APPLICATION OF A HIGH COORDINATION LATTICE MODEL IN PROTEIN STRUCTURE PREDICTION

A. KOLINSKI, P. ROTKIEWICZ
Department of Chemistry, University of Warsaw, ul. Pasteura 1
02-093 Warsaw, Poland
E-mail: kolinski@chem.uw.edu.pl

J. SKOLNICK
Department of Molecular Biology, The Scripps Research Institute (TCP-5)
10550 North Torrey Pines Rd. La Jolla, CA-92037, USA
E-mail: skolnick@scripps.edu

An efficient reduced model of protein structure, interaction and dynamics is proposed and evaluated. The model employs a high coordination lattice representation of protein conformational space. Only protein side chains are treated in an explicit way. Due to its high computational efficiency the new model enables studies of significantly larger protein systems at moderate resolution. We describe few applications in protein structure predictions based on sparse experimental data.

1 Introduction

Reduced models of proteins are very useful in studies of protein folding mechanisms, protein thermodynamics and dynamics.[1] The necessity to use a reduced level of representation is enforced by the enormous level of molecular complexities of protein-water systems. When a denatured polypeptide chain, under proper conditions, adopts a unique three-dimensional native structure, the protein folding process takes milliseconds to seconds.[2] The contemporary art of computing allows all-atom representation simulations of such systems within a time frame on the order of nanoseconds.[3] Replacing groups of atoms by "united atoms", while neglecting some degrees of freedom (for instance those within the side group chains) makes the problem tractable to some extent.

Various levels of simplification were found in the reduced models of proteins studied during the last 25 or so years.[1,4,5] Most commonly, a single interacting unit per residue or two interacting units per residue (one for main chain segment and one for the side group, where applicable) had been employed. All standard and the more special techniques of simulations were employed as an engine for conformational search, including various realizations of molecular dynamics, Monte Carlo methods, genetic algorithms and others.

In many cases, when a reduced model approach is adopted, it is quite reasonable to go one step further towards computational simplicity by assuming a discrete structure of protein conformational space. This leads to an entire spectrum of lattice models of polymers and proteins.[1] The lattice approach has several advantages. It enables much more efficient computations due to the smaller number of conformations to be considered. Moreover, lattice coarse-graining of rotational degrees of freedom leads to considerable smoothing of the energy landscape. A lot of local energy barriers, hopefully of little importance for the basic physics, are a priori

dismissed this way. It is the same for long-range interactions and the handling of excluded volume. A properly designed Monte Carlo scheme is a natural choice of sampling methodology for a given lattice model. For very simple models, full enumeration of conformational states (or of a relevant subset of conformations) is sometimes possible, leading to a complete description of the physics of such models.[5]

In this contribution, we present a short overview of protein lattice models, characterized by various levels of generalization and, consequently, by various levels of structural detail. Then, we propose a new type of lattice protein models. The reasons for designing yet another model are discussed. The new model is tested in some applications that could be useful in protein structure determination from sparse experimental data. First, a structure assembly from known secondary structure (in a three-letter code) and a few long-range restraints is presented, followed by a comparison of the fidelity and applicability of the new tool with the results of previous work. Then, the derivation of the short-range conformational propensities based on evolutionary information is described. This allows protein structure calculations from knowledge of only a few long-range contacts between polypeptide units. Other possible applications in protein structure prediction, including a purely de novo approach, and the study of long-time protein dynamics and protein thermodynamics are briefly discussed.

2 Lattice Protein Models

Lattice models of proteins and proteinlike systems can be built assuming various levels of resolution. So-called "simple exact models" employ simple cubic lattice chains, and each lattice bead represents a single amino acid residue.[5] These models could be studied in great detail due to the possibility of an exact enumeration of all compact states, which would lead to a good description of the model energy landscape. Extensive exploration of the sequence space is also possible for these models. Studies of simple exact models provide a general insight into the nature of hydrophobic collapse, explain the conditions necessary for stability of the folded state and explain some kinetic phenomena associated with protein folding. [6,7] This has been shown in other contributions to this book.

