Acta Biochimica Polonica

Vol. 44 No. 3/1997

389–422

QUARTERLY

High coordination lattice models of protein structure, dynamics and thermodynamics*

Andrzej Koliński[1]✉ and Jeffrey Skolnick[2]

[1]*Department of Chemistry, University of Warsaw, L. Pasteura 1, 02-093 Warsaw, Poland*
[2]*Department of Molecular Biology, The Scripps Research Institute La Jolla, California 92037, U.S.A.*

Key words: protein folding, protein models, lattice proteins, Monte Carlo method, protein dynamics, protein thermodynamics, protein structure prediction

A high coordination lattice discretization of protein conformational space is described. The model allows discrete representation of polypeptide chains of globular proteins and small macromolecular assemblies with an accuracy comparable to the accuracy of crystallographic structures. Knowledge based force field, that consists of sequence specific short range interactions, cooperative model of hydrogen bond network and tertiary one body, two body and multibody interactions, is outlined and discussed. A model of stochastic dynamics for these protein models is also described. The proposed method enables moderate resolution tertiary structure prediction of simple and small globular proteins. Its applicability in structure prediction increases significantly when evolutionary information is exploited or/and when sparse experimental data are available. The model responds correctly to sequence mutations and could be used at early stages of a computer aided protein design and protein redesign. Computational speed, associated with the discrete structure of the model, enables studies of the long time dynamics of polypeptides and proteins and quite detailed theoretical studies of thermodynamics of nontrivial protein models.

Amino-acid sequence of a globular protein determines its three dimensional structure [1, 2] and thereby its function [3–5]. Due to human genome project [6] (and other studies of the genetic code) the number of known protein sequences grows rapidly and now this number is of a range of a couple of hundred thousand. At the same time the three dimensional structures are known only for a small fraction of protein sequences [7, 8]. The reason is quite simple. Gathering of protein sequence information is relatively simple and experimental procedures could be to a large extent automated [5]. On the contrary, experimental determination of new structures is very expensive and time consuming [9]. At present the number of proteins for which the crystallographic structures are known is smaller than one thousand. Moreover, a majority of known three-dimensional

*A. Koliński acknowledges support of University of Warsaw (Grant BST 34/97) and of The Howard Hughes Medical Institute (International Scholar Grant No. 75195-543402). J. Skolnick was supported by NIH Grants No. GM-37408 and GM-38794.

✉Author and address for correspondence: phone: (+48+22)+822-02-11, ext. 20; fax (+48+22)+822-59-96; e-mail: Kolinski@chem.uw.edu.pl.

Abbreviations: BD, Brownian dynamics; BPTI, bovine pancreatic trypsin inhibitor; ESMC, entropy sampling Monte Carlo method; r.m.s., root mean square; ROP, ColE1 repressor of primer.

structures could be clustered into groups, or families, of very high sequence homology [10]. Assuming that proteins with higher than 30% sequence similarity have usually essentially the same structure the number of different protein folds known at this moment is only about two to five hundreds. At present, a substantial fraction of new protein structures are based on NMR experiments [11–13]. NMR techniques can be employed in many structural studies that are sometimes difficult for standard crystallography, they provide valuable information about protein structure in solution and allow studies of some dynamic aspects of protein structure formation [11, 13, 14]. There are, however, some disadvantages of the NMR based structural studies with respect to protein crystallography. First, only structures of rather small proteins or macromolecular assemblies could be solved using contemporary experimental and theoretical tools. Second, the NMR based structures are on average of a poorer quality than the crystallographic ones. This is due to a different character of the distance restraints provided by the two methods. The restraints from the crystallographic experiments are of a global nature. Namely, approximate positions of all heavy (other than hydrogen) atoms are determined. The NMR based restraints are local, experiments provide only approximate distances between pairs of atoms that are close in space in the protein structure [11]. Consequently, the errors during computational molecular modeling and structure refinement could propagate [15]. It is also known that contemporary tools for molecular modeling can be used only for a refinement of local aspects of protein structure [16–24]. They can not find the native state just on the basis of a subset (or approximations) of molecular interactions encoded in semiempirical force fields employed by the computational models [25]. Thus the quality of the solved structures relies on the quality and completeness of the experimental data [15].

Our present understanding of immunological mechanisms, rational developments of new drugs and design of biotechnological processes is to a large extent limited by insufficient knowledge of protein structure, dynamics and thermodynamics. In the context of the above mentioned gap between the number of known protein sequences and the number of known protein three dimensional structures it is not surprising that the solution to so called protein folding problem is one of the most important objectives of contemporary theoretical molecular biology. On the most general level the solution to the protein folding problem means development of theoretical methods for prediction of protein native structure and folding pathway (or pathways) from a given sequence of amino acids. In spite of numerous attempts only very limited progress has been achieved to date. Why the protein folding problem is so difficult? First, the conformational space of polypeptide chains is enormous. Only due to rotations around the main chain bonds (next to the alpha carbon atoms) there are about five distinct conformations per single residue. Thus a small protein built from hundreds of amino acids can in principle adopt 5^{100} different conformations. Additional degrees of conformational freedom are associated with rotations within the side chains and with configuration of the surrounding solvent. The native conformation is more or less unique. Due to a variety of interactions and because of intervening topological restrictions associated with the chain connectivity a search (by theoretical methods or exercised by proteins in nature) of such conformational space is a very complex task [26, 27]. Indeed, protein folding is a slow process. It takes 10^{-3} s to 10^{2} s to assemble the native conformation [28]. Thus the entire process can not be simulated by standard molecular dynamics (MD) tools. Contemporary computing technology allows MD simulations of a single protein molecule surrounded by a few layers of water that are equivalent to about 10 nanoseconds of the real time [21]. Thus, the simulation time is too short by about six orders of magnitude. Another reason why the prediction of protein structure by means of standard molecular mechanics tools is today not practical is associated with extreme complexity of molecular interactions in proteins. Using standard semiempirical force fields it is very difficult to predict crystal structure formed by small organic molecules. Proteins consist of twenty

various amino acids, some of them having themselves much more complex internal structure than these small organic molecules. Consequently, in order to identify a unique structure of polypeptide chains, the requirements for the applied potentials are much higher [25, 29].

In order to make simulations of the protein folding tractable it is necessary to reduce somehow the number of explicitly treated degrees of freedom and to simplify the functional form of potentials [30–33]. Many reduced models of protein structure and various simulation algorithms have been proposed in the past [31, 33–37]. The majority of these models assumed a united atom representation for entire amino-acid residues [33, 38] or two united atoms per each residue [31, 32, 39]; one for the main chain segment and one for the side chain. Molecular dynamics [19, 32] or Monte Carlo methods [40, 41] have been used as tools of conformational search. Low resolution structures, having some features of the native state, of small globular proteins have been found in these studies [30, 40–42]. Recently, genetic algorithms have been also employed as an energy minimization procedure, allowing quite accurate prediction of low energy structures of some small proteins [43]. Further simplification of the protein conformational space could be achieved by assuming a discrete set of rotational isomeric states. This leads to lattice models of protein chains [44]. Studies employing lattice models can be roughly divided into two categories. The first one deals with very simple lattice models of polymers [45–47] or heteropolymers [48] that have some basic features of polypeptide chains. Such models of protein-like systems can be studied in great detail. In spite of sometimes drastic simplifications, the work by Chan & Dill [47–49], Skolnick *et al.* [34, 50–57], Dill *et al.* [33, 58], Sali *et al.* [38, 59] and others [60–69] provided a valuable general insight into protein folding thermodynamics, folding pathways and possible factors determining uniqueness of the folded state. Lattice models of the second category attempt to mimic specific geometric features of real proteins and employ knowledge based potentials [44]. These models are conceptually closer to the continuous reduced models of protein structure and dynamics. Previous work along this direction will be briefly outlined in the next section.

Lattice discretization of protein geometry has several computational advantages [44]. Local conformational transitions could be enumerated and successively used in very fast Monte Carlo algorithms. Also computations of conformational energy could be considerably speeded up by *in front* calculations of various energy contributions for a discrete set of distances and/or orientations. However, immediately a question arises: does not the lattice representation lead to an unphysical distortion of the protein chain geometry? In this review we will show that it is too a large extent possible to have good peptide-like geometry and yet to exploit all advantages of the lattice approach.

During the last few years we have developed a series of high coordination lattice models of protein conformation and dynamics [36, 44, 70–75]. Good accuracy of representation of protein geometry, which was in the range of accuracy of crystallographic data, has been achieved [73, 76, 77]. Various variants of the Monte Carlo methodology has been used as sampling tools [44, 73, 78]. Force-fields for these models are knowledge-based [71, 79, 80]. Particular potentials are of a statistical origin and have been derived from analysis of structural regularities seen in known protein structures [81, 82]. Dynamics of denatured proteins, pathways of protein folding process and folding transition thermodynamics has been investigated. At present, the model can be used for structure prediction of very simple folding motifs [72, 74, 75, 83–89]. Predictive strength of the method could be considerably increased by implementation of approximate restraints derived from analysis of evolutionary information encoded in protein sequences [90]. Other applications, limitations and possible future developments of this new tool of theoretical molecular biology are also discussed in this contribution.

LOW RESOLUTION MODELS OF PROTEINS AND PROTEIN-LIKE SYSTEMS

Continuous reduced models of proteins

About twenty years ago a number of good resolution protein structures was already large enough to make some generalizations. Repeating structural motifs such as helices and β-sheets were well characterized. Regular character of dense packing of hydrophobic protein interior became obvious. These observations strongly suggested that the rules governing protein folding should be rather robust and that prediction of three dimensional structures of folded proteins should be not so difficult. Because of the large number of degrees of conformational freedom of polypeptide chains, numerous attempts to build simplified models were undertaken. Selection of only some degrees of freedom to be treated in an explicit way seemed to be justified due to the expected robustness of the rules governing the protein folding process.

Work by Levitt & Warshel [30] is now a classical example of such approach. In their model polypeptide chain, representation was reduced to two united atoms per residue. One united atom was centered on the Cα position and represented a segment of the polypeptide main chain, while the second represented the corresponding side chain (except glycine). A constant value of the planar angle between the two consecutive pseudobonds of the alpha carbon trace was assumed. This approximation contradicted rather wide and bimodal distribution of this angle seen in real proteins. In the model the only local degree of rotational freedom was associated with the dihedral angles for the main chain defined in this manner. The torsional potentials for the corresponding degrees of rotational freedom were derived from conformational analysis of several "representative" dipeptides. These were the only short range interactions, reflecting some secondary structure propensities of polypeptide chains. Long range interactions were limited to interactions between the model side chains. Simple semiempirical potential in a form of the Lennard-Jones function has been used as an approximation of these tertiary interactions. The model has been employed in folding simulations of bovine pancreatic trypsin inhibitor (BPTI) polypeptide. Brownian dynamics (BD) has been used as a sampling method. In a majority of simulations, the obtained structures had some features resembling the native fold with the distance root mean square, r.m.s., deviation from the native structure of about 6.5 Å. This demonstrated that even such simplified model has some properties of real proteins. A related study on BPTI has been also performed by Kuntz *et al.* [32] and Hagler & Honig [39] who demonstrated that results of similar quality could be obtained using a sequence code reduced to just two types of amino acids. Later, Wilson & Doniach [41] developed a somewhat similar model. Importance of their work lies in the application of a knowledge based force field. The potentials controlling short range (secondary propensities) and long range (tertiary interactions) were derived from statistical analysis of regularities seen in known crystallographic structures of globular proteins. Sampling procedure was based on a simulated thermal annealing protocol in the framework of the Metropolis type Monte Carlo scheme. Overall accuracy of the predicted structure of crambin was very low. However, the secondary structure was to a large extent in agreement with that of the native protein and elements of protein-like hydrophobic core were formed in the simulation experiments. Also the native-like pattern of cystine crosslinks was observed.

Accuracy of a reduced representation of protein could be improved by taking account of some internal degrees of freedom of protein side chains [91]. For example, larger side chains could be represented by two united atoms [92]. Alternatively, all atom representation of the main chain could be employed with a reduced representation of the side chains [43]. Using this kind of representation and more elaborated statistical potential structures of small proteins, such as melittin, pancreatic polypeptide inhibitor, apamin [43], PPT and PTHrP [92] have been predicted with accuracy from 1.7 Å r.m.s.

(measured for alpha carbon positions) for small single helix melittin to 4.5 Å for larger peptides.

Reduced continuous models were also used in studies of various aspects of folding of real proteins [93, 94] or polypeptides [95] and idealized folding motifs [96–101].

Exploration of conformational space of protein models could be done by various methods. The above mentioned works employed molecular dynamics [102] or its version, the Brownian dynamics [30, 96], Monte Carlo methods [41, 94, 98, 103] and even genetic algorithms [43, 104].

Simple lattice models of protein-like systems and proteins

When employing reduced representation, which is necessary if one wants to investigate entire protein folding, it seems reasonable to discretize the conformational space in order to facilitate a more effective sampling. This leads to lattice models of proteins and protein-like systems. The term protein-like systems means models that do not attempt to reproduce geometry and native structure of specific proteins but rather try to elucidate general rules of protein folding dynamics and thermodynamics.

Our previous Monte Carlo studies of semiflexible diamond lattice homopolymers [45, 46] demonstrated that a specific balance between short range and long range interactions could be responsible for the character of collapse transition being different in polypeptides and in other more flexible (like most of synthetic polymers) polymers. The random coil-globule transition of flexible homopolymers was always continuous [105], while finite length homopolymers of a limited flexibility underwent all-or-none (pseudo first order) transition [45]. In homopolymeric systems the structure of the globular state was not unique [45, 46, 106]. Protein-like uniqueness of the low energy state could be enforced by some differentiation of the short range and the long range interactions along the model chain. This way sequence-dependent secondary structure propensities and some sequence patterns of hydrophobic and hydrophilic residues could be introduced. This enabled modeling of the folding process of various simple protein motifs, including all possible topologies of four-helix bundles, Greek-key motif of β-type proteins and other [34, 50–52, 54–57, 107]. Somewhat more elaborated geometrical representation and potentials were needed to model mixed α/β motifs and to study conditions necessary for the all-or-none folding transition and uniqueness of the folded state [108–113]. These studies of simple protein-like models showed that, in order to reproduce the main features of protein folding thermodynamics and the uniqueness of the globular state, it is necessary to account for some secondary and tertiary preferences for the native conformations. Locally, they do not need to be fully consistent with the native state, however on average such consistency appeared to be necessary. A different point of view has been explored by Dill and coworkers [33, 58], Shakhnovich *et al.* [62–64], Sali *et al.* [38, 59], and others [26, 65, 112, 114–123]. They investigated very simple cubic lattice polymers and heteropolymers, implicitly assuming that just the long range interactions could be used as a folding driving force. Indeed, for some conditions and for some sequences a unique folded state of 27-mer cubic structure (or similar simple compact motifs) could be obtained. Thus, these models seem to reproduce the nature of hydrophobic collapse, neglecting all detailed aspects of secondary structure of globular proteins. Simplicity of such models allowed full exploration of their conformational and, to some extent, sequence space. The results provided essentially an exact description of an extremely reduced picture of the globular protein folding process.

