contribution of ~18% (Eramian et al. 2008). In SVMod, secondary structure is accounted for in terms of two scores, PSIPRED% and PSIPRED-weight. In the optimal case, the model secondary structure is in perfect compliance with the predicted secondary structure; however, predicted secondary structure has at most about ~75% accuracy to the true secondary structure (Frishman and Argos 1997; Rost 1999). While the PSIPRED-weight score resulted in an impressive average correlation of 0.86 to RMSD, this value represents an average of Pearson correlations calculated for each model target of the SVMod training set. A correlation drawn from the entire training set was not reported, but should be considerably less than 0.87 as the regression slopes and intercepts reported for each model target vary significantly. Furthermore, the SVMod model testing set (MODPIPE) is more comparable with our model testing set due to its variability in sequence and alignment length. SVMod had a significantly diminished average correlation for the MODPIPE testing set (0.68), and predicted RMSDs of ~5 Å for nearly all models between 5 and 20 Å RMSD. Thus, using secondary structure as a means to validate models is not unprecedented, and the increased accuracy of SSNMR secondary structure observations translates into higher overall correlations with RMSD and a score better predictive of RMSD, even when a limited subset of amino acids are evaluated.

While an enhanced predictive power of secondary structure based model scoring is an important finding, from a practical standpoint, it is imperative to discover partial assignments or type assignments that result in correlations similar to the case where every amino acid is unambiguously site specifically assigned by investigating the importance of individual amino acids to the score. When examining scores based on amino acid type, correlations are highest for those types that propagate secondary structure in a cooperative manner. For instance, branched-chain amino acids (e.g., Val, Ile and Thr among the VITAL subset) have coupled backbone and sidechain conformational preferences that propagate along the peptide backbone (Engelman et al. 2000; Senes et al. 2004; Swindells et al. 1995). When one surveys the relative propensities of amino acids for secondary structure type, one finds that the amino acids comprising VITAL possess amongst the strongest propensities for either α-helical (A and L) or β-strand (I, T, and V) secondary structure (Chou and Fasman 1974). Indeed, the residue most likely to be found in α-helices, Glu, also provides the best single addition to VITAL as measured by correlation to RMSD (−0.72, Table S1). Additional residues beyond VITAL have an almost negligible benefit to the predictive ability of the model score according to correlation (−0.71 for VITAL to −0.75 for VITLHRMK), the ΔRMSD and SEE are reduced by ~15% by switching from VITAL (SEE = 16.1, ΔRMSD = 2.2) to the auxotroph VITLHRMK (SEE = 13.1, ΔRMSD = 1.9, Table 1). It is promising that the inherent limit of correlation between secondary structure and RMSD can be approached using VITAL, a subset of the amino acids (~25–35% of a given protein sequence); and surprisingly, leads to a correlation slightly higher than when all amino acids are site specifically scored using SHIFTX/CSI. We attribute that this effect is in part due to the fact that amino acids with ionizable sites and/or aromatic rings have more complex chemical shift dependencies on conformation, as well as electrostatic and ring current phenomena, (e.g., cysteine). It is also likely that certain amino acids are highly prone to certain conformations (e.g., proline in random coil) and thus add little information while other amino acids appear to have a deleterious effect on correlation (e.g., serine, glycine) (Fig. 5).

The ability of an amino acid type to report on secondary structure varies dramatically as illustrated in Fig. 5, and thus the selection of an amino acid subset to be scored is important. Here, we conclude that if one can at least assign pairs of VITAL, one can obtain a reasonable estimate of the model RMSD from its native structure. This assumption appears to be robust when VITAL pairwise assignments are incomplete, as demonstrated for DsbB. Since several subsets of amino acids performed within range of VITAL pairwise assignments (see Tables 1, S1), this methodology also provides a measure of inherent flexibility when applied to a particular protein. Accordingly, scoring schemes can be tailored to the target protein sequence, the preponderance of assigned residues, regions of structural and chemical interest, and to certain auxotrophic labeling patterns. The ability to tune the score promises more successful applications of this method, and permits one to search according to a number of different schemes to identify the most consistently high-ranking models.

