Protein Folding: Flexible Lattice Models

Andrzej Kolinski,[1,2,*)] Piotr Rotkiewicz,[1] Bartosz Ilkowski[1]
and Jeffrey Skolnick[2]

[1]*Department of Chemistry, University of Warsaw*
ul.Pasteura 1, 02-093 Warsaw, Poland
[2]*Donald Danforth Plant Science Center*
7425 Forsyth Boulevard, Box 1098, St. Louis, Missouri 63105, USA

(Received October 11, 1999)

In the post genomic era a possibility of theoretical prediction of protein structure from sequence of amino acids is one of the most important and challenging goals of molecular biology. High complexity of the problem requires simplification of molecular models and very efficient computational tools. Proposed here model of protein structure, dynamics and interaction scheme assumes a single interaction center per amino acid residue. This highly simplified representation is supplemented by a number of build-in implicit packing rules that enable a reasonable modeling of protein geometry that is compatible with detailed atomic models. Preliminary applications to *ab initio* protein folding and distant homology comparative modeling are described and discussed.

§1. Introduction

Systematic sequencing of entire genomes of numerous organisms provides enormous volume of protein sequences.[1)] Only a small fraction of these proteins have three-dimensional structure solved by crystallographic or NMR techniques.[2)] Somewhat larger fraction (15-20%, depending on genome) of new proteins have a close homologue in the database of already known protein structures.[3)] For these proteins molecular models could be build using standard tools of comparative modeling.[4)] Quality of such models depends on the level of sequence similarity between the query and template proteins. When the sequence similarity drops below 25-30 % (depending on protein size) the quality of models obtained from comparative modeling decreases rapidly. New methods of comparative modeling that could generate reasonable models using templates that are structurally quite different from the true structure of query protein need to be developed.[5)]

Homology modeling approach is limited to these proteins that are structurally close to one or more of already known protein structures. As the number of experimentally solved structures increases, the applicability of the comparative modeling techniques will also increase. Interestingly, majority of the newly solved structures has examples of similar structures solved previously. As a result, a substantial fraction (probably more that 50%, however there are various estimates of the number) of possible protein folds have no examples in the structural databases. Thus, the *ab initio* structure prediction attempts are very important. New types of protein folds are most likely to be identified by an *ab initio* approach.

*) E-mail address: kolinski@chem.uw.edu.pl

Recently, we have developed an efficient lattice model of protein structure and dynamics.[6)-8)] The model polypeptide chain is confined to the simple cubic lattice with the lattice spacing equal to 1.45 Å. The chain beads correspond to centers of mass of side chains (including alpha carbons). Thus, the length of virtual bonds connecting the chain beads can assume a wide range of values, accommodating various sizes of amino acids, different conformations of the main chain and different rotational isometric states of the side chains. Consequently, the emphasis of the model is on the side chain packing instead of geometry of the main chain.[9)] It is probably reasonable to assume that the most specific interactions that define protein structure are between the side groups. The main chain is treated in an implicit way as a derivative of the side chain positions.

Previously, this model was applied in study of protein dynamics[6)] and thermodynamics,[10)] assembly of protein structure from sparse experimental data,[7)] in refinement of threading based protein models[11)] and in *ab initio* protein structure predictions. The *ab initio* applications were tested in the framework of CASP3 assessment of protein structure prediction approaches.[12)] Here, we describe a refined version of this methodology for *ab initio* structure prediction and for distant homology comparative modeling. Possible implications for a genomic scale protein structure prediction are briefly discussed.

§2. Lattice protein model

2.1. *Protein representation*

Each model protein unit, comprising the side chain atoms and the alpha carbon atom, is represented by a cluster of 19 points of the underlying cubic lattice. This cluster constitutes a hard core of a residue. For larger amino acids it is supplemented by a soft-core repulsive envelope. There are also soft repulsive interactions for some mutual orientations of the spatially close side groups which accommodate the main chain volume in an implicit way. The model is schematically depicted in Fig. 1. The excluded volume clusters associated with the interaction units are shown in Fig. 2. Coordination number for the closest approach of two clusters is equal to 30. Since the average packing distance is somewhat larger than the cluster diameter local lattice anisotropy could be safely neglected.