It is possible to study low resolution models of real proteins using simple lattice representation. A classic example is the work done by Go and coworkers [124–127]. Perhaps the most interesting are the results of their simulation study of folding process of lysozyme [128]. Native structure of this 129-residue protein has been represented by a 116 unit simple cubic lattice chain mimicking conformation of the polypeptide main chain, and additional 15 lattice vertices for some larger side chains. Consequently, there was no one-

to-one correspondence between the model and the structural units of the real protein. Nevertheless, the overall geometry and close packing of globular proteins were reproduced with a reasonable accuracy. The goal of their Monte Carlo experiments was to elucidate the factors responsible for the uniqueness of the folded structure and for the cooperativity of the folding process. It has been found that long range interactions, consistent with the target native structure, increase cooperativity of the folding process. Tertiary interactions inconsistent with the target structure decreased folding cooperativity and, when the "wrong" interactions were predominant, the native structure could not be obtained at all. The short range secondary interactions, consistent with the target structure, always decreased folding cooperativity. However, they increased stability of the native structure. Such general picture seems to be in a reasonable agreement with known experimental facts. Simple lattice models of real proteins were also studied by Covell [40], Krigbaum & Lin [129], Dashevskii [130], Covell & Jernigan [131] and Hinds & Levitt [132]. These studies were aimed at prediction of low resolution three dimensional structures of small globular proteins from their sequences of amino acids. Accuracy of the predicted structures was rather low, however of a similar quality as those predicted by the continuous reduced models.

Skolnick & Kolinski [133], and Godzik *et al.* [134] employed moderate resolution "chess knight" models of protein structure in folding simulations of plastocyanin and TIM barrels (the α subunit of tryptophan synthase and triose-phosphate isomerase [134]). It has been shown that using an amino-acid-dependent pairwise interactions and secondary propensities consistent with the native structure, a very fast folding into unique native-like structures of these complex folding motifs could be achieved by means of the Monte Carlo dynamics. Some insight into the protein folding mechanism could be perhaps gained from these studies. The folding process appeared to proceed along a loosely defined pathway, by the on-site mechanism, where already assembled fragments of the native structure served as a folding scaffold for the remainder of the model chain. This way, the folding funnel could be rapidly narrowed as the dense nucleus of globular structure emerged. Due to weak target contributions to the short range interactions these simulations can not be considered as examples of structure prediction, but rather as a plausible theoretical demonstration of a fast folding mechanism.

This short overview shows that it is possible to gain a valuable insight into the nature of the protein folding process by computer studies of reduced models of proteins and idealized protein-like systems. It proved also possible to predict the lowest energy conformations of small polypeptides and small proteins, albeit with rather low overall accuracy. Only few, very simple proteins were investigated, suggesting that the applicability of the proposed methods was not general. Is the problem associated with reduced representation or rather with inadequate potentials? Most likely the answer is that both factors contributed to low predictive power of these studies. As it will be shown later, in order to reproduce protein-like structural regularities it is necessary to use a quite complex set of knowledge based potentials reflecting various interactions that apparently control the unique behavior of globular proteins.

HIGH COORDINATION LATTICE MODELS OF PROTEINS

The models described here base on a high coordination lattice representation of the polypeptide main chain [44]. A set of vectors restricted to an underlying simple cubic lattice is used to represent the alpha carbon trace of protein. The lattice representation of main chain backbone can be then used as a convenient reference frame for very fast building of a reduced representation of protein side chains and very rapid computation of various interactions. Three models of increasing resolutions [76] have been investigated. The chess-knight model is of moderate resolution, however it allows to model various aspects of protein secondary and tertiary structure [44, 70, 108, 109, 133, 134]. The

alpha carbon trace has been approximated by a chain of virtual bonds from the set of vectors of the type |2,1,0|. With all possible permutations of the coordinates and the signs, the number of basis vectors is equal to 24, i.e., there are 24 possible orientations of virtual bonds between successive alpha carbon atoms. The model protein backbone resembles a three dimensional chess-knight path. The best representation of protein structures is obtained when the mesh size of the underlying cubic lattice is assumed to be 1.7 Å. Then the length of Cα–Cα vectors is 3.8 Å, and the accuracy of Cα-trace representation is in the range of 1.0–1.8 Å, depending on protein size. A crude representation of side chains by proper sets of occupied lattice points could be easily defined using the reference frame of the main chain. The chess knight model enables crude representation of helices [70], β-sheets [109] and mixed structural motifs [108].

Geometrical accuracy of the chess knight model, while much better than accuracy of the previously mentioned simple lattice models, is rather moderate. Even more disturbing is its lattice related anisotropy. For small protein motifs the fidelity of lattice approximation depends on the orientation with respect to the Cartesian coordinate system. This may have a dangerous effect on the Monte Carlo dynamics of such model chains. The problem of lattice anisotropy has been to a large extent eliminated in the hybrid 210 lattice model [71]. Here the set of basis vectors consists of 56 entries of the following form {|2,1,0|,... |2,1,1|,... |1,1,1|,...}. Allowing for length fluctuation of the model Cα–Cα pseudobonds the overall accuracy of representation has been improved to 1.0 Å r.m.s. [76], and the dynamics of the model chains became more physical. This model allowed prediction of moderate accuracy three dimensional structures of very simple folding motifs using sequences of amino acids as unique protein-specific information [72].

The model which will be discussed in the remainder of this work employs 90 basic vectors for Cα-trace representation [71]. An immediate question is why not a more exact representation? Certainly, it is easy to design more accurate discretizations of protein conformational space. However, at some point the advantages of lattice approach, i.e. the simplicity and the speed of computations will be lost. Moreover, it appears that at this point the accuracy of representation is not a bottleneck of the model [44]. Further improvements are rather expected from a better design of the interaction schemes. Figure 1 illustrates example fragments of various lattice chains showing increasing fidelity of the main chain representation.

Protein representation

Geometry of alpha carbon trace strictly corresponds to protein secondary structure. Starting from the Cα-coordinates it is relatively easy to define approximate positions of other atoms and groups of atoms [36, 135, 136]. That is a main reason why reduced modeling of protein conformations so frequently employs the alpha carbon chains. In the present model the virtual Cα–Cα bonds are restricted to a set of discrete orientations. The set consists of 90 vectors of the type {|3,1,1|,... |3,1,0|,... |2,2,1|,... |2,2,0|,...}. The best mesh size for the underlying cubic lattice is 1.22 Å. Alpha carbon traces of high resolution crystallographic structure could be fitted to this lattice with an average accuracy of 0.6–0.7 Å of the coordinate r.m.s. deviation after the best superposition [76]. The quality of fit essentially does not depend on the protein size or chain orientation with respect to the lattice coordinate system [76].

The Cα vertices of the model chain serve as interaction centers of the main chain units. The side chains are also represented as single united atoms. However, in order to account for internal rotations of the side chains a multiple rotamer, single sphere, library of the centers of mass of the side chains has been generated basing on analysis of protein crystallographic structures. The number of model rotamers depends on amino acid type and actual conformation of the main chain defined by two backbone vectors. The idea is explained in Fig. 2.

It could be shown that three consecutive Cα vectors define orientation of the central peptide bond plate (*trans* conformation assumed) with rather good accuracy [36, 136].

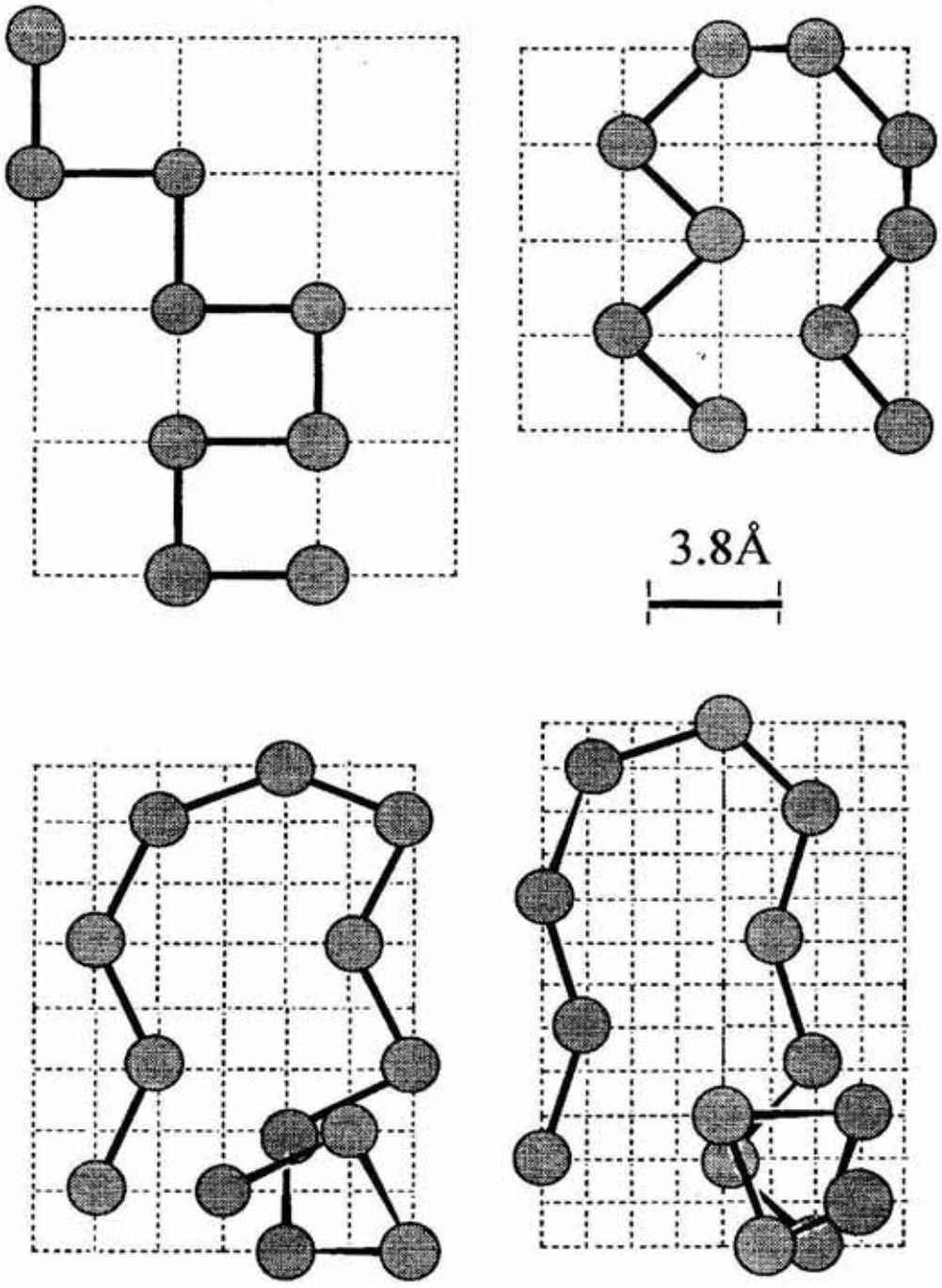

Figure 1. Example conformations of short fragments of lattice chains.

The first panel shows a polymer of a simple cubic lattice. The second one shows a chain on face centered cubic lattice (coordination number 12). The third panel shows a chain on the chess knight lattice. The last panel shows a chain restricted to the 310-hybrid lattice. Increasing resolution is illustrated by decreasing mesh size of the underlying simple cubic lattice, which should be compared to the length of virtual alpha carbon bond.

Thus, the positions of the main chain atoms with respect to the Cα atoms could be defined only once (again the numerical data are derived from the statistics of the known structures) and subsequently used during the simulations in computationally very effective way.

Monte Carlo modeling of protein dynamics

Many years ago Orwoll & Stockmayer [137] have shown that a random sequence of local (few chain units involved) conformational transitions of a lattice chain constitutes a numerical solution to stochastic equation of

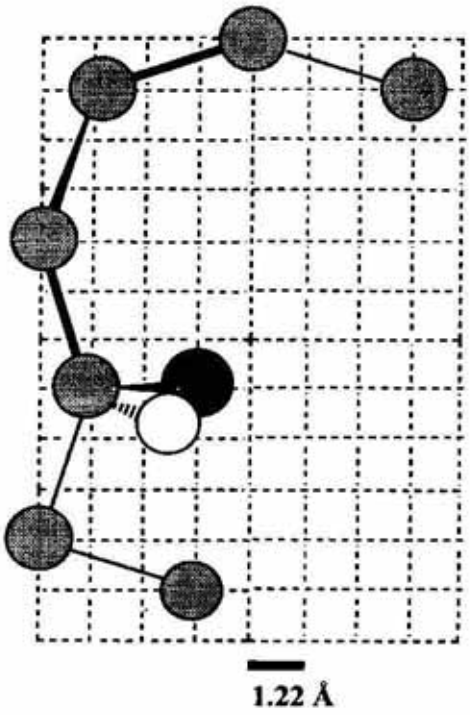

Figure 2. Example conformation of a short fragment of 310-lattice main chain backbone.

Two from several possible rotamers are shown for a selected residue, illustrating rotamer updating, which is one of the elementary conformational transitions incorporated into the Monte Carlo dynamics procedure.

motion. Consequently, a trajectory of a properly defined Monte Carlo process simulates polymer chain dynamics [36, 70, 71, 138–142]. Such dynamics has a well defined physical meaning only for processes that are long in comparison with characteristic time of the local conformational transitions. The following set of local transitions has been designed for the present model.

- (i) Random change of a side chain rotamer for a randomly selected residue.
- (ii) Random rearrangement of two main chain bonds with a proper rearrangement of three side chains involved (chain ends have to be treated in a somewhat different way, the end segments possessing more conformational freedom)
- (iii) Random rearrangement of a three-bond fragment (and four rotamers).
- (iv) Random rearrangement of a four-bond fragment (and five rotamers).
- (v) Random motion of a longer chain fragment by a small distance.

The time unit of the model process is defined as the time required for N (on average due to the random selection of particular moves) attempts to rotamer moves, N-1 attempts to two-bond jumps, N-2 three-bond jumps, N-3 four-bond jumps and a single attempt at the large fragment small-distance displacement. The last type of conformational transitions may slightly distort the time scale and should not be used in simulation when the dynamic aspects is of a major interest. Some examples of the conformational transition of the model chain are schematically shown in Figs. 3–6. Of course, each attempted conformational update is subject to geometry tests and the Metropolis criterion [143].

It has been shown that the proposed model of dynamics leads to Rouse-type dynamics at denatured state [36], a type of polymer dynamics proper for the polymeric systems if hydrodynamic effects could be neglected [144]. At low temperatures the model reproduces qualitatively the dynamic aspects of the molten globule-native state transition [44, 73]. Thus, the model of long time protein dynamics in various conditions is at least qualitatively correct.