Slightly more complex models, based on the diamond lattice representation of protein chains, have been employed in studies of an interplay between short-range conformational propensities and tertiary interactions.[1,8] Such models enabled simulations of simple β-type and helical protein motifs. Some necessary conditions for the all-or-none folding transition to the unique globular state have been formulated in these studies. Other models of protein and proteinlike systems, assuming a similar level of resolution (fcc or bcc lattices for instance), also have been proposed and their analysis provided a convincing picture of the origins of cooperativity of protein folding.[9]

Over last few years, a series of "high-coordination lattice" models have been developed and used in studies of various aspects of the protein folding problem. An intermediate model of this type assumed a 24-member set of lattice vectors for representation of virtual bonds between protein alpha carbons. Such models allowed a crude representation of protein side chains with a proper chirality of the alpha car-

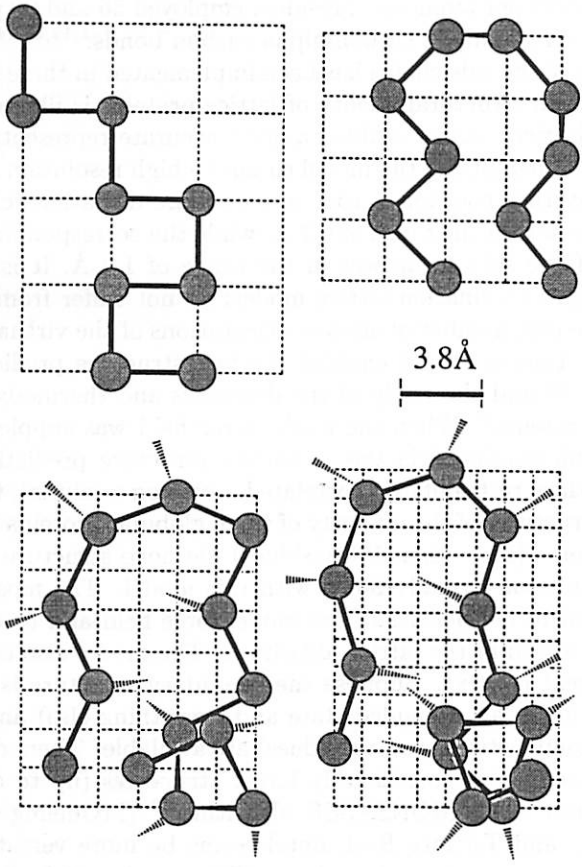

Figure 1: Schematic illustration of various lattice models (from the left top to the right bottom of the figure). The scale of all drawings with respect to protein geometry is maintained for easy comparison. (a) Simple cubic lattice polymer. Secondary structure could be defined only on a symbolic level. The beads of the chain correspond to amino acid residues and could be "colored", reflecting the various identities of amino acids. (b) A planar projection of a short fragment of the fcc (face centered cubic) chain. The model enables low resolution representation of some simple secondary structure motifs. (c) three-dimensional "Chess knight"- 24-vector protein model. This model could be used for modeling low resolution structures of all types of proteins, but with some distortions of secondary structure elements. Alpha carbon vertices are chiral and the side chains can be placed on-lattice or off-lattice. The definition of the side chain positions is provided by main chain geometry. (d) "310 Hybrid lattice" model of protein conformations. The set of virtual bonds between alpha carbons consists of 90-lattice vectors ($[3,1,1],\ldots[3,1,0],\ldots[2,2,1],\ldots[3,0,0],\ldots[2,2,0]\ldots$). Large number of the basis vectors and fluctuation bond length break-down lattice anisotropy and enable good flexibility in packing of the side chains that are defined with respect to the main chain. Internal degrees of freedom of the side chains is modeled by multiple allowed positions of the single-sphere model side groups.

