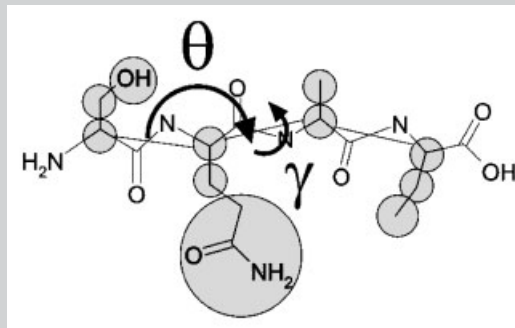


**Summary:** A reduced high-coordination lattice protein model and the Replica Exchange Monte Carlo sampling were employed in de novo folding simulations of a set of representative small proteins. Three distinct situations were analyzed. In the first series of simulations, the folding was controlled purely by the generic force field of the model. In the second, a bias was introduced towards the theoretically predicted secondary structure. Finally, we superimposed soft restraints towards the native-like local conformation of the backbone. The short-range restraints used in these simulations are based on approximate values of  $\phi$  and  $\psi$  dihedral angles, which may simulate restraints derived from inaccurate experimental measurements. Incorporating such data into the reduced model required developing a procedure, which transforms the  $\phi$  and  $\psi$  coordinates into coordinates of the protein alpha carbon trace. It has been shown that such limited data are sufficient for de novo determination of three-

dimensional structures of small and topologically not too complex proteins.



Protein folding based on secondary structure prediction and simulated torsion angles data.

# Protein Folding with a Reduced Model and Inaccurate Short-Range Restraints

Dorota Plewczynska, Andrzej Kolinski\*

Faculty of Chemistry, Warsaw University, Pasteura 1, 02-093 Warsaw, Poland

Fax: (+48) 22 8225996; E-mail: kolinski@chem.uw.edu.pl

Received: March 21, 2005; Revised: June 14, 2005; Accepted: July 7, 2005; DOI: 10.1002/mats.200500020

**Keywords:** dihedral angles; Monte Carlo simulation; proteins; reduced models; structure prediction

## Introduction

Experimental determination of protein structures remains far behind the recent rapid increase of the number of known protein sequences. In contrast to very fast and almost fully automated procedures for genome sequencing the experimental methods of structural proteomics are very expensive and time-consuming. Therefore, theoretical methods for determining low and moderate resolution protein structures become valuable tools of increasing range of applicability. Precise determination of protein structures often requires the implementation of restraints derived either from experiments or from the database searching. In this work, we analyze the effect of short-range inaccurate restraints of various types on the quality of predicted structures.

Restraints of the first type implemented in the folding algorithm are loosely defined geometrical biases consistent with the theoretically predicted secondary structure. The prediction of the secondary structure, usually based on computational neural network models and multiple se-

quence alignments, is typically 70–80% accurate for the three-letter code: H, a helix; E, an extended state; and C, a coil (less regular structures).<sup>[1]</sup> Thus, by no means the tertiary structure is defined by the restraints derived from such predictions.

Restraints of the second type employed in this work are inaccurate angular restraints for the main chain rotations based on  $\phi$  and  $\psi$  torsion angles. Given a fixed geometry of peptide bond, exact values of the  $\phi$  and  $\psi$  angles define a unique conformation of the protein backbone.<sup>[2]</sup> However, direct determination of the entire protein structure using approximate, experiment-based values of these torsion angles alone is impossible due to the rapid propagation of the experimental errors along the chain. Thus, the  $\phi$  and  $\psi$  angles encode only short-range conformational propensities.<sup>[3]</sup> Consequently, such restraints are commonly used in the molecular dynamic refinement of experimental models of proteins but occasionally used in de novo theoretical determination of protein structures, which often require global information about the fold.<sup>[4,5]</sup> Nevertheless, such

restraints could be helpful when no other structural information is collected as they could be obtained from relatively simple NMR experiments.<sup>[6]</sup>

Basing on the Ramachandran plot, the angular restraints could be used in the same manner as the restraints of the first type: to determine the location of the regular secondary structure elements.<sup>[7]</sup> However, such angular restraints go beyond the definition of the secondary structure, providing also approximate information about local geometry of loops and coil regions. Consequently, implementing them into the folding procedure should be more useful for the structure prediction than the knowledge of the secondary structure alone.

In this work, we introduce a set of restraints of both types into a recently developed lattice protein model CABS ( $C\alpha$ - $C\beta$ -side group)<sup>[8]</sup> and perform series of folding simulations to examine the range of applicability of short-range information alone in computational determination of protein structures. We use simulated data of the angular restraints extracted from the Brookhaven Protein Data Bank (PDB) structure files in order to evaluate precisely the effect of errors or inaccuracies of the restraints on the accuracy of the obtained models. The model which is used in simulations employs the C-alpha backbone trace representation and, therefore, a transformation of the  $\phi$ - and  $\psi$ -based restraints into corresponding angles between  $C\alpha$ - $C\alpha$  virtual bond vectors is required.

