Protein Structure Prediction: Combining De Novo Modeling with Sparse Experimental Data

DOROTA LATEK, DARIUSZ EKONOMIUK,* ANDRZEJ KOLINSKI
Faculty of Chemistry, Warsaw University, Pateura 1, 02-093 Warsaw, Poland

Received 28 July 2006; Accepted 2 November 2006
DOI 10.1002/jcc.20657
Published online 6 March 2007 in Wiley InterScience (www.interscience.wiley.com).

Abstract: Routine structure prediction of new folds is still a challenging task for computational biology. The challenge is not only in the proper determination of overall fold but also in building models of acceptable resolution, useful for modeling the drug interactions and protein–protein complexes. In this work we propose and test a comprehensive approach to protein structure modeling supported by sparse, and relatively easy to obtain, experimental data. We focus on chemical shift-based restraints from NMR, although other sparse restraints could be easily included. In particular, we demonstrate that combining the typical NMR software with artificial intelligence-based prediction of secondary structure enhances significantly the accuracy of the restraints for molecular modeling. The computational procedure is based on the reduced representation approach implemented in the CABS modeling software, which proved to be a versatile tool for protein structure prediction during the CASP (CASP stands for critical assessment of techniques for protein structure prediction) experiments (see http://predictioncenter/CASP6/org). The method is successfully tested on a small set of representative globular proteins of different size and topology, including the two CASP6 targets, for which the required NMR data already exist. The method is implemented in a semi-automated pipeline applicable to a large scale structural annotation of genomic data. Here, we limit the computations to relatively small set. This enabled, without a loss of generality, a detailed discussion of various factors determining accuracy of the proposed approach to the protein structure prediction.

 J Comput Chem 28: 1668–1676, 2007

Key words: chemical shifts; protein structure prediction; de novo protein folding; reduced models; Monte Carlo simulations

Introduction

Basically, there are two distinct situations in the computational determination of protein structure. The classical one is the comparative modeling, where the new structure is build on the scaffold of a protein structure due to homology or analogy, resulting from convergent evolution, is expected to be similar to the query structure. However, quite frequently (we do not know exact percentage of cases for more complex organisms) it is impossible to find a proper template for the comparative modeling. In such cases the structure prediction needs to be done in a de novo fashion. In principle, de novo protein folding does not require any information about the homologues, or structural analogs, of the query protein. The prediction is based only on the protein-like biases derived from the database of known protein structures and the appropriate search algorithm of the protein conformational space. Despite the success of the CABS algorithm in the last CASP6 experiment,[1] de novo structure determination still remains the challenging task, and in general is limited to relatively small proteins of not too complex topology. Moreover, the resulting structures are of a rather low resolution. There are exceptions from this statement, although they still seem to remain in a status of "proof of the principle" than a routine computational methodology.[2–4]

In this work we enhanced de novo prediction with the application of the sparse experimental data. Here, the term de novo means folding simulations without any specific structural templates, although the preceding prediction of secondary structure employs local sequence similarity to known protein structures. The bias from possible homologous proteins introduced this way into the CABS force field is negligible (due to the averaging with thousands of unrelated proteins in the database). The additional, experiment based (in this case NMR) information helps in the protein structure determination in two ways. First, it accelerates the prediction due to significant reduction of the con-

*Present address: Department of Biochemistry, University of Zurich, Winterthurerstrasse 190, CH-8057 Zurich, Switzerland

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

Contract/grant sponsor: Polish Committee for Scientific Research (KBN); contract/grant number: PBZKBN-088/P04/2003

formational space which is being explored. Second, it allows assessing the obtained protein models and choosing the one which is the most similar to the native structure.[5] In this work we focus mainly on the first issue, although it is also clearly shown that the experimental restraints, even of the fuzzy nature, increase significantly the resolution of the modeled structures.

In the last few years, a number of novel experimental techniques, which provide structural data in relatively short time and with relatively small effort, have been developed. Namely, two types of NMR measurements were extensively used in the protein structure determination: chemical shifts and residual dipolar couplings.[6] In this work we focused on the chemical shifts data alone, which carry the information about the local conformation of the main chain backbone.[6] Generally, chemical shifts were used in the protein structure determination only with some additional sparse experimental data which completed the lack of the global information about the fold (RDC or NOEs).[5,7] Here, we present the novel method for exploiting the sparse information from the chemical shifts alone in the framework of de novo structure determination using restrained folding simulations.

The chemical shifts data was processed with the TALOS program[8] and the PsiCSI server[9] to obtain the constraints appropriate for the Monte Carlo simulations with the CABS algorithm. As it is shown later, this approach (experiment based and evolutionary) are highly complementary and significantly increase the precision and the information content of the restraints for the molecular modeling.

The PsiCSI server for the fast secondary structure prediction combines the homology based predictions with the experimental approach. It uses the neural network which is trained on the two kinds of the input data: the secondary structure derived from the chemical shifts and the secondary structure based on the profiles analysis from the PSIPRED server.[10] The average accuracy of this combined approach is about 86%,[9] which is significantly better than in case of using the chemical shifts data or the PSIPRED alone. It has to be stressed out that a difference of few percentage points for the accuracy of secondary structure predictions in the zone of 80–90% is significant, frequently crucial for a proper prediction of the three-dimensional structure.

