CABS-NMR—De Novo Tool for Rapid Global Fold Determination From Chemical Shifts, Residual Dipolar Couplings and Sparse Methyl-Methyl NOEs

DOROTA LATEK, ANDRZEJ KOLINSKI
Faculty of Chemistry, University of Warsaw, Pasteura 1, 02-093 Warsaw, Poland

Received 9 November 2009; Revised 27 June 2010; Accepted 27 June 2010
DOI 10.1002/jcc.21640
Published online 30 August 2010 in Wiley Online Library (wileyonlinelibrary.com).

Abstract: Recent development of nuclear magnetic resonance (NMR) techniques provided new types of structural restraints that can be successfully used in fast and low-cost global protein fold determination. Here, we present CABS-NMR, an efficient protein modeling tool, which takes advantage of such structural restraints. The restraints are converted from original NMR data to fit the coarse grained protein representation of the C-Alpha-Beta-Side-group (CABS) algorithm. CABS is a Monte Carlo search algorithm that uses a knowledge-based force field. Its versatile structure enables a variety of protein-modeling protocols, including purely de novo folding, folding guided by restraints derived from template structures or, structure assembly based on experimental data. In particular, CABS-NMR uses the distance and angular restraints set derived from various NMR experiments. This new modeling technique was successfully tested in structure determination of 10 globular proteins of size up to 216 residues, for which sparse NMR data were available. Additional detailed analysis was performed for a S100A1 protein. Namely, we successfully predicted Nuclear Overhauser Effect signals on the basis of low-energy structures obtained from chemical shifts by CABS-NMR. It has been observed that utility of chemical shifts and other types of experimental data (i.e. residual dipolar couplings and methyl-methyl Nuclear Overhauser Effect signals) in the presented modeling pipeline depends mainly on size of a protein and complexity of its topology. In this work, we have provided tools for either post-experiment processing of various kinds of NMR data or fast and low-cost structural analysis in the still challenging field of new fold predictions.

 J Comput Chem 32: 536–544, 2011

Key words: protein structure prediction; de novo protein folding; methyl-methyl NOEs; chemical shifts; residual dipolar couplings; backbone reconstruction; reduced models; Monte Carlo simulations

Introduction

In the recent years, we have observed steady growth of a number of nuclear magnetic resonance (NMR) structures in the PDB. It is mainly due to new, high-throughput techniques that place NMR as a complementary approach to X-ray crystallography in protein structure determination. Many proteins that are difficult or even impossible to crystallize can be examined without serious obstacles by modern NMR methods. That important issue is often raised by experimentalists from large-scale protein structure production centers of National Institutes of Health - National Institute of General Medical Sciences (NIH - NIGMS) Protein Structure Initiative.[1] Recent development of Transverse Relaxation Optimized Spectroscopy and new techniques of selective protonation and deuteration, which focus on such observables as methyl-methyl Nuclear Overhauser Effect signals (NOEs), has extended the size limit of protein NMR.[2] A significant improvement in the solid state spectra resolution by Magic Angle Spinning techniques and weak alignment of samples made possible to examine even such difficult objects as membrane proteins.[3]

An important issue regarding NMR techniques is extremely high diversity of the recorded data, in the sense of quality and sort of structural information. This requires efficient computational methods for the data processing and analyzing. It is especially inevitable in the case of such sparse NMR data as residual dipolar couplings, chemical shifts, or methyl-methyl NOEs, which are relatively easy to obtain but carry incomplete information about a protein fold. Such data, which are usually translated to a limited set of structural restraints, need to be thoroughly included in high-throughput computational pipelines to exploit its full potential and significantly decrease the time of NMR structure determination. In

Correspondence to: D. Latek; e-mail: pledor@chem.uw.edu.pl

Contract/grant sponsor: Polish Ministry of Science and Higher Education; contract/grant numbers: N N204 239934 and NN301465634

Contract/grant sponsor: NIH; contract/grant number: 1R01GM081680

Contract/grant sponsor: Computing Centre of the Department of Chemistry, University of Warsaw and Michal Jamroz

the recent years, several new algorithms have been developed to take advantage of sparse NMR data, such as RDCs and chemical shifts, namely: Side-Chain-Only (SICHO)—a coarse grained structure prediction algorithm,[8] Rosetta-NMR[4] and other molecular fragment replacement approaches,[5–7] several interesting fold-recognition approaches[9–11] and some pipelines that involved the conventional NMR software such as X-plor.[12–14] In the above-mentioned algorithms, residual dipolar couplings (RDC) data have been included either in its original form, i.e. values of dipolar couplings in Hertz, or in a converted form, i.e. ranges of internuclear vectors projection angles computed by e.g. DipoCoup program.[9,15] The latter approach is especially useful in simulated annealing protocols, because it does not require complicated orientating an alignment tensor with respect to a protein molecule during simulations. Chemical shifts, implemented in a modified version of a CABS coarse-grained algorithm,[16] by Rosetta-Chemical Shifts (CS),[17] a CS23D server,[18] CHESHIRE[19] and a homology search in SimShiftDB,[20] are used explicitly or after the transformation into dihedral angles Φ and Ψ by a TALOS program.[21]

Those of the above-mentioned computational methods that are based on molecular fragments assembly approaches (i.e. Rosetta-CS, CS23D, CHESHIRE) encounter a time-limiting step requiring parallel computations for comparing experimental data with each structure generated during simulations. Moreover, performance of a majority of those methods depends on the similarity of a target protein to known protein folds.[18] For a couple of reasons, the CABS-NMR tool is free of those limitations. During CABS-NMR simulations, experiment-based restraints are computed only for short protein fragments that have been modified in a simulation step. Restraints for the entire protein model are computed during a Monte Carlo (MC) simulation only twice, at the beginning and at the end of a run. Consequently, the total simulation time is quite short in comparison with conventional NMR software, i.e. less than 1 day for a small globular protein (2gb1).[16] Then, because CABS is a de novo method it does not necessarily require any information about structural templates derived from known protein folds. Here, we use the term de novo in the meaning of protein structure prediction without any tertiary information from specific structural templates. Nevertheless we use prediction of secondary structure and torsion angles that uses some local sequence similarity to known protein structures. However the bias from possible homologous proteins introduced in that way into the CABS force field is negligible in comparison with purely template-based structure prediction algorithms. It is the consequence of the fact that the information about the local protein structure is averaged over the thousands of unrelated proteins in the database.