FORCE FIELD FOR HIGH COORDINATION LATTICE MODEL

In principle, there are two qualitatively different ways of designing on interaction scheme for reduced models. One possibility is to perform a projection of a detailed, all-atom, force field onto interactions of united atoms in the reduced models. There are, how-

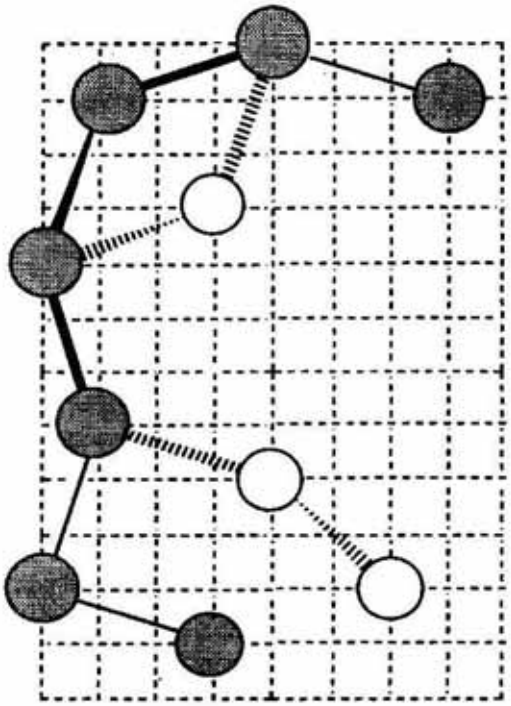

Figure 3. Example of a two-bond conformational transition, and a chain end move.

ever, serious problems with such approach. Reduced models have a different set of degrees of freedom and a different underlying distribution of the conformational entropy between these degrees of freedom (which has to be taken into account). Consequently, such projection is poorly defined [44]. The alternative avenue bases on knowledge based potentials [44, 79, 80, 93, 145]. Assuming that the native structure of a globular protein is a realization of the lowest conformational energy state (i.e. assuming validity of the so called thermodynamic hypothesis [146, 147]) one may compare the interactions in folded structures with the interactions that would exist in random conformations. The former have to be at a minimum of conformational energy. This way a proper statistical analysis of the structural regularities seen in known crystallographic (and NMR) structures [8] leads to a set of semiempirical potentials which reflect various interactions controlling protein stability. Below, a force field resulting from such considerations and analysis of thermodynamics and dynamics of various reduced models is outlined. Details of the derivation and numerical data can be found elsewhere [44, 73, 78]. Short range interactions are the interactions controlling conformational correlations between close (along the chain) peptide units. Abbreviation "long range" means the tertiary interactions between the model united atoms that are usually far away from each other along the chain.

Short range interactions

The first contribution describes rotamer energy [44, 73]. Various conformations of the side chains have various energy and therefore a different thermodynamic probability. The corresponding potential can be defined as follows:

$$E_r = -k_B T \ln(f_{observed}/f_{uniform}) \qquad (1)$$

Where $f_{observed}$ is the frequency observed in the structural database for a given rotamer of a particular amino acid and $f_{uniform}$ is the average frequency of a uniform distribution of rotamers. The corresponding energy parameters are in $k_B T$ units. The distribution of rotamers is a function of main chain geometry.

In a similar way potentials that reflect secondary structure propensities (observed in protein structures with respect to a random distribution) could be derived [44, 73, 78]. We assumed two types of sequence specific short range interactions. There is a contribution from main chain conformation and four con-

tributions from mutual orientations of the side chains:

$$E_{\text{short}} = 4E_b(A_i,A_j,\mathbf{v}_{i-1},\mathbf{v}_i,\mathbf{v}_{i+1}) + \\ + E_{sg}(A_i,A_{i+k},\cos(\mathbf{a}_i,\mathbf{a}_{i+k})) \\ k = 1,2,3,5 \qquad (2)$$

The numerical values of the potential have been derived for coarse grained bins that approximately corresponded to the local geometry of various secondary structure motifs. Six classes of main chain geometry defined by three consecutive Cα vectors ($\mathbf{v}_{i-1},\mathbf{v}_i,\mathbf{v}_{i+1}$), and five ranges of angle between the side chain vectors (vectors from Cα to the center of corresponding side group) $\mathbf{a}_i$ and $\mathbf{a}_{i+k}$ have been considered. The idea is explained in Figs. 7–8.

The sequence specific short range interactions given in Eqn. 2 are defined with respect to an average protein conformation. The distribution of conformations of the model lattice chain without any interaction is, however, not identical to that averaged over all proteins. Therefore two kinds of generic (amino-acid sequence independent) biases have been introduced. The corresponding potentials enforce a bimodal distribution of distances between i-th and $i + 4$th alpha carbons and strong orientational correlations of the peptide bond plates which are typical for all globular proteins [36].

Models of hydrogen bond network

Hydrogen bonds play a very important structure-regularizing role in globular proteins. Energy difference between the protein–water hydrogen bonds and the intraprotein hydrogen bonds is perhaps moderate. However, conformational energy cost for having not hydrogen bonded residues inside protein globule is large.

Structure regularization by hydrogen bonds is even more important in the framework of reduced models. This has not been sufficiently appreciated in previous studies of protein models and protein-like systems. Simplification of conformational space always introduces some additional conformational freedom that does not exactly mimic the conformational space of real polypeptide chains. Consequently, the entropy of model systems could be distorted. In the previous paragraph we have shown that properly designed generic terms can correct for entropy associated with short range geometrical correlations. Hydrogen bonds play a similar role for medium range and short range correlations.

Two distinct models of hydrogen bonds have been designed for the high coordination lattice models of proteins [44]. In spite of quite different design, the two schemes are almost equivalent, although the accuracy of the second scheme appears to be slightly better.

The first model of hydrogen bonds bases on Levitt-Greer method [148] of secondary structure assignment. They have shown that protein secondary structure can be assigned with good accuracy and reproducibility just by using the knowledge of alpha carbon coordinates. We followed this idea [73] building a geometric model of hydrogen bond. Only hydrogen bonds within the main chain were taken into account. Two units i and j of the model polypeptide chain were considered to be "hydrogen-bonded" when the following set of geometrical criteria has been satisfied:

$$R_{\min} < \mathbf{r}_{ij} < R_{\max} \\ |(\mathbf{v}_{i-1} - \mathbf{v}_i) \bullet \mathbf{r}_{ij}| < a_{\max} \\ |(\mathbf{v}_{j-1} - \mathbf{v}_j) \bullet \mathbf{r}_{ij}| < a_{\max} \\ \text{and } |i-j| \geq 2 \qquad (3)$$

where $\mathbf{r}_{ij}$ is the distance between the alpha carbon atoms i and j, and $\mathbf{v}_j$ is the j-th backbone vector connecting the j-th and $j + 1$st alpha carbons. The constants $R_{\min}$, $R_{\max}$ and $a_{\max}$ are equal to 4.6 Å, 7.3 Å and 13.4 Å^2, respectively. This reflects geometrical correlations within helices, within β-sheets and, to a large extent, also the main chain packing preferences in turns and in loop regions. Assignment of the model hydrogen bond instead of the peptide bonds, to the alpha carbons changes registration of the hydrogen bond network. The model bond between i and $i + 3$ alpha carbons is equivalent to the hydrogen bond between i-th and $i + 4$th amino acids in an α-helix. Each hydrogen bond contributes $E^H < 0$ to the total confor-

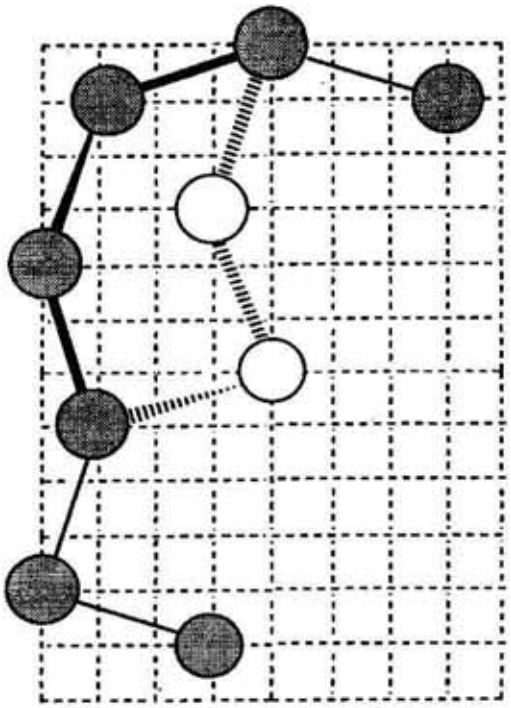

Figure 4. Example of a three-bond conformational transition.

mational energy of the model system. Further regularization of model structures that accounts for known aspects of physics of real proteins could be achieved by making the model hydrogen bond network cooperative in an explicit way. Total energy associated with cooperative network of hydrogen bonds can be then written as [73]:

$$E_{\text{H-bond}} = \Sigma\Sigma E^{\text{H}}\,\delta(i,j) + \\ + \Sigma\Sigma E^{\text{HH}}\,\delta(i,j)\delta(i\pm1,j\pm1) \qquad (4)$$

where $\delta(i,j)$ is equal to one when residues i and j are hydrogen bonded, and zero when they are not bonded. The values of the energy parameters, $E^{\text{H}} = 0.5\ k_{\text{B}}T$ and $E^{\text{HH}} = 0.75\ k_{\text{B}}T$, were selected by trial and error method such that the secondary structure in the denatured states was marginal and the folded states had a well defined network of hydrogen bonds. This scheme reproduces about 90% of the main chain hydrogen bonds assigned to the lattice models of native structures when compared to the classical Kabsch-Sander method [149] executed for the all atom structures.

Another model [36, 44, 78] of hydrogen bonds is somewhat more explicit. As mentioned before, three consecutive alpha carbon backbone vectors define orientation of the central (for this fragment) peptide bond (see Fig. 9). Thus approximate coordinates of the carbonyl oxygens and the amide hydrogens could be calculated and stored as a function of identity of these three backbone vectors. Consequently, Coulomb-like interactions could be computed rapidly.

$$\varepsilon_{\text{H-bond}} = q_{\text{H}}(1 - f_{\text{H}})/(r_{\text{O,H}} + 2 \bullet \\ \bullet \exp(-r^2_{\text{O,H}})) \qquad (5)$$

where: q_{H} is an arbitrary scaling factor, equal to one. The angular factor, f_{H}, reflects the average geometry of protein hydrogen bonds and has been assumed to have the following form:

$$f_{\text{H}} = (0.77 - \cos(\mathbf{r}_{Oj,\text{H}i}, \mathbf{r}_{Oi,\text{H}i}))^2 + \\ + (0.77 - \cos(\mathbf{r}_{Oi,\text{H}j}, \mathbf{r}_{Oj\text{H}j}))^2. \qquad (6)$$

The strength of the model hydrogen bond accommodates also partial charges, local dielectric constant, etc. The indices i and j denote peptide bonds which are numbered sequentially along the chain. The i-th peptide bond is the bond between the i-th and i + 1st alpha carbons. Thus, $\mathbf{r}_{Oi\text{H}j}$ is the vector between the oxygen in peptide bond i and the hydrogen in peptide bond j, and $\mathbf{r}_{Oj\text{H}j}$ is the vector across the j-th peptide bond plate. The

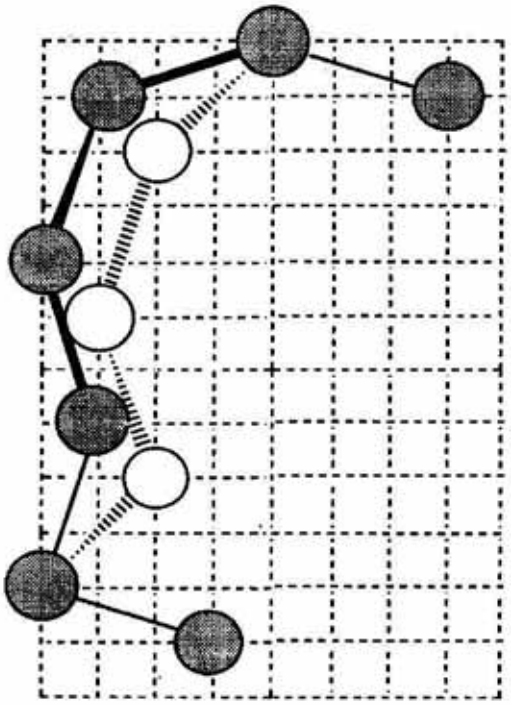

Figure 5. Example of a four-bond conformational transition.

numerical value 0.77 corresponds to the most probable geometry of the main chain hydrogen bonds seen in globular proteins. These simplified Coulombic interactions were cut-off at a distance range of 4.0 Å. Like the previously described model, this model of the hydrogen bond agrees well with the Kabsch-Sander DSSP assignments [149]. Almost all main chain hydrogen bonds are reproduced by the model definition. Explicit cooperativity of the hydrogen bond network has been introduced for this model in a similar fashion as described in the case of the previous model.

Both above outlined models of hydrogen bonds propagate secondary structure of the model chains that mimics very well the geometry of secondary structure motifs seen in real proteins [44].

Long range interactions

Three types of tertiary interactions have been considered: one body burial interactions, pairwise interactions between united atoms, and multibody interactions of the model side chains.

Single domain monomeric globular proteins have a well defined hydrophobic core and a polar, hydrophilic surface. These proteins have native shapes that are almost always close to spherical. Packing density is very similar for all single domain proteins [150]. Thus, assuming an average amino-acid composition, the radius S_N of a globular protein could be estimated [73] basing on number of residues N.

$$\text{with: } S_N = 2.2\ N^{0.38} \text{ (in ångstroms)} \quad (7)$$

From the statistics of known structures of single domain proteins a one body potential could be derived which depends only on $r(A_i)$, the distance of the center of mass of the i-th side group from the center of mass of the entire chain. The one body contribution to the total conformational energy can be then expressed as follows:

$$E_1 = \Sigma\varepsilon_i(r(A_i)/S_N) \quad (8)$$

This contribution is small in a folded state, however it assumes large positive values for expanded random coil states. Consequently, this potential has negligible influence on a specific folding pattern, although it facilitates rapid chain collapse. Alternatively, the burial interactions could be modeled on a more local level. For instance, target numbers (also derived from the statistical analysis of protein structures) of nearest neighbor

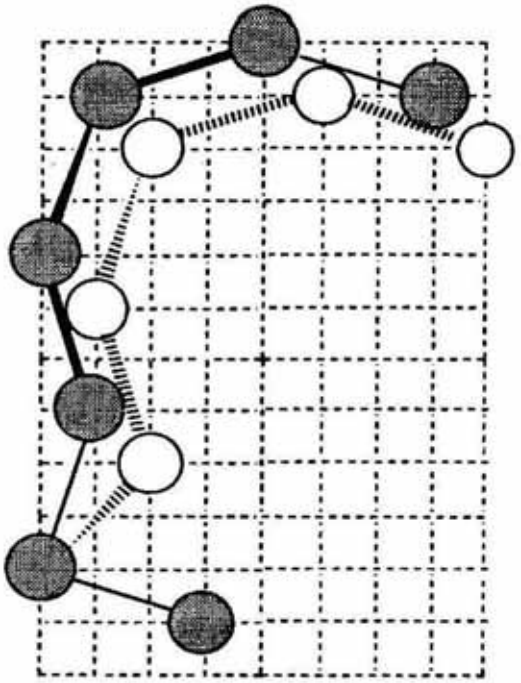

Figure 6. Example of small, rigid body, move of a large part of the model chain.

side chains for particular types of residues can be used for this purpose [85].