## Model and Method

### A Protein Structure Prediction Algorithm

Both types of the restraints are implemented into the force field of the recently published CABS protein model.<sup>[8]</sup> The model is based on a lattice representation of the C-alpha backbone trace. Coordinates of  $C\beta$  atoms and centers of side groups are estimated from  $C\alpha$  positions. Interaction scheme includes knowledge-based potentials such as: generic protein-like biases, statistical potentials for the short-range conformational propensities, a model of hydrogen bonds, and potentials describing interactions between side groups of amino acids. The conformational space is sampled using a variant of the Replica Exchange Monte Carlo (REMC) method with number of replicas equal to 20 in all tested cases. Simulations were carried out in two steps. First, a randomly expanded polypeptide chain was subjected to a thermal annealing REMC procedure with the temperature (in other words, the scaling factor of the model energy) of the replica of the lowest energy ranging from 3.0 (4.0 in cases of larger proteins, e.g., 5mba) to 1.0. The second stage was the REMC simulation in which the temperature of the lowest energy replica was 1.0. Trajectories obtained in the last stage of simulations were subjected to a clustering procedure, called the Hierarchical Clustering of Protein Models (HCPM), which grouped the most similar

(in the sense of the cRMSD measure) structures and provided the final, most likely model of the protein as the representative of the largest cluster. More details of the HCPM method are described in the recent publication.<sup>[9]</sup> Although the predicted structures are  $C\alpha$ , only models with approximate positions of the  $C\beta$  atoms and the centers of the side chains, it is possible to rebuild the all-atom protein chain using a simple modification of a published previously reconstruction procedure.<sup>[10]</sup> The CABS modeling tool has been extensively tested during the last round of the community-wide protein structure prediction experiment (CASP6) and proven to be one of the two best algorithms for a large-scale protein modeling and de novo protein structure prediction from sequence of amino acids alone (see the CASP6 homepage <http://predictioncenter.llnl.gov/casp6/Casp6.html> or a summary of the results on our website <http://biocomp.chem.uw.edu.pl/files/casp>).

### Implementation of Restraints

We used the PSIPRED server for the secondary structure prediction.<sup>[11]</sup> The predicted secondary structure enters into the force field of the model in three ways. First, for the H and E regions the potential of the short-range interactions is biased towards the average geometry of these regular secondary structure elements. Second, the model of the main chain hydrogen bonds employs the predicted secondary structure in a set of "mixing" rules. Namely, long-range hydrogen bonds are excluded for pairs of residues assigned as HH and HE. Finally, for longer series of H and E states a soft potential favors the proper distances between  $i$ th and  $(i + 7)$ th, and between  $i$ th and  $(i + 6)$ th alpha carbons for the H and E series, respectively. The details of the implementation could be found in recent publications.<sup>[8,11]</sup>

Due to the usage of the reduced protein model, the restraints of the second type based on the  $\phi$  and  $\psi$  angles had to be applied into the folding algorithm in the form of corresponding  $\gamma$  and  $\theta$  angles, where  $\gamma$  is a torsion angle between three successive  $C\alpha$ - $C\alpha$  virtual bond vectors and  $\theta$  is a bond angle between two successive  $C\alpha$ - $C\alpha$  vectors [see Figure 1(a) and (b)]. According to Hubbard and Oldfield, these two pseudoangles define the protein conformation in a

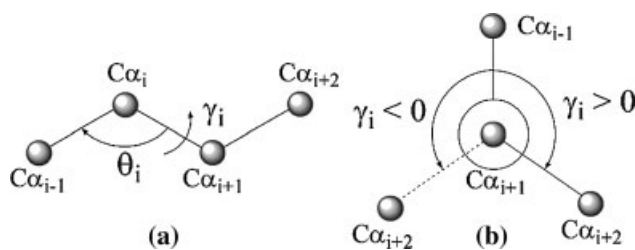


Figure 1. (a) The definitions of the  $\gamma$  pseudotorsion angle and the  $\theta$  pseudobond angle. (b) Explanation of the chirality of the  $C\alpha$  backbone in the definition of the  $\gamma$  angle.

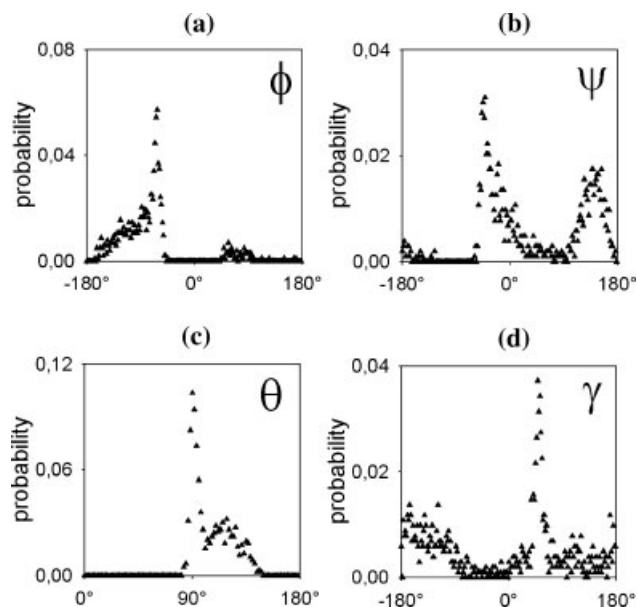


Figure 2. The probability distribution of the  $\phi$  and  $\psi$  angles (a, b) in comparison to the distribution of the  $\theta$  and  $\gamma$  pseudoangles (c, d). The main difference between the plots could be observed in the regions of the regular secondary structure. The highest peak, indicating helical conformations, is more diffuse in the case of dihedral angles than it is for the  $\theta$  and  $\gamma$  angles. Consequently, identification of the helical conformation could be more precise on the basis of pseudoangles. For this analysis we used a set of 11 high-resolution protein structures (see *Results and Discussion*).