Developing pipelines and algorithms for global protein fold determination from sparse NMR data are important also in the case of high-molecular mass proteins. Complexity of NMR spectra of such systems needs to be reduced by perdeuteration, which, on the other hand, causes serious loss of many NOE signals.[22] Structure determination with a low-density restraints set, including only methyl and backbone H^N NOEs, although possible with standard programs such as CNS,[23] is still of low accuracy and could be enhanced with sophisticated structure prediction algorithms such as CABS-NMR.

In this study, we present CABS-NMR—a new high-throughput tool for protein structure prediction supported by different kinds of sparse NMR data that complementarily provide structural information about a protein fold. CABS-NMR combines an efficient method for de novo protein structure prediction (CABS) together with energy terms associated with distance and orientation restraints based on chemical shifts, residual dipolar couplings, and methyl-methyl NOEs. The main aim of this de novo approach, which does not use tertiary information from templates and is based solely on protein sequence information, is to reduce time and cost of the experimental determination of protein structures by the simultaneous application of the MC search algorithm (CABS) and high-throughput NMR techniques. The computationally efficient CABS algorithm uses a coarse-grained representation based on the trace of C-alpha atoms. Coordinates of C-alpha atoms are restricted to lattice vertices, whereas coordinates of other atoms: C beta and united atoms representing side groups of amino acids (named here Side Group [SG]), are off-lattice. In the CABS-NMR modeling pipeline, NMR data have to be converted into the suitable form based on a reduced, not all-atom, protein representation. For example, chemical shifts are incorporated into CABS-NMR in the form of the C-alpha-based pseudoangles computed from Φ and Ψ dihedral angles using a procedure called TRANSFORM[24] and the protocol described elsewhere.[16] Furthermore, methyl-methyl NOEs are converted into distances between C-beta atoms that are defined explicitly in the CABS algorithm in contrast to methyl groups. The only kind of NMR data incorporated into CABS-NMR using a computationally expensive all-atom representation of a protein backbone is RDCs data. The CABS tool and the above-mentioned NMR data were combined together in an efficient modeling pipeline. This modeling pipeline was successfully tested in structure modeling of several globular proteins including new folds for which evolutionary information about the structure is weak or nonexisting. Such structurally new proteins lack reliable templates and require de novo modeling approach such as the one presented here.

The algorithm for using RDCs data explicitly is completely different from the one that uses RDCs only implicitly, which was described in Ref. 8 by one of the authors of this manuscript.

Results

Before testing the CABS-NMR pipeline, it was necessary to see what is the final loss (if any) of the structural information resulted from the transformation of experimental data fitting it to the CABS-NMR algorithm. In Table 1, we compared the accuracy of the original experimental data (NOEs) or data partly transformed by TALOS or DipoCoup (i.e., Φ, Ψ, φ angles) with the accuracy of the final restraints generated by the CABS-NMR toolkit and used in the modeling (i.e. θ, γ angles, C-beta interatomic distances). In the case of chemical shifts and residual dipolar couplings, fitting to the CABS model involved minor (CS) or none (RDCs) drop of accuracy in comparison with the TALOS or DipoCoup transformations. This has already been inferred in our previous study,[24] in which details of the Φ/Ψ to θ/γ transformation were also described. The only noticeable loss of the structural information was observed in the case of NOEs (drop of accuracy: 3% in the 1p89 case—from 98.1% to 95%, to

Table 1. The Impact of the Restraints Accuracy on the Protein Structure Modeling Results.

PDB id	BMRB id	Nres	α or β (%)	Experimental data and converted data	Data accuracy[a]	Modeling results—cRMSD[b] model ranked 1 (best model) (Å)	
						Restrained folding	Nonrestrained folding
2p81	7386	44	59 (α)	511 CS converted to:			8.5 (8.5)
				Φ, Ψ	79.4, 73.5		
				θ, γ	82.3, 75.9	5.5 (3.9)	
				43 $D_{NH\text{-}RDC}$ converted to:			
				500 φ	73.8	8.1 (7.7)	
				All exp. data (CS+$D_{NH\text{-}RDC}$)		4.2 (4.2)	
1sf0	6187	68	19 ($\alpha+\beta$)	277 CS converted to:			9.7 (8.3)
				Φ, Ψ	50.9, 49.1		
				θ, γ	62.3, 48.8	9.5 (7.1)	
				68 $D_{NH\text{-}RDC}$ converted to:			
				494 φ	82.6	7.5 (6.6)	
				All exp. data (CS+$D_{NH\text{-}RDC}$)		6.5 (6.5)	
2f40	7073	74	20 ($\alpha+\beta$)	250 CS converted to:			9.7 (7.2)
				Φ, Ψ	54.2, 44.1		
				θ, γ	57.6, 46.9	12.7 (5.0)	
				113 $D_{NH\text{-}RDC}$ converted to:			
				500 φ	80.2	10.4 (7.5)	
				All exp. data (CS+$D_{NH\text{-}RDC}$)		6.2 (5.5)	
2js1_A	15350	74	61 (α)	654 CS converted to:			7.9 (6.4)
				Φ, Ψ	81.3, 81.3		
				θ, γ	85.7, 80.0	5.3 (4.8)	
				52 $D_{NH\text{-}RDC}$ converted to:			
				472 φ	47.0	6.4 (5.5)	
				All exp. data (CS+$D_{NH\text{-}RDC}$)		4.6 (4.5)	
2fe9	6922	86	56 (α)	984 CS converted to:			11.0 (9.6)
				Φ, Ψ	88.0, 80.0		
				θ, γ	93.0, 79.4	9.2 (6.3)	
				60 $D_{NH\text{-}RDC}$ converted to:			
				500 φ	67.4	7.3 (6.5)	
				All exp. data (CS+$D_{NH\text{-}RDC}$)		9.5 (5.9)	
2ea9	15088	103	36 (α/β)	900 CS converted to:			14.4 (10.9)
				Φ, Ψ	72.9, 80.2		
				θ, γ	76.9, 58.6	5.3 (4.5)	
				80 $D_{NH\text{-}RDC}$ converted to:			
				496 φ	86.1	14.3 (8.7)	
				All exp. data (CS+$D_{NH\text{-}RDC}$)		9.0 (7.1)	
1y8b domain 3	5471	139	52 (α/β)	222 Φ, Ψ converted to:	81.8, 76.4		15.0 (13.2)
				θ, γ	86.1, 73.7	16.8 (7.9)	
				74 CH_3-CH_3 NOE converted to:	100		
				Cβ-Cβ	74.3	7.6 (4.9)	
				All exp. data (Φ, Ψ+ CH_3-CH_3 NOE)		6.4 (5.1)	
1hbg	4038	147	66 (α)	1528 CS converted to:			14.7 (10.9)
				Φ, Ψ	94.6, 92.4		
				θ, γ	94.6, 89.2	9.9 (4.7)	
1zwm domain 1	410051[c]	87	48 (β)	158 Φ, Ψ converted to:	69.6, 67.1		11.9 (8.7)
				θ, γ	85.5, 79.2	12.9 (9.1)	
				26 CH_3-CH_3 NOE converted to:	92.3		
				Cβ-Cβ	76.9	7.7 (7.7)	
				All exp. data (Φ, Ψ+ CH_3-CH_3 NOE)		7.9 (6.5)	
1p89	4848	216	53 ($\alpha+\beta$)	2330 CS converted to:			17.2 (17.2)
				Φ, Ψ	70.8, 66.7		
				θ, γ	70.4, 58.5	18.9 (18.9)	
				161 CH_3-CH_3 NOE converted to:	98.1		
				Cβ-Cβ	95.0	3.4 (3.4)	
				All exp. data (CS+ CH_3-CH_3 NOE)		3.2 (3.2)	