bon units. Such representation enabled the modeling of all protein motifs and the low resolution simulation of the folding process of such complex real protein models as plastocyanin[10] and TIM-motifs.[11] Higher resolution models, based on the same idea of two interacting centers per amino acid residue, employed 56 and 90 possible lattice orientations of the virtual alpha carbon-alpha carbon bonds.[1,12,13,14] Multiple rotamer representation of the side chains has been implemented in these models. The increasing resolution and structural fidelity of lattice proteins is illustrated in Fig. 1. The "310-hybrid lattice" model enabled a quite accurate representation of the protein conformation. When fitting the model chains to high resolution crystallographic structures the average coordinate root mean square deviation (cRMSD) for the alpha carbon atoms was in the range of 0.7 Å, while the corresponding error for the centers of mass of the side group was in the range of 1.0 Å. It is important to note that such high coordination lattice models do not suffer from lattice anisotropy due to the large (90) number of allowed orientations of the virtual bonds of the model chain. This type of model enabled the test structure prediction of small and simple proteins, [15] and the study of the dynamics and thermodynamics of designed and natural proteins.[16] When the model force field was supplemented with some evolutionary information (via the secondary structure prediction and prediction of plausible tertiary restraints via correlated mutation analysis), the prediction of low resolution structures of the majority of small globular proteins became possible.[17] For larger proteins (more than 100 residues), de novo structure predictions by Monte Carlo simulations usually failed with this model. The most likely reasons could be attributed to deficiencies of the model force field and to the still insufficient computational speed of the lattice algorithm. The model also could be used in a somewhat different context. Suppose the secondary structure is known (in a three-letter code: a helix, an expanded state and everything else) and some long-range restraints (contacts between some residues) are available. Then, dependable and relatively fast assembly of considerably larger structures (up to ca. 150 residues) could be simulated. This MONSSTER algorithm [18] (MOdeling of New Structures from Secondary and Tertiary Restraints) seems be more versatile and more accurate than other corresponding algorithms. This may be useful for early stages of structure determination based on the results of NMR experiments.

3 Why Yet Another Lattice Model?

Briefly outlined high-coordination lattice models of proteins provide quite accurate representation of protein structure. With respect to corresponding continuous models, they are at least 100 times faster in computer simulations due to the implementation of a "prefabricated" conformational transition and the simplified computation of various elements of the interaction scheme. However, in such models, the motion of side chains and the main chain segments are to some extent decoupled. On one hand, it is a rather physical feature that enables the crude modeling of the various stages of protein folding: from topology assembly to collective adjustment of the structure resulting in side chain fixation, very much in the spirit of our view of the molten globule-native transition in real proteins. [1] There is, however, a price that has to be paid for such a defined structure of the model. The side chains are placed

in a continuous space; thus, a rather expensive algorithm detects overlapping and "contacts" of the side chains and main chain segment units (in all combinations). This is actually very far from the elegant simplicity of the low resolution proteinlike models, where excluded volume could be handled by a simple occupancy test of the lattice points (the contacts could be detected in a similar way). Is it possible to build a protein model that possesses virtues of both approaches, i.e., reasonable resolution that enables the modeling of some details of protein conformation and the simplicity of pure lattice models? The model described in this contribution apparently meets these requirements, at least to some extent. We start from the assumption that the dense packing within a protein is a very important feature of all protein systems. This actually is inspired by other studies of the simple exact models. Dense packing and amino acid specific interactions involve mostly side chains. The main chain segments are somewhat generic and could be perhaps treated in some models in an implicit way. Interactions and close packing of the side groups determine protein structure, provided that the main chain geometry satisfies proteinlike conformational restrictions. Moreover, the principles of the interplay between the main chain conformation and the side chain packing seem be so well defined in protein structures that they could be actually reversed. Provided a proper, proteinlike packing of the side chains, the geometry and interactions within the main chain are preserved. The above provides a conceptual framework for the definition of the model described in this work. The model employs single united-atom representation for the side groups of polypeptide chains. They are, however, bonded by a set of restrictions that accounts for a specific protein geometry in an implicit way, including conformational restrictions for the main chain conformations. An outline of such protein representation is given below. As shown later, such a very simple model is about two orders of magnitude faster in protein structure assembly with respect to a high coordination 310-hybrid lattice model with no apparent loss of accuracy.

4 Side Chain-Only Representation of Protein Conformational Space

Side chains of proteins possess their internal degrees of freedom. Let us assume that the centers of mass of the side chains in their actual rotational isomeric state serve as a center of interaction for side chains. For Gly residues, the center of mass is placed in the backbone Cα atoms. This defines a chain with virtual bonds connecting the centers of mass of the polypeptide side chains. The distance between two successive units of such a chain depends on the identity of the corresponding residues and on the actual rotameric state of their side chains. The distribution of these distances is quite broad. The shortest distance is observed between two successive Gly residues and that would be almost three times larger for a pair of residues with long side chains in an extremely expanded conformation. For a protein consisting of N amino acids, the model chain is defined by $n+1$ vectors $\{\mathbf{v}_i\}$, $i = 0, 1, \ldots n$, connecting $n+2$ united atoms. Two dummy residues are added for a convenient definition of "conformation" of the N- and C-terminal residues. The set of these virtual bonds $\{\mathbf{v}\}$ could be defined as $\{\mathbf{v}\} = \{\mathbf{a} \bullet \mathbf{q}\}$ where vectors $\mathbf{q}$ belong to the following set of lattice vectors:

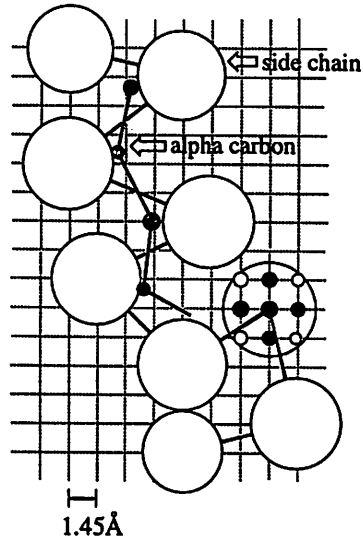

Figure 2: An example of a conformation of a short chain connecting centers of mass of a protein side groups. The upper part of the figure shows an example of extended conformation. The bottom part of the chain illustrates a helical turn. The large spheres illustrate the effective hard core of the model chain. Within one such sphere, the cluster of occupied lattice points is shown. The black dots correspond to three occupied sites along the axis orthogonal to the figure plane, the open ones to a single point in the plane. Exclusion of double occupancy of the lattice points leads to the effective excluded volume radius shown for all model residues. For a given position of a selected chain bead the number of points at which a second bead could be found, providing the closest distance approach is equal to 24. At slightly longer distances, the number of possible positions grows rapidly. Consequently, any effects of the lattice anisotropy are practically nonexistent in such a defined model. Approximate positions of the model alpha carbons could be easily computed as a linear combination of the coordinates of three successive side chains. A fragment of the resulting Cα chain is also shown in the figure. These coordinates could used for comparison with other models that handle Cα-chains in an explicit way.

$$\{\mathbf{q}\} = \{[\pm \mathrm{k}, \pm \mathrm{l}, \pm \mathrm{m}]\}$$
$$\text{with: } k,l,m = 0,1,2,3,4, \text{ or } 5 \text{ and } 11 \leq \mathbf{q}^2 \leq 30 \qquad (1)$$

The number of the lattice basis vectors $\mathbf{q}$ is equal to 592. With the assumed lattice spacing (the scaling constant with respect to the lattice) parameter a= 1.45 Å, the resulting distance between an arbitrary pair of side chains can change from 4.81 Å to 7.94 Å. This nicely covers the main portion of the distance distribution seen in real proteins that have an average value of about 6.6 Å and a standard deviation of about 1.3 Å. The wings of the distribution have been arbitrarily cut off. Such model chains can be fitted to the high resolution crystallographic structures with an accuracy of about 0.7-0.8 Å cRMSD (for the side chains centers of mass). The excluded volume of the chain units could be defined in the form of a cluster of 19 points of the underlying cubic lattice closest to a given chain bead. The cluster includes the central point, 6 simple cubic lattice points and 12 face-centered cubic lattice points. When each pair of two clusters is not allowed to overlap, the resulting hard core is characterized by the closest approach distance equal to 3 lattice units,

which corresponds to 4.35 Å. This is a somewhat underestimated representation of the excluded volume of real polypeptide chains; for larger residues, a soft core repulsive envelope has been added. Fig. 2. explains the geometric properties of the model chain.

A conformational updating scheme for such a chain could be designed in a very straightforward way. A single cycle of the Monte Carlo algorithm consists of several simple operations. First, the random number generator selects the chain bead. For the selected chain fragment, a random conformational transition is attempted and accepted provided the new conformation does not violate the excluded volume or assumed boundaries of allowed chain geometry. Transitions in models with energy parameters are additionally subject to the Metropolis criterion. The set of local moves consists of one- and two-bead motions, and rigid-bodylike small translations of somewhat longer fragments of the chains (of randomly selected lengths of up to 20 residues). These are illustrated in Fig. 3. The moves of a larger portion of the model chain are attempted less frequently. When the dynamics of the model chain are of interest, a model time scale can be easily set up. A conventional time unit corresponds to $f \cdot n$ attempts to various conformational transitions, where n is the chain length and f is a constant (range of 2) associated with the number of types of conformational transitions built into the algorithm.

