Pol. J. Chem., **75**, 587–599 (2001)

Structure of Proteins: New Approach to Molecular Modeling*

by A. Kolinski[1,2], P. Rotkiewicz[1] and J. Skolnick[2]**

[1]*Faculty of Chemistry, Warsaw University, ul. Pasteura 1, 02-923 Warsaw, Poland*
E-mail: kolinski@chem.uw.edu.pl
[2]*Laboratory of Computational Genetics, Department of Molecular Biology, The Scripps Research Institute, Miami, USA*

(Received December 15th, 2000)

The force field and Monte Carlo sampling method of our recently developed reduced model of proteins is described. Recent applications of the models include *ab initio* structure prediction for small globular proteins, modeling of protein structure based on distantly homologous (or analogous) structural templates, assembly of protein structure from sparse experimental data, and computational studies of protein folding dynamics and thermodynamics. The newest application, described in this paper, enables the prediction of low-to-moderate resolution coordinates of the parts of protein structure that are missed in incomplete PDB files.

Key words: protein folding, lattice protein models, comparative modeling, loop modeling, Monte Carlo simulations, protein structure prediction

In this genomic era, there is an urgent need to annotate the structure and function [1] of the thousands of new protein sequences that have arisen from the DNA sequencing of various organisms, including humans [2]. Knowledge of protein structures, functions, and the nature of protein–protein interactions will result in the elucidation of known metabolic pathways and the discovery of new pathways. As a consequence of our better understanding of immunological responses, it will be easier to discern and implement effective mechanisms for drug delivery, genetically modify plants to enhance their nutritional value and resist viruses, and develop new, less invasive ways of treating of animal and human diseases while advancing biology-based technologies.

For a fraction of cases (probably about 50% of the newly discovered proteins) sequence comparison alone [3] can teach us a lot about the structure and function of a new protein [4]. When a new protein is sequentially similar to another protein of known structure and function, it can consequently be considered strongly evolutionarily related. This is the domain of traditional bioinformatics and comparative modeling [4–7]. The number of new protein structures is growing much faster than the number of experimentally determined protein structures (by methods such as crystal-

* Dedicated to Prof. Jan Stecki on the occasion of his 70th birthday.
** To whom correspondence should be addressed.

lography and NMR) in spite of an enormous effort to increase the speed of large-scale structure determination [2]. Thus, the necessity to be able to theoretically predict protein structures from their sequences of amino acids becomes paramount [1].

Traditional tools of molecular modeling [8,9] can now simulate short-time (and short conformational distance) protein dynamics, on the order of tens to hundreds of nanoseconds of the real time. The time needed for a protein to fold into its unique native state from a random denatured state is much longer, and ranges from milliseconds to minutes [10], depending on protein size, solvent conditions, and other factors. This is true for both *i vivF* and *i vitF* protein structure assembly (in the rest of this paper the word "structure" will be used synonymously with the term "three-dimensional protein structure," meaning tertiary or quaternary). Consequently, with the exception of small peptides, large structural changes, including the protein folding process leading to structure prediction, can not be simulated by means of molecular dynamics, Monte Carlo methods, or other conformational sampling tools on the level of the detailed atomic representation of these systems.

In order to facilitate the study of protein structure dynamics and thermodynamics, numerous attempts were undertaken to simplify the problem by reducing the number of degrees of conformational freedom treated in an explicit way [11–29]. Some of these protein models use an alpha carbon trace (a virtual chain connecting alpha carbon atoms) to mimic the conformation of the main chain. This simplification can be rationalized by the fact that the peptide bonds are relatively rigid molecular fragments and they are the same for all amino acids (except the relatively rare case of the cis-proline conformation). To further speed up the simulation process, the conformational space is frequently discretized, either by discretization of the valence angles and dihedral angles of the alpha carbon trace or (more frequently) by assuming a lattice representation of the model chain [11]. Some reduced models neglect the protein side chains, while other models assume a single sphere or a multiple united atom representation of the side chains. The internal degrees of freedom of the side chains can be also treated at different levels of generalization.

As a result of various levels of simplification, the emerging models differ qualitatively in their potential applicability to the side group representation.