To accommodate the distribution of distances between consecutive side chains the length of virtual bonds of the chain is allowed to fluctuate between 4.35 Å (lattice vectors type $[\pm 2, \pm 2, \pm 1]$ or $[\pm 3, 0, 0]$) and 7.95 Å (lattice vectors type $[\pm 5, \pm 2, \pm 1]$. Thus the wings of the distribution have been cut-off. The corresponding error is below the accuracy of the lattice representation. The set of allowed bond orientations consists of 646 vectors. Consequently, in spite of its simplicity, the discretization of the protein conformational space is very flexible.

2.2. *Model of dynamics and sampling scheme*

Stochastic dynamics of the model chain consists of long series of randomly selected local rearrangements of the chain conformation. Examples of such local moves

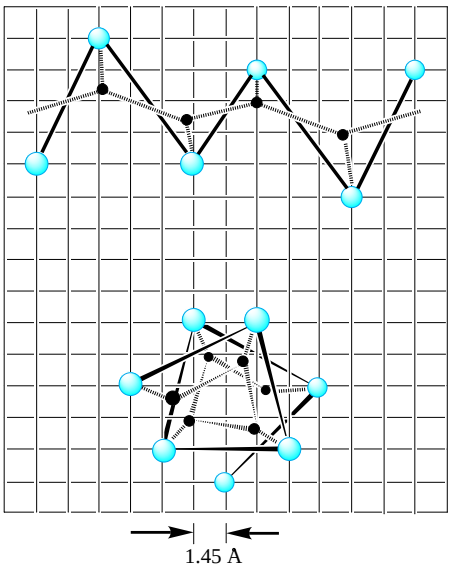

Fig. 1. Schematic illustration of the lattice model design. The larger gray spheres indicate the lattice points that correspond to the center of mass of the model side chains. The solid thick lines are the virtual bonds of the model chain. The broken lines indicate the main chain in alpha carbon (solid dots) representation. The main chain is treated in an implicit way; the approximate Cα positions are computed from the coordinates of three consecutive side chains. The upper part of the drawing shows an extended conformation and the lower part a helical fragment. The spacing of the underlying cubic lattice is equal to 1.45 Å. As a result pdb structures can be represented with the average accuracy of 0.7-0.8 Å in respect to Cα or side chain positions.

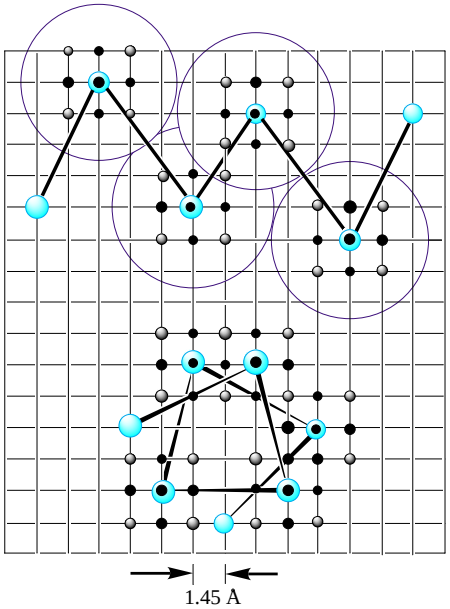

Fig. 2. Illustration of the model chain excluded volume. A cluster of 19 lattice points represents each side chain. The black dots indicate three lattice points along the Z-axis (orthogonal to the drawing plane), the gray dots indicate single points of the same z coordinate as the coordinate of the central point of the cluster. Large open spheres in the upper part of the figure correspond to the volume excluded for other unit's center of interaction.

are given in Fig. 3. Asymmetric Metropolis scheme was employed in the Monte Carlo algorithm. The transition probabilities were calculated according to the force field designed for this model.