more comprehensive way than the  $\phi$  and  $\psi$  angles do.<sup>[12]</sup> Namely, the ranges of the allowed  $\gamma$  and  $\theta$  angles typical for helices and beta sheets are more restrictive than it is in the case of the corresponding  $\phi$  and  $\psi$  angles (see also Figure 2, which presents our results). Thus, the protein conformation is more precisely defined in terms of the secondary structure. Moreover, analysis of the pseudoangles enables to distinguish different conformations of beta sheets and turns, which cannot be identified on the basis of Ramachandran plot.<sup>[12]</sup>

Although we used simulated  $\gamma$  and  $\theta$  restraints extracted directly from the PDB files, a procedure which transforms  $\phi$  and  $\psi$  angles into  $\gamma$  and  $\theta$  angles is needed for the future applications to experimental data. Levitt demonstrated that  $\gamma$  and  $\theta$  are related to  $\phi$  and  $\psi$  according to the following formulas:<sup>[13]</sup>

$$\begin{aligned} \gamma &\approx 180^\circ + \phi(i+2) + \psi(i+1) + 20^\circ [\sin\{\phi(i+1)\} \\ &\quad + \sin\{\psi(i+2)\}] \\ \theta &\approx 106^\circ + 13^\circ \cos(\gamma - 45^\circ) \end{aligned} \quad (1)$$

In this work we propose a different procedure (see Figure 3), which is based on a reconstruction of the short fragment of the protein main chain on the basis of its dihedral angles. Such reconstruction is insufficient for obtaining the model of the whole protein, due to inaccuracies of the dihedral angles (see *Introduction*). However, its accuracy is

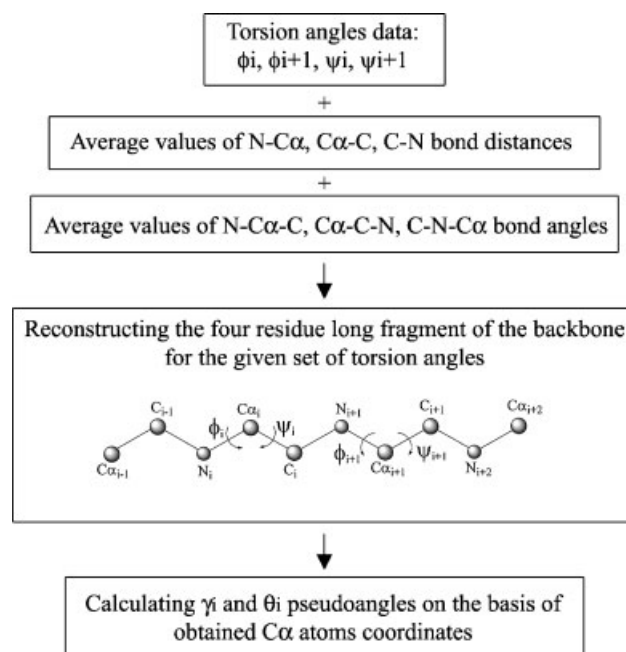


Figure 3. The flowchart of the procedure transforming the  $\phi$  and  $\psi$  coordinates into the  $\gamma$  and  $\theta$  coordinates.

acceptable for approximate evaluation of the  $\gamma$  and  $\theta$  angles, as we proved in this work.

In order to evaluate the  $\gamma_i$  and  $\theta_i$  angles of the  $i$ th residue employing our transformation procedure the coordinates of four successive  $C\alpha$  atoms ( $i-1, i, i+1, i+2$ ) are used [see Figure 1(a)]. For the given coordinates of the first  $C\alpha$  atom (that is  $C\alpha_{i-1}$ ) the rest of the four-residue fragment of the protein chain could be rebuilt using: average bond angles, average bond lengths (see Table 1), and only two successive pairs of  $\phi$  and  $\psi$  angles ( $\phi_i, \psi_i, \phi_{i+1}, \psi_{i+1}$ ), as the rest of them do not have any impact on the  $\gamma_i$  and  $\theta_i$  values. We made here an additional assumption that all peptide bonds

Table 1. Average values of bond angles and lengths evaluated from a set of 11 high-resolution protein structures. These values were used in the procedure translating the torsional angles of the main chain into the  $\gamma$  and  $\theta$  angles.

Atoms of the main chain	Bond lengths
	Å
C $\alpha$ -C	1.520 $\pm$ 0.031 <sup>a)</sup>
C-N	1.330 $\pm$ 0.320
N-C $\alpha$	1.455 $\pm$ 0.030
	Bond angles
	degree
C $\alpha$ -C-N	116.91 $\pm$ 3.61 <sup>a)</sup>
C-N-C $\alpha$	121.60 $\pm$ 4.13
N-C $\alpha$ -C	111.18 $\pm$ 4.84

<sup>a)</sup> Error ranges which include more than 95% of predicted values, evaluated as two standard deviations.

are planar and of the *trans*-type. After rebuilding the whole, four-amino acid fragment of the main chain, the  $\gamma_i$  and  $\theta_i$  angles could be evaluated on the basis of three virtual bond vectors, which connect four successive C $\alpha$  atoms. The procedure is repeated for all pairs of the  $\gamma_i$  and  $\theta_i$  angles using each time the same bond angles and bond lengths but different values of the  $\phi$  and  $\psi$  angles.