In this contribution, we describe a new approach to low and moderate resolution modeling of protein dynamics, thermodynamics, and structure prediction. Then we outline various applications, including the study of protein folding pathways and thermodynamics [15], *ab i itF* folding of small proteins [20], the assembly of multimeric proteins, the prediction of protein structures from sparse experimental data [13], and application of the model with an extension of comparative modeling to remotely related pairs of proteins [30]. Yet another new application is described here in more detail. Namely, a fraction of solved protein structures are incomplete; they miss substantial parts of the structure for various reasons – usually because it is not visible in X ray data or because the chain tracking during the data processing was incomplete in the protein structure prediction.

Before we do so, let us recount the states of the field. The simplest possible model of protein-like copolymers is most likely the cubic lattice chain with two types of residues mimicking two types of amino acids, polar and non-polar hydrophobic [31]. This model was studied in great detail and it probably reproduces some of the most general properties of water-soluble proteins. On the other hand, some very basic features of proteins are completely neglected in such simple models. For instance, the important interplay between local conformational propensities (resulting in the extreme stiffness of polypeptides and in the formation of secondary structure) and long-range interactions is neglected in simple lattice models [11]. Consequently, the application of very simple models in protein structure prediction is rather problematic.

Recently, we proposed a qualitatively different approach to the reduced modeling of protein structure and dynamics. It combines the simplicity of representation with a relatively accurate model of protein packing. Instead of modeling the main chain in an explicit way, we adopted the side chain representation with an implicit (not simulated in a straightforward way) representation of the backbone. The model protein is confined to a lattice and the virtual chain (with fluctuating bond length) connects the centers of mass of the side chains in their actual rotameric state [12,13]. Such an approach has some potential advantages with respect to main-chain based models. First, it is commonly accepted that the specificity of the intra-protein (as well as inter-protein) interactions that determine protein three-dimensional structure is encoded in side chain interactions. The interactions between the main chain units are rather generic and sequence independent. Second, having the positions of the side chains, the reconstruction of the main chain constitutes a very simple and well-defined task. On the contrary, rebuilding the side chains having just the alpha carbon trace involves complicated and expensive optimization of the packing. The model chain connecting the centers of mass of the side groups is, however, less regular and the potentials controlling this chain have to be designed in a somewhat more elaborate way. The benefit is that there is a single degree of conformational freedom per amino acid that describes a convoluted motion of the main chain unit and the corresponding side group, including the internal flexibility of the side chains. The proposed model is applied here to rebuild and optimize these fragments using the known part of the structure as a scaffold for the assembly of the complete structure. To estimate the plausibility of obtained models, we also perform a test experiment on known structures from which we removed parts of the chain and then rebuilt it for comparison with the crystallographic data. While real "blind" predictions await experimental verification, such a test procedure should provide some measure of the precision and accuracy of the modeling protocol.

Methods

Protein representation. The conformations of model polypeptides are represented by strings of virtual bonds connecting the interaction centers that correspond to the center of mass of the side chains, including the α-carbons [12,13]. For instance, the