The TALOS program predicts the most probable values of φ and ψ angles for the set of the chemical shifts data. It uses both, the sequence and the chemical shifts similarity of triplets of residues of the query protein to the database of the solved structures and the corresponding chemical shifts. Typically, TALOS predicts very accurate torsion angles for about 40% of residues, but this value can be pushed up to about 70% after optimization by a human expert.[11]

The main advantage of the chemical shifts is the simplicity of the required NMR measurements. However, in some cases (see the Results section), the accuracy of the angular data derived from the chemical shifts is insufficient for the unambiguous determination of the structure.[7] What is perhaps more important, chemical shifts carry only the local (and incomplete) information about the backbone geometry, thus the structure determination lacks some global, long-range constraints.[6] In the context of these shortcomings of the chemical shifts data, we also attend briefly two NMR based methods which could enhance the chemical shifts based information. These can supplement the chemical shift data. Moreover, they provide some information on the nature of the global fold. This is not the main subject of the present work; however, we think that it could be useful to show how various additional data, when available, could be easily included into the proposed approach to the structure determination. The first kind of these additional experimental data is based on the NMR measurements of three bond coupling constants, which are widely used in the determination of different dihedral angles (for example $^3J_{NHH\alpha}$).[12,13] The second method is based on the long-range distances between Cα atoms and between side chains, which can be obtained in the experiments based on the nuclear overhauser effect (NOE). Both of these methods are commonly used in the protein structure determination but, on the other hand, they are much more difficult in processing and analyzing than the data obtained from chemical shifts measurements. It should be, however, pointed out that in this work we use far less J-couplings and NOEs (40% in the case of J-couplings and 5% in the case of NOEs) than it is needed in the NMR based structure determination (typically, a few NOEs per residue and at least two different J-couplings per residue are needed).[12,14] Therefore, such data could be collected with a smaller effort, sometimes comparable to the cost of the chemical shifts data.

Methods

De Novo Protein Structure Prediction Algorithm

The conformational space of query proteins is sampled using the recently published simplified lattice model CABS with the Replica Exchange Monte Carlo (REMC) method (also referred as the parallel tempering MC).[15] The model is based on three centers of interactions per residue: Cα atom, Cβ atom, and a united atom which represents a side group of an amino acid (see Fig. 1). The model force field includes several knowledge-based potentials, namely protein-like biases, potentials for the short and long range interactions, and a model of hydrogen bonds. The details of the force field implementation could be found in recent publications.[15]

Experimental Data

Chemical shifts are affected by the local conformation of the protein backbone.[16,17] On the other hand, the local conformation is coded by the type of the secondary structure and, more precisely, by the ϕ and ψ torsion angles. This is why we decided to use chemical shifts based constraints in the form of the secondary structure type (using the PsiCSI server) and the torsion angles (using the TALOS program). This is also more straightforward for the input files for the CABS modeling tool, although it appears that the translation into the secondary structure code could be beneficial for other approaches to the protein modeling. It seems to be evident that such translation eliminates a significant fraction of wrong predictions from the NMR data. Of course, sometimes the data are diffused, though the most contemporary tools for the protein structure prediction are less sensible to the inaccuracy of data than to their limited coverage of the modeled structure.

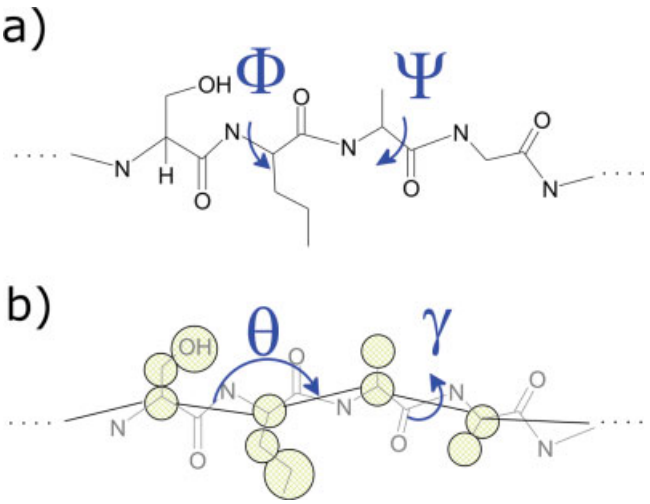

Figure 1. The full atom representation of a protein with marked constraints derived from chemical shifts data (a). The reduced representation of the protein backbone in the CABS model (b). NMR based constraints defined in a full atom representation (a) had to be converted into corresponding constraints in the CABS model, which are defined only with the use of the Cα trace (b). [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

Both types of the constraints, the secondary structure type and the torsion angles, are implemented into the CABS algorithm (see Fig. 2). The additional transformation procedure[18] was inevitable for the torsion angles data due to the requirement of the reduced representation of the protein conformational space in the CABS model. Namely, to determine ϕ and ψ angles one needs coordinates of Cα, N and C$'$ atoms of the backbone, but the latter two are not explicitly defined in the CABS model. Therefore, these two torsion angles ϕ and ψ were transformed into the corresponding θ and γ angles between the consecutive Cα—Cα pseudobond vectors (see Fig. 1). What is very important, this transformation procedure does not decrease the accuracy of the short range angular data.[18,19] Noteworthy, it appears an almost universal statement that the reduced backbone pseudoangles define the local secondary structure much more precisely than the conventional ψ-ϕ angles, based on the detailed backbone coordinates.[19]