The  $\gamma$  and  $\theta$  angular restraints are introduced into the model force field in the form of two separate linear potentials, each given by the same Equation (2):

$$E_i = \varepsilon_{\text{restraints}}(\Delta\xi_i - \Delta\xi_{\text{max}}) \quad \text{for } \Delta\xi_i > \Delta\xi_{\text{max}} \\ E_i = 0 \quad \text{for } \Delta\xi_i < \Delta\xi_{\text{max}} \quad (2)$$

Here  $\xi_i$  is either the  $\gamma_i$  or the  $\theta_i$  angle,  $\Delta\xi_i = \xi_i - \xi_{\text{real}}$  ( $\xi_i$  is the current value of the  $\xi$  angle, and  $\xi_{\text{real}}$  is the value of  $\xi_i$  extracted from PDB file). Both parameters of the potential,  $\varepsilon_{\text{restraints}}$  (a scaling factor) and  $\Delta\xi_{\text{max}}$  (half of the width of the potential well), were optimized to minimize the cRMSD value (root mean square deviation of the corresponding C $\alpha$  coordinates of the final model from the native structure). As the  $\gamma$  angles are more crucial for the backbone conformation than the  $\theta$  angles, we set the scaling factor  $\varepsilon_{\text{restraints}}(\gamma)$  twice as large as  $\varepsilon_{\text{restraints}}(\theta)$ . As for the  $\Delta\xi_{\text{max}}$  parameters, which refer to the largest allowed difference between the current values of the angles and the real values extracted from the PDB file, we used the following values:  $\Delta\xi_{\text{max}}(\gamma) = 20^\circ$  and  $\Delta\xi_{\text{max}}(\theta) = 10^\circ$ .

The range of typical  $\gamma$  and  $\theta$  angles for helices is tighter than for beta sheets.<sup>[3]</sup> Consequently, assuming the secondary structure dependent width of the potential well could improve the results. However, such improvement turned to be insignificant and not worth applying.

## Results and Discussion

### Calculation of $\gamma$ and $\theta$ Angles

A procedure which transforms  $\phi$  and  $\psi$  coordinates into  $\gamma$  and  $\theta$  was tested on a set of small 11 proteins (1a6m, 1b9o, 1bq8, 1brf, 1d4t, 1f94, 1iqz, 1ir0, 1j0p, 1kth, 1pm1). We used only high-resolution protein structures (at least 1.2 Å) in order to minimize uncertainties of the experimental data, which could influence the final results.<sup>[10]</sup> Average bond angles and bond lengths (see Table 1) were also derived from this set of proteins.

For each structure of the tested group all its  $\gamma$  and  $\theta$  angles were predicted on the basis of its  $\phi$  and  $\psi$  torsion angles (which were extracted from the PDB files using the DSSP program)<sup>[14]</sup> and compared with the real values. Summary results for all proteins (see Figure 4) prove the high correlation with the real values. As it is shown in Figure 5, more than 95% of the  $\gamma$  and  $\theta$  angles are in the  $\pm 10^\circ$  error range. Our transformation procedure seems to be more precise than the Levitt's calculations (see Figure 5), though

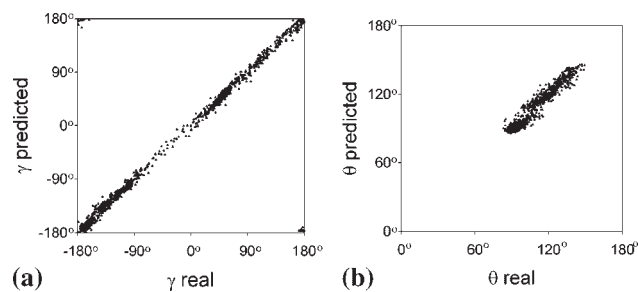


Figure 4. Results of the transformation procedure for 11 high-resolution protein structures (1a6m, 1b9o, 1bq8, 1brf, 1d4t, 1f94, 1iqz, 1ir0, 1j0p, 1kth, 1pm1). (a) The prediction of the  $\gamma$  angles. (b) The prediction of the  $\theta$  angles.

some further testing on a larger group of proteins should be performed. In particular, Levitt's approximation of the  $\theta$  angle enables to distinguish helical and extended regions,

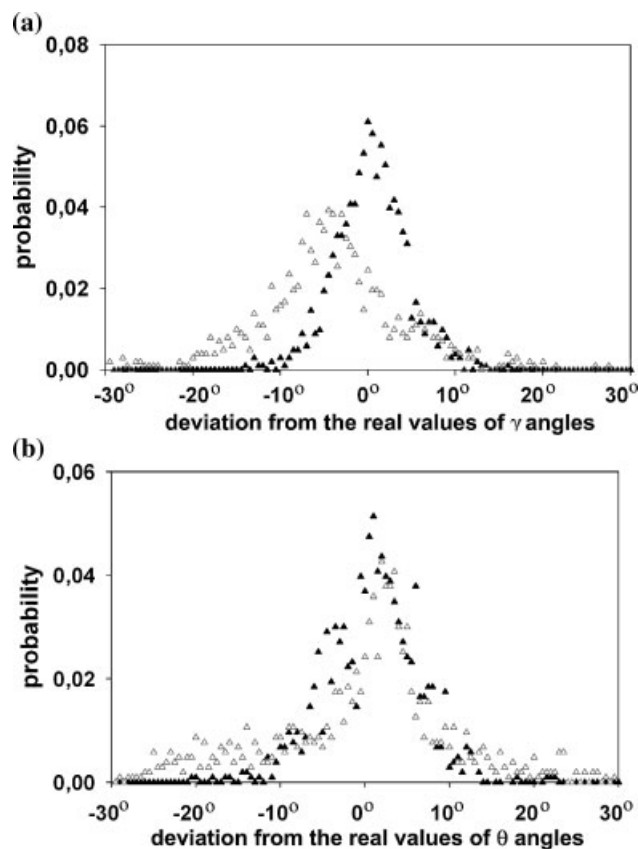


Figure 5. Histograms comparing the results obtained using our transformation procedure (black solid triangles) and Levitt's approximations (gray open triangles). Because in the original Levitt's paper the  $\gamma$  angle is differently defined (its values are in the range 0–360° instead of the range –180 to 180°, which is used in this paper) we made an appropriate modification of the Equation (1), consistent with the convention assumed in the present work. The transformation of coordinates performed using our procedure seems to be more precise in comparison with the Levitt's method. (a) The prediction of the  $\gamma$  pseudoangles. (b) The prediction of the  $\theta$  pseudoangles.