center of glycine coincides with its C_α, the center of alanine is located in the middle of the C_α-C_β bond, the center of valine coincides with the position of the C_β atom of the side group, *etc*. For the larger side chains that possess internal degrees of freedom, the interaction centers correspond to the center of mass (all heavy atoms are treated as being the same) of the actual rotamer. These interaction centers (beads) are confined to the underlying cubic lattice with a lattice spacing of 1.45 Å. The lattice spacing parameter defines the spatial resolution of the model. The virtual bonds resulting from such a projection are of various lengths, depending on the identity of the two successive amino acids, the main chain conformation, and the actual rotameric state of the side chain. In proteins, the distances between two such defined residues have a quite broad distribution, ranging from 3.8 Å between a pair of glycines to about 10 Å for some pairs of large side chains in their anti-parallel orientation and expanded conformations. The corresponding set of lattice vectors covers this distribution with good accuracy. The shortest vectors are in the form of (±2, ±2, ±1) or (±3,0,0) vectors, including all possible permutations of the coordinates corresponding to a distance of 4.35 Å in protein structures. The longest lattice vectors are of the (±5, ±2, ±1) type and their length corresponds to 7.94 Å; thus, the wings of the distribution are arbitrarily cut off. The number of observed extreme distances is small and neglecting them should not have any significant affect on the model accuracy. The set of the allowed "bonds" consists of 646 vectors. To mimic a part of the hard core of the chain, each residue occupies a cluster of the lattice points of the underlying simple cubic lattice. Each cluster consists of 19 lattice points: the central one, six points at the positions (±1,0,0), (0, ±1,0) and (0,0,±1) with respect to the central one, and twelve points at the positions type of (±1, ±1,0). The distance of the closest approach (3 lattice units, *i.e.*, 1.45 Å) of two clusters nicely corresponds to the smallest values of the inter-residue distances in real proteins. The number of possible orientations between the contacting clusters is equal to 30 (vectors type (±2, ±2, ±1) and (±3,0,0) between their centers). Since the average "contact distances" between the side groups in folded proteins are somewhat larger than the distance of the closest approach, there are much more than 30 spatial orientations of two residues in contact. Consequently, the lattice anisotropy effects are negligible. All PDB protein structures could be represented with an average root mean square accuracy, (coordinate root-mean-square deviation from the crystallographic coordinates) RMSD, of about 0.8 Å.

Interaction scheme. The force field for the protein model described above has been explained in detail in our recent publications [30,32]. Here we limit ourselves to a concise outline of various contributions to the interaction scheme. First there are generic potentials that do not depend on the sequence of amino acids. These are designed to bias the model chain towards protein-like local conformational stiffness and protein-like packing of residues that are in contact (but separated along the chain). Thus there is a bias towards either very expanded (β-type) or compact (helix, or turn-type) conformations [33] of the four residue fragments. Indeed, in proteins the distribution of the corresponding distances are bimodal, reflecting the effect of secondary structure. Long-range (interactions between the residues separated along the

chain but close in space) packing correlations are enforced by a model of the main chain hydrogen bond network. The orientational effect of the hydrogen bonds is translated into equivalent correlations between alpha carbon contacts (the positions of the alpha carbons are estimated from the shape of the model chain and the average distances between the center of interactions for a given type of amino acids and their alpha carbons). The hydrogen bond scheme is made explicitly cooperative. There is an additional energy gain for propagation of protein-like patterns. When taken alone, these generic interactions will fold the model chain into a compact structure with fluctuating (and not structurally unique) secondary structure (helices and β-sheets).

Sequence specific interactions are modeled by knowledge-based potentials of mean force extracted from the statistical correlations seen in the database of known protein structures. For example, a contact potential for two amino acids of a given type could be calculated as – log ("number of observed contacts"/"number of randomly expected contacts"). Other statistical potentials can be derived in a similar way. For the short-range interactions, there are four potentials controlling the distances between residues n and m = n + k (with k = 1,2,3, and 4) with a given identity for the flanking residues n and m. The potential between the n-th and n + 3th residue has a chiral character; left-handed and right handed conformations are treated separately [12]. For the long-range interactions [32,34], there are pairwise potentials of mean force (square well, contact type), and multibody potentials simulating the hydrophobic effect, and thereby the effect of solvent in an averaged implicit fashion.

Sampling method. The Replica Exchange Monte Carlo (REM) sampling method [35] is used to search for the lowest energy conformation, which should correspond to the native (or near-native) state of the model protein. As demonstrated recently [36] the REM method is superior to the classical simulated annealing or generalized ensemble sampling techniques. During the REM simulations a number of copies (replicas) of the system are simulated at various temperatures spanning the range between a temperature above the folding transition and a temperature below the folding transition. The neighboring (according to the temperature) replicas are occasionally compared and exchanged, according to a Metropolis type criterion. The replica exchange process allows the copies that are trapped in local minima of the energy landscape to go to the higher temperature, where they surmount the barriers easily. The sampling technique for each replica between the exchange events is a standard asymmetric Metropolis scheme with a proper weighting of the different states. The sampling employs a set of small local perturbations of the system conformations. The trial moves involve one residue, two residues, or three model residues and are controlled by a pseudorandom mechanism.