We envision that the primary application of our method will be to filter out models in poor agreement with available experimental data. Based on our analyses, the scoring threshold for selecting the most representative models is dependent on the type of scoring implemented. Thus, for site-specific scoring of all amino acids, models scoring greater than 70% should be compared for structural convergence. The average RMSD for models above 80% was 2.91 ± 4.16 Å and above 85% was 2.1 ± 3.19 Å. The threshold increases for scoring schemes that incorporate less experimental data. For instance, when scoring VITAL pairs of amino acids, a threshold of 90% selects models on

average within 3.1 ± 2.46 Å of their actual structure and increased to 1.23 ± 1.5 Å when a threshold of 95% was used. The relatively high standard deviation is due to the range of scores that accompany the 25 models from a given template due to propagation of error of assigning secondary structure, and thus could be improved even further by using the average score of each model for a given template. Additionally, we believe that the VITAL NMR method can be used synergistically with programs such as CS-ROSETTA and CHESHIRE, which improve upon the structure generation process when complete chemical shift assignments are available. This will help to eliminate the obvious deficiency of our score to discern when a protein adopts an improper conformation but retains the correct secondary structure. By focusing on the most representative models, computationally expensive energetic analyses can be utilized more productively, and molecular dynamics and/or simulated annealing-based structure refinement seeded with improved initial conditions. In principle, higher-quality model structures may be generated than produced via homology modeling alone, because templates and alignments can be iterated based on the predicted RMSD while none of the currently available model quality assessment scores possess an absolute relationship with RMSD, as demonstrated here. Instead, correlations are reported on a case-by-case basis because the model scores can only order models in a relative sense.

Conclusions

We have demonstrated for the first time that unassigned and/or incomplete NMR data can be used as a coarse prediction of RMSD, and that a limited pairwise assignment of the amino acids comprising VITAL is as predictive of RMSD as all amino acids site specifically assigned. The correlation between secondary structure, as measured by SSNMR, and RMSD can be exploited in the determination of templates for remotely homologous sequences bearing less than 30% sequence identity. Better templates and better models can in turn expedite structure elucidation. Future work includes testing more cases where actual SSNMR data is available and defining a prescriptive limit on the usefulness of fewer partial assignments. Work will also be done to pair the CSI derived secondary structures with PSI-PRED to more accurately and confidently assign secondary structure to isolated pairs of amino acids as opposed to the current method which utilizes all available assignments to determine secondary structure. Other metrics such as solvent accessible surface area will be explored, but only utilized if they retain the absolute predictive ability of the current site-specific and pairwise scoring functions. As we continue to develop our model score, we envision that our method will be applicable to experimental SSNMR data from larger proteins, and open the door to solving the structure of DsbB de novo.

Acknowledgments The authors thank the National Institute of Health for funding through R01GM79530, R01GM75937, NRSA (F32 GM095344), the Ruth L. Kirschstein National Research Service Award to AEN and the Chemical Biology Interface Training Program (GM070421-06) to MCB and the Department of Homeland Security Fellowship Program to MCB, as well as Dr. Ying Li, Dr. Aleksandra Kijac, and Dr. Andrew Nieuwkoop for early assistance on this project.

References

Alber F, Förster F, Korkin D, Topf M, Sali A (2008) Integrating diverse data for structure determination of macromolecular assemblies. Annu Rev Biochem 77:443–477. doi:10.1146/Annurev.Biochem.77.060407.135530

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215(3):403–410

Baranov VI, Morozov IY, Ortlepp SA, Spirin AS (1989) Gene-expression in a cell-free system on preparative scale. Gene 84(2):463–466

Baudry J, Rupasinghe S, Schuler MA (2006) Class-dependent sequence alignment strategy improves the structural and functional modeling of P450. Protein Eng Des Sel 19(8):345–353

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

Bertini I, Bhaumik A, De Paëpe G, Griffin RG, Lelli M, Lewandowski JR, Luchinat C (2010) High-resolution solid-state NMR Structure of a 17.6 kDa protein. J Am Chem Soc 132(3):1032–1040. doi:10.1021/Ja906426p