but due to the quite large error range ( $\pm 30^\circ$ ) it is rather not sufficient to apply the experimental  $\phi$  and  $\psi$  data to the computational determination of the protein structure, which is the aim of this work. In the case of the Levitt's approximation of the  $\gamma$  angle, which falls into the  $\pm 20^\circ$  error range, it could be possible to use it for the transformation of the experimental dihedral angles.

In both approximations, using the Levitt's method and ours, the prediction of the  $\theta$  angle is worse than the prediction of the  $\gamma$  angle. In the case of Levitt's method it is probably the consequence of the fact that the  $\theta$  angle is evaluated on the basis of the  $\gamma$  angle using a too approximate relation. During the testing of our procedure we discovered that  $\theta$  angles were far more sensitive to slight differences in the bond angles and the bond lengths, which were used in the reconstruction of the backbone, than  $\gamma$  angles. Most likely, it is the consequence of the fact that the values of  $\theta$  angles observed in proteins fall into a very limited range [see Figure 2(c)] in contrast to the wider distribution of the  $\gamma$  angle.

Apart from a set of crystallographic resolution protein structures we also tested our procedure on a group of 10, NMR-derived protein structures of low resolution. The results (see Figure 6) proved that the procedure could be also applied to the data of low accuracy. Therefore, it could be applied to the transformation of the experimentally derived, uncertain dihedral angles into the pseudoangles.

As it was mentioned before, experimental NMR data for  $\phi$  and  $\psi$  angles are not accurate. The minimal error is of the  $10\text{--}15^\circ$  range.<sup>[6,15,16]</sup> We estimated the corresponding  $\gamma$  and  $\theta$  errors using the procedure similar to the one used for the  $\gamma$  and  $\theta$  angles evaluations. First, an expanded conformation of the main chain, four residues long, was built. Then, the values of the  $\phi_2, \phi_3, \psi_2, \psi_3$  angles were randomly perturbed

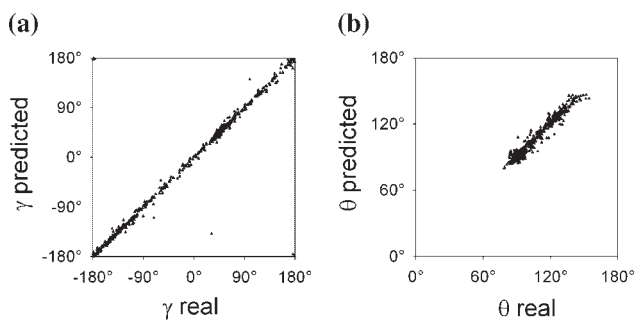


Figure 6. Results of the transformation procedure performed for 10 low-resolution protein structures (1aiw, 1bw5, 1d3z, 1imq, 1jrt, 1mkn, 2cnp, 1ns1, 2spz, 2dvh). (a) The prediction of the  $\gamma$  angles. (b) The prediction of  $\theta$  angles. The number of totally wrong prediction (predictions which exceed the  $\pm 20^\circ$  range for  $\theta$  angle and the  $\pm 60^\circ$  range for  $\gamma$  angle) is below 1%.

many times in the  $\pm 10^\circ$  range and the new values of  $\gamma_2$  and  $\theta_2$  were calculated each time. The obtained error ranges are equal to  $\pm 20^\circ$  for the  $\gamma$  angle and  $\pm 5^\circ$  for the  $\theta$  angle, respectively. Thus, instead of defining the four-peptide conformation by the values of four dihedral angles, each with at least  $\pm 10^\circ$  error range, we could use only three ( $\gamma_2, \theta_2, \theta_3$ ) angles, with the error ranges:  $\pm 20^\circ, \pm 5^\circ, \pm 5^\circ$ . These calculations have shown that significant errors of  $\phi$  and  $\psi$  angles lead to relatively small errors of corresponding  $\gamma$  and  $\theta$  values. Thus, the  $\gamma$  and  $\theta$  angles more precisely define the protein conformation than  $\phi$  and  $\psi$  dihedral angles.

The resulting values of the  $\gamma$  and  $\theta$  angles should not exceed the evaluated error bars. Otherwise, the new restraints would not exactly agree with the experimental dihedral angles. This is the reason why the Levitt's approximation seems to be suitable only for obtaining the  $\gamma$  angle. However, our transformation procedure could be used for both of the pseudoangles (see the error ranges in Figure 5).

The width of the angular restraints potential well characterized by the optimized values of the  $\Delta\xi_{\max}$  parameters (see the Model and Method) reflects the ambiguity of the  $\gamma$  and  $\theta$  angles resulting from the experimental errors and the transformation procedure.