Previous applications of the model. The applications include *ab initio* folding, folding with a small number of experimental restraints [13], the refinement of threading models [30], and the study of protein folding kinetics and thermodynamics [15].

The present status of the force field associated with the described lattice model enables the *ab initio* folding of some small and structurally simple single domain proteins [20]. The efficiency of the folding algorithm increases when the statistical

pairwise potential and the short-range potentials are enhanced by a weighting procedure based on the sequence similarity of protein fragments [32]. Due to the possibility of becoming trapped in local minima of the very complex conformational energy landscape, the yield of the correctly folded three-dimensional structures is never equal to 100%. Nevertheless, for a fraction of small proteins, the proper clustering procedure is capable of identifying the correct fold.

When some restraints obtained from various experiments (NMR, fluorescence, crosslink experiments, *etc.*) are available, then the applicability of the model to structure prediction increases significantly [13]. Proteins up to 250 residues could be assembled with as few as N/7 long-range restraints (known side chain contacts). The accuracy of the obtained structures depends on the number of restraints and the size of the protein and ranges from 2 Å to 6 Å. This could be a very useful tool for speeding up the process of structure determination from NMR data. At the beginning, very few signals can be used for building the low to moderate resolution model. Subsequently, such a model may aid in the identification of other (usually very convoluted) signals.

The restraints (approximate in this case) can be also derived in a theoretical fashion, *via* a so-called correlated mutation analysis [17], or by extraction of consensus contacts in a threading procedure [32]. Since these restraints are never exact, the obtained structures are usually of a lower resolution.

When attempting structure predictions for new proteins, three different situations may emerge. The simplest is the situation when the protein of interest (the target protein) is highly homologous (sequence identity of 35% or more) to another protein of already known structure (template). Since during evolution the three-dimensional structure of proteins is more strongly conserved than the sequence, this level of sequence similarity means high similarity of the structures of the target and template proteins. Thus, the template protein can be used as a scaffold for modeling the target structure by the standard tools of comparative modeling. For very distant homologs or when two unrelated evolutionary proteins have similar folds, it is sometimes possible to identify their similarity by a so-called threading [37] (or inverse folding) procedure where the query sequence is threaded throughout the known structures and appropriate scoring functions (usually knowledge-based potentials) are used to rank-order the structure-sequence compatibility. When a high compatibility is detected, then the resulting template can be used to build an approximate model of the target (query) protein [38]. When neither the sequence methods nor the threading methods are successful one must rely on less certain *ab initio* approaches [1,34].

The threading methods usually lead to rather poor molecular models [38–41]. The template usually differs significantly from the true structure of the target proteins. Moreover, the alignment of the query sequence on the structure of the template is often very far from the structurally optimal alignment. The standard tools of comparative modeling [6,7] almost never improve the starting threading models [38,39]. The obtained model structures are closer to the structure of the template than to the structure of the target proteins. Deviations from the template are essentially random instead of being directed into the target structure. In a recent paper we demonstrated

that the proposed lattice model can be successfully used for restrained folding in the spatial vicinity of the threading models. This sometimes led to qualitative improvement of the model accuracy [30]. Consequently, this "generalized comparative modeling approach" is expected to qualitatively extend the possibility of structural and functional annotations of new protein sequences obtained from sequencing the genomes of various organisms. Applications are now in progress on a genomic scale [1].

Completing incomplete protein structures. In this paper we describe and test a new application of the lattice modeling tool described above. For various reasons, a fraction of the structures deposited in the Protein Data Bank [42] (PDB) is incomplete. Coordinates of parts of the polypeptide chains are left undefined. Such a situation may be related to the crystallization problems, to difficulties in the chain tracking procedure after the X-ray experiments, or to a combination of these and other factors. Alternately, some of these cases may actually reflect physics – the "missed" fragment could be more structurally mobile than the rest of the molecule. Although, even in such a situation the more mobile fragment should have a single preferred conformation or a small number of energetically plausible conformations. Complete structures can be very useful for the subsequent study of ligand docking, protein-protein interactions, protein redesign, and so on [33].