2.3. *Interaction scheme*

The force field of the model consists of three types of contributions. Interactions of the first type are designed to mimic generic (sequence independent) structural regularities seen in globular proteins and to compensate for excessive flexibility of the model chains. Several potentials were introduced in order to achieve this goal. In proteins the distribution of the distance between i-th and i+4th side chains (a

similar effect exists for alpha carbon distances) is bimodal. The low value peak corresponds to helical and tight turn conformations while the longer distance, more diffused, peak comes from expanded, β-type conformations. To mimic this effect a bias is introduced that favors energetically such conformations. To model the hydrogen bond interactions between the main chain units a bias was introduced towards mutual orientations of interacting chain fragments that are characteristic for regular elements of secondary structure (helices or β-sheets). Additionally, a generic packing cooperativity potential provides secondary structure propagation effect. Namely, the system gains an energy price when a series of side chain contacts typical for proteins occurs. The pattern is of the same type for helices and sheets. When residues i and j are in contact it is more likely to see contacts between the residues $i \pm 1$ and $j \pm 1$ (the order is fixed for helices, while for β-sheets it depends on their topology). Such generic scheme of interactions induces protein-like chain stiffness and cooperativity of packing interactions. This leads to spontaneous formations of short ordered fragments of secondary structure upon the chain collapse.

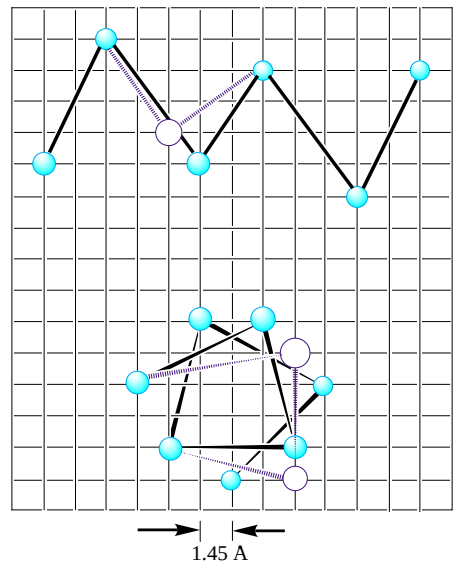

Fig. 3. Examples of local micro modifications of the model chain. Two-bond move is shown in the upper part and three-bonds transition is shown in the lower part of the figure. The chain ends need to be treated in a separate fashion.

Sequence specific interactions consist of short range, long range pairwise and multibody interactions that simulate the hydrophobic effects. The short range statistical potentials are pairwise dependent and were derived from statistics of distances between various pairs of amino acids, separated by one, two, three and four virtual bonds along the chain. For i, i+3 distance the chirality (or handness) was accounted for in the potential. Long range pairwise interactions were also derived from statistical analysis of known protein structures. The pairwise potentials are orientation dependent (in respect to the angle between the vectors calculated as difference of two consecutive chain bond vectors); separate interactions tables were derived for parallel, antiparallel and acute/orthogonal orientations of contacting residues. For instance, parallel contacts are typical for residues belonging to two strands of the same β-sheets, while contacts between two sheets tend to be antiparallel. Residues of opposite charges are strongly attractive in a parallel orientation, while repulsive or inert in antiparallel orientations. Indeed, charged residues in globular proteins are located mainly on protein surface and (locally) their point in the same direction.

Multibody, hydrophobic interactions account for particular tendencies of various amino acids to have a specific number of parallel, antiparallel and acute contacts,

regardless of identity of their partners. Detailed description and numerical values of these potentials may be found in our previous publications. [7), 8), 10), 11)]

Specificity of the short range and the long range pairwise interactions could be enhanced by weighting the statistics by sequence similarity (to the query sequence) of corresponding fragments of polypeptide chains from the structural database. Such local-homology-enhanced potentials have to be derived separately for each protein. [15)] They are not only sequence-specific, but also protein-specific (the same pairs of amino acids in various sequence contexts may interact differently).