Pairwise interactions involve interactions between side chains, between alpha carbon units and between side chains and alpha carbons. They have been assumed in the form of square-well potentials, with the cut-off distances that are pairwise specific. This non-additivity partially accounts for some side chain to side chain packing details. The short distance cut-offs represent a finite strength hard core interactions. All pairwise energy parameters and the cut-off distances have been derived from a proper statistics [44, 79, 80, 151, 152] of a set of non-homologous, high resolution crystallographic structures. The scaling of the cut-off distances has been adjusted in such a way that the side chain contact maps [153] of the model representations of the native proteins were as similar as possible to the contact maps obtained from the corresponding all atom structures. The reference state in potential derivation procedure was a randomly folded compact polypeptide chain of the average (for all proteins) amino-acid composition.

The above set of tertiary interactions, together with short range interactions and with the model hydrogen bonds can distinguish between native-like folded structures and other (compact or expanded) misfolded states of small and structurally relatively simple globular proteins. However, the above set is not capable of reproducing final structure fixation to a unique pattern of the side chain contacts, typical for native proteins. For this purpose it is necessary to introduce multibody potentials that mimic packing preferences seen in protein structures. The most common side chain contact repeat pattern, typical for β-sheets and helix-to helix packing, could be facilitated by the following pseudo four-body potential [44, 72, 73, 78].

$$E_4 = \Sigma[(\varepsilon_{ij}+\varepsilon_{i+k,j+n}) \bullet C_{ij} \bullet \\ \bullet C_{i+k,j+n}] \text{ with } |k|=|n| \qquad (9)$$

where: $C_{ij} = 1$ for side groups being in contact (otherwise $C_{ij} = 0$), ε_{ij} is the pairwise interaction parameter for side groups i and j, and the summation is over all protein-like repeat patterns of side chain contacts, i.e. $|k|=1$, $|k|= 3$ and $|k|= 4$. The last two values of repeat spacing are of a long range nature in all structural motifs [44, 81]. The $|k|=1$ repeat reflects tertiary interactions in β-sheets and the short range packing correlation within a single helix. The idea is illustrated in Fig. 10. This potential has an *ad hoc* structure. Namely, the strength of these four body interactions has been taken arbitrarily

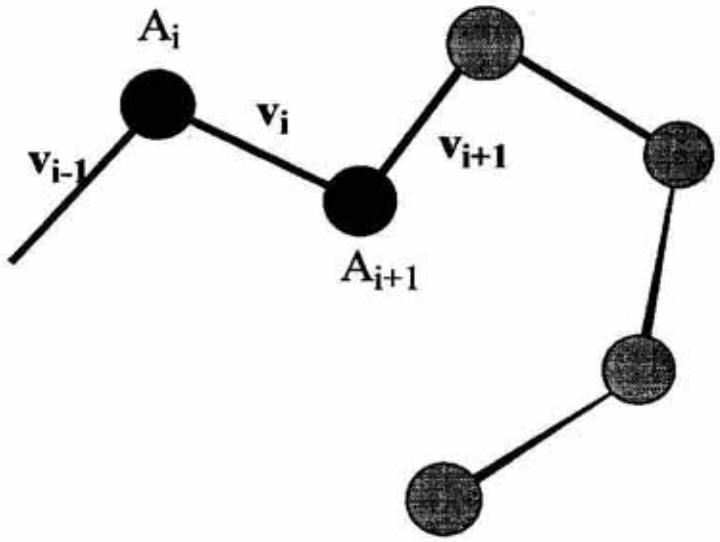

Figure 7. Illustration of the geometry employed in definition of sequence specific short range potential controlling the main chain conformation.

as a sum of two pairwise interactions. Unfortunately, the present size of the crystallographic database is too small for a statistical derivation of the four body terms that would be explicitly dependent on the identity of a four residue set.

It is also possible to account for more complex patterns of the side chain packing, employing an artificial intelligence approach. An experienced human eye can easily distinguish between a native protein side chain contact map and a contact map of more or less randomly folded polypeptide chain [81]. If so, a computational model of neural network could be trained to distinguish between the two classes [82]. A back propagation for suitably trained neural net algorithm is computationally very fast. Consequently, a potential that biases toward protein-like packing patterns generalized in a neural net could be designed and used in the Monte Carlo simulation algorithms [88, 89]. Relatively large fragments of the side chain contact maps (7×7) could be treated. Such potential accounts for very complex multibody interactions typical for real proteins. This is perhaps an avenue which opens a lot of new possibilities in solving the protein folding problem. The neural network approach resulted in the most accurate [154–156] to date method of secondary structure prediction [157, 158]. Pattern recognition methods for predictions of tertiary templates [82, 88, 89] could be equally beneficial.

APPLICATIONS IN STRUCTURE PREDICTION

The proposed methodology that employs lattice discretization of protein conformational space and knowledge based potentials has several features that allow applications that are not attainable with standard methods of molecular modeling. First, the lattice approach and Monte Carlo dynamics enables simulations that correspond to the real folding time of small proteins. Second, the knowledge based force field, at least for subsets of relatively small and structurally simple globular proteins, has its global energy minimum at conformations that are close to the native fold. Finally, the methodology is simple enough for easy encoding of a fragmentary experimental information. In such a case it is possible to extend possibilities of structure prediction to more complex cases. Below, we outline some straightforward examples of globular protein structure prediction and some applications where the model could be used for structural predictions based on sparse experimental data and on evolutionary information. The described work constitutes a small step towards the solution of

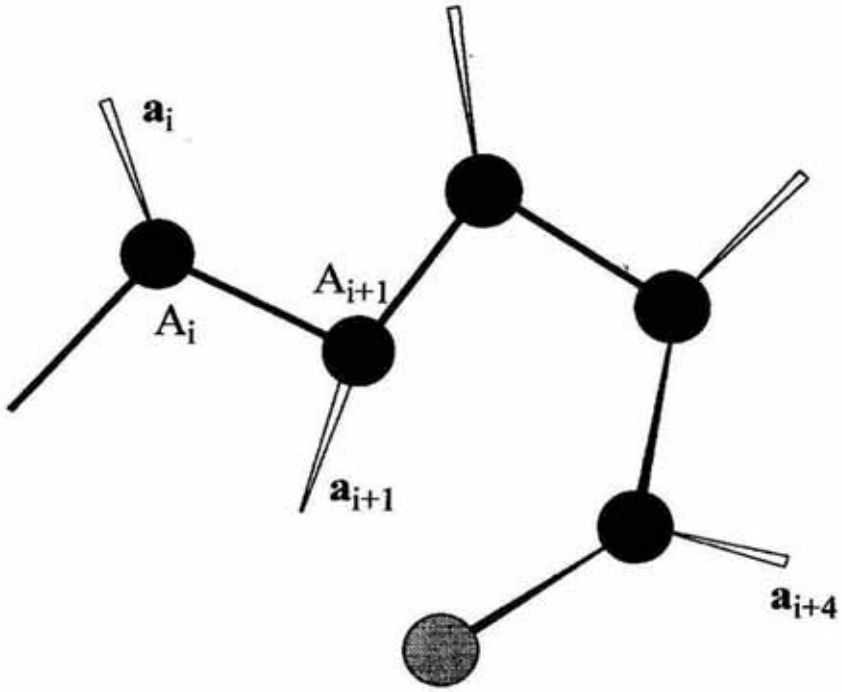

Figure 8. Illustration of the geometry employed in definition of sequence specific short range potential controlling mutual orientations of the side chains.

the protein folding problems and shows how analysis of simulation experiments can contribute to our understanding of general principles of the protein folding process.

De novo structure predictions

The most straightforward application of structure prediction uses sequence of amino acids as the only protein-specific information. The methodology [44] could be summarized as follows. A set of random coil conformations of the tested protein is generated by a separate algorithm. Then, starting from such denatured states the Monte Carlo simulated thermal annealing procedure, using the model described in this contribution, is performed several times. Then, the obtained low energy structures are analyzed. If the results of many simulations are well clustered around a single or few well defined folds they are subjected to further analysis which selects the lowest energy, native-like conformation. Otherwise, when the results of simulated annealing are not reproducible, the structure prediction is not possible. A frequent problem with structure prediction is that an algorithm generates a proper fold together with its topological mirror image.

A typical example are prediction experiments made with, designed by DeGrado and coworkers [159–163], sequences that have been expected to adopt a four helix bundle conformation. We analyzed two sequences [72, 84]. One of them had a leucine based hydrophobic core. In the second sequence several mutations in the core of the designed protein have been introduced in order to break down the degeneracy of the side chain packing. Simulation experiments showed that for the first sequence it was impossible to distinguish between the right-handed and the left-handed topology of a four α-helix bundle. Later, the experiments of Raleigh *et al.* [161] confirmed this result. The fold was energetically very stable, however the packing of the hydrophobic core was poorly defined and the two above mentioned forms existed in equilibrium. For the redesigned sequence the simulation algorithm properly selected a unique structure. Moreover, the analysis of the model folding trajectories have shown that the first sequence produced a molten globule-like structure oscillating between two types of fold while in the second case a native-like unique structure formed, with a typical for real proteins transition from a molten globular to the crystal-like structure, with a fixed pattern of the side chain packing.

A similar structure prediction has been performed for a redesigned (Sander, private communication) monomeric ROP (ColE1 repressor of primer) sequence [164]. The four

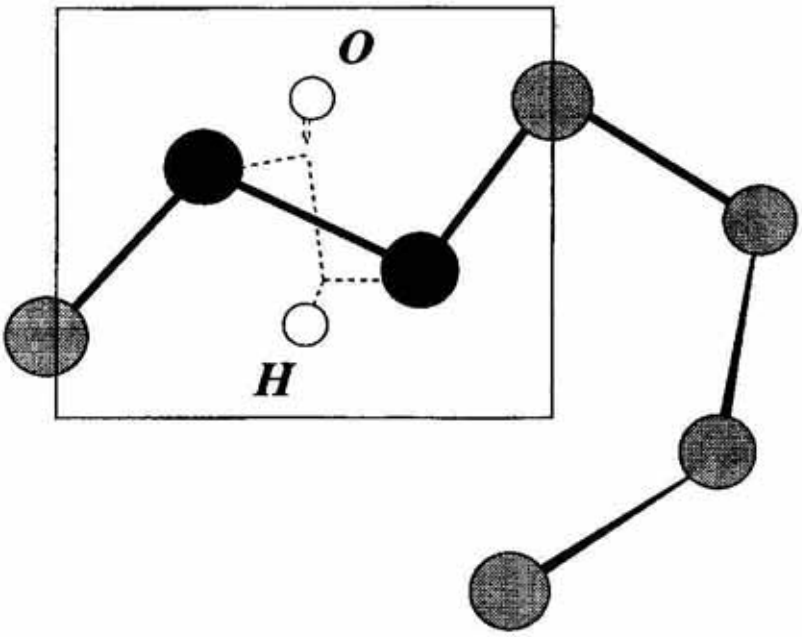

Figure 9. Geometry employed in definition of peptide bond orientation and in model of hydrogen bonds.

helix structure of the protein has been predicted [74]. The predicted structure had 3.6–4.2 Å r.m.s. deviation from the equivalent residues of the native (dimeric) ROP protein [165]. For individual 30 residue helices the error was within the range of 1.0 Å to 2.4 Å r.m.s. [74]. The redesigned sequence has been predicted to have a slightly smaller supertwist of the helices and a unique native-like packing of the side chains.

The test prediction have been also performed for known structures of simple natural globular proteins [74]. A couple of typical examples are protein A domain [166] and crambin [167]. Protein A is a three helix bundle, and its structure has been determined by NMR spectroscopy. The folding algorithm [73, 74] predicted correctly the right handed topology of the bundle and the Cα-trace deviation from the native structure was in the range of 2.25 Å r.m.s.

Very interesting is the case of crambin [167]. The straightforward algorithm tended to predict the native structure, however, with rather low reproducibility. Very frequently the folding algorithm has been trapped in local energy minima when some (not necessarily native) disulfide crosslinks have been formed. The crosslink interactions were very strong and when the first one formed, the folding algorithm had a very hard time to escape from corresponding local minimum of conformational energy. This became clear after a brief inspection of folding trajectories. To remedy this problem we applied a two stage folding protocol. In the first series of simulations (which were performed at a relatively high temperature) we extracted the secondary structure propensities from set of manifold compact structures. These propensities were then used as a secondary structure biases in the proper folding experiments. In this way the short range interactions have been augmented and as a result the native structure formed reproducibly. In such a way very strong cysteine interactions have been thermalized, providing kinetic channels for proper folding. Using this two stage procedure moderate resolution (3.6 Å r.m.s. deviation from native) structures were obtained in majority of simulation runs [74]. The proper structure was easy to select basing on the criterion of the lowest conformational energy.

The most accurate were predictions of the native structure of leucine zipper fragment of GCN4 transcriptional activator [168] and its mutants [169, 170]. In these cases the simulation algorithm [85, 86] produced structures which after a proper all atom rebuilding procedure were indistinguishable from the experimentally determined high resolution structures. These test predictions [85, 86] were interesting for several reasons.

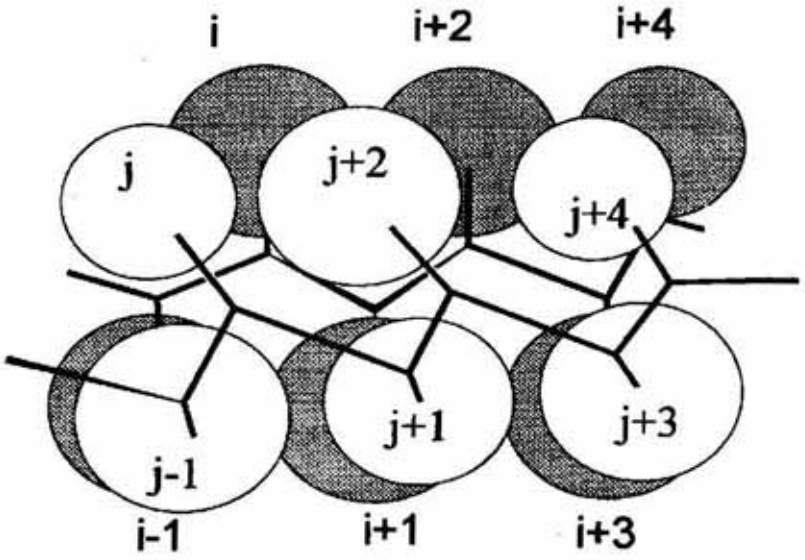

Figure 10. Illustration of side chain packing correlations in globular proteins for a parallel β-sheet fragment.

When the side chain of residues i and j are in contact then almost always the residues $i + 1$ and $j + 1$, residues $i + 2$ and $j + 3$, and residues $i + 4$ and $j + 4$ are also in contact, provided that the entire fragment is a part of regular secondary motif. Similar correlations exist for antiparallel β-sheets and α-helices.