Bowie JU, Luthy R, Eisenberg D (1991) A method to identify protein sequences that fold into a known 3-dimensional structure. Science 253(5016):164–170

Carugo O, Pongor S (2001) A normalized root-mean-square distance for comparing protein three-dimensional structures. Protein Sci 10(7):1470–1473

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(6911):98–102. doi:10.1038/Nature01070

Chou PY, Fasman GD (1974) Prediction of protein conformation. Biochemistry 13(2):222–245

De Angelis AA, Howell SC, Nevzorov AA, Opella SJ (2006) Structure determination of a membrane protein with two transmembrane helices in aligned phospholipid bicelles by solid-state NMR spectroscopy. J Am Chem Soc 128(37):12256–12267. doi:10.1021/Ja063640w

Endo Y, Sawasaki T (2003) High-throughput, genome-scale protein production method based on the wheat germ cell-free expression system. Biotechnol Adv 21(8):695–713. doi:10.1016/S0734-9750(03)00105-8

Engelman DM, Brunger A, Cocco M, Fleming K, Gerstein M, Mackenzie K, Prestegard J, Russ W, Senes A, Zhou F (2000) Helix interactions in membrane proteins. Faseb J 14(8):A1506

Eramian D, Eswar N, Shen MY, Sali A (2008) How well can the accuracy of comparative protein structure models be predicted? Protein Sci 17(11):1881–1893. doi:10.1110/Ps.036061.108

Eswar N, Marti-Renom MA, Webb B, Madhusudhan MS, Eramian D, Shen M, Pieper U, Sali A (2000) Comparative protein structure