§3. *Ab initio* prediction of protein structure

Described above computational model was tested in *ab initio* folding of several small single domain proteins. The following procedure was applied to each test case:

1. Random conformation of the model chain of a proper length was generated.
2. Short Monte Carlo simulation at a very high temperature was performed to ensure sufficient randomization of the starting conformation.
3. Large number (200-300) of folding simulations (simulated annealing from $T = 2.0$ to $T = 1.0$) were executed.
4. The resulting structures (few snapshots from each trajectory at the final temperature) were clustered using energy filter (only the small fraction of the lowest energy conformations was taken into the clustering procedure) and pairwise rmsd (root-mean-square distance between the obtained structures).
5. When a well-defined cluster was found, the average structure was calculated and compared to crystallographic structure.

Since our clustering procedure is now in a preliminary form, here we report only a comparison between the structures obtained from simulations with the crystallographic structures. Representative results are given in Table I. Figure 4 shows a series of selected snapshots from a successful folding trajectory of ribosomal protein 1ctf. In spite of quite complex topology the final structure is 2.6 Å from the native structure (drms for alpha carbons). Such high accuracy is rare, however for some proteins (1ctf is a good example) such good quality final structures cluster nicely, and could be separated from the sea of misfolded structures.

Table I. Percentage of the successful folding experiments for a representative set of small proteins.

Protein	type	length	rmsd<6Å	rmsd<4.5Å
1gb1	α/β	56	16%	5%
3icb	α	75	34%	8%
1ctf	$\alpha+\beta$	68	11%	7%
1hmd	α	113	5%	3%
6pti	$\alpha+\beta$	58	6%	0%
1c5a	α	73	36%	13%
1tlk	β	103	4%	1%
1cis	$\alpha+\beta$	66	4%	0%
1shg	β	62	5%	0%
2pcy	β	99	1%	0%

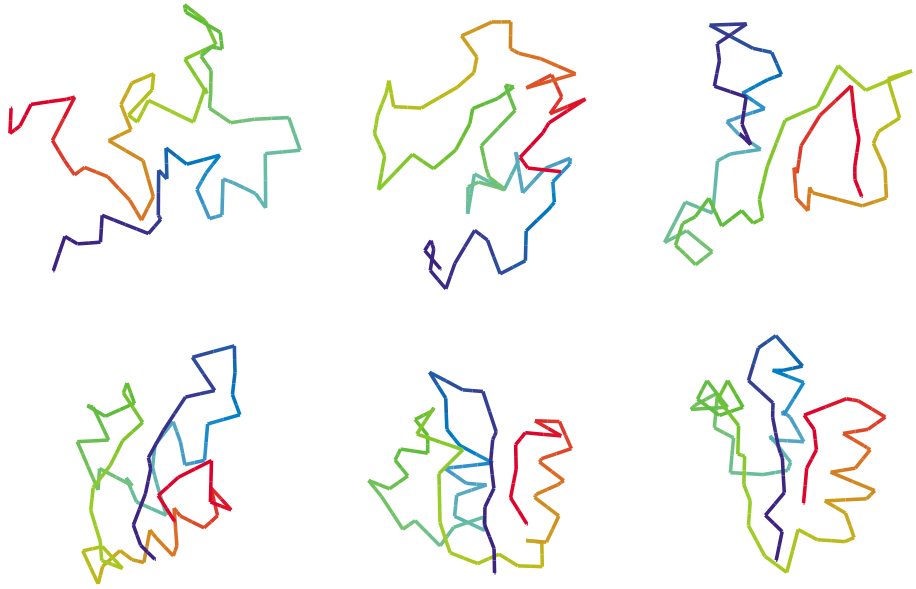

Fig. 4. A series of snapshots from the simulated annealing folding trajectory of ribosomal protein 1ctf. For clarity, only the Cα trace (estimated from the positions of the side groups) is shown in each snapshot.

These preliminary results show that it is possible to predict low-resolution structure of a fraction of small single domain proteins. The results are always far from random, even in the worst case of 2 pcy. However, in order to make a genuine prediction the cluster of near-native structures has to be well defined in respect to more uniformly distributed misfolded structures. This is clearly the case in about half of the investigated examples.