[a]The data accuracy is defined as Acc = N_{true}/N_{all}, where $N_{true} = N_{all} - N_{false}$; N_{false}—number of angular or distance intervals that do not include the real value extracted from the PDB structure. N_{all}—number of all distance or angular intervals computed by DipoCoup, TALOS, the CABS-NMR toolkit or, provided in BMRB file. Acc is the measure of the data transformation accuracy of different protocols (TALOS—for Φ, Ψ; DipoCoup for φ; CABS-NMR for θ, γ, and Cβ-Cβ distances) and the original experimental data accuracy in the case of methyl-methyl NOEs. In the case of 1y8b and 1zwm, no TALOS computations were performed, and Φ and Ψ angles were obtained directly from the NMR Restraints Grid (http://restraintsgrid.bmrb.wisc.edu/).
[b]cRMSD—root mean square deviation of C-alpha atoms coordinates of protein models with respect to native structures.
[c]mrblock id—id of the data file from the NMR Restraints Grid repository of converted NMR restraints that were parsed with entries in the PDB database.

more than 25% in the 1y8b case—from 100% to 74,3%). Fortunately, that drop of accuracy of the NOEs-based restraints is still small due to high information content of these data.

The second important observation involved the effect of each kind of the experimental data on the modeling results. In Table 1, we compared results of folding simulations without any restraints with the simulations supported by each type of experimental restraints separately and by the entire set. Provided the protein structure had achieved the native-like fold in the simulations, in most tested cases we observed better results of restrained than of nonrestrained simulations, regardless of a type of experimental restraints used. This condition was not fulfilled in the case of the high molecular mass proteins (1zwm, 1y8b, 1p89). Such proteins could not be folded without any restraints or with the medium-accuracy (see below) local, CS-based restraints. In the 2f40 case, restrained MC folding provided the best model with the lower cRMSD then in the nonrestrained folding. But when comparing the cRMSD of the models ranked as 1 we observed that nonrestrained folding provided the better protein model. Only with the whole, diverse set of NMR restraints (RDCs and CS) we obtained both, the best and the model ranked as 1, of the lowest cRMSD than in the case of nonrestrained folding. The 2f40 case is an example of a difficult novel fold with high content of coils that requires the high level of diversity of the experimental data to cover uncertainties of different prediction methods.

The most efficient seemed to be distance restraints based on the NOEs data. If we take into consideration the specific construction of the CABS force field, which is already based on distance-dependent potential functions as the most efficient computationally,[25] that observation is not unexpected. The mean improvement rate for NOEs-restrained simulations over the best model from nonrestrained simulations is 7.7Å and over the model ranked as 1:8.5Å. In the case of chemical shifts, to compare the mean improvement rates are: 2.87Å (the best model) and 1.4Å (the model ranked as 1). The improvement was computed as the mean difference of cRMSD of compared protein models.

As for the local orientation restraints based on the chemical shifts, they are useful in our modeling pipeline as long as their accuracy does not drop below 80% (for small proteins up to 100 amino acids). In the case of proteins of higher molecular mass with more complex topology, the minimum accuracy should be at least 90% (compare the results for proteins 1hbg and 1y8b in Table 1). Noteworthy, those accuracy cutoffs are only approximate because the modeling results depend mostly on the complexity of the target protein topology, which cannot be fully defined by only local, chemical shifts based restraints. Nonetheless, the greatest advantage of chemical shifts in contrast to NOEs, which is the simplicity of the spectra recording and assignment process,[19] encourages to incorporate them in the routine NMR global fold determination. Another kind of orientation restraints, which are based on φ angles (obtained from RDCs), improved the structure prediction to much lesser extent in comparison with chemical shifts (see Table 1). It is mainly due to the limited accuracy of φ angles, ranging from 47% to 86% (according to a typical accuracy definition, see Table 1). What is more, the average 20% of false φ angles (computed as: N_{false}/N_{all}, where N_{false}—number of angular intervals that do not include the real value extracted from the PDB structure; N_{all}—number of all intervals computed by either DipoCoup or TALOS), which define the global arrangement of protein fragments, have more effect on the modeling results than the same 20% of inaccurate θ and γ pseudoangles that define only local geometry (e.g. the 2p81 case, see Table 1).

In Figure 1, we illustrated the effects of the restraints type on the modeling results, i.e. the protein model quality. We presented there a graphical interpretation of the modeling results instead of numerical similarity scores such as cRMSD shown in Table 1. Total structures similarity score such as RMSD, GDT-TS, or TM-score, though widely used in fast structure-to-structure comparison, does not seem to be sufficient in the detailed structural analysis of protein models.[26] It is difficult to assess from a numerical score if whole model is totally wrong or just a small protein fragment is mispredicted (e.g., a loop or a terminal fragment). Instead, graphical evaluation of structure predictions involves a more detailed analysis of fragments similarity.[26] The useful example of such graphical illustration is a Hubbard's plot, which is commonly used for evaluating CASP results.[26] In our work, Hubbard's plots helped us to assess if the model improvement due to experimental restraints involved only a fragment or a whole protein structure. We prepared Hubbard's plots for two proteins (see Fig. 1). Expectedly, we observed the global improvement of the whole protein model when all available experimental restraints were used in the modeling (green lines in Fig. 1). It confirmed our hypothesis that sparse experimental data from NMR could be highly complementary (see the Introduction section). In case of the modeling supported by only one type of experimental restraints, we observed irregular sudden increasing of lines on the Hubbard plots indicating that some fragments are significantly less accurate than the others. Those major deviations from the native structures were observed for the fragments with no experimental data available (see the right part of Fig. 1). Those deviations were more distinct when the fragments were not incorporated into any secondary structure unit (an alpha helix or a beta sheet) and were forming irregular loops. It can be explained by the fact that the CABS algorithm is able to rebuilt regular secondary structure units without serious obstacles even without experimental restraints, whereas loops regions are still a challenge (likewise for any other, available currently, de novo algorithm based on knowledge-based or physical potentials).