Completing protein structures is somewhat similar to the generalized comparative modeling outlined in the previous section. The incomplete structure is used as a template. The model chain is fit to the known part of the structure and the fragment of unknown structure (usually on the surface of the protein) is generated in a random fashion. This provides the starting conformation that is subject to the subsequent search for the energy minima by means of the Replica Exchange Monte Carlo algorithm. During the optimization procedure, these residues that belong to the known part of the structure are kept in close proximity to their crystallographic coordinates. The optimized part moves without any restrictions except those related to excluded volume and other interactions.

The simulations were performed for two sets of proteins. The first is the training set. For five complete protein structures we ignored coordinates of a part of the chain and compared the resulting optimized structures with the crystallographic coordinates. This provides a measure of the accuracy of our algorithm. Then we performed a similar procedure on a set of proteins that have incomplete coordinates of the folded chain. These are "blind" predictions that await experimental verification. Recently developed procedures allow the rebuilding of all atomic details from coordinates of the reduced lattice models.

Results

First, the test simulations were done for a small set of globular proteins of various size. Fragments of various length were removed from these structures and treated as unknown. The obtained results are compiled in Table 1. There are two aspects of the

quality of the rebuilt fragments. First, is the fidelity of the fragment itself, which is measured by RMSD (root-mean-square deviation) from the native structure of the fragment after the best superimposition of the modeled piece. Second, is the fidelity of the location of the "docked" fragment in respect to the entire structure. The data given in Table 1show that the structure of the fragment itself (see Fig. 1) is reproduced better than its location in the entire structure (see Figs. 2–3). Figures 2–3 show examples of the predicted fragment location (in gray) with respect to the entire structure of the protein (in black). Here the accuracy is somewhat worse. Nevertheless, the low-resolution structure was reproduced properly for most of the test cases.

Table 1. Compilation of the simulation data.

	PROTEIN LENGTH	GAP LENGTH	GAPPED RESIDUES	FRAG RMSD SEPARATED	FRAG RMSD IN PROTEIN*
BENCHMARK					
1ctf_	68	11	28–38	1.06	2.32
3cd4_	178	17	128–144	2.51	3.64
1fts_	295	25	133–157	4.30	7.43
2azaA	129	33	50–82	5.74	6.51
1ubq_	76	15	24–38	2.77	5.52
BLIND PREDICTIONS					
1ax8_	146	13	46–59		
1dekB	241	10	200–209		
1dhs_	344	17	76–92		
1maz_	221	53	28–80		

* When entire structure is superimposed (all values of RMSD in Angstroms).

In the next stage, a blind prediction was done for proteins, for which indeed the fragment of structure is unknown. The completed structures are shown in Figs. 4–7, where the gray lines correspond to the known parts of structures, while the bold gray lines correspond to the reconstructed fragments. The obtained models await experimental verification. The model coordinates in the PDB format could be read from our homepage (http://biocomp.chem.uw.edu.pl).

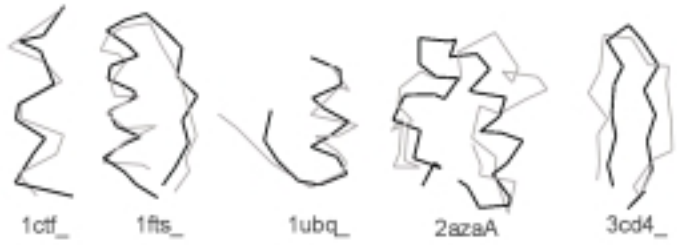

Figure 1. Schematic drawing of the alpha carbon traces of the rebuilt fragments (in gray) superimposed onto the corresponding fragments of the crystallographic structures for five test proteins. See Table 1 for details.

Figure 2. An example of fragment rebuilding for the ribosomal protein 1ctf. The black line corresponds to the alpha carbon trace of the crystallographic structure, the gray line the reconstructed fragment.

Figure 3. An example of fragment rebuilding for the N-terminal domain of the T-cell surface glycoprotein 3cd4. The black line corresponds to the alpha carbon trace of the crystallographic structure, the gray line the reconstructed fragment.

Figure 4. Alpha carbon trace for the 1ax8 structure. The bold fragment corresponds to the predicted part of the structure. See Table 1 for details.

Figure 5. Alpha carbon trace for the 1dekB structure. The bold fragment corresponds to the predicted part of the structure.