modeling with MODELLER. Curr Protoc Bioinform Suppl 15:5.6.1–5.6.30
Fasnacht M, Zhu J, Honig B (2007) Local quality assessment in homology models using statistical potentials and support vector machines. Protein Sci 16(8):1557–1568
Fischer MW, Losonczi JA, Weaver JL, Prestegard JH (1999) Domain orientation and dynamics in multidomain proteins from residual dipolar couplings. Biochemistry 38(28):9013–9022
Forrest LR, Tang CL, Honig B (2006) On the accuracy of homology modeling and sequence alignment methods applied to membrane proteins. Biophys J 91(2):508–517. doi:10.1529/Biophysj.106.082313
Franks WT, Kloepper KD, Wylie BJ, Rienstra CM (2007) Four-dimensional heteronuclear correlation experiments for chemical shift assignment of solid proteins. J Biomol NMR 39(2):107–131. doi:10.1007/S10858-007-9179-1
Franks WT, Wylie BJ, Schmidt HLF, Nieuwkoop AJ, Mayrhofer RM, Shah GJ, Graesser DT, Rienstra CM (2008) Dipole tensor-based atomic-resolution structure determination of a nanocrystalline protein by solid-state NMR. Proc Natl Acad Sci USA 105(12):4621–4626
Frericks HL, Zhou DH, Yap LL, Gennis RB, Rienstra CM (2006) Magic-angle spinning solid-state NMR of a 144 kDa membrane protein complex: *E. coli* cytochrome bo_3 oxidase. J Biomol NMR 36(1):55–71. doi:10.1007/S10858-006-9070-5
Frishman D, Argos P (1997) Seventy-five percent accuracy in protein secondary structure prediction. Proteins 27(3):329–335
Grishaev A, Tugarinov V, Kay LE, Trewhella J, Bax A (2008) Refined solution structure of the 82-kDa enzyme malate synthase G from joint NMR and synchrotron SAXS restraints. J Biomol NMR 40(2):95–106. doi:10.1007/S10858-007-9211-5
Grzesiek S, Bax A (1993) Amino-acid type determination in the sequential assignment procedures of uniformly C-13/N-15-enriched proteins. J Biomol NMR 3(2):185–204
Hanson MA, Stevens RC (2009) Discovery of new GPCR biology: one receptor structure at a time. Structure 17(1):8–14. doi:10.1016/J.Str.2008.12.003
Heinig M, Frishman D (2004) STRIDE: a web server for secondary structure assignment from known atomic coordinates of proteins. Nucleic Acids Res 32:W500–W502. doi:10.1093/Nar/Gkh429
Hiller S, Garces RG, Malia TJ, Orekhov VY, Colombini M, Wagner G (2008) Solution structure of the integral human membrane protein VDAC-1 in detergent micelles. Science 321(5893):1206–1210. doi:10.1126/Science.1161302
Hooft RWW, Vriend G, Sander C, Abola EE (1996) Errors in protein structures. Nature 381(6580):272
Humphrey W, Dalke A, Schulten K (1996) VMD: visual molecular dynamics. J Mol Graph 14(1):33–38
Ikura M, Kay LE, Krinks M, Bax A (1991) Triple-resonance multidimensional NMR-study of calmodulin complexed with the binding domain of skeletal-muscle myosin light-chain kinase—indication of a conformational change in the central helix. Biochemistry 30(22):5498–5504
Jehle S, Rajagopal P, Bardiaux B, Markovic S, Kühne R, Stout JR, Higman VA, Klevit RE, van Rossum BJ, Oschkinat H (2010) Solid-state NMR and SAXS studies provide a structural basis for the activation of alphaB-crystallin oligomers. Nat Struct Mol Biol 17(9):1037–1042. doi:10.1038/Nsmb.1891
Kelly K (1999) Multiple sequence and structure refinement in MOE. Chemical Computing Group Inc. http://www.chemcomp.com/journal/align.htm
Laskowski RA, Rullmann JAC, MacArthur MW, Kaptein R, Thornton JM (1996) AQUA and PROCHECK-NMR: programs for checking the quality of protein structures solved by NMR. J Biomol NMR 8(4):477–486