### Simulation Results

The test set employed in the series of folding simulations consists of ten proteins of various secondary structure and topology and with size ranging from 45 to 146 residues. For each protein the complete data set was used (the single  $\theta$  angle and the  $\gamma$  angle per each residue with the exception of the last constrained residue for which only the  $\theta$  angle is defined). However, in NMR study of protein structures not all angles can be determined from simple experiments—typically ca. 15% angles remain undefined. This would decrease somewhat the accuracy of predicted structures. This issue will be studied in details in the forthcoming work. Here, we would like to provide a general insight into the role of the short-range restraints for *in silico* protein folding. For this reason, we present in detail only the results of the simulations with complete data sets, although the effect of incompleteness of the restraints on the results is demonstrated on a few examples (see Table 3).

The final results of the isothermal REMC simulations of all three series of simulations: without any restraints, with the secondary structure predicted by the PSIPRED server, and finally with both types of restraints (the predicted secondary structure and the short-range, angular restraints) are presented in Table 2. Introducing the bias towards the predicted secondary structure noticeably enhance the prediction in a few cases, but still insufficiently. Only applying both kinds of the restraints significantly improves the results.

For small proteins (<80 residues) implementation of restraints decreases the lowest cRMSD value by  $1.58 \text{ \AA}$

Table 2. The results of the REMC simulations.

PDB id	Length	Type	Without any restraints		With the secondary structure		With the secondary structure and angular restraints	
2gb1	56	$\alpha + \beta$	6.75 <sup>a)</sup>	7.41 <sup>b)</sup>	3.49 <sup>a)</sup>	6.39 <sup>b)</sup>	0.95 <sup>a)</sup>	1.07 <sup>b)</sup>
1ubq	76	$\alpha + \beta$	10.70	12.41	10.46	12.16	1.82	2.33
5nll	138	$\alpha/\beta$	13.27	14.98	13.64	16.07	9.28	14.61
2spz	58	$\alpha$	3.28	5.70	3.01	4.46	1.70	3.48
1bw5_A	66	$\alpha$	6.77	12.96	10.43	11.47	1.61	4.46
1imq	86	$\alpha$	9.73	11.90	7.76	11.06	2.86	5.09
5mba	146	$\alpha$	13.88	15.52	9.12	10.80	4.82	5.05
1ed7	45	$\beta$	6.76	8.86	5.73	7.75	1.15	1.55
2pcy	99	$\beta$	10.71	12.02	8.73	11.93	3.33	3.60
1a3k	137	$\beta$	11.67	12.51	11.01	11.12	8.61	11.44

<sup>a)</sup> The lowest cRMSD (Å) from the crystallographic structures obtained in the REMC simulations.

<sup>b)</sup> The cRMSD (Å) of the representative structure of the best cluster (the largest of all clusters).

(2spz)–8.88 Å (1ubq) as compared with the predictions without any restraints. As it is shown in Figure 7, applying short-range restraints systematically improves the global orientation of the three helices in the 2spz model. The quality of the most likely model of 2gb1 structure (see Figure 8) is even comparable with the accuracy of the experimentally determined models.

For larger proteins, the structure prediction also improves. The predicted models have correctly assigned secondary structure and the global fold, although the overall topology is slightly distorted, mainly in irregular regions (see Figure 9). Only the 1a3k and 5nll models have wrongly predicted global folds.

The CABS model used in this study has built-in the tendency of a protein to fold into regular secondary structure likewise the previously developed SICH0 lattice model.<sup>[17]</sup> As a result, fragments of beta sheets and helices agree better with the native structure in all three series of simulations, while loops and undefined regions are less accurate. For example, inaccuracy of the prediction of the 1bw5 structure (Figure 10) is observed mainly in N-terminus and C-terminus regions, both representing the coil

structure. On the contrary, the helical core of the protein is well predicted as the cRMSD for residues 10–50 is only 1.85 Å, while for the whole chain it is equal to 4.46 Å. Moreover, proteins with a substantial fraction of the chain classified as loops or coil seem to fold less accurately. For instance, 5mba and 1imq are both  $\alpha$  proteins, but the content of helical regions in the 5mba structure is larger (72%) than in the 1imq (only 48%), according to the PDB database. The most likely models of both protein structures have the similar cRMSD value of the best predicted model obtained in the folding simulations with short-range restraints, despite the fact that 5mba, nearly twice as large as 1imq, should fold less accurately.<sup>[17]</sup>

The prediction of the 1a3k and 5nll structures does not improve noticeably while applying restraints. Both of these proteins have complex topology with a large number of building blocks. 5nll is  $\alpha/\beta$  protein, with five beta strands forming a sheet surrounded by five helices and 1a3k is composed of two beta sheets containing five and six strands, respectively. Implementation of some additional long-range distance restraints would certainly improve predictions of these two protein structures. However, such restraints are not used in this work. Here, we examined the effect of applying the short-range restraints alone on the quality of the predicted structures.

The results of the simulations with incomplete set of restraints are presented in Table 3. We used about 85% of all restraints, which were randomly chosen. Remaining 15% of the restraints were located mainly in coil and loop regions, as it is more difficult to obtain precise experimental data for the less regular secondary structure fragments. As it is demonstrated in Table 3 the increase of the cRMSD values of the obtained structures (due to the incomplete restraints) is observed in all cases, although the effect is significant only in the case of 1ubq (3.03 Å). The secondary structure assignment and the overall topology are correctly reproduced in all cases. As expected, the main difference is

Table 3. Results of REMC simulations with incomplete (85%) set of restraints.