Helical proteins are easier to predict, β-proteins are more difficult due to more complex topology. On the other hand, for simpler topologies, one needs higher reproducibility of simulations than for complex ones. That is because for simple and very small proteins the number of alternative structures of comparable energy is relatively small. Thus, in some cases, the folds of wrong topology may form also relatively well defined clusters. In most cases they could be rejected due to a higher average value of conformational energy.

These results are very encouraging, however further work is needed to improve prediction accuracy and fidelity. Performing a larger number of longer simulations (especially for larger proteins) should be beneficial. Also, as mentioned before, a better method of selecting the native-like structures from these that have been trapped in local energy minima needs to be developed.

§4. Lattice simulations in homology modeling

Using sequence alignment methods or threading methods it is possible to detect even remotely related proteins. A match between a query sequence and a protein of known structure enables building a plausible molecular model. When sequence iden-

tity is large (say more than 30 %) the situation is relatively simple and conventional methods of comparative modeling usually can provide a good quality model of the query protein. However, it is possible to detect some level of similarity between proteins that have much lower sequence identity. In such cases, models resulting from conventional automated modeling are typically much more similar to the structure of the template than to the true structure of the target (query) protein. For these poor quality threading-based models the lattice simulations can be quite useful. The following modeling procedure (we omit some less relevant details) has been proposed and tested on a set of proteins.

1. A structural template was detected for the query protein. This was done by one of the standard threading methods.[14)]
2. Standard automated comparative modeling was performed (we used MODELLER[16)]) providing molecular model of the test protein.
3. The initial model was projected onto the lattice, with appropriate restriction build-in into the reduced modeling tool described in this paper.
4. Monte Carlo simulations were performed, using the initial model as a weak target.
5. The lowest energy lattice structure was selected and used as a template for standard comparative modeling.

The results are shown in Table II. In most cases the application of the lattice simulations improved the models obtained from the automated comparative modeling. Only in two cases the final models are marginally worse. At the same time, in about half of cases the improvement was large. In four cases the initial 6-8 Å models were replaced by 4 Å models. This is a qualitative change that could be crucial for identification of protein function from the analysis of the obtained molecular models.[13), 17), 18)] Figure 5 shows an example of the structures obtained from automated modeling and *via* method employing lattice Monte Carlo simulations as an

Table II. Comparison of the results of an automated comparative modeling and the results obtained using lattice simulations near the template structure. The numbers in the last two columns denote rmsd in Å from the native structure for Cα atoms.

Target protein	length	Template protein	Threading +MODELLER	Threading-MODELLER -Lattice-MODELLER
1aba_	87	1ego_	4.43	4.86
1bbhA	131	1ccy_	6.77	6.82
1cewI	108	1molA	14.96	14.38
1hom_	68	1lfb_	7.82	3.70
1stfI	98	1molA	6.40	5.95
1tlk_	103	2rhe_	7.23	4.17
256bA	106	1bbhA	6.09	4.36
2azaA	129	1paz_	21.95	10.77
2pcy_	99	2azaA	6.56	4.41
2sarA	96	9rnt_	10.28	7.83
3cd4_	178	2rhe_	6.74	6.39
5fd1_	106	2fxd_	25.67	12.40

Note: The values of the rmsd are given for the structured parts of the target molecules (1hom_ : residues 7-59, 1tlk_ : residues 9-103, 3cd4_ : residues 1-97 i.e., the first domain).

intermediate optimization step.

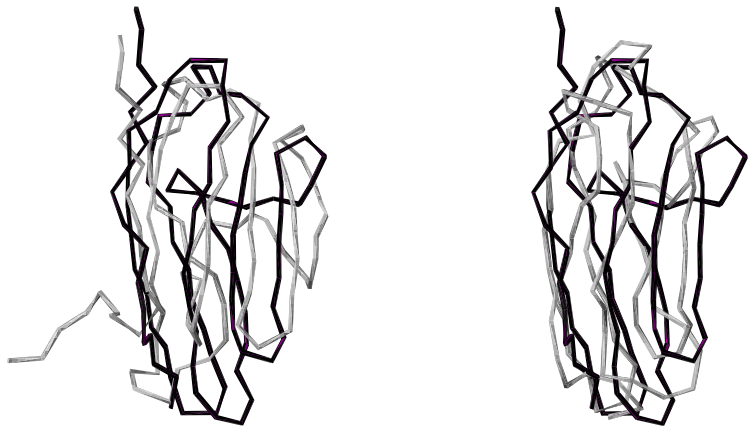

Fig. 5. Molecular models (gray, Cα-trace) of 1tlk superimposed on the crystallographic structure (black). Left side - MODELLER, right side - MODELLER followed by lattice Monte Carlo simulations, followed by all atom reconstruction *via* MODELLER.