Li Y, Berthold DA, Frericks HL, Gennis RB, Rienstra CM (2007) Partial ^{13}C and ^{15}N chemical-shift assignments of the disulfide-bond-forming enzyme DsbB by 3D magic-angle spinning NMR spectroscopy. Chembiochem 8(4):434–442. doi:10.1002/Cbic.200600484
Li Y, Berthold DA, Gennis RB, Rienstra CM (2008) Chemical shift assignment of the transmembrane helices of DsbB, a 20-kDa integral membrane enzyme, by 3D magic-angle spinning NMR spectroscopy. Protein Sci 17(2):199–204
Lin MT, Sperling LJ, Frericks Schmidt HL, Tang M, Gennis RB, Rienstra CM (2011) A rapid and robust method for selective isotope labeling of proteins for solid-state NMR and pulsed EPR studies. Methods. doi:10.1016/j.ymeth.2011.08.019
Loquet A, Bardiaux B, Gardiennet C, Blanchet C, Baldus M, Nilges M, Malliavin T, Bockmann A (2008) 3D structure determination of the Crh protein from highly ambiguous solid-state NMR restraints. J Am Chem Soc 130(11):3579–3589. doi:10.1021/Ja078014t
Manolikas T, Herrmann T, Meier BH (2008) Protein structure determination from ^{13}C spin-diffusion solid-state NMR spectroscopy. J Am Chem Soc 130(12):3959–3966
Marassi FM, Opella SJ (2003) Simultaneous assignment and structure determination of a membrane protein from NMR orientational restraints. Protein Sci 12(3):403–411. doi:10.1110/Ps.0211503
Marti-Renom MA, Stuart AC, Fiser A, Sánchez R, Melo F, Sali A (2000) Comparative protein structure modeling of genes and genomes. Annu Rev Biophys Biomol Struct 29:291–325
Melo F, Feytmans E (1997) Novel knowledge-based mean force potential at atomic level. J Mol Biol 267(1):207–222
Mercier KA, Baran M, Ramanathan V, Revesz P, Xiao R, Montelione GT, Powers R (2006) FAST-NMR: functional annotation screening technology using NMR spectroscopy. J Am Chem Soc 128(47):15292–15299. doi:10.1021/Ja0651759
Mobarec JC, Sanchez R, Filizola M (2009) Modern homology modeling of G-protein coupled receptors: which structural template to use? J Medicinal Chem 52(16):5207–5216. doi:10.1021/Jm9005252
Monleon D, Colson K, Moseley HN, Anklin C, Oswald R, Szyperski T, Montelione GT (2002) Rapid analysis of protein backbone resonance assignments using cryogenic probes, a distributed Linux-based computing architecture, and an integrated set of spectral analysis tools. J Struc Funct Genomics 2(2):93–101
Nakashima H, Nishikawa K (1994) Discrimination of intracellular and extracellular proteins using amino-acid-composition and residue-pair frequencies. J Mol Biol 238(1):54–61
Neal S, Nip AM, Zhang HY, Wishart DS (2003) Rapid and accurate calculation of protein ^{1}H, ^{13}C and ^{15}N chemical shifts. J Biomol NMR 26(3):215–240
Oberai A, Ihm Y, Kim S, Bowie JU (2006) A limited universe of membrane protein families and folds. Protein Sci 15(7):1723–1734
Pearson WR (1996) Effective protein sequence comparison. Method Enzymol 266:227–258
Pieper U, Eswar N, Webb BM, Eramian D, Kely L, Barkan DT, Carter H, Mankoo P, Karchin R, Marti-Renom MA, Davis FP, Sali A (2009) MODBASE, a database of annotated comparative protein structure models and associated resources. Nucleic Acids Res 37(Database issue):D347–D354
Raman S, Lange OF, Rossi P, Tyka M, Wang X, Aramini J, Liu GH, Ramelot TA, Eletsky A, Szyperski T, Kennedy MA, Prestegard J, Montelione GT, Baker D (2010) NMR Structure determination for larger proteins using backbone-only data. Science 327(5968):1014–1018. doi:10.1126/Science.1183649
Randazzo A, Acklin C, Schafer BW, Heizmann CW, Chazin WJ (2001) Structural insight into human Zn^{2+}-bound S100A2 from