PDB id	With the secondary structure and 85% of angular restraints	
2gb1	1.06 <sup>a)</sup>	1.25 <sup>b)</sup>
1ubq	3.08	5.36
1bw5_A	4.21	4.68
2pcy	4.37	5.44

<sup>a)</sup> The lowest cRMSD (Å) from the crystallographic structures obtained in the REMC simulations.

<sup>b)</sup> The cRMSD (Å) of the representative structure of the best cluster (the largest of all clusters).

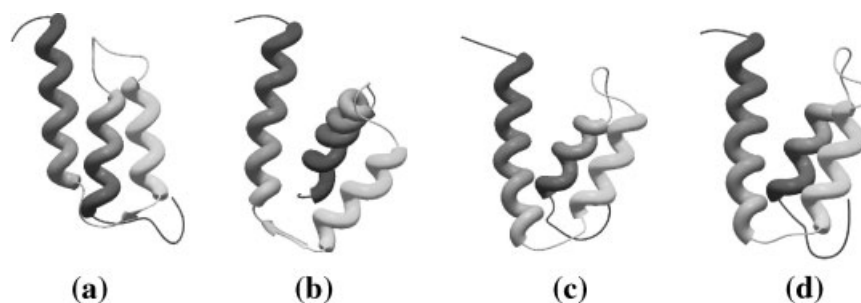


Figure 7. Prediction of the 2spz structure. The structure prediction systematically improves while applying short-range restraints. (a) The native structure. (b) The best model predicted without any short-range restraints (5.70 Å from the native structure). (c) The best model obtained using only the secondary structure information (cRMSD = 4.46 Å). (d) The best model, which was predicted using both types of short-range restraints: the secondary structure assignment and the angular restraints (cRMSD = 3.48 Å).

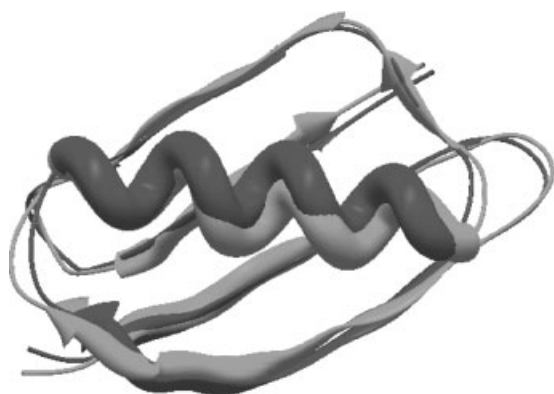


Figure 8. Comparison of the model obtained in simulations with both types of short-range restraints (light gray) and the native structure of protein G (2gb1) (dark gray). The cRMSD value is equal to 1.07 Å. Slight distortions are observed in loops regions, while extended and helical fragments are well superimposed. For the sake of clarity only the alpha carbon trace is shown.

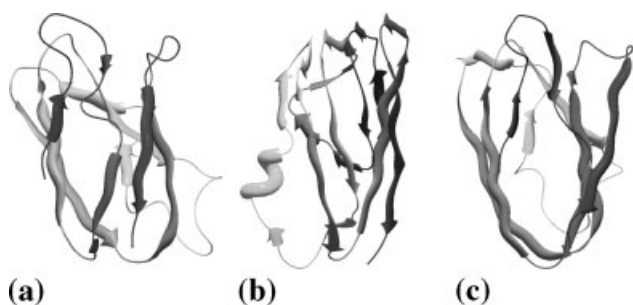


Figure 9. Comparison of the predicted models and the crystallographic structure of the apoplastocyanin (2pcy). (a) The native structure. (b) The best model obtained in the simulations without any restraints. The folding algorithm predicted more fragments of the regular secondary structure than it is observed in the native structure. Consequently, loops and coils, which are more difficult to predict, are distorted. (c) The representative structure of the best cluster obtained using both types of the local restraints (cRMSD = 3.60 Å).

observed in the irregular regions, which were less constrained (see Figure 11). In the forthcoming work it will be tested what is the minimal set of the short-range orientational restraints, which has still noticeable impact on the protein folding simulations.

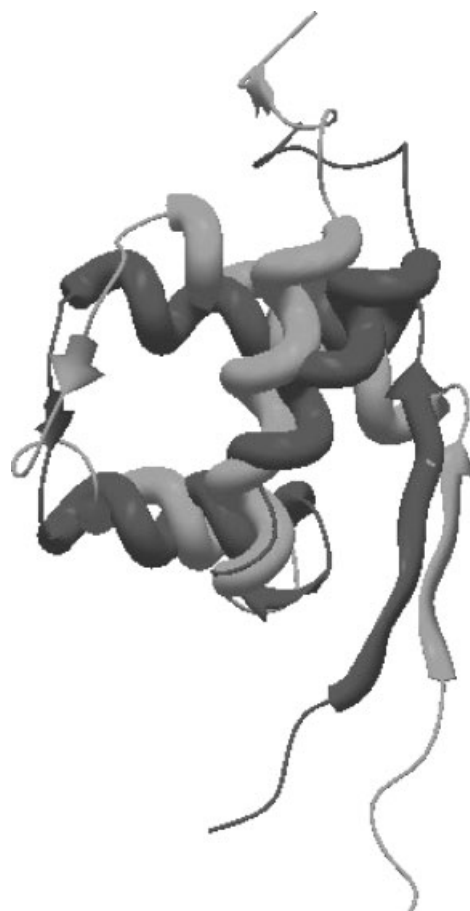


Figure 10. Superimposition of the predicted (light gray) and the native structure (dark gray) of 1bw5. The N-terminal and C-terminal coil regions are poorly predicted while the helical core is well reconstructed.