§5. Conclusion

In this work we described a reduced model of protein structure and dynamics and some of its applications. The model was employed in the test *ab initio* folding of a number of small globular proteins. In about half of tested cases the fraction of successful folding experiments and the accuracy of obtained models are probably sufficient for low resolution prediction of protein structure. A better method for evaluation of plausibility of obtained models needs to be developed. Work in this direction is now in progress.

The proposed simulation method could be also used for optimization of crude threading-based (or obtained from sensitive sequence comparison methods [19]) models of proteins. This may expand significantly range of applicability of comparative modeling in structure and function prediction. Since the methodology is relatively simple, it could be applied in automated structure prediction procedures (refinement of threading models and *ab initio* folding) for large sets of target proteins, i.e. for entire genomes of simpler organisms. This possibility is now being explored.

References

1) S. Rastan and L. Beeley, Curr. Opin. Genet. Devel. **7** (1997), 777.
2) L. Holm and C. Sander, Science **273** (1996), 595.
3) F. C. Bernstein, T. F. Koetzle, G. J. B. Williams, E. F. Meyer Jr, M. D. Brice, J. R. Rodgers, O. Kennard, T. Simanouchi and M. Tasumi, J. Mol. Biol. **112** (1977), 535.
4) A. Aszodi and W. R. Tylor, Folding & Design **1** (1996), 325.

5) L. Jaroszewski, K. Pawlowski and A. Godzik, J. Molecular Modelling (1998).
6) A. Kolinski, L. Jaroszewski, P. Rotkiewicz and J. Skolnick, J. Phys. Chem. **102** (1998), 4628.
7) A. Kolinski and J. Skolnick, Proteins **32** (1998), 475.
8) A. Kolinski, P. Rotkiewicz and J. Skolnick, in *Monte Carlo Approach to Biopolymers and Protein Folding*, ed. P. Grassberger, G. T. Barkema and W. Nadler (World Scientific, Singapore/London, 1998), p. 100.
9) A. Kolinski and J. Skolnick, *Lattice models of protein folding, dynamics and thermodynamics* (R. G. Landes, Austin, TX, 1996).
10) A. Kolinski, B. Ilkowski and J. Skolnick, Biophys. J. (1999), in press.
11) A. Kolinski, P. Rotkiewicz, B. Ilkowski and J. Skolnick, Proteins (1999), in press.
12) A. R. Ortiz, A. Kolinski, P. Rotkiewicz, B. Ilkowski and J. Skolnick, Proteins (1999), in press.
13) A. Bairoch, *Prosite: A dictionary of protein sites and patterns* (Department de Biochimie Medicale, Universite de Geneva, Geneva, (1990).
14) A. Godzik, J. Skolnick and A. Kolinski, J. Mol. Biol. **227** (1992), 227.
15) J. Skolnick, A. Kolinski and A. R. Ortiz, Proteins (1999), in press.
16) A. Sali, MODELLER. *A program for protein structure modeling by satisfaction of spatial restraints* (http://guitar.rockefeller.edu/modeller/modeller.html).
17) J. Fetrow, A. Godzik and J. Skolnck, J. Mol. Biol. (1998), in press.
18) J. S. Fetrow and J. Skolnick, J. Mol. Biol. (1998), in press.
19) S. F. Altschul, T. L. Madden, A. A. Schaefer, J. Zhang, Z. Zhang, W. Miller and D. J. Lipman, Nucleic Acid Res. **25** (1997), 3389.