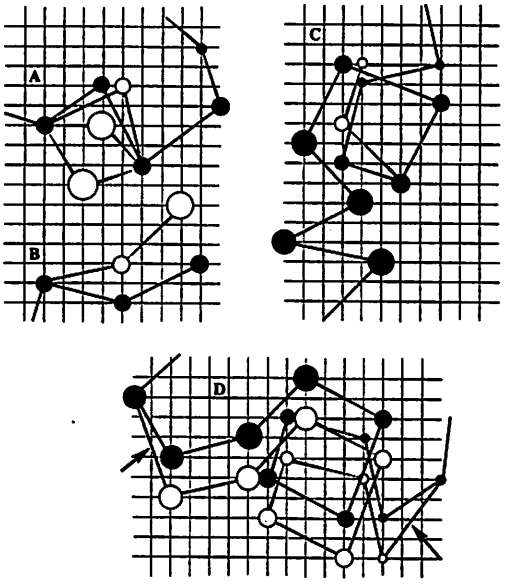

Figure 3: Elementary moves employed in the Monte Carlo sampling scheme. A - The single unit move, the number of alternative "new" conformations depicted by the open circles is actually many times larger than it is shown in the figure. B - An example of a random change of chain end conformation. C - An example of a two-bead (three-bond) conformational transition. D - A rigid-body type small translation of a larger fragment of the model chain. The arrows indicate the two virtual bonds that change upon this kind of conformational update. It is worthwhile to note that due to the large "coordination number" of the model chain, the described Monte Carlo dynamics avoid various ergodicity problems typical for low coordination lattice chains.

5 Modeling Generic "Proteinlike" Conformational Bias

The model chain defined in the previous section is very flexible. On other hand, proteins are relatively rigid copolymers with a specific chirality and strong short-range conformational restrictions. For example, the distribution of the distance between $i-th$ and $i+4^{th}$ side chains in globular proteins is bimodal, reflecting the existence of helices and expanded states. These and other properties of polypeptides have to be built into the model 20.

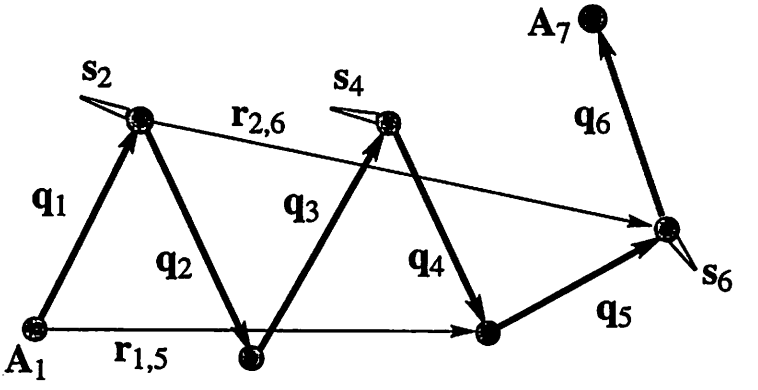

Figure 4: Illustration of the geometrical parameters employed in the design and definition of the short-range generic "proteinlike" conformational biases.

Let us consider a small fragment of the model chains consisting of beads numbered from 1 to 7 and the connecting vectors numbered from 1 to 6 (for reference see Fig. 4). Let us also define a set of vectors that are almost perpendicular to the planes defined by three successive chain beads according to the formula given below:

$$\mathbf{s}_i = \frac{\mathbf{q}_{i-1} \otimes \mathbf{q}_i - \mathbf{q}_{i-1} - \mathbf{q}_i}{|\,\mathbf{q}_{i-1}\,| \bullet |\,\mathbf{q}_i\,|} \tag{2}$$

A small deviation from orthogonality accommodates the super-twist of the secondary structure of protein chains. As a result, the vectors are exactly parallel along the long axis of a canonical helix; thus, these vectors are parallel to each other for all the residues within a helix. For regular expanded states, every second vector (and every fourth) along the polypeptide chains is parallel to one another. Consequently, it is very easy to introduce a bias toward some regular secondary structure conformations of the model chains. Of course, the bias cannot favor any specific secondary structure; rather, it should act as a bias against non-proteinlike irregular conformations. The following generic (sequence independent) potentials seem to serve this purpose very well:

$$\begin{aligned} &E_{2,4} = -\varepsilon_s \text{ for } \mathbf{s}_2 \bullet \mathbf{s}_4 > 0.25, \\ &\text{otherwise } E_{2,4} = 0 \end{aligned} \tag{3}$$