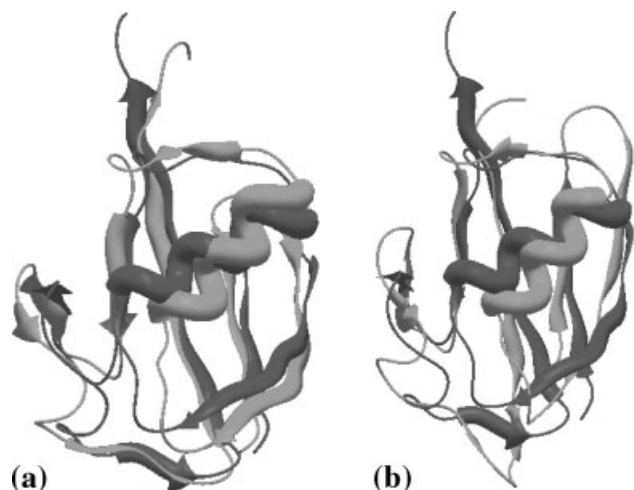


Figure 11. Superimposition of the native structure (dark gray) and two final models of Iubq (light gray): (a) the result of the simulation with complete set of restraints, (b) the result of the simulation with 85% of all restraints. As it can be seen in both figures, the overall topology of the protein is maintained. However, decreasing the number of applied restraints leads to somewhat worse prediction of the loops, and consequently it caused slight inaccuracies in the orientation of the regular structure elements.

## Conclusion

De novo determination of the protein structure within the resolution that allows study of the protein function often requires implementation in the folding algorithm some distance or angular restraints derived either from experimental data or from the structures of homologous proteins.<sup>[11]</sup> Introducing even soft conformational biases derived from theoretically predicted secondary structure improves reproducibility and average accuracy of the folding simulations. Implementing additional short-range, orientational restraints based on torsion angles improves accuracy of the prediction more significantly, not only in the case of the local geometry but also the overall topology of the models. What is perhaps more important, small and topologically simple protein structures could be predicted with the accuracy approaching the accuracy of experimentally derived structures. Nevertheless, implementing long-range restraints is inevitable for a precise prediction of larger proteins with more complex tertiary structure. The

work in progress is aimed on implementation of additional information about the global orientation of protein building blocks, which could be applied in the form of distance and/or orientational restraints. Such restraints could be obtained with much less effort than the complete experimental determination of protein structures.<sup>[18–20]</sup>

*Acknowledgements:* This work was partially supported by the Polish Committee for Scientific Research (KBN) Grant No. PBZ-KBN-088/P04/2003. All protein snapshots were prepared with the help of the Biodesigner program (freely available via the following homepage: <http://www.pirx.com/biodesigner>). Computational part of this work was done using the computer cluster at the Computing Center of Faculty of Chemistry, Warsaw University.

- [1] D. T. Jones, *J. Mol. Biol.* **1999**, *292*, 195.
- [2] C. Branden, J. Tooze, "Introduction to Protein Structure", Garland Publishing, Inc., New York, London 1991, p. 8.
- [3] I. Bahar, M. Kaplan, R. L. Jernigan, *Proteins Struct. Funct. Genet.* **1997**, *29*, 292.
- [4] A. Bax, *Protein Sci.* **2003**, *12*, 1.
- [5] J. P. Priestle, *J. Appl. Cryst.* **2003**, *36*, 34.
- [6] G. Cornilescu, F. Delaglio, A. Bax, *J. Biomol. NMR* **1999**, *13*, 289.
- [7] G. N. Ramachandran, V. Sassiakharan, *Adv. Protein Chem.* **1968**, *28*, 283.
- [8] A. Kolinski, *Acta Biochim. Pol.* **2004**, *51*, 349.
- [9] D. Gront, A. Kolinski, *Bioinformatics* **2005**, *21*, 3179.
- [10] M. Feig, P. Rotkiewicz, A. Kolinski, J. Skolnick, C. L. Brooks, *Proteins Struct. Funct. Genet.* **2000**, *41*, 86.
- [11] A. Kolinski, J. Skolnick, *Polymer* **2004**, *45*, 511.
- [12] T. J. Oldfield, R. E. Hubbard, *Proteins Struct. Funct. Genet.* **1994**, *18*, 324.
- [13] M. Levitt, *J. Mol. Biol.* **1976**, *104*, 59.
- [14] W. Kabsch, C. Sander, *Biopolymers* **1983**, *22*, 2577.
- [15] J. Hu, A. Bax, *J. Am. Chem. Soc.* **1997**, *119*, 6360.
- [16] P. V. Bower, N. Oyler, M. A. Mehta, J. R. Long, P. S. Stayton, G. P. Drobny, *J. Am. Chem. Soc.* **1999**, *121*, 8373.
- [17] A. Kolinski, J. Skolnick, *Proteins Struct. Funct. Genet.* **1998**, *32*, 475.
- [18] J. W. Back, L. de Jong, A. O. Muijsers, C. G. de Koster, *J. Mol. Biol.* **2003**, *331*, 303.
- [19] T. Haliloglu, A. Kolinski, J. Skolnick, *Biopolymers* **2003**, *70*, 545.
- [20] C. A. Rohl, D. Baker, *J. Am. Chem. Soc.* **2002**, *124*, 2723.