Methods for Improvement of the Chemical Shifts Accuracy

The high accuracy of the constraints in the folding simulations is crucial for the final results (see the Results section). For this reason, we also presented two methods for improving the accuracy of the chemical shifts data. In the first method presented later we use additional experimental data to filter the TALOS prediction and in this way we improve the accuracy of the backbone geometry prediction. In the second method, the accuracy of the chemical shifts based constraints is not changed, but the additional distance constraints compensate the influence of the wrong TALOS predictions. The first method is based on the prediction of torsion angles from three-bond coupling constants. We used the MULDER program[20] to determine the possible ranges of ϕ angle for each $^3J_{NHH\alpha}$ coupling constants (for unambiguous determination of the torsion angle more than one type of the ^{3}J constant is needed[12]). The obtained ranges were used to filter out the wrong predictions from the TALOS program. The final set of torsion angles were then subjected to the transformation procedure and the θ and γ based constraints were obtained and implemented into the protein folding simulation. The second method is based on the NOE-like distance constraints which are implemented together with the θ and γ based constraints into the folding simulations. The simulated distance constraints were obtained from PDB structure files of the three proteins. Generally, about $N/7$ (N is the length of the protein sequence) distance constraints are needed for reliable prediction with the CABS algorithm.[1,21] In this case we used only $N/12$ distance constraints. Both of the presented improvement methods were applied to four (out of all tested proteins) structures for which the quality of the final models was the worst with respect to their size (see Table 2): 1ed7, 1imq, 1a3k, 1g1.

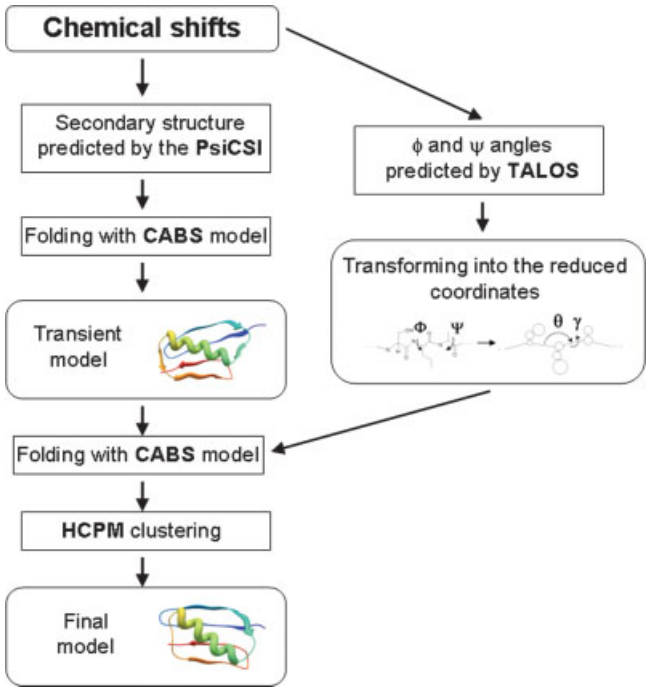

Figure 2. A simplified flowchart of the implementation of the NMR based constraints in the protein folding. In the first step, chemical shifts from BMRB files are used in the prediction of the secondary structure by the PsiCSI server. The predicted secondary structure is implemented into the folding algorithm and the extended protein chain is folded without any additional constraints. In the second step, the angular constraints are employed. The chemical shift data from BMRB files are analyzed by the TALOS program and a set of torsion angles is predicted and converted into the corresponding θ and γ pseudoangles. Local, angular constraints are implemented in the second stage of simulations. The obtained models are subjected to HCPM procedure and the final models are selected. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

Simulated Chemical Shifts Data

Apart from the experimental data we also tested our approach on the simulated data, extracted directly from the PDB structure files. First of all, we wanted to assess if more accurate, chemical shifts-type based, data improve the prediction. For this reason we constructed a simple algorithm which extracts the θ and γ angles directly from the PDB structure file. The algorithm mimics the real NMR experiment with typical ranges of the θ and γ ambiguities. Only 80% of all θ and γ angles are extracted, leaving out data for residues in loops, which are often less accurately defined in the experiment and poorly predicted by the TALOS program.

Constraints Potentials in the Folding Simulations

The predicted secondary structure, obtained from the PsiCSI server, was employed into the CABS algorithm in the form of short-range conformational biases and specific hydrogen bonding pattern biases. The θ and γ based constraints, were implemented in the form of simple linear potentials tested extensively in the recent work.[18]

Imposing even soft conformational constraints during the protein folding simulations may severely handicap sampling of the conformational space, especially in the case of large and medium-size proteins. The reason is that such local restrictions of the backbone may block the proper mutual approach and binding of the secondary structure elements. There is a possibility that various conformations may be trapped in local energy minima in which majority of the constraints (which are not accurate) are satisfied and thus a better conformation, closer to the native one, may not be reached. The optimal situation would be if the majority of the constraints were satisfied and global minima in the CABS force field was reached at the same time. The application of the parallel tempering Monte Carlo in the CABS model certainly improves the sampling,[22] though it does not guarantee the final success of the search process.

For this reason we decided to carry out the simulations in three steps. First, an expanded protein chain was folded only with the predicted secondary structure. Second, soft angular constraints were imposed. Finally, the conformations which were quite close to the native structure were subjected to the kind of the refinement simulations, in which strong constraints were imposed. The last stage should be omitted if the experimental data is of poor quality.