It has been shown that the lattice model and its force field could be applied to the problem of macromolecular assembly [87]. The obtained lattice models provided adequate restraints for successive all atom model rebuilding using CHARMM [171] force field (other standard all atom force fields could be perhaps used as well). The obtained all atom structures were highly reproducible and essentially the same as the structures obtained in the refinement process which started from the crystallographic data. Namely, the coordinate r.m.s. deviation from the crystal structure of GCN4 leucine zipper for the backbone atoms was 0.8 Å, and for the atoms of the side chains from the hydrophobic interface between helices it was 1.31 Å, and for all heavy atoms — 2.29 Å. Moreover, the lattice model responded properly to a subtle sequence mutations and properly predicted the state of association of various coiled coils.

Structure prediction based on fragmentary experimental data

As mentioned before, for more complex folding motifs, the *de novo* approach to the three dimensional structure prediction failed. For example, simulated annealing of plastocyanin and flavodoxin led to compact structures of various topology. Native fold was never obtained in about 20 folding experiments. What is interesting, large fragments of structure were correct and the observed secondary structure was mostly in agreement with the secondary structure of the native state. As it is well known, the folding process is fastest at a close vicinity of the folding transition [28]. For purely practical reasons simulated thermal annealing must be relatively fast and has to cover a wide range of temperature. Thus, in a single simulation, the model system spends only a small fraction of time within the range of the best folding conditions. Since the exact folding temperature is very difficult to estimate *a priori*, it is possible that the simulation experiments are simply too short. It is, however, also possible that the present status of the model force field is too ambiguous and the conformational energy differences between the model native state and various misfolds are too small. Nevertheless, various simulations show that the model force field must be very close to a proper one. If so, even a small additional bias toward native folds should be sufficient for a successful folding of more complex motifs. The above considerations suggested designing of folding algorithms that use some restraints taken from experimental data for a specific sequence on top of the general purpose force field.

The method of folding globular proteins with partially known secondary structure and a small number of long range restraints could be useful at early stages of model building from NMR based experimental data. The restraints have been implemented in a very

general form [172]. The short range restraints assume partial knowledge of protein secondary structure. They are implemented as a weak bias toward an extended backbone conformation (a broad range assumed) for chain fragments that are known to be part of a β-structure and a bias towards compact and right handed conformations for fragments known to be helical. The remaining fragments of the chain have been controlled by the short range interactions of the original force field as described before. Long range restraints have been superimposed on pairs of side chains in the form of broad (3–7 Å) and rather shallow square well potentials. At large distances these restraints had a form of weak harmonic force. The method has been tested on several proteins of various size and topology. Using a moderate number of the long range restraints (many times smaller than the number required by the standard methods of structure building) low to moderate resolution structures were always obtained in a substantial fraction of folding experiments (usually in most of them). The remaining structures obtained in the folding experiments were grossly misfolded or had the mirror image topology of the native fold. The wrong folds could be dismissed in all studied cases basing on the conformational energy criterion. The lowest energy conformations were always native-like. Additionally, the grossly misfolded structures could be rapidly dismissed due to violation of several long range restraints. It should be, however, pointed out that proper folds dominated also in the cases when a fraction of the long restraints was incorrect, i.e. inconsistent with the native structure. Consequently, the method allows structure predictions from not only fragmentary but also from partially incorrect experimental data.

Some representative examples of globular protein folding calculated using a small number of long distance restraints have been recently published [44, 172]. Here we list these examples.

B1 domain of protein G [173] (native structure consists of a four-member β-sheet and has a helix) has been folded many times using various (randomly selected) sets of long range restraints. The number of restraints ranged from 23 to 7. The accuracy of the folded structures changed only slightly in this range. The lowest energy structures were within 3–4 Å r.m.s distance from the native one. Experiments with smaller than N/8 (for N = 56 residues of B1 domain of the protein G, N/8 means 7 pairwise restraints distributed more or less randomly along the polypeptide chain) failed in majority of simulations.

68 Residues α/β fold of 1ctf [174] has been reproducibly obtained using 10 restraints (r.m.s range of 3–4.5 Å). For 8 restraints the lowest energy structures were also native, however the fraction of failed folding experiments became large.

Moderate resolution structures of 99-residues eight-member β-barrel of plastocyanin [175] (1pcy) have been obtained with 23–25 restraints. With 17 restrains the obtained folds were correct, however r.m.s from the native structure was large — about 6.0 Å.

For α/β barrels of thioredoxin [176] (108 residues, 2trx) and flavodoxin [177] (138 residues, 3fxn) 3.0–4.5 Å r.m.s folds were obtained with 25 and 35 restraints, respectively. Good low energy structure has been observed for thioredoxin even with 15 restraints.

For 146 residues myoglobin [178] (1mba) 5.5 Å r.m.s. structures were obtained with 20 long range restraints. With 40 restraints the coordinate r.m.s dropped to 4 Å and the distance r.m.s deviation to about 3 Å.

These experiments show that it is always possible to determine a low resolution structure having N/4 long range restraints. In this respect the lattice based method is much better than the recently proposed methods employing continuous models [179, 180]. Typically, the number of restraints required for dependable structure prediction was the lowest for helical proteins (range of N/7), somewhat higher for α/β type structures, and β-proteins required the strongest set of restraints. This suggests that the lattice model described here works better for helical proteins. Perhaps, the proposed description of main chain conformation and of the short range interactions is better than description of the side chain packing and the corresponding model of the long range interactions.

However, it is also possible that a loose definition of the long range restraints is satisfactory for restricting mutual packing of helices, while for β-structures more precise restrictions are required. Possibly, if some restraints were superimposed on alpha carbons the predictions of behavior of β-proteins would improve. This aspect of the restrained structure prediction technique is now being explored.

Structure prediction based on evolutionary information

It is known that many protein sequences can be grouped into structurally conserved families. Often, two proteins have very similar structures in spite of a very low sequence similarity [153, 181–184]. This opens a possibility to exploit evolutionary information in protein structure prediction [155]. From the multiple sequence alignment of even marginally similar sequences it is possible to find a small set of strongly conserved residues. This could be used in various ways. For example, implementation of the evolutionary information allowed substantial improvement of secondary structure prediction. Rost & Sander [155] estimated that application of a successful multiple sequence alignment leads to about 9% improvement in secondary structure prediction by a computational model of neural network as compared to a corresponding prediction carried out for a single sequence. This way better than 70% accuracy of secondary structure prediction has been achieved [154]. In an implicit way evolutionary information is also exploited in the so called inverse folding method [153, 182–185], where structure similarity is detected by threading (and a proper scoring) of protein sequences trough known structural templates.

Recently, prediction of protein packing patterns has been attempted using the evolutionary information encoded in sequence database [186]. Analyzing correlated mutations [187] observed in multiple sequence alignment it is possible to predict a set of the most probable side chain contacts. These sparse side chain contact maps could be then filtered by a threading-based technique. The purpose of the threading procedure is to remove some false predictions and to extract a more self-consistent set of restraints. The entire methodology, that has been developed recently [90], can be outlined as follows:

- (i) Prediction of secondary structure. For this purpose the Rost & Sander PHD method [155, 156] that employs multiple sequence alignment and neural network technique is used. Only high reliability predictions are taken into consideration.
- (ii) Prediction of chain reversal regions, i.e. surface turns (or loops) where polypeptide chain changes its average direction of propagation. A very accurate algorithm developed recently is used for this purpose [188]. Predictions of protein surface turns override secondary structure predictions, eliminating a substantial fraction of false predictions and improving the overall accuracy of the secondary structure prediction.
- (iii) Using the same multiple sequence alignment the correlated mutations analysis [187] is used to predict side chain contacts. Again, only the strongest predictions are taken into consideration and all the predictions that involve previously predicted turn region residues are neglected. The aim of this filtration procedure is to restrict the predictions only to the protein core, where the residue conservation is expected to be the highest, and consequently the side chain contact prediction the most reliable. The steps (i–iii) result in a reasonable estimation pattern of contacts between the strongly predicted regular elements of secondary structure (helices and β-strands).
- (iv) Enrichment of the predicted contact map. Conservative application of correlated mutations analysis leads only to very few predicted side chain contacts. These provide seeds for enrichment procedure. Since there are limited number of packing patterns of secondary structure elements in globular proteins, the threading procedure [90, 153] could be used for a very dependable detection of the most plausible fragments of side chain contact maps (packing patterns) that contain already predicted contacts. The entire pro-

cedure leads to five-fold increase of the number of predicted contacts.

- (v) Monte Carlo folding simulations with original force field of the lattice model supplemented with predicted short range (secondary structure) and long range (predicted side chain contacts) restraints. This is done by the simulated thermal annealing procedure described in the previous section.
- (vi) Clustering of structure prediction and determination of lowest energy state after long Monte Carlo isothermal refinement runs. Here, mirror image folds and misfolded structures can be eliminated basing on their conformational energy.

The above described procedure has been successfully applied to a set of 16 globular proteins representing all known structural classes of single domain proteins (Ortiz, A. R., Hu, W.-P., Kolinski, A., Skolnick, J., unpublished results). In all cases correct low resolution (coordinate r.m.s deviation from the native structures within the range of 4–6 Å) structures have been predicted. The necessary condition for applicability of the above outlined method is the existence of a sufficient pool (ten or more) of homologous sequences. With rapidly increasing number of known protein sequences this requirement becomes easy to be met for vast majority of proteins. The method is, however, relatively expensive. Many long folding simulations are necessary in order to select the lowest energy fold with a high reliability. The current work focuses on the development of more precise methods of prediction of secondary structure elements and their contact maps and on designing a faster lattice algorithm for the structure assembly. This work is expected to decrease substantially the computational cost (now dependable structure prediction for a 100 residue protein requires about 20 days CPU of fast work-station) and to improve accuracy of prediction to a 3–4 Å r.m.s.

LATTICE SIMULATIONS IN PROTEIN DESIGN AND REDESIGN

Computer aided protein design and redesign is as old as old are computational methods in structural biology [162]. The best known methodology is perhaps the homology modeling [5, 189], where computer graphics and molecular modeling tools are used for relatively limited redesign of known protein structures [190]. This technique is commonly used in rational drug design and other related areas. On the other end of the spectrum of the protein design are theoretical studies of very simple models of protein-like systems for which the sequence space could be extensively explored [162, 191].

Applications based on the high coordination lattice models of proteins have several features, both above mentioned approaches. On one hand, rather realistic geometry of these lattice models enables modeling of some details of protein structure which are completely missed in very simple reduced models. On the other hand, dynamics of model lattice structures faster by several orders of magnitude in comparison with the all-atom models allows for many computational experiments involving large conformational changes. In other words, a meaningful game "what-if" could be played with help of these models. Below we describe a couple of example studies which illustrate such kinds of applications.

De novo design of β-barrel proteins

Even small β-globular proteins have relatively complex folds, as compared with the α/β type proteins and especially with the all-alpha proteins [5]. This could be one of the reasons why the structure prediction with the reduced models described here is more difficult (and less accurate) for β-type proteins. Perhaps, in reality, folding pathways of β-proteins are more complex than folding pathways of helical motifs [192]. Helical proteins always have some residual helical structure in the denatured state, which is partially consistent with secondary structure of the native state [193]. Consequently, during the folding process the three dimensional structure can propagate along the scaffolds provided by already assembled helices. Alternatively, larger helical fragments can coalesce and the number of folding possibilities can narrow rapidly. Early folding intermedi-

ates are easier to observe in various experiments for helical proteins than for β-type proteins [194–196].

In order to gain some insight into possible early folding events of β-type protein folding and also to better understand limitations of the reduced models we attempted a *de novo* design of several β-type folding motifs. We focus here on the example of a six member β-barrel having a G-key folding motif common for many globular proteins. The design of this sequence has been intentionally exaggerated. The β-strands have an ideal odd/even pattern of hydrophobic and hydrophilic residues. Only these amino acids that have strong or moderate propensity for extended conformations have been taken into consideration during the designing process. After several more or less failed attempts we arrived at the following sequence [75]:

(Strand No. 1) Gly-Val-Asp-Val-Asp-Val-

(turn and strand No. 2) Gly-Gly-Gly-Val-Asp-Val-Asp-Val-

(turn and strand No. 3) -Gly-Gly-Phe-Arg-Phe-Arg-Val-

(turn and strand No. 4) Gly-Gly-Gly-Val-Arg-Phe-Arg-Phe-

(turn and strand No. 5) -Gly-Gly-Val-Asp-Val-Asp-Val-

(turn and strand No. 6) Gly-Gly-Gly-Val-Asp-Val-Asp-Val

In the designed structure the strands No. 1, No. 4 and No. 5 were expected to form the first β-sheet, while strands No. 2, No. 3 and No. 6 were supposed to assembly the second sheet of the barrel. In majority of simulated annealing Monte Carlo simulations the designed sequence folded into the desired structure. Misfolded structures were rare, non-reproducible and could be dismissed basing on the conformational energy criterion. The properly folded structures were well defined [44, 75]. The alpha carbon r.m.s deviation between pairs of independently folded structures oscillated between 2.5 and 3.0 Å. In all cases the packing of the hydrophobic core of the folded structures was well defined with the same network of interactions between the strongly hydrophobic Phe side groups. The overlap between the side chain contact maps for independently folded structures was in the range of 70%. This level of packing uniqueness and the range of the distance r.m.s deviations are typical for the resolution of the lattice model. Consequently, it may be assumed that in the range of the model fidelity the designed sequence folded to a unique native-like structure.

Several interesting conclusions can be derived from the designing process and from the folding simulations. First, the successful design has shown that the model force field can easily drive a well defined folding process of relatively complex motifs. Consequently, at least in the cases of exaggerated sequences, the potentials constituting the model force field must be at least qualitatively correct. The minimum of the conformational energy landscape corresponds to a unique compact structure. During the designing process it became obvious that it is not sufficient to have a sequence that favors the native packing. It was also necessary to design against alternative folding motifs. This requirement of the design resulted in a specific pattern of charged surface residues, which destabilized possible alternative arrangements of β-strands. Actually, more effort was directed to the design of the model protein surface than was necessary to design its hydrophobic core. The turn residues in the above listed sequence are all glycines. Attempts with more β-turn specific sequences failed. A plausible explanation is that with more rigid turns it is difficult to achieve good consistency between the geometry of these turns and specific packing of the hydrophobic core. Flexible Gly-type turns accommodate more easily some deficiencies (or rather inconsistencies) of the entire design. Interestingly, similar Gly-based turns have proven to be the most efficient in recent experimental design of a monomeric version of ROP [190]. Most likely, the reasons for this rather surprising result are similar in these experimental studies and in our theoretical design.

Redesign of simple helical folds

The native structure of B domain of protein A from *Staphylococcus aureus* is a three helix bundle [166]. The lattice model has been used in simulation experiments that aimed for a sequence redesign that would reverse the

topology of the native fold. Multiple mutations of the hydrophobic core and the turn regions have been introduced into the protein sequence and tested in simulation experiments [89]. All introduced mutations were rather conservative. Numerous mutations of the hydrophobic core did not induce the desired change of fold topology, however majority of these mutations destabilized the native fold. A complex sieve procedure enabled selection of some mutations in the turn regions that led to the inverted topology of the three helix bundle. The all atom model has been reconstructed for the predicted new structure [89]. A careful refinement procedure showed that the change in handedness of the turns induced by the mutations enabled energetically favorable repacking of the protein hydrophobic core.