Figure 6. Alpha carbon trace for the 1dhs structure. The bold fragment corresponds to the predicted part of the structure.

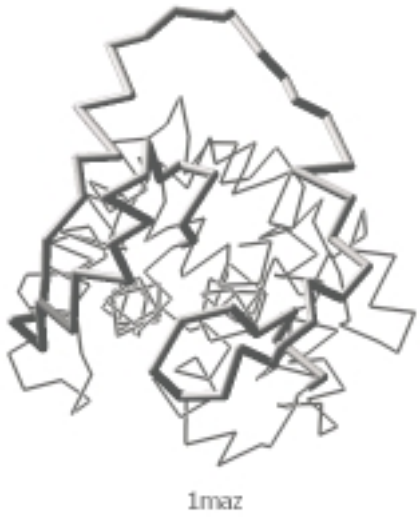

Figure 7. Alpha carbon trace for the 1maz structure. The bold fragment corresponds to the predicted part of the structure.

Conclusions

In this paper we described a new approach to the modeling of protein structure dynamics and thermodynamics. The lattice model developed during the last two years has been already applied to a variety of problems. In particular, the new method proved to be a very efficient tool for assembly protein structure from sparse experi-

mental data, to *ab initio* prediction of three dimensional structure of small proteins, to building more accurate models from crude threading-based models and for study of dynamics and thermodynamics of protein folding.

A newest application of the method, described in this work, enables a completion of incomplete protein structure. In a number of protein structures deposited in the Protein Data Bank, parts of structures are missed for various reasons. The lattice modeling tool can be applied to the fast reconstruction of the approximate coordinates of missed elements of structure. The procedure consists of several steps. First, a chain building algorithm generates lattice approximation of the protein of interest. Then, an approximate conformation of the missed fragment (or fragments) is built by a random mechanism. The main part of the procedure, is the folding simulation using the Replica Exchange Monte Carlo technique, where the known parts of the structure are kept very close to their crystallographic coordinates and the missed part moves freely, subject to chain connectivity and the force field of the model. The method was tuned and then tested on a set of known (and complete) structures from which parts of the structures were removed and then rebuilt by the procedure outlined above. The removed parts were always located on the protein surface, since this is a typical situation for incomplete structures. Then the structure completion was done for proteins with gaps in the structure. These predictions await experimental confirmation.

REFERENCES

1. Skolnick J., Fetrow J.S. and Kolinski A., *Nature Biotech.*, **18**, 283 (2000).
2. Montelione G.T. and Anderson S., *Nature Struct. Biol.*, **6**, 11 (1999).
3. Altschul S.F., Madden T.L., Schaefer A.A., Zhang J., Zhang Z., Miller W. and Lipman D.J., *Nucleic Acid Res.*, **25**, 3389 (1997).
4. Clark M.S., *Bioessays*, **21**, 121 (1999).
5. Rastan S. and Beeley L., *Curr. Opin. Genet. Devel.*, **7**, 777 (1997).
6. Sali A., MODELLER, A program for protein structure modeling by satisfaction of spatial restraints (http://guitar.rockefeller.edu/modeller/modeller.html)
7. Sali A., Overington J.P., Johnson M.S. and Blundell T.L., *TIBS*, **15**, 235 (1990).
8. Brooks C.L.I., Karplus M. and Pettitt B.M., Proteins: A theoretical perspective of dynamics structure and thermodynamics, Wiley, N Y, 1988.
9. Brooks C.L.I., *Curr. Opin. Struct. Biol.*, **3**, 92 (1993).
10. Creighton T.E., Proteins: structures and molecular properties, W.H. Freeman and Company, N Y, 1993.
11. Kolinski A. and Skolnick J., Lattice models of protein folding, dynamics and thermodynamics, R.G. Landes, Austin, TX, 1996.
12. Kolinski A., Jaroszewski L., Rotkiewicz P. and Skolnick J., *J. Phys. Chem.*, **102**, 4628 (1998).
13. Kolinski A. and Skolnick J., *Polymer*, **32**, 475 (1998).
14. Kolinski A., Rotkiewicz P. and Skolnick J., Application of high coordination lattice model in protein structure prediction. in *Monte Carlo Approach to Biopolymers and Protein Folding*, P. Grassberger G.T. Barkema and W. Nadler, Eds., World Scientific, Singapore/London, 1998.
15. Kolinski A., Ilkowski B. and Skolnick J., *Biophys. J.*, **77**, 2942 (1999).
16. Levitt M., *Current Opin. Struct. Biol.*, **1**, 224 (1991).
17. Ortiz A.R., Kolinski A., Rotkiewicz P., Ilkowski B. and Skolnick J., *Proteins Suppl.*, **3**, 117 (1999).
18. Hinds D. and Levitt M., *J. Mol. Biol.*, **243**, 668 (1994).
19. Hoffmann D. and Knapp E.W., *Phys. Rev. E*, **53**, 4221 (1996).