Three-stage simulations improved the sampling of the conformational space for the majority of the proteins of a medium size.

However, in the case of larger proteins (more than 100 residues) we also had to impose a specific, less restrictive potential of the constraints. This specific potential is a slight modification of the simple linear potential used in the previous work:[18]

$$
\begin{aligned}
E_i &= \varepsilon_{\text{restraints}}[f(d\xi_i\text{-}d\xi_{\text{cut-off}}) \\
&\qquad + (d\xi_{\text{cut-off}} - d\xi_{\max})] && \text{for } d\xi_i > d\xi_{\text{cut-off}} \\
E_i &= \varepsilon_{\text{restraints}}(d\xi_i - d\xi_{\max}) && \text{for } d\xi_i > d\xi_{\max} \\
E_i &= 0 && \text{for } d\xi_i < d\xi_{\max}
\end{aligned}
$$

Here, ξ_i is either the γ_i or θ_i angle, $d\xi_i = \xi_i - \xi_{\text{real}}$ (ξ_i is the current value of the ξ angle, ξ_{real} is the value of ξ_i extracted from a PDB file), $\varepsilon_{\text{restraints}}$ is a scaling factor, $d\xi_{\max}$ is a half of the width of the potential well, $d\xi_{\text{cut-off}}$ is a cut-off value of $d\xi_i$ for the less restrictive potential, f is a scaling factor for the less restrictive potential. The optimized values for the above set of parameters are as follows: $d\xi_{\max}(\theta) = 10^\circ$, $d\xi_{\text{cut-off}}(\theta) = 30^\circ$, $d\xi_{\max}(\gamma) = 20^\circ$, $d\xi_{\text{cut-off}}(\gamma) = 60^\circ$, $f(\theta) = 0.3$, $f(\gamma) = 0.24$. The optimization of the parameters was carried out on a small set of proteins (5mba, 1a3k, 5nll), which includes three SCOP classes of proteins: α, β, and α/β. We tested different sets of parameters during the folding simulations and selected the one which provided the most accurate final model.

The modified potential is far less restrictive especially for the states which significantly differ from the native conformation. Namely, values of the potential for θ and γ angles which greatly differ from the restraints (more than $d\xi_{\text{cut-off}}$) are much lower than in the previous version of the potential—the values are rescaled by the factor f. Consequently, the θ and γ based restraints have minor influence on the protein folding at the beginning of the simulation, when the conformation is still quite far from the native structure.

The simultaneous application of the parallel tempering, three-stage simulations, and finally less restrictive potentials turn to be the most profitable way to sample the conformational space of the large proteins limited by the local restraints, because in this case we obtained the final models with the lowest RMSD values with respect to the native structures.

Selection of the Final Protein Model

All structures obtained in the last stage of simulations were grouped using a clustering procedure, called the HCPM (hierarchical clustering of protein models),[23] and several clusters of similar structures were obtained. The HCPM method stands for a clustering procedure, in the first step of which each structure forms a separate cluster and in the last step all the structures belong to a single large cluster. Hierarchical grouping of the cluster stops at a specific merging distance provided by user. In this work, the clustering procedure was stopped when at least five clusters of more than 30 structures were formed (out of several hundreds).

In practice, when the native structure is not known, it is difficult to choose the cluster which includes the model that is the most similar to the native structure. In some cases, the cluster selection could be improved with some additional information from metaservers.[1] In the present work, choosing the most populated cluster from clusters provided by the HCPM seems to be the method that provides models of acceptable accuracy. For comparison, we also analyzed the best model observed in simulations (in the sense of the cRMSD measure)—not necessarily belonging to the most populated cluster. As it is shown in Table 2, in majority of the studied cases the final protein model (from the most populated cluster) is very similar to or identical with the best model (the most similar to the native structure).

Noteworthy, in the case of the restrained protein folding, structures from different clusters are often much closer one to another (in the sense of the cRMSD measure) than it could be

observed in the simulations without restraints (data not shown). It is the consequence of the significant reduction of the conformational space by the restraints. Consequently, the selection of the best protein model is easier in the protein structure prediction supported by the sparse experimental data.

Results and Discussion

Application to the Experimental Data

For the purpose of this work, we extracted data for Cα, Cβ, Hα, and N chemical shifts from several BMRB (Biological Magnetic Resonance Bank) entries.[24] We selected proteins of different topology and size and with different set of chemical shifts data, often incomplete (see Tables1 and 2). First, we used chemical shifts for predicting the secondary structure of the query protein with the PsiCSI server. The average accuracy of this prediction, which is the number of residues with the properly predicted secondary structure (which is in agreement with the DSSP definition of secondary structure[25]) divided by the number of all residues in the protein, for all tested proteins was 84%. Second, the BMRB files were converted into the TALOS input files and torsion angles were computed. We used the recent version of the TALOS program with the database of 78 structures[26] instead of the older version which used only 20 high resolution structures. According to our tests, for the newer version the coverage of the predicted torsion angles seems to be larger than that for the previous version of the TALOS.