Another computational redesign experiment [88] involved a retroprotein, i.e. a protein obtained by reading the protein A sequence backwards. Retroprotein has the same amino-acid composition and the same pattern of hydrophobic/polar residues along the chain. The lattice folding experiments followed by the all atom rebuilding and refinement procedure showed that the retroprotein had the same fold as the native sequence, however with slightly different secondary structure elements (the length of helices changed by one or two residues) and consequently with some, albeit minor, differences in packing of hydrophobic core. Nevertheless, the packing pattern of the hydrophobic core was essentially preserved.

The redesign experiments with protein A fold have shown that the lattice model and its force field responded to sequence mutations in an apparently reasonable way. The predicted structures were consistent with their all atom models, which have been built using the restraints extracted from the lattice structures, i.e. low energy structures have been generated during the refinement procedure. According to energetical and structural criteria the new folds had all features of a native protein. These predictions await experimental verification.

STUDIES OF PROTEIN FOLDING THERMODYNAMICS

Folding transition of globular proteins is very cooperative and exhibits several features of a first-order phase transition. At the transition temperature the population of folding intermediates is negligible [28]. Almost all protein molecules are completely folded or their conformation could be described as essentially random. Thus the folding transition is frequently abbreviated as an all-or-none process [192]. What kind of interactions could be responsible for such striking behavior of a rather small physical system consisting of only a few thousand atoms? Computer simulations could bring valuable insight into this puzzling problem. As it is known, the protein folding process is very slow [192]. It takes milliseconds to a second to assemble a native-like structure. Full atom models are capable of covering only a very fast relaxation of protein structure with characteristic time range of at most several nanoseconds. Even the equilibrium Monte Carlo simulations employing reduced models described here are not efficient enough to provide quantitative characteristics of the folding thermodynamics. In a single long isothermal Monte Carlo run at the folding temperature it is possible to observe but few folding transitions. At lower temperatures the sampling process slows down even more, due to many intervening barriers on the conformational energy landscape. Consequently, relative population of various states is difficult to estimate.

A couple of years ego, Hao & Scheraga [109–111] proposed the so called entropy sampling Monte Carlo method (ESMC) for investigation of thermodynamics of protein models. This method has been previously used by Lee [197] in studies of a simple Ising model. A general idea of ESMC method is similar to the multicanonical MC technique of Berg & Neuhaus [198] which was recently employed by Hansmann & Okamoto [199] in studies of folding of several small peptides. ESMC method generates a nonequilibrium

ensemble of conformations. Instead of energy criterion used in Metropolis scheme, ESMC is controlled by entropy, or rather density of states. Consequently, the energy barriers could be easily surmounted by the sampling process. Moreover, the ESMC method is quasi deterministic. The results from series of simulations could be always used to improve estimation of conformational entropy in the following simulations. In a limit of sufficiently long series of runs an arbitrary accuracy level could be in principle achieved. Single series of simulations lead to a full thermodynamic description (entropy, energy and free energy) in a broad range of temperatures. A disadvantage of the ESMC lies in its relatively high computational cost.

Hao & Scheraga [109, 110] employed the ESMC method in studies of a chess knight model of a simple β-barrel protein. Due to a simple design of the protein model, with a fixed conformation of the side chains with respect to the main chain backbone, the native state was represented by a single, well defined conformation. Various sequences of amino acids have been studied assuming a simple scheme of short range and long range interactions. They showed that some sequences with well defined pattern of hydrophilic and hydrophobic residues exhibited a first order folding transition with very high free energy barrier between the folded and the random coil states. More random sequences undergo continuous transition. It has been shown that the all-or-none transition of model protein is of entropic origin. The changes in the system conformational entropy were the smallest between energy levels corresponding to the vicinity of the transition state. It has been also shown that the main contribution to the folding cooperativity comes from tertiary interactions. Essentially, the short range interactions contributed only to increased stability of the folded state. These findings provided a very convincing picture of protein folding thermodynamics.

There is a substantial body of experimental evidence, and theoretical arguments that the final, however the slowest, stage of the protein folding process is associated with a structural fixation of the side chain packing [200–203]. The transition state, so called molten globule [203], has overall topology of the native fold and has most of the native secondary structure, it is, however, rather mobile and swollen as compared to the native state. The model studied by Hao and Scheraga neglected conformational freedom of the side chains, and the conformational entropy of their model main chain was probably underestimated. Consequently, the issue of the nature of the molten globule state could not be properly addressed. This inspired our ESMC studies on higher resolution reduced model of proteins. As an example we used the designed Greek-key β-barrel sequence [75] (described above) and the knowledge based potentials comprising short range interactions, cooperative model of hydrogen bonds, and tertiary interactions. We investigated three models of tertiary interactions [44, 78]. The first had only one-body burial energy and pairwise interactions. The second model of the tertiary interactions has been supplemented by the pseudo four-body term described in previous sections assuming $|k| = 3$ and $|k| = 4$ repeat period for packing cooperativity. In the third model the cooperative tertiary interactions accounted additionally for $|k| = 1$ repeat of the side chain packing. The scaling of the relative strength of the pairwise interactions in these three models has been adjusted in such way that the balance between tertiary and secondary interaction was the same in all cases. Consequently, the estimated folding temperature was almost the same for all three models. The results of the ESMC experiments could be summarized as follows:

- (i) In all cases the lowest energy states have Greek-key native-like conformation. The main chain conformation and packing of the side chains were the same at low energy region for all three models of interactions.
- (ii) At the same time the folding transition was continuous for the first model of tertiary interactions, much sharper for the second model and clearly all-or-none for

the third model, with a high free energy barrier between the folded and unfolded state.

- (iii) The transition state (at the free energy maximum) had all qualitative features of molten globule state, as outlined above.

As in Hao & Scheraga studies [109–111] we found that the first-order folding transition is of entropic origin. It was necessary to account for tertiary multibody interactions in order to reproduce the all-or-none folding transition. Exaggerated design of the protein (an ideal hydrophobic pattern and strong β-type secondary propensities of amino-acid sequence) and a strong cooperativity of the hydrogen bond network did not result in a sufficiently cooperative transition in the absence of the multibody terms. What is then the mechanism associated with the multibody interactions that leads to the cooperative all-or-none transition? As mentioned before, the characteristics of the folded state were the same for all models. In the random coil state the tertiary interactions are weak regardless of model of interactions. However, the multibody interactions caused relative energetic destabilization of partly folded states, decreasing their thermodynamic probability and consequently increasing the free energy gap between the folded and unfolded states. Recent extensive simulations of globular proteins within the framework of all atom potentials suggest that analogous multibody interactions could be also important in detailed molecular models.

CONCLUSION

High coordination lattice models of protein conformation described here allow to study protein systems at moderate resolution. Computational speed of the lattice algorithms enables simulation of the entire folding process. The force fields of these lattice models are knowledge based. Potentials of mean force [204] have been derived from statistical analysis of known protein structures. Currently, the model force field is specific enough to fold very simple globular proteins using sequence of amino acids as the only protein specific input data. Applicability in protein structure prediction can be considerably extended when evolutionary information of the model is available or/and when sparse distance restraints are available from experiment. In a semiquantitative way the lattice models reproduce the dynamic and thermodynamic properties of globular proteins. Consequently, the methodology presented here can be considered a useful alternative or a complementary approach to the standard molecular modeling tools in structural studies of proteins, studies of long time protein dynamics and in computer aided protein design and redesign.

We wish to express our thanks to our collaborators: Dr. Adam Godzik, Dr. Michał Vieth, Dr. Mariusz Milik, Dr. Angel R. Ortiz, Dr. Krzysztof Olszewski, Dr. Wei-Ping Hu, Dr. Andrzej Sikorski, Wojtek Gałązka and Łukasz Jaroszewski, who have contributed to the studies described in this review.

REFERENCES

1. Creighton, T.E. (1990) Protein folding. *Biochem. J.* **270,** 131–146.
2. Creighton, T.E. (1993) *Proteins: Structures and Molecular Properties.* W.H. Freeman and Company, New York.
3. Davies, D.R. & Metzger, H. (1983) Structural basis of antibody function. *Annu. Rev. Immunol.* **1,** 87–117.
4. Fersht, A. (1984) *Enzyme Structure and Mechanism.* W.H. Freeman, New York.
5. Branden, C. & Tooze, J. (1991) *Introduction to Protein Structure.* Garland Publishing Inc., New York, London.
6. McKusick, V.A. (1991) Current trends in mapping human genes. *FASEB J.* **5,** 12–20.
7. Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, E.F., Jr., Brice, M.D., Rodgers, J.R., Kennard, O., Simanouchi, T. & Tasumi, M. (1977) The protein data bank: A computer-based archival file for macromolecular structures. *J. Mol. Biol.* **112,** 535–542.

8. PDB (1995) Quarterly Newsletter, No. 71, January 1995.

9. Haliwell, J.R. (1992) *Macromolecular Crystallography.* Cambridge University Press, Cambridge.

10. Doolittle, R.F. (1981) Similar amino acid sequences: Chance or common ancestry? *Science* **214,** 149–159.

11. Wutrich, K. (1986) *NMR of Proteins and Nucleic Acids.* J. Wiley, New York.

12. Kaptein, R., Boelens, R., Scheek, R.M. & van Gunsteren, W.F. (1988) Protein structures from NMR. *Biochemistry* **27,** 5389–5395.

13. Clore, G.M., Robien, M.A. & Gronenborn, A.M. (1993) Exploring the limits of precision and accuracy of protein structures determined by nuclear magnetic resonance spectroscopy. *J. Mol. Biol.* **231,** 82–102.

14. Wright, P.E., Dyson, H.J. & Lerner, R.A. (1988) Conformation of peptide fragments of proteins in aqueous solution: Implications for initiation of protein folding. *Biochemistry* **27,** 7167–7175.

15. Brunger, A.T. & Karplus, M. (1991) Molecular dynamics simulations with experimental restraints. *Acc. Chem. Res.* **24,** 54–61.

16. Meirovitch, H., Vasquez, M. & Scheraga, H.A. (1988) Stability of polypeptide conformational states. II. Folding of a polypeptide chain by the scanning simulation method, and calculation of the free energy of the statistical coil. *Biopolymers* **27,** 1189–1204.

17. Meirovitch, H. & Meirovitch, E. (1996) New theoretical methodology for elucidating the solution structure of peptides from NMR data. 3. Solvation effects. *J. Phys. Chem.* **100,** 5123–5133.

18. DeLisi, C. (1988) Computers in molecular biology: Current applications and emerging trends. *Science* **240,** 47–52.

19. Karplus, M. & Petsko, G.A. (1990) Molecular dynamics simulations in biology. *Nature* **347,** 631–639.

20. Skolnick, J. & Kolinski, A. (1997) Protein Modeling; in *Encyclopedia of Computational Chemistry* (Rauge Schleyer, P., Allinger, N.L., Clark, T., Gasteiger, J., Kollman, P.A., Schaefer, H.F. III, eds.) John Wiley & Sons, London, New York.

21. Brooks, C.L. III. (1993) Molecular simulations of peptide and protein unfolding: In quest of a molten globule. *Curr. Opin. Struct. Biol.* **3,** 92–98.

22. Brooks, C.L. III., Karplus, M. & Pettitt, B.M. (1988) Protein: A theoretical perspective of dynamics, structure, and thermodynamics. *Adv. Chem. Phys.* **71,** 1–259.

23. Bruccoleri, R.E. & Karplus, M. (1987) Prediction of the folding of short polypeptide segments by uniform conformational sampling. *Biopolymers* **26,** 137–168.

24. Elber, R. & Karplus, M. (1987) Multiple conformational states of proteins: A molecular dynamics analysis of myoglobin. *Science* **235,** 318–321.

25. Elofsson, A. & Nilsson, L. (1993) How consistent are molecular dynamics simulations? Comparing structure and dynamics in reduced and oxidized *Escherichia coli* thioredoxin. *J. Mol. Biol.* **223,** 766–780.

26. Bryngelson, J.D., Onuchic, J.N., Socci, N.D. & Wolynes, P.G. (1995) Funnels, pathways and the energy landscape of protein folding: A synthesis. *Proteins* **21,** 167–195.

27. Zwanzig, R., Szabo, A. & Bagchi, B. (1992) Levinthal's paradox. *Proc. Natl. Acad. Sci. U.S.A.* **89,** 20–22.

28. Anfinsen, C.B. & Scheraga, H.A. (1975) Experimental and theoretical aspects of protein folding. *Adv. Prot. Chem.* **29,** 205–300.

29. DeBolt, S. & Skolnick, J. (1996) Evaluation of atomic level mean force potentials *via* inverse folding and inverse refinement of protein structures: Atomic burial position and pairwise nonbonded interactions. *Protein Eng.* **9,** 637–655.

30. Levitt, M. & Warshel, A. (1975) Computer simulation of protein folding. *Nature* **253,** 694–698.

31. Levitt, M. (1975) A simplified representation of protein conformations for rapid simulation of protein folding. *J. Mol. Biol.* **104,** 59–107.

32. Kuntz, I.D., Crippen, G.M., Kollman, P.A. & Kimelman, D. (1976) Calculation of protein tertiary structure. *J. Mol. Biol.* **106,** 983–994.

33. Dill, K.A., Bromberg, S., Yue, K., Fiebig, K.M., Yee, D.P., Thomas, P.D. & Chan, H.S. (1995) Principles of protein folding — A perspective from simple exact models. *Protein Sci.* **4,** 561–602.

34. Skolnick, J. & Kolinski, A. (1989) Computer simulations of globular protein folding and tertiary structure. *Annu. Rev. Phys. Chem.* **40,** 207–235.

35. Levitt, M. (1991) Protein folding. *Curr. Opin. Struct. Biol.* **1,** 224–229.

36. Kolinski, A., Milik, M., Rycombel, J. & Skolnick, J. (1995) A reduced model of short range interactions in polypeptide chains. *J. Chem. Phys.* **103,** 4312–4323.

37. Karplus, M. & Shakhnovich, E. (1992) Thermodynamics of protein folding; in *Protein Folding* (Creighton, T.E., ed.) pp. 127–196, W.H. Freeman, New York.

38. Sali, A., Shakhnovich, E. & Karplus, M. (1994) Kinetics of protein folding. A lattice model study of the requirements for folding to the native state. *J. Mol. Biol.* **235,** 1614–1636.

39. Hagler, A.T. & Honig, B. (1978) On the formation of protein tertiary structure on a computer. *Proc. Natl. Acad. Sci. U.S.A.* **75,** 554–558.

40. Covell, D.G. (1992) Folding protein α-carbon chains into compact forms by Monte Carlo methods. *Proteins* **14,** 409–420.

41. Wilson, C. & Doniach, S. (1989) A computer model to dynamically simulated protein folding: Studies with crambin. *Proteins* **6,** 193–209.

42. Crippen, G.M. (1991) Prediction of protein folding from amino acid sequence over discrete conformation spaces. *Biochemistry* **30,** 4232–4237.

43. Sun, S. (1993) Reduced representation model of protein structure prediction: Statistical potential and genetic algorithms. *Protein Sci.* **2,** 762–785.