The test case: S100A1 protein

The developed toolkit was used in the structure modeling of a calcium-binding S100A1 homodimer protein (available now in PDB as 2jpt). The protein was in its apo form, modified chemically by the disulfide bond formation between Cys 85 and BME (β-mercaptoethanol). That modification induced structural changes that resulted in a dramatic increase in the protein affinity for calcium.[27] Namely, the apo-S100A1 structure became more similar to that of holo-S100A1 and thus much better adjusted for the calcium binding. Structural changes in a single subunit of S100A1 involve elongation of a helix IV (red in Fig. 2) and increasing an angle between helices I (blue in

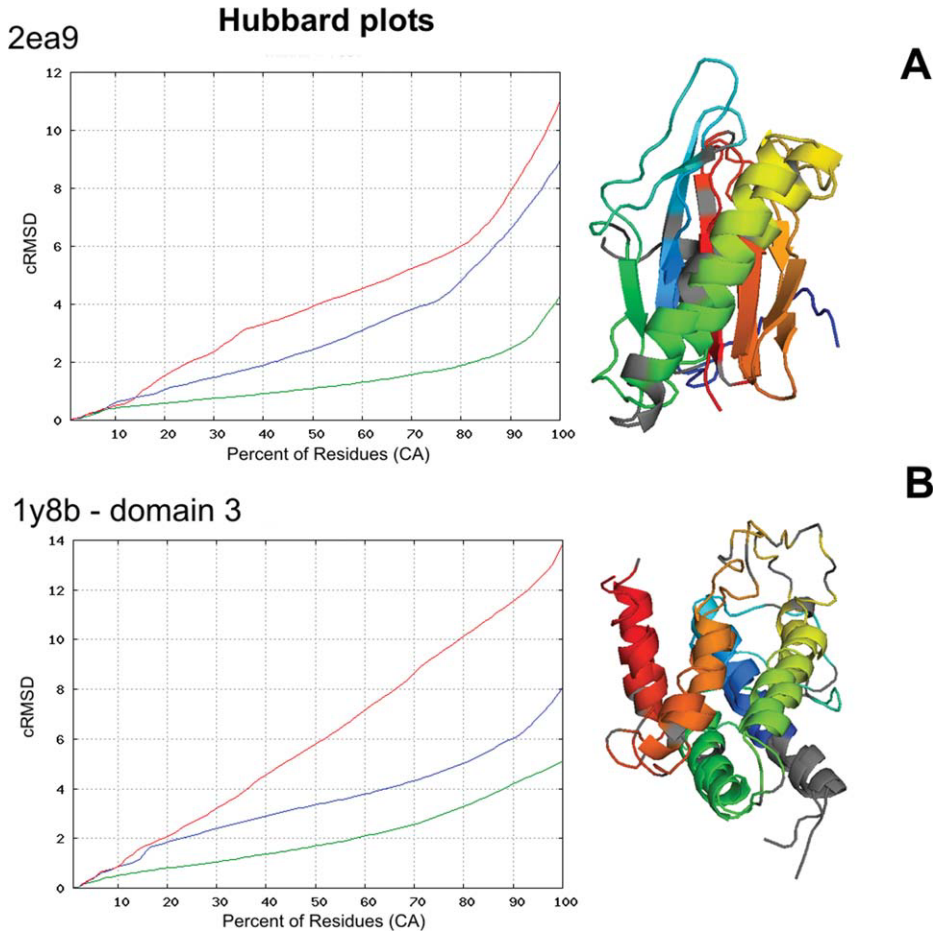

Figure 1. Graphical evaluation of CABS-NMR modeling results. Left, Hubbard's plots representing the nonrestrained (red lines) and restrained modeling results: blue lines (modeling supported by only one type of NMR data: RDCs [A] or CS [B]) and green lines (modeling supported by all available types of experimental data: RDCs, CS, and NOEs). Right, final protein models of 2ea9 (A) and the third domain of 1y8b (B) superimposed on the experimental structures from the PDB. Grey regions indicate these protein fragments for which at least one type of the experimental data (CS, RDC, or NOE) is missing.

Fig. 2) and IV. In such case, the comparative modeling studies were not appropriate because any available templates would bias the protein model to a typical apo-type structure and the structural change toward the holo form after the chemical modification would not be noticed. For example, bioinfo.pl, a protein structure prediction metaserver,[28] produced a 1k2h_A template as the closest homologous structure, which is an apo-S100A1 structure with shorter helix IV and the interhelical angle I-IV typical for the apo-form. The SimShiftDB[20] homology search, which uses chemical shifts information, also suggested that structure as the best template (see Fig. 2). In contrast, our de novo method, which does not depend on homology search, produced the modified structure of S100A1 structure with novel features. Here, we performed modeling of a S100A1 monomer because oligomeric structure prediction, although possible with the other versions of the CABS algorithm,[29] is not supported in the current version of CABS-NMR.

The chemical shifts data were provided by authors of the modified apo-S100A1 NMR structure[27] (Biological Magnetic Resonance Bank accession number: 4982). The modeling procedure performed for the subunit 1 of S100A1 is described in Methods section. We obtained a near-native structure (see Fig. 2), with an elongated helix IV (to residue 90 ± 1) but a rather weakly adjusted angle between helices I and IV.

Next, the obtained S100A1 protein models were used to predict all possible NOE-type contacts between H^N and H^α hydrogen atoms. In that way, we wanted to confirm the reliability of our modeling procedure and its utility as a part of NMR structure determination, i.e. in the assignment of the NOESY spectra. In Table 2, we compared the PDB-derived NOE contacts with the H^N-H^N, H^α-H^α, and H^N-H^α contacts derived either from a single protein model or a set of six, low energy models. In this work, we named two hydrogen atoms as contacting when they were closer than 5 Å (a NOE-type cutoff).