NMR and homology modeling. Biochem Bioph Res Co 288(2): 462–467
Ray A, Lindahl E, Wallner B (2010) Model quality assessment for membrane proteins. Bioinformatics 26(24):3067–3074
Robustelli P, Kohlhoff K, Cavalli A, Vendruscolo M (2010) Using NMR chemical shifts as structural restraints in molecular dynamics simulations of proteins. Structure 18(8):923–933. doi: 10.1016/J.Str.2010.04.016
Rost B (1999) Twilight zone of protein sequence alignments. Protein Eng 12(2):85–94
Sali A, Blundell TL (1993) Comparative protein modeling by satisfaction of spatial restraints. J Mol Biol 234(3):779–815
Sawasaki T, Ogasawara T, Morishita R, Endo Y (2002) A cell-free protein synthesis system for high-throughput proteomics. Proc Natl Acad Sci USA 99(23):14652–14657. doi:10.1073/Pnas.232580399
Schröder GF, Levitt M, Brunger AT (2010) Super-resolution biomolecular crystallography with low-resolution data. Nature 464(7292):1218–1222. doi:10.1038/Nature08892
Schwarz D, Dotsch V, Bernhard F (2008) Production of membrane proteins using cell-free expression systems. Proteomics 8(19):3933–3946. doi:10.1002/Pmic.200800171
Senes A, Engel DE, DeGrado WF (2004) Folding of helical membrane proteins: the role of polar, GxxxG-like and proline motifs. Curr Opin Struct Biol 14(4):465–479. doi:10.1016/J.Sbi.2004.07.007
Shen Y, Bax A (2007) Protein backbone chemical shifts predicted from searching a database for torsion angle and sequence homology. J Biomol NMR 38(4):289–302. doi:10.1007/S10858-007-9166-6
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(1):13–22. doi: 10.1007/S10858-010-9433-9
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(12):4685–4690. doi: 10.1073/Pnas.0800256105
Shen Y, Delaglio F, Cornilescu G, Bax A (2009) TALOS + : a hybrid method for predicting protein backbone torsion angles from NMR chemical shifts. J Biomol NMR 44(4):213–223. doi: 10.1007/S10858-009-9333-Z
Sippl MJ (1993) Recognition of errors in 3-dimensional structures of proteins. Proteins 17(4):355–362
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(14):5490–5492
Swindells MB, Macarthur MW, Thornton JM (1995) Intrinsic phi, psi propensities of amino-acids, derived from the coil regions of known structures. Nat Struct Biol 2(7):596–603
Tang MT, Sperling LJ, Berthold DA, Schwieters CD, Nesbitt AE, Niuwkoop AJ, Gennis RB, Rienstra CM (2011) High-resolution membrane protein structure by joint calculations with solid-state NMR and X-ray experimental data. J Biol NMR. doi: 10.1007/s10858-011-9565-6
Traaseth NJ, Shi L, Verardi R, Mullen DG, Barany G, Veglia G (2009) Structure and topology of monomeric phospholamban in lipid membranes determined by a hybrid solution and solid-state NMR approach. Proc Natl Acad Sci USA 106(25):10165–10170. doi:10.1073/Pnas.0904290106
Tycko R, Hu KN (2010) A Monte Carlo/simulated annealing algorithm for sequential resonance assignment in solid state NMR of uniformly labeled proteins with magic-angle spinning. J Magn Reson 205(2):304–314
Van Horn WD, Kim HJ, Ellis CD, Hadziselimovic A, Sulistijo ES, Karra MD, Tian CL, Sönnichsen FD, Sanders CR (2009) Solution nuclear magnetic resonance structure of membrane-integral diacylglycerol kinase. Science 324(5935):1726–1729. doi:10.1126/Science.1171716
Van Melckebeke H, Wasmer C, Lange A, Eiso AB, Loquet A, Böckmann A, Meier BH (2010) Atomic-resolution three-dimensional structure of HET-s(218-289) amyloid fibrils by solid-state NMR spectroscopy. J Am Chem Soc 132(39):13765–13775. doi: 10.1021/Ja104213j
Vila JA, Aramini JM, Rossi P, Kuzin A, Su M, Seetharaman J, Xiao R, Tong L, Montelione GT, Scheraga HA (2008) Quantum chemical ^{13}Cα chemical shift calculations for protein NMR structure determination, refinement, and validation. Proc Natl Acad Sci USA 105(38):14389–14394. doi:10.1073/Pnas.0807105105
Wallner B, Elofsson A (2003) Can correct protein models be identified? Protein Sci 12(5):1073–1086. doi:10.1110/Ps.0236803
Waugh DS (1996) Genetic tools for selective labeling of proteins with alpha-^{15}N-amino acids. J Biomol NMR 8(2):184–192
Weichenberger CX, Sippl MJ (2006) NQ-Flipper: validation and correction of asparagine/glutamine amide rotamers in protein crystal structures. Bioinformatics 22(11):1397–1398. doi: 10.1093/Bioinformatics/Bt/128
White SH (2009) Biophysical dissection of membrane proteins. Nature 459(7245):344–346. doi:10.1038/Nature08142
Wishart DS, Sykes BD (1994) The ^{13}C chemical-shift index—a simple method for the identification of protein secondary structure using ^{13}C chemical-shift data. J Biomol NMR 4(2): 171–180
Wylie BJ, Schwieters CD, Oldfield E, Rienstra CM (2009) Protein structure refinement using ^{13}Cα chemical shift tensors. J Am Chem Soc 131(3):985–992. doi:10.1021/Ja804041p
Yang YH, Ramelot TA, McCarrick RM, Ni SS, Feldmann EA, Cort JR, Wang HA, Ciccosanti C, Jiang M, Janjua H, Acton TB, Xiao R, Everett JK, Montelione GT, Kennedy MA (2010) Combining NMR and EPR methods for homodimer protein structure determination. J Am Chem Soc 132(34):11910–11913. doi: 10.1021/Ja105080h
Yarnitzky T, Levit A, Niv MY (2010) Homology modeling of G-protein-coupled receptors with X-ray structures on the rise. Curr Opin Drug Discov Devel 13(3):317–325