If we had used the TALOS program in a simple automatic way we would have obtained torsion angles for about 50–60% of all residues (for the proteins from our test set, see Table 1). To transform the TALOS output data into the constraints used in the reduced model, we need two consecutive pairs of ϕ and ψ angles for one γ angle and the corresponding pair of θ angles. There are many breaks in the TALOS prediction, and so the actual coverage of the θ and γ pseudoangles data obtained by our method would be only 40–50%. That is because a sequential pair of ϕ and ψ angles is needed for the single γ angle determination.[18] An additional comment is needed for the transformation of ϕ and ψ into θ angle. The coverage of the θ angle should be the same as the coverage of pairs of ϕ and ψ obtained by the TALOS. However, our method does not take into consideration the distinct conformation of the proline amino acids. Hence, we reject all pairs of ϕ and ψ for the proline residues during the transformation into θ angles. The reason for such choice is that proline amino acids are rare in proteins and a detailed reconstruction of the backbone is not the aim of this work. Though, in the future implementation of our pipeline the angular data for proline residues could be certainly added.

Such limited set of θ and γ based constraints from the automatic TALOS prediction turned out to be insufficient for a good structure prediction with our method. For proper identification of the best protein model we need at least 60% of γ angles and 70% of θ angles defined (it corresponds to at least 70% of ϕ and ψ angles). Therefore, we had to optimize the prediction from the TALOS to obtain a larger set of data (with coverage 70–80% for our test proteins, see Table 1). We optimized the TALOS prediction by adding to the default predictions several pairs of ϕ and ψ angles for which at least 7 out of 10 database matches were in the same high-populated region of the Ramachandran map (for the TALOS default prediction it should be at least 9 from 10 matches converged). There were very few predictions, which could be described as completely wrong and most of the predicted angles felt in the range of $\pm 20^\circ$ (γ angles) and $\pm 10^\circ$ (θ angles) for all tested proteins (see results in Table 1). It appears to be an important observation for experimentally supported protein modeling in general.

As it is presented in the Table 2 (see also Fig. 3), final results of the folding simulations correlate with the accuracy and the coverage of the angular constraints. The accuracy and the coverage of the θ and γ constraints are defined as follows:

$$\mathrm{ACC} = N_{\mathrm{good}}/N_{\mathrm{all}}$$
$$\mathrm{COV}(\theta) = N_{\mathrm{all}}/(N_{\mathrm{RES}} - 2)$$
$$\mathrm{COV}(\gamma) = N_{\mathrm{all}}/(N_{\mathrm{RES}} - 3)$$

Here, ACC is the accuracy of the prediction of either θ or γ angles, N_{good} is the number of good predictions ($\pm 20^\circ$ for γ and $\pm 10^\circ$ for θ), N_{all} is the number of all predictions, COV(θ) is the

Table 1. Summary of Experimental Data and TALOS Predictions.

PDB id	Length	Number of chemical shifts	Coverage of ϕ and ψ obtained automatically from TALOS (%)	Accuracy of the automatic TALOS prediction of ϕ and ψ (%)	Coverage of ϕ and ψ from optimized TALOS prediction (%)	Accuracy of the optimized TALOS prediction of ϕ and ψ (%)
1ed7	45	79	44	58	73	52
2gb1	56	109	68	66	80	67
1bw5	66	250	65	77	82	72
1tiz	67	320	79	84	85	82
1ubq	76	374	66	97	79	97
4icb	76	373	75	96	83	94
1imq	86	325	53	72	81	65
1gm1	94	411	48	74	67	72
1a3k	137	530	54	84	72	81

Table 2. Summary of the Restraints and Accuracy of the CABS Models.

			The pseudoangles data used as the restraints							
			Coverage (%)		Accuracy (%)		Completely wrong restraints (%)		Results of the simulations	
PDB id	Length	SCOP class	θ	γ	θ ($<10^\circ$)[a]	γ ($<20^\circ$)	θ ($>30^\circ$)[b]	γ ($>60^\circ$)	cRMSD[c] (Å)	cRMSD[d] (Å)
1ed7	45	β	71	56	75	44	0	24	3.4	3.4
2gb1	56	$\alpha + \beta$	80	64	82	58	4	14	2.4	2.4
1bw5	66	α	80	68	80	71	13	16	7.2	4.7
1tiz	67	α	85	78	82	86	0	2	3.2	2.9
1ubq	76	$\alpha + \beta$	75	63	86	94	4	1	3.7	2.8
4icb	76	α	80	71	90	87	0	2	4.1	4.0
1imq	86	α	79	66	82	65	1	19	9.9	7.6
1gm1	94	β	67	50	81	62	2	19	10.4	10.4
1a3k	137	β	67	53	73	67	1	12	11.6	11.6

[a]The accurate θ angle differs from the real data, extracted from the PDB structure file, by at most 10° (in the case of the γ angle, by at most 20°).
[b]The totally wrong prediction of the θ angle means that it differs from the real data, extracted from the PDB structure file, by at least 30° (in the case of the γ angle, by at least 60°). Such error ranges resemble differences between θ and γ angles typical for helices and beta sheets, respectively.
[c]The model from the most populated cluster of the structures.
[d]The model from the cluster of the structures, which is the most similar to the native structure.