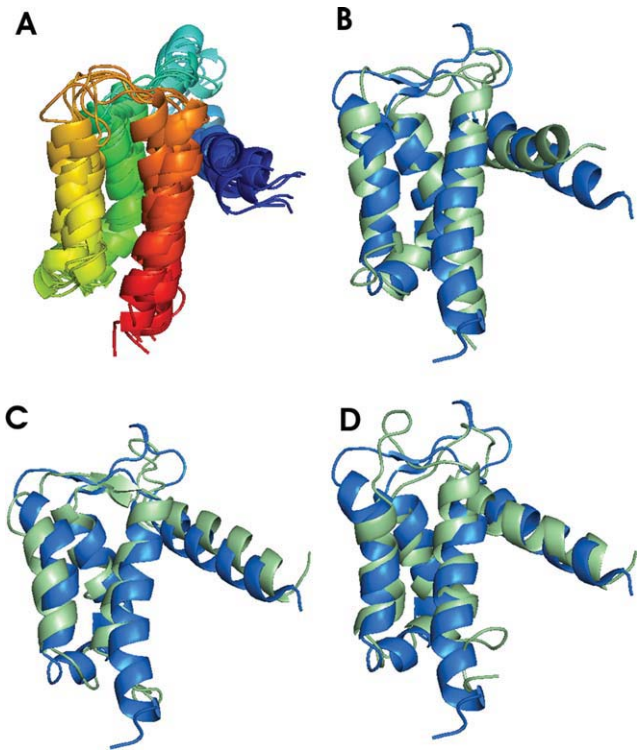

Figure 2. The S100A1 protein model generated from chemical shifts only. (A) Superposition of six low-energy structures obtained in restrained MC folding simulations with the CABS-NMR algorithm. The cRMSD of the best model with respect to the first model in the PDB entry (2jpt) is 4.3Å. The average pairwise cRMSD for the six conformers is 4.6Å. (B) Superposition of the best CABS-NMR protein model and the native S100A1 structure (PDB: 2jpt_A). (C) Superposition of 2jpt_A and the best model generated by SimShiftDB on the 1k2h_A template structure (cRMSD = 5.71Å, E-value = 8.80e-85) (D) Superposition of 2jpt_A and the SimShiftDB model generated as the second (template: 2k2f_B, cRMSD = 5.15Å, E-value = 1.11e-82). In (B), (C), and (D) the native structure of 2jpt is colored with blue. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Typically, the final result of the CABS-based structure modeling is a set of low-energy protein structures (medoids of the most populated clusters) instead of a single protein model. Usually, we perform the clustering analysis and use Model Quality Assessment Programs to choose the best predicted structure (with the lowest cRMSD vs. a native structure). However, in de novo structure prediction tasks, we often take an insight into the whole set of medoid structures of several, the most populated clusters. In that way, we are able to detect the most probable core of a protein by discarding the diverged loops (see Fig. 2). Therefore, we supposed that the NOEs prediction would be better if performed on a whole set of structures and not on the most probable, yet single, protein model. Our presumption was right especially in the case of H^{α}–H^{α} contacts (see Table 2). False NOEs, placed at the end of Table 2, which were predicted from a single protein model were almost completely discarded by ranking based on the occurrence frequency in a whole set of protein models. The set of protein models where chosen by calculating an interhelical angle for helices I and IV in each model by a Interhlx program.[30] We discarded all conformations with the angle sign opposite to a typical value observed for S100 proteins,[27] in which the EF hand motif did not appear. Unfortunately, such conformations, although in minority (4 of 10 structures), were inevitable because we folded a monomer not a dimer structure in the implicit-solvent simulation. Without the presence of the second monomer structure, in such a folding simulations, the protein may fold into a more compact bundle of four vertical helices instead of three vertical helices and one set horizontally (see Fig. 2A). The final obtained six low-energy structures are superposed on each other in Figure 2A.

Discussion

The main purpose of the proposed modeling scheme is to assist the experiment-based protein structure determination. The application of the robust coarse-grained structure prediction algorithm (CABS) accelerates the structure determination process. On the other hand, usage of sparse and relatively easy to obtain experimental data in the modeling procedure protects it from severe

Table 2. The S100A1 Protein—Long-Range NOEs Prediction.

Residue No.[a]		PDB-derived[b]	A single protein model[c]	Average of 6 protein models	
i	j (>i+3)	distance (Å)	Distance (Å)	Distance (Å)	Rank[d]
			H^N-H^N NOEs		
29	**33**	**4.17**	**3.64**	**4.29**	**1**
			Hα-H^N NOEs		
30	**68**	**4.57**	**3.86**	**4.30**	**1**
28	**70**	**4.93**	**3.61**	**3.73**	**2**
			Hα-Hα NOEs		
38	**42**	**3.13**	**4.34**	**4.63**	**5**
87	**91**	**4.20**	**6.24**	**6.07**	**-**
16	**20**	**4.44**	**4.84**	**4.25**	**4**
29	**68**	**4.47**	**2.95**	**2.98**	**1**
27	**70**	**4.58**	**3.16**	**2.91**	**2**
59	**63**	**4.81**	**7.33**	**6.28**	**-**
30	**61**	**4.86**	**7.27**	**7.51**	**-**
45	49	6.06	3.62	3.92	3
15	20	8.40	4.95	4.95	35
27	71	5.61	4.42	4.42	31
29	67	8.29	4.81	7.08	-
30	67	8.55	4.79	4.79	29

The prediction were based on the S100A1 protein models obtained in the CABS-NMR procedure from chemical shifts data. The real NOE-type contacts, which were observed in the PDB structure, are listed in bold. The most accurate predictions were performed for H^N—H^N Hα—H^N NOE-type contacts using a set of six protein models instead of one.

[a]Residue sequence number.

[b]Distance between hydrogen atoms in the first model of a PDB structure.

[c]The model ranked as the first, i.e. the most probable model out of six proposed. Typically, it is the model from the most populated cluster of protein structures obtained from the folding simulation trajectory.