44. Kolinski, A. & Skolnick, J. (1996) *Lattice Models of Protein Folding, Dynamics and Thermodynamics.* R.G. Landes, Austin, Texas.

45. Kolinski, A., Skolnick, J. & Yaris, R. (1986) The collapse transition of semiflexible polymers. A Monte Carlo simulation of a model system. *J. Chem. Phys.* **85,** 3585–3597.

46. Kolinski, A., Skolnick, J. & Yaris, R. (1987) Monte Carlo studies on equilibrium globular protein folding. I. Homopolymeric lattice models of β-barrel proteins. *Biopolymers* **26,** 937–962.

47. Chan, H.S. & Dill, K.A. (1989) Compact polymers. *Macromolecules* **22,** 4559–4573.

48. Chan, H.S. & Dill, K.A. (1991) "Sequence space soup" of proteins and copolymers. *J. Chem. Phys.* **95,** 3775–3787.

49. Chan, H.S. & Dill, K.A. (1990) Origins of structure in globular proteins. *Proc. Natl. Acad. Sci. U.S.A.* **87,** 6388–6392.

50. Skolnick, J., Kolinski, A. & Yaris, R. (1988) Monte Carlo simulations of the folding of β-barrel globular proteins. *Proc. Natl. Acad. Sci. U.S.A.* **85,** 5057–5063.

51. Skolnick, J., Kolinski, A. & Yaris, R. (1989) Monte Carlo studies on equilibrium globular protein folding. II. β-Barrel globular protein models. *Biopolymers* **28,** 1059–1095.

52. Skolnick, J., Kolinski, A. & Yaris, R. (1989) Dynamic Monte Carlo study of a six stranded Greek key globular protein. *Proc. Natl. Acad. Sci. U.S.A.* **86,** 1229–1233.

53. Skolnick, J., Kolinski, A. & Sikorski, A. (1990) Dynamic Monte Carlo simulations of globular protein and structure. *Chemical Design Automation News* **5,** 1–20.

54. Sikorski, A. & Skolnick, J. (1989) Monte Carlo studies on equilibrium globular protein folding. III. The four helix bundle. *Biopolymers* **28,** 1097–1113.

55. Sikorski, A. & Skolnick, J. (1989) Monte Carlo simulation of equilibrium globular protein folding. α-Helical bundles with long loops. *Proc. Natl. Acad. Sci. U.S.A.* **86,** 2668–2672.

56. Sikorski, A. & Skolnick, J. (1990) Dynamic Monte Carlo simulations of globular protein folding/unfolding pathways. II. α-Helical motifs. *J. Mol. Biol* **212,** 819–836.

57. Sikorski, A. & Skolnick, J. (1990) Dynamic Monte Carlo simulations of globular protein folding model studies of *in vivo* assembly of four helix bundles and four member β-barrels. *J. Mol. Biol.* **215,** 183–198.

58. Dill, K.A. (1993) Folding Proteins: Finding a needle in a haystack. *Curr. Biol.* **3,** 99–103.

59. Sali, A., Shakhnovich, E. & Karplus, M. (1994) How does a protein fold? *Nature* **369,** 248–251.

60. Shakhnovich, E.I. & Finkelstein, A.V. (1989) Theory of cooperative transitions in protein molecules. I. Why denaturation of globular protein is a first-order phase transition. *Biopolymers* **28,** 1667–1680.

61. Shakhnovich, E.I. & Gutin, A.M. (1989) Formation of unique structure in a polypeptide chain. *Biophys. Chem.* **34,** 187–199.

62. Shakhnovich, E., Farztdinov, G. & Gutin, A.M. (1991) Protein folding bottlenecks: A lattice Monte Carlo simulation. *Phys. Rev. Lett.* **67,** 1665–1668.

63. Shakhnovich, E.I. & Gutin, A.M. (1993) Engineering of stable and fast-folding sequences of model proteins. *Proc. Natl. Acad. Sci. U.S.A.* **90,** 7195–7199.

64. Shakhnovich, E.I. (1994) Proteins with selected sequences fold into unique native conformations. *Phys. Rev. Lett.* **72,** 3907–3910.

65. Socci, N.D. & Onuchic, J.N. (1994) Folding kinetics of protein-like heteropolymers. *J. Chem. Phys.* **100,** 1519–1528.

66. Lau, K.F. & Dill, K.A. (1989) A lattice statistical mechanics model of the conformational and sequence space of proteins. *Macromolecules* **22,** 3986–3997.

67. Brower, R.C., Vasmatiz, G., Silverman, M. & DeLisi, C. (1993) Exhaustive conformational search and simulated annealing for models of lattice peptides. *Biopolymers* **33,** 329–334.

68. Camacho, C.J. & Thirumalai, D. (1993) Kinetics and thermodynamics of folding in model proteins. *Proc. Natl. Acad. Sci. U.S.A.* **90,** 6369–6372.

69. Onuchic, J.N., Wolynes, P.G., Luthey-Schulten, Z. & Socci, N.D. (1995) Toward an outline of the topography of a realistic protein-folding funnel. *Proc. Natl. Acad. Sci. U.S.A.* **92,** 2626–3630.

70. Kolinski, A., Milik, M. & Skolnick, J. (1991) Static and dynamic properties of a new lattice model of polypeptide chain. *J. Chem. Phys.* **94,** 3978–3985.

71. Kolinski, A. & Skolnick, J. (1992) Discretized model of proteins. I. Monte Carlo study of cooperativity in homopolypeptides. *J. Phys. Chem.* **97,** 9412–9426.

72. Kolinski, A., Godzik, A. & Skolnick, J. (1993) A general method for the prediction of the three dimensional structure and folding pathway of globular proteins. Application to designed helical proteins. *J. Chem. Phys.* **98,** 7420–7433.

73. Kolinski, A. & Skolnick, J. (1994) Monte Carlo simulations of protein folding. I. Lattice model and interaction scheme. *Proteins* **18,** 338–352.

74. Kolinski, A. & Skolnick, J. (1994) Monte Carlo simulations of protein folding. II. Application to protein A, ROP, and crambin. *Proteins* **18,** 353–366.

75. Kolinski, A., Galazka, W. & Skolnick, J. (1995) Computer design of idealized β-motifs. *J. Chem. Phys.* **103,** 10286–10297.

76. Godzik, A., Kolinski, A. & Skolnick, J. (1993) Lattice representation of globular proteins: How good are they? *J. Comput. Chem.* **14,** 1194–1202.

77. Rykunov, D.S., Reva, B.A. & Finkelstein, A.V. (1995) Accurate general method for lattice approximation of three-dimensional structure of a chain molecule. *Proteins* **22,** 100–109.

78. Kolinski, A., Galazka, W. & Skolnick, J. (1996) On the origin of the cooperativity of protein folding. Implications from model simulations. *Proteins* **26,** 271–287.

79. Godzik, A., Kolinski, A. & Skolnick, J. (1995) Are proteins ideal mixtures of amino acids?

Analysis of energy parameter sets. *Protein Sci.* **4,** 2107–2117.

80. Skolnick, J., Jaroszewski, L., Kolinski, A. & Godzik, A. (1997) Derivation and testing of pair potentials for protein folding. Beyond the quasichemical approximation. *Protein Eng.* (in press).

81. Godzik, A., Skolnick, J. & Kolinski, A. (1993) Regularities in interaction patterns of globular proteins. *Protein Eng.* **6,** 801–810.

82. Milik, M., Kolinski, A. & Skolnick, J. (1995) Neural network system for the evaluation of side chain packing in protein structures. *Protein Eng.* **8,** 225–236.

83. Skolnick, J., Kolinski, A., Brooks, C.L. III, Godzik, A. & Rey, A. (1993) A method for prediction of protein structure from sequence. *Curr. Biol.* **3,** 414–423.

84. Godzik, A., Kolinski, A. & Skolnick, J. (1993) De novo and inverse folding predictions of protein structure and dynamics. *J. Comput. Aided Mol. Design* **7,** 397–438.

85. Vieth, M., Kolinski, A., Brooks, C.L. III. & Skolnick, J. (1994) Prediction of the folding pathways and structure of the GCN4 leucine zipper. *J. Mol. Biol.* **237,** 361–367.

86. Vieth, M., Kolinski, A., Brooks, C.L. III. & Skolnick, J. (1995) Prediction of the quaternary structure of coiled coils. Application to mutants of the GCN4 leucine zipper. *J. Mol. Biol.* **251,** 448–467.

87. Vieth, M., Kolinski, A. & Skolnick, J. (1996) Method for prediction the state of association of discretized protein models. Application to leucine zippers. *Biochemistry* **35,** 955–967.

88. Olszewski, K.A., Kolinski, A. & Skolnick, J. (1996) Does a backwardly read sequence have a unique native state? *Protein Eng.* **9,** 5–14.

89. Olszewski, K.A., Kolinski, A. & Skolnick, J. (1996) Folding simulations and computer redesign of protein A three helix bundle motifs. *Proteins* **25,** 286–299.

90. Ortiz, A.R., Hu, W.-P., Kolinski, A. & Skolnick, J. (1997) Method for low resolution prediction of small protein tertiary structure. *J. Mol. Graphics.* (in press).

91. Fogolari, F., Esposito, G., Viglino, P. & Cattarinussi, S. (1996) Modeling of polypeptide chains as C_α chains, C_α chains with C_β, and C_α chains with ellipsoidal lateral chains. *Biophys. J.* **70,** 1183–1197.

92. Wallqvist, A. & Ullner, M. (1994) A simplified amino acid potential for use in structure prediction of proteins. *Proteins* **18,** 267–289.

93. Miyazawa, S. & Jernigan, R.L. (1985) Estimation of effective interresidue contact energies from protein crystal structures: Quasichemical approximation. *Macromolecules* **18,** 534–552.

94. Lee, B., Kurochkina, N. & Kang, H.S. (1996) Protein folding by a biased Monte Carlo procedure in the dihedral angle space. *FASEB J.* **10,** 119–125.

95. Hoffmann, D. & Knapp, E.W. (1996) Polypeptide folding with off-lattice dynamics: The method. *Eur. Biophys. J.* **24,** 387–403.

96. Honeycutt, J.D. & Thirumalai, D. (1990) Metastability of the folded states of globular proteins. *Proc. Natl. Acad. Sci. U.S.A.* **87,** 3526–3529.

97. Garrett, D.G. & Kastella, K. (1992) New results on protein folding from simulated annealing. *J. Am. Chem. Soc.* **114,** 6555–6556.

98. Knapp, E.W. (1992) Long time dynamics of a polymer with rigid body monomer unit relating to a protein model: Comparison with the Rouse model. *J. Comput. Chem.* **13,** 793–798.

99. Rey, A. & Skolnick, J. (1991) Comparison of lattice Monte Carlo dynamics and Brownian dynamics folding pathways of α-helical hairpins. *Chem. Phys.* **158,** 199–219.

100. Rey, A. & Skolnick, J. (1993) Computer modeling and folding of four-helix bundles. *Proteins* **16,** 8–28.

101. Guo, Z. & Thirumalai, D. (1995) Kinetics of protein folding: Nucleation mechanism, time scales, and pathways. *Biopolymers* **36,** 83–102.

102. Zhou, Y., Hall, C.K. & Karplus, M. (1996) First-order disorder-to order transition in an isolated homopolymer model. *Phys. Rev. Lett.* **77,** 2822–2825.

103. Hoffmann, D. & Knapp, E.W. (1996) Protein dynamics with off-lattice Monte Carlo moves. *Phys. Rev. E* **53,** 4221–4224.

104. Rabow, A.A. & Scheraga, H.A. (1996) Improved genetic algorithm for protein folding problem by use of a Cartesian combination operator. *Protein Sci.* **5,** 1800–1815.

105. Kolinski, A., Skolnick, J. & Yaris, R. (1987) Dynamic Monte Carlo study of the conformational properties of long flexible polymers. *Macromolecules* **20,** 438–440.

106. Kolinski, A. & Skolnick, J. (1986) Monte Carlo simulations on an equilibrium globular protein folding model. *Proc. Natl. Acad. Sci. U.S.A.* **83,** 7267–7271.

107. Skolnick, J. & Kolinski, A. (1989) Dynamic Monte Carlo simulations of globular protein folding/unfolding pathways. I. Six member, Greek key β-barrels. *J. Mol. Biol.* **212,** 787–817.

108. Skolnick, J. & Kolinski, A. (1991) Dynamic Monte Carlo simulations of a new lattice model of globular protein folding, structure, and dynamics. *J. Mol. Biol.* **221,** 499–531.

109. Hao, M.-H. & Scheraga, H.A. (1994) Monte Carlo simulations of a first-order transition for protein folding. *J. Phys. Chem.* **98,** 4940–4948.

110. Hao, M.-H. & Scheraga, H.A. (1994) Statistical thermodynamics of protein folding: Sequence dependence. *J. Phys. Chem.* **98,** 9882–9893.

111. Hao, M.-H. & Scheraga, H.A. (1995) Statistical thermodynamics of protein folding: Comparison of mean-field theory with Monte Carlo simulations. *J. Chem. Phys.* **102,** 1334–1348.

112. O'Toole, E., Venkataramani, R. & Panagiotopoulous, A.Z. (1995) Simple lattice model of proteins directional bonding and structural solvent. *AIChE J.* **41,** 954–958.

113. Galzitskaya, O.V. & Finkelstein, A.V. (1994) Folding of chains with random and edited sequences: Similarities and differences. *Protein Eng.* **8,** 883–892.

114. Dinner, A.R., Sali, A. & Karplus, M. (1996) The folding mechanism of larger proteins: Role of native structure. *Proc. Natl. Acad. Sci. U.S.A.* **93,** 8356–8361.

115. Chan, H.S. & Dill, K.A. (1994) Transition states and folding dynamics of proteins and heteropolymers. *J. Chem. Phys* **100,** 9238–9257.

116. O'Toole, E.M. & Panagiotopoulos, A.Z. (1992) Monte Carlo simulation of folding transitions of simple model proteins using a chain growth algorithm. *J. Chem. Phys.* **97,** 8644–8652.

117. Hao, M.-H. & Scheraga, H.A. (1996) How optimization of potential functions affects protein folding. *Proc. Natl. Acad. Sci. U.S.A.* **93,** 4984–4989.

118. Socci, N.D., Onuchic, J.N. & Wolynes, P.G. (1996) Diffusive dynamics of the reaction coordinate for protein folding funnels. *J. Chem. Phys.* **104,** 5860–5868.

119. Bryngelson, J.D. (1994) When is a potential accurate enough for structure prediction? Theory and application to a random heteropolymer model of protein folding. *J. Chem. Phys.* **100,** 6038–6045.

120. Li, H., Helling, R., Tang, C. & Wingreen, N. (1996) Emergence of preferred structures in a simple model of protein folding. *Science* **273,** 666–669.

121. Beutler, T.C. & Dill, K.A. (1996) A fast conformational search strategy for finding low energy structures of model proteins. *Protein Sci.* **5,** 2037–2043.

122. Hao, M.-H. & Scheraga, H.A. (1996) Optimizing potential functions for protein folding. *J. Phys. Chem.* **100,** 14540–14548.