coverage of the θ predictions, COV(γ) is the coverage of the γ predictions, N_{RES} is the number of all residues in a protein. The influence of the coverage of the angular data is clearly visible in the cases of the 1imq and 1gm1 proteins, both of comparable size. For the 1gm1 protein we obtained smaller set of restraints than for the 1imq, what affected the results of the folding simulations. The influence of the accuracy could be seen from inspection of the final models for 1imq and 4icb proteins, for which the coverage of the data was similar. The quality of the final model of 1imq (measured by cRMSD from the crystallographic structure) was significantly worse than in the 4icb case. Moreover, the fraction of the completely wrong constraints is also important. For example, the 1bw5 and 1tiz protein structures, both of the same SCOP class, with the similar coverage of

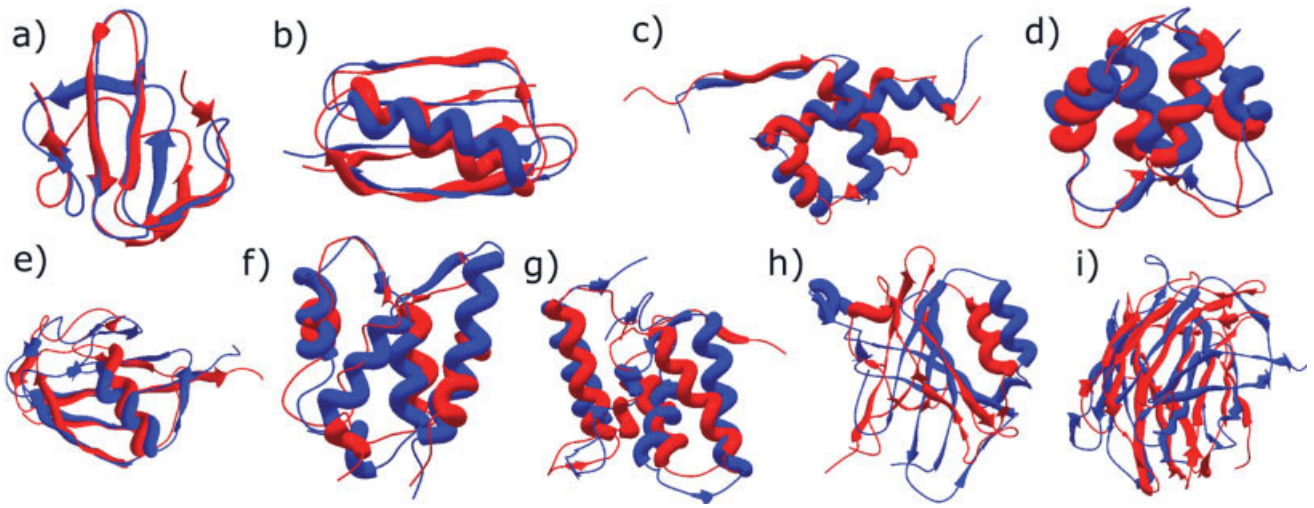

Figure 3. The results of the folding simulations with the PsiCSI predicted secondary structure and the local angular constraints. The native structures, which are colored blue, are superimposed on the predicted models (red). Models correspond to the centroids of the best clusters of the obtained structures (in the sense of the cRMSD from native). The model for the (a) 1ed7 protein, (b) 2gb1, (c) 1bw5, (d) 1tiz), (e) 1ubq, (f) 4icb, (g) 1imq, (h) 1gm1, (i) 1a3k. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

Table 3. Summary of Simulations with Chemical Shifts Based Constraints and Sparse Distance Constraints.

PDB id	Length	Number of distance constraints	cRMSD[a] (Å)	cRMSD[b] (Å)
1imq	86	6	5.4	5.4
1gm1	94	7	11.7	8.9
1a3k	137	10	4.5	4.5

[a]The model from the most populated cluster of the structures.
[b]The model from the best cluster of the structures, which is the most similar to the native structure.

the data and nearly the same size, were predicted with different resolution, because the fraction of wrong constraints was significantly larger for the 1bw5 protein.

The quality of the final protein model depends also on the length of the protein sequence. In general, small protein structures are more accurately predicted. We did not observe significant influence of the type of the protein secondary structure on the final results, but increasing the number of loops and irregular fragments of the backbone hampered the TALOS prediction and consequently the final results also.

The most similar to the native structures were models of proteins of small and medium size (1tiz, 1ubq, 4icb), for which the coverage of θ and γ angular constraints reaches about 70–80%, the accuracy is at least 80% and there is no or nearly no false constraints. In the cases of two small proteins 1ed7 and 2gb1, despite the fact that the accuracy is low, final models are still close to the native structures. It is worth to notice that for the 1ed7 protein we did not have any carbon chemical shifts. Consequently, the accuracy of the TALOS prediction was extremely poor. Despite this, we obtained the model in which the overall location of the secondary structure elements is the proper one (especially for the central β-hairpin). Slight distortions are located only in the loops. It seems obvious that for small proteins the angular constraints could be less accurate than in the case of large proteins.

The only protein model with the high cRMSD value, despite the high accuracy of the TALOS prediction, is the model of the 1bw5 protein. Both the C-terminus and the N-terminus fragments of the protein are ambiguously determined by NMR data (see the 50 NMR structures of the 1bw5 deposited in PDB).[27] Therefore, instead of the cRMSD of the whole protein model, we should take into consideration only the helical core (residues from 12 to 55). Then, the cRMSD of the best model is only 2.7 Å.

For the larger and structurally more complex proteins (1imq, 1gm1, 1a3k) the accuracy of the TALOS prediction is not sufficient for an acceptable structure prediction. We carried out some tests for the 1imq protein to prove the influence of the accuracy of the data on final results. It turned out that increase in the number of the correct constraints by about 10% leads to decrease of the best model cRMSD from 7.6 to 5.5 Å.