[d]Rank based on the frequency of occurrence of the given NOE-type contact in the set of six protein models. Dashes indicate that a contact was not detected at all in the set of six protein models.

loss of resolution of the generated protein models. The benefits from using the coarse-grained algorithm, which is the reducing of computation time, outweigh a minor loss of structural information resulted from the data transformation. The resolution of the obtained protein structures is sufficient for using them in the afterward NMR data analysis, for example in the verification of the NOEs assignment. The all atom reconstruction module based on the BBQ method provide the useful way to incorporate to CABS-NMR detailed all-atom force field based on either knowledge-based or physical potential functions. It also enables to include other kinds of experimental data, especially from ^{1}H NMR experiments, which will be investigated in future. In contrast to other structure prediction algorithms, CABS-NMR is not biased toward template-based modeling tasks and therefore is the most useful in new folds prediction, especially as a part of NMR-based structure determination in high-throughput pipelines followed by molecular dynamics refinement. The CABS-NMR Toolkit package provides also important tools for the NMR data formats conversion. It enables to perform useful transformations of experimental data into the angular or distance restraints that are easy to incorporate in MC or MD simulations.

Methods

CABS-NMR is an extension of a previously developed algorithm[16] dealing with sparse chemical shifts data. The chemical shifts-based module in the previous and in the current algorithm is based on a simple transformation of NMR data to angular relations which are easy to apply in MC simulations. The first stage of the transformation is carried out by the TALOS program from the NMRPipe package.[31] Although TALOS uses sequence homology for prediction of dihedral angles, the actual protein folds similarity derived from such prediction is negligible. It is due to small window size (only three residues) and small number of proteins in the database (only 186) which is far not enough to find a reliable structural homolog to any of the proteins tested in this work.

The resulted Φ and Ψ dihedral angles from TALOS predictions are then translated by the TRANSFORM program[24] to pseudoangles θ and γ. Those angles θ and γ are the reduced coordinates, defined in the C-α-based coordinates system. The described data transformations from chemical shifts to angular restraints are performed without serious loss of structural information.[21,24] Particularly, in the test set of 10 proteins used in this work, such θ and γ restraints described in details 70%–80% of the main chain degrees of freedom. Chemical shifts were also used in this work for approximate secondary structure assignment performed by a PsiCSI server,[32] which combines a well-known PSIPRED method[33] with a Chemical Shift Index (CSI) procedure.[34] That rough prediction of secondary structure before folding simulations is inevitable in our modeling pipeline. However, it is an external procedure (see Fig. 3) and thus could be performed by any other tool, not necessary the PsiCSI server.

The procedure described above, dealing with chemical shifts data, is the same as in Ref. 16. A novel module in CABS-NMR is a procedure for a backbone atoms reconstruction from a C-alpha trace that is needed for the incorporation of ^{15}N-^{1}H RDCs based restraints. The original CABS algorithm[25] uses a reduced representation of a protein (see Introduction section) and therefore backbone N and H atoms are not defined explicitly. Substitution of an N—H bond vector by some other interatomic vector, defined in the reduced representation, although possible,[8] could propagate errors and uncertainties of RDCs, which are already of low resolution. For those reasons, we add a separate backbone reconstruction procedure to the CABS-NMR algorithm despite the unavoidable elongation of simulation time (4 times according to nonrestrained simulations and 2 times with respect to CS or NOEs-restrained simulations). Our backbone reconstruction procedure is based on a robust algorithm called BBQ.[35] The BBQ program reconstructs a protein backbone using a database of average atoms positions in a local coordinates system. The local coordinates system is defined by three interatomic vectors of a quadrilateral (a consecutive fragment of four C-alpha atoms). Originally, the statistics for backbone atoms positions in the BBQ program was derived from the nonredundant database of known protein structures. That statistics had to be customized to fit protein-like, but not observed in PDB, conformations of a C-alpha trace obtained in CABS simulations. That customization was done following suggestions of Gront et al.,[35] by replacing a protein-like quadrilateral fragment by its closest neighbor from the original BBQ program database.

The backbone reconstruction procedure included in the CABS-NMR algorithm was combined with the simple geometrical rebuilding of amide H atoms from coordinates of three consecutive atoms: carbonyl C, amide N, and C-alpha. Giving the exact positions of N and H atoms in the amide group, one could finally define an angle between N—H bond vectors. A set of such angles is a typical output of the DipoCoup program.[9,15] DipoCoup converts residual dipolar couplings into allowed ranges of projection angles between intramolecular N—H bonds vectors. Such translation of dipolar couplings was originally developed to overcome convergence problems in RDCs-restrained simulated annealing protocols, possibly related to determination of the alignment tensor.[9] We observed that CABS could benefit from the DipoCoup translation of RDCs data in the same way as standard simulated annealing protocols despite being based on a more efficient search algorithm, i.e. a parallel tempering MC method.[25]

Intramolecular projection angles obtained from DipoCoup, named here φ, are incorporated into CABS-NMR in the form of ambiguous restraints with two allowed ranges for each angle. The potential based on such angular restraints, graphically presented in Figure 3, is described by the following formula:

$$E_{ij} = -k\ \min(|\varphi_{ij} - \varphi^1_{ij}|, |\varphi_{ij} - \varphi^2_{ij}|) \quad \text{for} \quad \begin{cases} \varphi_{ij} \in (\varphi^1_{ij} - \Delta\varphi^1_{ij}, \varphi^1_{ij} + \Delta\varphi_{ij}) \\ \varphi_{ij} \in (\varphi^2_{ij} - \Delta\varphi_{ij}, \varphi^2_{ij} + \Delta\varphi_{ij}) \end{cases}$$

$$E_{ij} = 0 \quad \text{for} \quad \begin{cases} \varphi_{ij} < \varphi^1_{ij} - \Delta\varphi_{ij} \\ \varphi_{ij} \in (\varphi^1_{ij} - \Delta\varphi_{ij}, \varphi^2_{ij} + \Delta\varphi_{ij}) \\ \varphi_{ij} < \varphi^2_{ij} + \Delta\varphi_{ij} \end{cases} \qquad (1)$$

where: i and j are intramolecular N—H bond vectors; k is a scaling factor; φ^1_{ij} is the mean of the projection angle range in