20. Kolinski A., Rotkiewicz P., Ilkowski B. and Skolnick J., *Progress Toward the Prediction of Tertiary Structure of Small Globular Proteins*, **138**, 292 (2000).
21. Liwo A., Kazimierkiewicz R., Czaplewski C., Groth M., Oldziej S., Wawak R.J., Rackovsky S., Pinkus M.R. and Scheraga H.A., *J. Comput. Chem.*, **19**, 259 (1988).
22. Osguthorpe D.J., *Proteins Struct.*, **3**, 186 (1999).
23. Simons K.T., Bonneau R., Ruczinski I. and Baker D., *Proteins Struct.*, **3**, 171 (1999).
24. Sun S., *Protein Sci.*, **2**, 762 (1993).
25. Sun Z., Xia X., Guo Q. and Xu D., *J. Protein Chem.*, **18**, 39 (1999).
26. Vieth M., Kolinski A., Brooks III C.L. and Skolnick J., *J. Mol. Biol.*, **237**, 361 (1994).
27. Vieth M., Kolinski A., Brooks III C.L. and Skolnick J., *J. Mol. Biol.*, **251**, 448 (1995).
28. Vieth M., Kolinski A. and Skolnick J., *Biochem.*, **35**, 955 (1996).
29. Samudrala R., Xia H., Huang E. and Levitt M., *Proteins Struct.*, **3**, 194 (1999).
30. Kolinski A., Rotkiewicz P., Ilkowski B. and Skolnick J., *Proteins*, **37**, 592 (1999).
31. Dill K.A., Bromberg S., Yue K., Fiebig K.M., Yee D.P., Thomas P.D. and Chan H.S., *Protein Sci.*, **4**, 561 (1995).
32. Skolnick J., Kolinski A. and Ortiz A.R., *Proteins*, **38**, 3 (2000).
33. Branden C. and Tooze J., Introduction to protein structure, Garland Publishing, Inc., NY and London, 1991.
34. Skolnick J. and Kolinski A., Protein Modelling. in *Encyclopedia of Computational Chemistry*, Vol. 3, N.L.A.P.V. Schleyer T. Clark, J. Gasteiger P.A. Kollman H.F. Schaefer III and P.R. Shreiner, Eds., John Wiley & Sons, Chichester, U.K., 1997.
35. Swendsen R.H. and Wang J.S., *Phys. Rev. Lett.*, **57**, 2607 (1986).
36. Gront D., Kolinski A. and Skolnick J., *J. Chem. Phys.*, **113**, 5065 (2000).
37. Godzik A., Skolnick J. and Kolinski A., *J. Mol. Biol.*, **227**, 227 (1992).
38. Jaroszewski L., Rychlewski L., Zhang B. and Godzik A., *Protein Sci.*, **7**, 1431 (1998).
39. Jones D.T., *J. Mol. Biol.*, **287**, 797 (1999).
40. Madej T., Gibrat J.F. and Bryant S.H., *Proteins*, **23**, 356 (1995).
41. Miller R.T., Jones D.T. and Thornton J.M., *FASEB*, **10**, 171 (1996).
42. Bernstein F.C., Koetzle T.F., Williams G.J.B., Meyer Jr, E.F., Brice M.D., Rodgers J.R., Kennard O., Simanouchi T. and Tasumi M., *J. Mol. Biol.*, **112**, 535 (1977).