The average running time of the folding simulations for the protein 2gb1 was 1 day. The time of the restrained folding simulations scales linearly with the number of residues in a protein.

Improvement of the Accuracy of the Experimental Data

The accuracy of the chemical shifts based constraints was not sufficient for the prediction of the three-dimensional structure of two out of all tested proteins (1a3k and 1gm1). We obtained protein models quite far from the native structures (10–11 Å). In the other two cases (1ed7 and 1imq) the quality of the final models is better (correct overall topology of the fold) but still not satisfactory. For these four proteins we applied two methods to improve the quality of the final models. The first method, applied to the 1ed7 protein, is based on the prediction of torsion angles from three-bond coupling constants. In this case we had only 36 $^3J_{NHH\alpha}$ constants available. The ranges for the ϕ angle values from the MULDER program improved the accuracy of the constraints by a few percentage points. The cRMSD of the best model of the protein structure improved from 3.7 to 2.9 Å. The second method, based on the NOE-like distance constraints, is applied to three proteins (1imq, 1a3k, and 1gm1). Folding simulations with only $N/12$ distance constraints superimposed on the contact of the side groups and the chemical shifts based constraints improved significantly the quality of the best models for all proteins (see Table3). The methods for improving the accuracy of the chemical shifts data, which were presented in this section, are promising, although some further research on a larger set of protein structures should be carried out. Again, here we focus on the application of the chemical shifts data alone in the protein structure determination. The main factor behind performing the additional simulations with supplementary experimental data is to illustrate the open character of the proposed methodology for the sparse data supported structure determination.

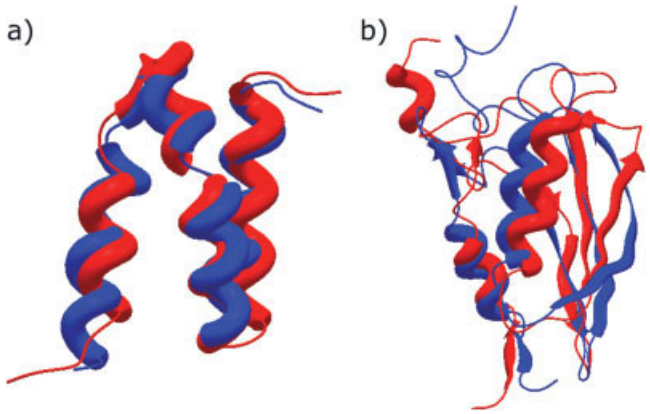

Figure 4. (a) Superimposition of the native structure (blue) and the model from the cluster most similar to the native structure (red) for the T0215 target (PDB id: 1 × 9b) protein and for the T0230 target (PDB id: 1wcj) (b). T0215 is 53 residues long and T0230 is 102 residues long. We used 223 chemical shifts to obtain angular constraints for the T0215 and 486 for the T0230. Accuracy of the θ angles prediction is 74% for T0215 and 52% for T0230. The γ angles were predicted with the accuracy 58% (T0215) and 38% (T0230). The cRMSD is 1.9 Å for the T0215 model and 4.5 Å for the T0230 model. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

Table 4. Summary of Models Obtained with Simulated θ and γ Based Constraints, Extracted from the PDB Structure Files.

PDB id	Length	Type	cRMSD[a] (Å)	cRMSD[b] (Å)
2gb1	56	$\alpha + \beta$	1.3	1.3
1ubq	76	$\alpha + \beta$	4.5	3.3
5nll	138	α/β	4.3	3.7
4icb	76	α	3.3	3.3
2spz	58	α	4.8	3.6
1bw5	66	α	4.2	3.8
1imq	86	α	3.8	3.8
5mba	146	α	14.7	8.7
1tiz	67	α	2.8	2.8
1gm1	94	β	8.4	7.4
1ed7	45	β	1.9	1.9
2pcy	99	β	3.7	3.7
1a3k	137	β	5.9	5.8

[a]The model from the most populated cluster of the structures.
[b]The model from the best cluster of the structures, which is the most similar to the native structure.

Application to the CASP6 Fold-Recognition-Analogy Targets

We have tested our approach in post-predictions of structures of two targets from the CASP6 experiment. The target proteins were classified in the fold-recognition-analogy category, meaning that there are no close homologues structures in the PDB database. This is the case in which sparse experimental data could be very helpful. Both protein structures for the target T0215 and T0230 were modeled with a high accuracy (see Fig. 4) using the method proposed in this work. What is more important, all clusters of structures obtained in the simulations were very similar (the cRMSD values for all pairs of the cluster's centroids did not exceed 1 Å). In such cases the selection of the best model from the simulation trajectory via hierarchical clustering could be omitted, thereby reducing computational cost of the structure determination process.

Obtained in this work models of T0215 and T0230 are qualitatively more accurate than the best models presented during the CASP6 (see the summary of the experiment on the CASP6 web site: http://predictioncenter/CASP6/org), where the most advanced methods for structure prediction were evaluated. The secondary structure prediction by PSIPRED done for the purpose of this work was almost identical with the predictions available during the CASP6 (with older protein database). Moreover, the folding was done without any restraints from possible homologous (or analogous) templates. Thus, the observed qualitative improvement of the accuracy of the T0215 and T0230 models is highly significant.