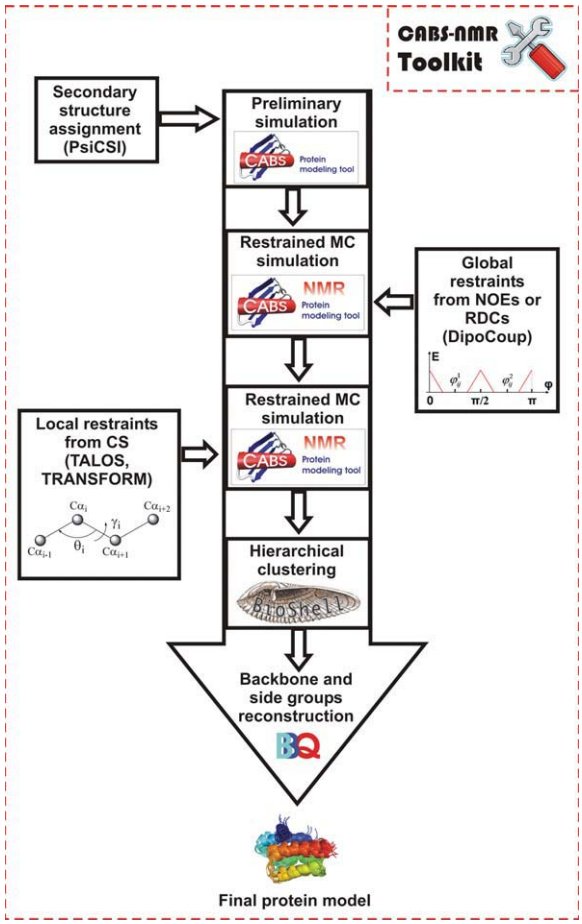

Figure 3. The CABS-NMR modeling pipeline. The core of the pipeline is the protein folding simulation divided into three stages varying experimental restraints used. The presented pre- and postsimulation data processing requires additional tools (CABS-NMR Toolkit, Bioshell and BBQ), which can be freely downloaded from: http://biocomp.chem.uw.edu.pl/services.php. The remaining external modules: TALOS, DipoCoup, and PsiCSI are also available from authors free of charge for academic purposes.

the 0–π/2 interval, φ_{ij}^2 is the mean of the projection angle range in the π/2–π interval; and $\Delta\varphi$, a half of the length of the projection angle range (the same for both ranges in intervals: 0–π/2 and π/2–π). The above potential favors protein conformations with most of the RDCs-based restraints fulfilled. The linear function was used in this potential as the most efficient in the MC sampling of the protein conformational space, following the research on the NMR-restrained structure prediction described previously.[16] Namely, comparing the typical harmonic-type potentials, the linear one less severely restricts the conformational space of a protein chain in these regions that are distant from the native global minimum.

The methyl-methyl NOE data were incorporated into CABS-NMR in the form of distance restraints. We did not use the distance between protons of methyl groups, because the CABS model lacks an explicit definition of methyl groups. Instead, we used the distance between C-beta atoms, which are defined explicitly in the CABS model (they are rebuilt from C-alpha coordinates in-flight). Thus, no additional computations had to be done during MC simulations. We chose the distance between C-beta atoms because it resembles the methyl-methyl distance better than the distance between other, defined explicitly atoms in the CABS model (C-alpha atoms or SG united atoms). That observation was done after generating histograms of C beta, C alpha, and SG distances only between these residues for which methyl-methyl NOEs were observed (see Fig. 4). Histograms were derived from the ASTRAL40 database of such protein structures that share less than 40% sequence identity with each other. The lowest standard deviation from the mean value was observed for the C-beta atoms distance and therefore we substituted NOEs by that interatomic distance. In the first approach, the mean value (6Å) from the C-beta distances histogram was chosen as a cutoff distance (equivalent to the 5Å interproton distances observed in NOESY spectra). However, in some preliminary tests, it turned out that the limited resolution of the CABS model required a larger value of the upper cutoff distance, i.e. 7Å. The short range cutoff distance between C-beta atoms was already defined in the CABS model as an excluded volume of C-beta atoms. The final restraints potential reproducing methyl-methyl NOEs is described as following:

$$E_{ij} = k\sqrt{d_{ij} - d^{\max}} \quad \text{for} \quad d_{ij} > d^{\max}$$
$$E_{ij} = 0 \quad \text{for} \quad d_{ij} > d^{\max} \qquad (2)$$

Here: i and j are a pair of amino acid residues for which a methyl-methyl NOE is detected, k is a scaling factor, d_{ij} is an

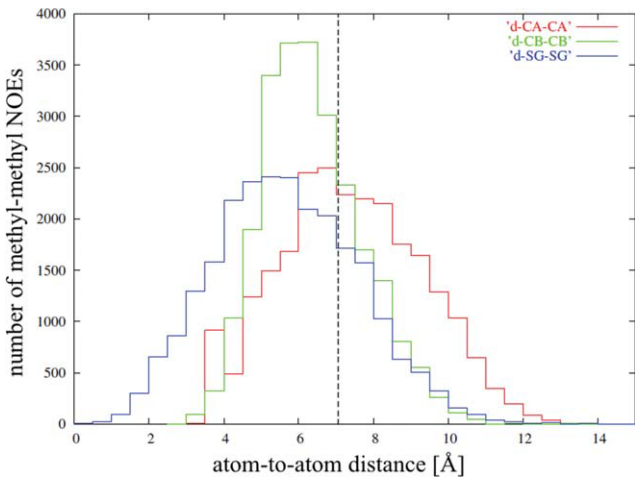

Figure 4. Distributions of different interatomic distances provided in CABS. Here, we compared histograms based on distributions of distances between C-alpha atoms (d-CA-CA), C-beta atoms (d-CB-CB) and united atoms representing side groups (d-SG-SG). All distances were computed only for these pairs of residues between which a methyl-methyl NOE were detected. The lowest standard deviation is observed for the C-beta distances histogram. The dotted line depicts the final cutoff for C-beta distances that reproduce methyl-methyl NOEs in the restraints potential used in CABS-NMR.

observed distance between C-beta atoms, d^{max} is a top cutoff for C-beta atoms distance (7Å). The above potential penalizes protein conformations that do not fulfill the distance restraints. The potential was adapted from the previous research, which was the protein folding with the simulated, not experimental, contact restraints between other types of atoms (Cα and pseudoatoms representing side groups of amino acids).[36] Namely, the potential function remained the same but the parameters were differently optimized.