123. Yue, K., Fiebig, K.M., Thomas, P.D., Chan, H.S., Shakhnovich, E.I. & Dill, K.A. (1995) A test of lattice protein folding algorithms. *Proc. Natl. Acad. Sci. U.S.A.* **92,** 325–329.

124. Go, N. & Taketomi, H. (1978) Respective roles of short- and long-range interactions in protein folding. *Proc. Natl. Acad. Sci. U.S.A.* **75,** 559–563.

125. Go, N., Abe, H., Mizuno, H. & Taketomi, H. (eds.) (1980) *Protein Folding.* Elsevier/ /North Holland, Amsterdam.

126. Taketomi, H., Kano, F. & Go, N. (1988) The effect of amino acid substitution on protein folding and unfolding transition studied by computer simulation. *Biopolymers* **27,** 527–559.

127. Taketomi, H., Ueda, Y. & Go, N. (1988) Studies of protein folding, unfolding and fluctuations by computer simulations. *Int. J. Pept. Protein Res.* **7,** 445–449.

128. Ueda, Y., Taketomi, H. & Go, N. (1978) Studies on protein folding, unfolding, and fluctuations by computer simulation. II. A three-dimensional lattice model of lysozyme. *Biopolymers* **17,** 1531–1548.

129. Krigbaum, W.R. & Lin, S.F. (1982) Monte Carlo simulation of protein folding using a lattice model. *Macromolecules* **15,** 1135–1145.

130. Dashevskii, V.G. (1980) Lattice model of three-dimensional structure of globular proteins. *Molekulyarnaya Biologiya (Translation from Russian)* **14,** 105–117.

131. Covell, D.G. & Jernigan, R.L. (1990) Conformations of folded proteins in restricted spaces. *Biochemistry* **29,** 3287–3294.

132. Hinds, D.A. & Levitt, M. (1992) A lattice model for protein structure prediction at low resolution. *Proc. Natl. Acad. Sci. U.S.A.* **89,** 2536–2540.

133. Skolnick, J. & Kolinski, A. (1990) Simulations of the folding of a globular protein. *Science* **250,** 1121–1125.

134. Godzik, A., Skolnick, J. & Kolinski, A. (1992) Simulations of the folding pathway of TIM type α/β barrel proteins. *Proc. Natl. Acad. Sci. U.S.A.* **89,** 2629–2633.

135. Rey, A. & Skolnick, J. (1992) Efficient algorithm for the reconstruction of a protein backbone from the α-carbon coordinates. *J. Comput. Chem.* **13,** 443–456.

136. Milik, M., Kolinski, A. & Skolnick, J. (1997) An algorithm for rapid reconstruction of a protein backbone from alpha carbon coordinates. *J. Comput. Chem.* **18,** 80–85.

137. Orwoll, R.A. & Stockmayer, W.H. (1969) Stochastic models for chain dynamics. *Adv. Chem. Phys.* **15,** 305–324.

138. Baumgartner, A. (1984) Simulation of polymer motion. *Annu. Rev. Phys. Chem.* **35,** 419–435.

139. Kolinski, A., Skolnick, J. & Yaris, R. (1987) Does reptation describe the dynamics of entangled, finite length polymer system? A model simulation. *J. Chem. Phys.* **86,** 1567–1585.

140. Kolinski, A., Skolnick, J. & Yaris, R. (1987) Monte Carlo studies on the long time dynamic properties of dense cubic lattice multichain systems. I. The homopolymeric melt. *J. Chem. Phys.* **86,** 7164–7173.

141. Kolinski, A., Skolnick, J. & Yaris, R. (1987) Monte Carlo studies on the long time dynamic properties of dense cubic lattice multichain systems. II. Probe polymer in a matrix of different degrees of polymerization. *J. Chem. Phys.* **86,** 7174–7180.

142. Skolnick, J. & Kolinski, A. (1990) Dynamics of dense polymer systems: Computer simulations and analytic theories. *Adv. Chem. Physics* **78,** 223–278.

143. Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H. & Teller, E. (1953) Equation of state calculations by fast computing machines. *J. Chem. Phys.* **51,** 1087–1092.

144. Zwanzig, R. (1974) Theoretical basis for the Rouse-Zimm model in polymer solution dynamics. *J. Chem. Phys.* **60,** 2717–2720.

145. Kyte, J. & Doolittle, R.F. (1982) A simple method for displaying the hydropathic character of protein. *J. Mol. Biol.* **157,** 105–132.

146. Anfinsen, C.B. (1973) Principles that govern the folding of protein chains. *Science* **181,** 223–230.

147. Finkelstein, A.V., Badretdinov, A.Y. & Gutin, A.M. (1996) Why do protein architecture have Boltzmann-like statistics? *Proteins* **23,** 142–150.

148. Levitt, M. & Greer, J. (1977) Automatic identification of secondary structure in globular proteins. *J. Mol. Biol.* **114,** 181–293.

149. Kabsch, W. & Sander, C. (1983) Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22,** 2577–2637.

150. Abad-Zapatero, C. & Lin, C.T. (1990) Statistical descriptors for the size and shape of globular proteins. *Biopolymers* **29,** 1745–1754.

151. Edelman, J. (1992) Pair distribution function in small systems: Implication for protein folding. *Biopolymers* **21,** 3–10.

152. Skolnick, J. & Kolinski, A. (1996) Monte Carlo lattice dynamics and prediction of protein folds; in *Computer Simulations of Biomolecular Systems. Theoretical and Experimental Studies* (van Gunsteren, W.F., We-

iner, P.K. & Wilkinson, A.J., eds.) ESCOM Science Publ., The Netherlands.

153. Godzik, A., Skolnick, J. & Kolinski, A. (1992) A topology fingerprint approach to the inverse folding problem. *J. Mol. Biol* **227,** 227–238.

154. Rost, B. & Sander, C. (1993) Prediction of secondary structure at better than 70% accuracy. *J. Mol. Biol.* **232,** 584–599.

155. Rost, B. & Sander, C. (1994) Combining evolutionary information and neural networks to predict protein secondary structure. *Proteins* **19,** 55–72.

156. Rost, B. & Sander, C. (1996) Progress of 1D protein structure prediction at last. *Proteins* **23,** 295–300.

157. Vieth, M., Kolinski, A., Skolnick, J. & Sikorski, A. (1992) Prediction of protein secondary structure by neural networks: Encoding short and long range patterns of amino acid packing. *Acta Biochim. Polon.* **39,** 378–392.

158. Vieth, M. & Kolinski, A. (1991) Prediction of protein secondary structure by an enhanced neural network. *Acta Biochim. Polon.* **38,** 335–351.

159. Handel, T. & DeGrado, W.F. (1992) A designed 4-helical bundle shows characteristics of both molten globule and native state. *Biophysical J.* **61,** A265.

160. Raleigh, D.P. & DeGrado, W.F. (1992) A De Novo designed protein shows a thermally induced transition from a native to a molten globule like state. *J. Am. Chem. Soc.* **114,** 10079–10081.

161. Raleigh, D.P., Betz, S.F. & DeGrado, W.F. (1995) A de novo designed protein mimics the native state of natural proteins. *J. Am. Chem. Soc.* **117,** 7558–7559.

162. Betz, S.F., Raleigh, D.P. & DeGrado, W.F. (1993) De novo protein design: From molten globules to native-like states. *Curr. Opin. Struct. Biol.* **3,** 601–610.

163. Betz, S.F., Bryson, J.W. & DeGrado, W.F. (1995) Native-like and structurally characterized designed α-helical bundles. *Curr. Biol.* **5,** 457–463.

164. Sander, C. (ed.) (1986) *Protein Design Exercises 86.* EMBL, Heidelberg.

165. Banner, D.W., Kokkinidis, M. & Tsernoglou, D. (1987) Structure of the ColE1 rop protein at 1.7 Å resolution. *J. Mol. Biol.* **196,** 657–675.

166. Gouda, H., Torigoe, H., Saito, A., Sato, M., Arata, Y. & Schimada, I. (1992) Three-dimensional solution structure of the B-domain of Staphylococcal Protein A: Comparisons of the solution and crystal structures. *Biochemistry* **40,** 9665–9672.

167. Hendrickson, W.A. & Teeter, M.M. (1981) Structure of the hydrophobic protein crambin. *Nature* **290,** 107–109.

168. Alber, T. (1992) Structure of the leucine zipper. *Curr. Opin. Genet. Develop.* **2,** 205–210.

169. Harbury, P.B., Zhang, T., Kim, P.S. & Alber, T. (1993) A switch between two-, three-, and four-stranded coiled coils in GCN4 leucine zipper mutants. *Science* **262,** 1401–1407.

170. Harbury, P.B., Kim, P.S. & Alber, T. (1994) Crystal structure of an isoleucine-zipper trimer. *Nature* **371,** 80–83.

171. Brooks, B.R., Bruccoleri, R., Olafson, B., States, D., Swaminathan, S. & Karplus, M. (1983) CHARMM: A program for macromolecular energy minimization, and molecular dynamics. *J. Comp. Chem.* **4,** 187–217.

172. Skolnick, J., Kolinski, A. & Ortiz, A.R. (1997) MONSSTER: A method for folding globular proteins with a small number of distance restraints. *J. Mol. Biol.* **265,** 217–241.

173. Gronenborn, A., Filpula, D.R., Essig, N.Z., Achari, A., Whitlow, M., Wingfield, P.T. & Clore, G.M. (1991) A novel, highly stable fold of the immunoglobulin binding domain of streptococcal protein G. *Science* **253,** 657–660.

174. Leijonmarck, M. & Liljas, A. (1987) Structure of the C-terminal domain of ribosomal protein 17/L12 from *Escherichia coli* at 1.7 Å resolution. *J. Mol. Biol.* **195,** 555–579.

175. Guss, J.M. & Freeman, H.C. (1983) Structure of oxidized poplar plastocyanin at 1.6 Å resolution. *J. Mol. Biol.* **169,** 521–563.

176. Katti, S.K., LeMaster, D.M. & Eklund, H. (1990) Crystal structure of thioredoxin from *Escherichia coli* at 1.68 Å resolution. *J. Mol. Biol.* **212,** 167–184.

177. Smith, W.W., Burnett, R.M., Darling, G.D. & Ludwig, M.L. (1977) Structure of semiqui-

none form of flavodoxin from *Clostridium* hp. Extension of 1.8 Å resolution and some comparisons of the oxidized state. *J. Mol. Biol.* **117,** 195–225.

178. Bolognesi, M., Onesti, S., Gatti, G., Coda, A., Ascenzi, P. & Brunori, M. (1989) Aplysia limacina myoglobin. Crystallographic analysis at 1.6 Å resolution. *J. Mol. Biol.* **205,** 529–544.

179. Smith-Brown, M.J., Kominos, D. & Levy, R.M. (1993) Global folding of proteins using a limited number of distance restraints. *Protein Eng.* **6,** 605–614.

180. Aszodi, A., Gradwell, M.J. & Taylor, W.R. (1995) Global fold determination from a small number of distance restraints. *J. Mol. Biol.* **251,** 308–326.

181. Bowie, J.U., Luethy, R. & Eisenberg, D. (1991) A method to identify protein sequences that fold into a known three dimensional structure. *Science* **253,** 164–170.

182. Jones, D.T., Taylor, W.R. & Thornton, J.M. (1992) A new approach to protein fold recognition. *Nature* **358,** 86–89.

183. Luethy, R., Bowie, J.U. & Eisenberg, D. (1992) Assessment of protein models with three dimensional profiles. *Nature* **356,** 83–85.

184. Madej, T., Gibrat, J.F. & Bryant, S.H. (1995) Threading a database of protein scores. *Proteins* **23,** 356–369.

185. Thornton, J.M., Flores, T.P., Jones, D.T. & Swindells, M.B. (1991) Prediction of progress at last. *Nature* **354,** 105–106.

186. Thomas, D.J., Casari, G. & Sander, C. (1996) The prediction of protein contacts from multiple sequence alignments. *Protein Eng.* **9,** 941–948.

187. Gobel, U., Sander, C., Schneider, R. & Valencia, A. (1994) Correlated mutations and residue contacts in proteins. *Proteins* **18,** 309–317.

188. Kolinski, A., Skolnick, J. & Godzik, A. (1997) A method for the prediction of surface "U"-turns and transglobular connections in small proteins. *Proteins* (in press).

189. Ponder, J.W. & Richards, F.M. (1987) Tertiary templates for proteins. Use of packing criteria in the enumeration of allowed sequences for different structural classes. *J. Mol. Biol.* **193,** 775–791.

190. Predki, P.F. & Regan, L.R. (1995) Redesigning the topology of a four helix bundle protein: Monomeric Rop. *Biochemistry* **34,** 9834–9839.

191. DeGrado, W.F., Wasserman, Z.R. & Lear, J.D. (1989) Protein design, a minimalist approach. *Science* **243,** 622–628.

192. Skolnick, J., Kolinski, A. & Godzik, A. (1993) From independent modules to molten globules: Observations on the nature of protein folding intermediates. *Proc. Natl. Acad. Sci. U.S.A.* **90,** 2099–2100.

193. Dyson, J.H. & Wright, P.E. (1993) Peptide conformation and protein folding. *Curr. Biol.* **3,** 60–65.

194. Eliezer, D., Jennings, P.A., Wright, P.E., Doniach, S., Hodgson, K.O. & Tsuruta, H. (1995) The radius of gyration of an apomyoglobin folding intermediate. *Science* **270,** 487–488.

195. Kim, P. & Baldwin, R.L. (1990) Intermediates in the folding reactions of small proteins. *Annu. Rev. Biochem.* **59,** 631–660.

196. Baldwin, R.L. & Roder, H. (1991) Characterizing protein folding intermediates. *Curr. Biol.* **1,** 219–220.

197. Lee, J. (1993) New Monte Carlo algorithm: Entropic sampling. *Phys. Rev. Lett.* **71,** 211–214.

198. Berg, B.A. & Neuhaus, T. (1991) Multicanonical ensemble: A new approach to simulate first-order phase transitions. *Phys. Rev. Lett.* **68,** 9–12.

199. Hansmann, U.H.E. & Okamoto, Y. (1993) Prediction of peptide conformation by multicanonical algorithm: New approach to the multiple minima problem. *J. Comput. Chem.* **14,** 1333–1338.

200. Kuwajima, K. (1989) The molten globule state as a clue for understanding the folding and cooperativity of globular protein structure. *Proteins* **6,** 87–103.

201. Kuwajima, K., Mitani, M. & Sugai, S. (1989) Characterization of the critical state in protein folding. Effects of guanidine hydrochloride and specific Ca^{2+} binding on the folding kinetics of α-lactalbumin. *J. Mol. Biol.* **206,** 547–561.

202. Ptitsyn, O.B. (1987) Protein folding: Hypotheses and experiments. *J. Protein Chem.* **6,** 273–293.

203. Ptitsyn, O.B., Pain, R.H., Semisotnov, G.V., Zerovnik, E. & Razgulyaev, O.I. (1990) Evidence for a molten globule state as a general intermediate in protein folding. *FEBS Lett.* **262,** 20–24.

204. McQuarrie, A.D. (1976) *Statistical Mechanics.* Harper & Row, New York.