Simulated Constraints

Chemical shifts are the relatively effortless way to obtain local, angular data which could be useful in the protein structure prediction. However, it is worth noticing that the method employed in this work could give much better results if the experimental data of dihedral angles was more accurate.[18] As it can be seen in Table4, 80% of θ and γ angles as the local constraints is enough for reasonable predictions of all tested proteins with an acceptable resolution. A comment is needed at this point. Why in these cases we did not need any global information about the fold even for the larger proteins? The answer is that when the angular, local constraints are very accurate the CABS force field is good enough to assembly the secondary structure elements in the proper mutual orientation and native-like registration of the side chains' interactions.

Conclusions

It has been shown that such limited experimental data as chemical shifts are sufficient for determination of three-dimensional structures of proteins as long as the accuracy of the data is satisfactory. The required accuracy of data is higher for larger proteins than for small proteins with not too many structural building blocks. In the case of poor quality chemical shifts data, the structure prediction could be improved by application of additional, sparse experimental data of different kinds. The method presented in this work has been tested for globular proteins with the sequence length below 150 amino acids. For proteins with more than 150 residues employing some global, distance, or orientational constraints could be inevitable for the proper prediction. The long range goal of this line of work is to provide a versatile method for protein structure determination from sparse experimental data that are easy to obtain rapidly, using very limited experimental resources. Applications to the chemical shifts data described here are the first step to achieve this goal. Finally, it should be pointed out that the reduced α-carbon trace representation of the CABS protein models is consistent with the all atom representation of the main chain and that the atomic details of the backbone could be easily reconstructed,[27,28] making the obtained models more useful.

Acknowledgment

Computational part of this work was done using the computer cluster at the Computing Center of Faculty of Chemistry, Warsaw University.

References

1. Kolinski, A.; Bujnicki, J. M. Proteins 2005, 61, 84.
2. Bradley, P.; Misura, K. M.; Baker, D. Science 2005, 309, 1868.
3. Simmerling, C.; Lee, M.; Ortiz, R.; Kolinski, A.; Skolnick, J.; Kollman, P. A. J Am Chem Soc 2000, 122, 8392.
4. Vieth, M.; Kolinski, A.; Brooks, C. L., III; Skolnick, J. J Mol Biol 1995, 251, 448.
5. Meiler, J.; Baker, D. Proc Natl Acad Sci USA 2003, 100, 15404.
6. Bax, A. Protein Sci 2003, 12, 1.
7. Bonvin, A. M.; Houben, K.; Guenneugues, M.; Kaptein, R.; Boelens, R. J Biomol NMR 2001, 21, 221.
8. Cornilescu, G.; Delaglio, F.; Bax, A. J Biomol NMR 1999, 13, 289.
9. Hung, L. H.; Samudrala, R. Protein Sci 2003, 12, 288.
10. McGuffin, L. J.; Bryson, K.; Jones, D. T. Bioinformatics 2000, 16, 404.
11. Ginzinger, S. W.; Fischer, J. Bioinformatics 2006, 22, 460.

12. Schmidt, J. M.; Blumel, M.; Lohr, F.; Ruterjans, H. J Biomol NMR 1999, 14, 1.
13. Malliavin, T. E. Curr Org Chem 2006, 10, 555.
14. Liu, G.; Shen, Y.; Atreya, H. S.; Parish, D.; Shao, Y.; Sukumaran, D. K.; Xiao, R.; Yee, A.; Lemak, A.; Bhattacharya, A.; Acton, T. A.; Arrowsmith, C. H.; Montelione, G. T.; Szyperski, T. Proc Natl Acad Sci USA 2005, 102, 10487.
15. Kolinski, A. Acta Biochim Pol 2004, 51, 349.
16. Spera, S.; Bax, A. J Am Chem Soc 1991, 113, 5490.
17. Wishart, D. S.; Sykes, B. D.; Richards, F. M. J Mol Biol 1991, 222, 311.
18. Plewczynska, D.; Kolinski, A. Macromol Theory Simul 2005, 14, 444.
19. Oldfield, T. J.; Hubbard, R. E. Proteins 1994, 18, 324.
20. Padrta, P.; Sklenar, V. J Biomol NMR 2002, 24, 339.
21. Kolinski, A.; Betancourt, M. R.; Kihara, D.; Rotkiewicz, P.; Skolnick, J. Proteins 2001, 44, 133.
22. Earl, D. J.; Deem, M. W. Phys Chem Chem Phys 2005, 7, 3910.
23. Gront, D.; Kolinski, A. Bioinformatics 2005, 21, 3179.
24. Seavey, B. R.; Farr, E. A.; Westler, W. M.; Markley, J. L. J Biomol NMR 1991, 1, 217.
25. Kabsch, W.; Sander, C. Biopolymers 1983, 22, 2577.
26. http://spin.niddk.nih.gov/NMRPipe/talos/.
27. Ippel, H.; Larsson, G.; Behravan, G.; Zdunek, J.; Lundqvist, M.; Schleucher, J.; Lycksell, P. O.; Wijmenga, S. J Mol Biol 1999, 288, 689.
28. Feig, M.; Rotkiewicz, P.; Kolinski, A.; Skolnick, J.; Brooks, C. L., 3rd. Proteins 2000, 41, 86.