In preliminary tests, we observed that all experimental restraints significantly modify the conformational energy landscape of a protein. Generally, the energy landscape is more rugged than in the case of non-restrained folding simulations. As a result, the conformational search during the folding process is less efficient and a protein conformation may be trapped in local minima more frequently. Noteworthy, local restraints based on chemical shifts are much more restrictive for protein conformational space than long range, global restraints obtained from RDCs or NOEs. It is the consequence of the fact, that any global change of a protein conformation in the CABS algorithm is performed by series of modifications of small protein fragments (from one to few residues long) and such local conformation changes may be hindered by imposing local restraints (e.g. based on chemical shifts). To overcome these problems, we applied different types of experimental restraints sequentially, introducing all of them in turn from the beginning of a simulation run. A similar approach was described elsewhere.[37] Global restraints and the biases resulting from the secondary structure assignment are incorporated as the first, and then followed by more restrictive, local restraints. Figure 3 shows the detailed description of the folding simulation protocol and the entire modeling pipeline. Our modeling pipeline uses not only well-known tools such as TALOS, PsiCSI or Bioshell (a package for structural analysis) but also a new package called CABS-NMR Toolkit, available from our website: http://biocomp.chem.uw.edu.pl/services.php. It includes CABS-NMR and TRANSFORM programs and simple shell scripts for converting experimental data formats (e.g. from NMRSTAR format to TALOS and DipoCoup programs formats). It also includes scripts for the protein models selection on the basis of fitting experimental NMR data or sparse evolutionary information (i.e. predicted contacts[38]).

References

1. Snyder, D. A.; Chen, Y.; Denissova, N. G.; Acton, T.; Aramini, J. M.; Ciano, M.; Karlin, R.; Liu, J.; Manor, P.; Rajan, P. A.; Rossi, P.; Swapna, G. V.; Xiao, R.; Rost, B.; Hunt, J.; Montelione, G. T. J Am Chem Soc 2005, 127, 16505.
2. Tugarinov, V.; Kanelis, V.; Kay, L. E. Nat Protoc 2006, 1, 749.
3. Opella, S. J.; Nevzorov, A.; Mesleb, M. F.; Marassi, F. M. Biochem Cell Biol 2002, 80, 597.
4. Rohl, C. A.; Baker, D. J Am Chem Soc 2002, 124, 2723.
5. Valafar, H.; Mayer, K. L.; Bougault, C. M.; LeBlond, P. D.; Jenney, F. E., Jr.; Brereton, P. S.; Adams, M. W.; Prestegard, J. H. J Struct Funct Genomics 2004, 5, 241.
6. Delaglio, F.; Kontaxis, G.; Bax, A. J Am Chem Soc 2000, 122, 2142.
7. Andrec, M.; Du, P.; Levy, R. M. J Biomol NMR 2001, 21, 335.
8. Haliloglu, T.; Kolinski, A.; Skolnick, J. Biopolymers 2003, 70, 548.
9. Meiler, J.; Blomberg, N.; Nilges, M.; Griesinger, C. J Biomol NMR 2000, 16, 245.
10. Qu, Y.; Guo, J. T.; Olman, V.; Xu, Y. Nucleic Acids Res 2004, 32, 551.
11. Annila, A.; Aitio, H.; Thulin, E.; Drakenberg, T. J Biomol NMR 1999, 14, 223.
12. Bax, A. Protein Sci 2003, 12, 1.
13. Tugarinov, V.; Kay, L. E. J Mol Biol 2003, 327, 1121.
14. Choy, W. Y.; Tollinger, M.; Mueller, G. A.; Kay, L. E. J Biomol NMR 2001, 21, 31.
15. Meiler, J.; Peti, W.; Griesinger, C. J Biomol NMR 2000, 17, 283.
16. Latek, D.; Ekonomiuk, D.; Kolinski, A. J Comput Chem 2007, 28, 1668.
17. Shen, Y.; Lange, O.; Delaglio, F.; Rossi, P.; Aramini, J. M.; Liu, G.; Eletsky, A.; Wu, Y.; Singarapu, K. K.; Lemak, A.; Ignatchenko, A.; Arrowsmith, C. H.; Szyperski, T.; Montelione, G. T.; Baker, D.; Bax, A. Proc Natl Acad Sci U S A 2008, 105, 4685.
18. Wishart, D. S.; Arndt, D.; Berjanskii, M.; Tang, P.; Zhou, J.; Lin, G. Nucleic Acids Res 2008, 36(Web Server issue), W496.
19. Cavalli, A.; Salvatella, X.; Dobson, C. M.; Vendruscolo, M. Proc Natl Acad Sci U S A 2007, 104, 9615.
20. Ginzinger, S. W.; Coles, M. J Biomol NMR 2009, 43, 179.
21. Cornilescu, G.; Delaglio, F.; Bax, A. J Biomol NMR 1999, 13, 289.
22. Grishaev, A.; Tugarinov, V.; Kay, L. E.; Trewhella, J.; Bax, A. J Biomol NMR 2008, 40, 95.
23. Tugarinov, V.; Choy, W. Y.; Orekhov, V. Y.; Kay, L. E. Proc Natl Acad Sci U S A 2005, 102, 622.
24. Plewczynska, D.; Kolinski, A. Macromol Theory Simul 2005, 14, 444.
25. Kolinski, A. Acta Biochim Pol 2004, 51, 349.
26. Hubbard, T. J. Proteins 1999, (Suppl 3), 15.
27. Zhukov, I.; Ejchart, A.; Bierzynski, A. Biochemistry 2008, 47, 640.
28. Ginalski, K.; Elofsson, A.; Fischer, D.; Rychlewski, L. Bioinformatics 2003, 19, 1015.
29. Kurcinski, M.; Kolinski, A. J Mol Model 2007, 13, 691.
30. Yap, K. L.; Ames, J. B.; Swindells, M. B.; Ikura, M. Methods Mol Biol 2002, 173, 317.
31. Delaglio, F.; Grzesiek, S.; Vuister, G. W.; Zhu, G.; Pfeifer, J.; Bax, A. J Biomol NMR 1995, 6, 277.
32. Hung, L. H.; Samudrala, R. Protein Sci 2003, 12, 288.
33. McGuffin, L. J.; Bryson, K.; Jones, D. T. Bioinformatics 2000, 16, 404.
34. Wishart, D. S.; Sykes, B. D. J Biomol NMR 1994, 4, 171.
35. Gront, D.; Kmiecik, S.; Kolinski, A. J Comput Chem 2007, 28, 1593.
36. Li, W.; Zhang, Y.; Kihara, D.; Huang, Y. J.; Zheng, D.; Montelione, G. T.; Kolinski, A.; Skolnick, J. Proteins 2003, 53, 290.
37. Wu, Z.; Delaglio, F.; Wyatt, K.; Wistow, G.; Bax, A. Protein Sci 2005, 14, 3101.
38. Latek, D.; Kolinski, A. BMC Struct Biol 2008, 8, 36.