Of course, the same holds for beads 4 and 6. Additionally:

$$E_{2,6} = -\varepsilon_s \text{ for } \mathbf{s}_2 \bullet \mathbf{s}_6 > 0.25,$$
$$\text{otherwise } E_{2,6} = 0 \tag{4}$$

Further generic bias could be introduced via an energy price for regular secondary structure local geometry. Here also the gain is not dependent on a specific type of secondary structure.

$$E_{H1} = -\varepsilon_s \text{ for } \mathbf{q}_1 \bullet \mathbf{q}_3 < -5$$
$$\text{and when the local conformation is "helical-compact"} \tag{5}$$

$$E_{H2} = -\varepsilon_s \text{ for } \mathbf{q}_2 \bullet \mathbf{q}_4 < -5$$
$$\text{and when the local conformation is "helical-compact"} \tag{6}$$

$$E_{E1} = -\varepsilon_s \text{ for } \mathbf{q}_1 \bullet \mathbf{q}_3 > 5$$
$$\text{and when the local conformation is expanded} \tag{7}$$

$$E_{E2} = -\varepsilon_s \text{ for } \mathbf{q}_2 \bullet \mathbf{q}_4 > 5$$
$$\text{and when the local conformation is expanded} \tag{8}$$

where "helical-compact" is assumed in all cases when: $\mathbf{r}_{1,5}^2 < 32$ and $\mathbf{q}_1 \bullet \mathbf{q}_4 > 0$, and "expanded" for $60 < \mathbf{r}_{1,5}^2 < 125$ (in lattice units).

The next contribution to the generic potential propagates stretches of regular secondary structure and also increases the chains persistent length:

$$Ep = \begin{cases} 0 & \text{for } \mid \mathbf{r}_{2,6} - \mathbf{r}_{1,5} \mid^2 \leq 15 \text{ (in lattice units)} \\ \varepsilon_s \frac{\mid \mathbf{r}_{2,6} - \mathbf{r}_{1,5} \mid^2}{40} & \text{for } 15 < \mid \mathbf{r}_{2,6} - \mathbf{r}_{1,5} \mid^2 \leq 40 \\ \varepsilon_s & \text{for } 40 < \mid \mathbf{r}_{2,6} - \mathbf{r}_{1,5} \mid^2 \end{cases} \tag{9}$$

where $\mathbf{r}_{i,j}$ is the vector from the i-th to j-th chain bead.

Finally, a small bias, E_R, is introduced toward the right-handed conformation of compact states. This facilitates formation of right-handed helices in the absence of the sequence specific potential. When the sequence specific part of the potential is employed, which of course contains chiral components (see the next sections), this contribution has been omitted, i.e., $E_R = 0$.

$$E_R = -\varepsilon_s \text{ for } \mathbf{r}_{i-1,i+2}^2 < 32, \text{ and } (\mathbf{q}_{i-1} \otimes \mathbf{q}_i) \bullet \mathbf{q}_{i+1} > 0 \tag{10}$$

The total energy of the chain is the sum of all contributions and the scaling factors for particular terms have been adjusted to reproduce the secondary structure of globular proteins with good fidelity (when the sequence-dependent potential is added).

$$E_{gen} = \sum (0.25E_{2,4} + 0.25E_{2,6} + E_{H1} + E_{H2} + E_{E1} + E_{E2} + 0.5E_p + E_R) \tag{11}$$

where $\sum$ means summation along the chain. It should be pointed out that this somewhat complex set of potentials is controlled by a single energetic parameter ε_s.

The above set of biases does indeed lead to proteinlike geometry of the model chain. At proper temperature (or with a proper value of the ε_s parameter, $\varepsilon_s/kT =$

1 is a reasonable value) the distribution of local distance matches almost exactly the distributions seen in globular proteins. What is very interesting is that not only chirality (or rather handness of the chain), bimodal distribution of $r_{i,i+4}$ distance or other short-range geometry is reproduced in a semiquantitaive way, but also intermediate distributions because the distribution of $r_{i,i+8}$ distances becomes very similar to that seen in globular proteins. This is illustrated in Fig. 5, where some selected distributions of the distances between side chains obtained from the statistics of representative database of protein structures are compared to the corresponding distributions in the model chain simulated at T=1.

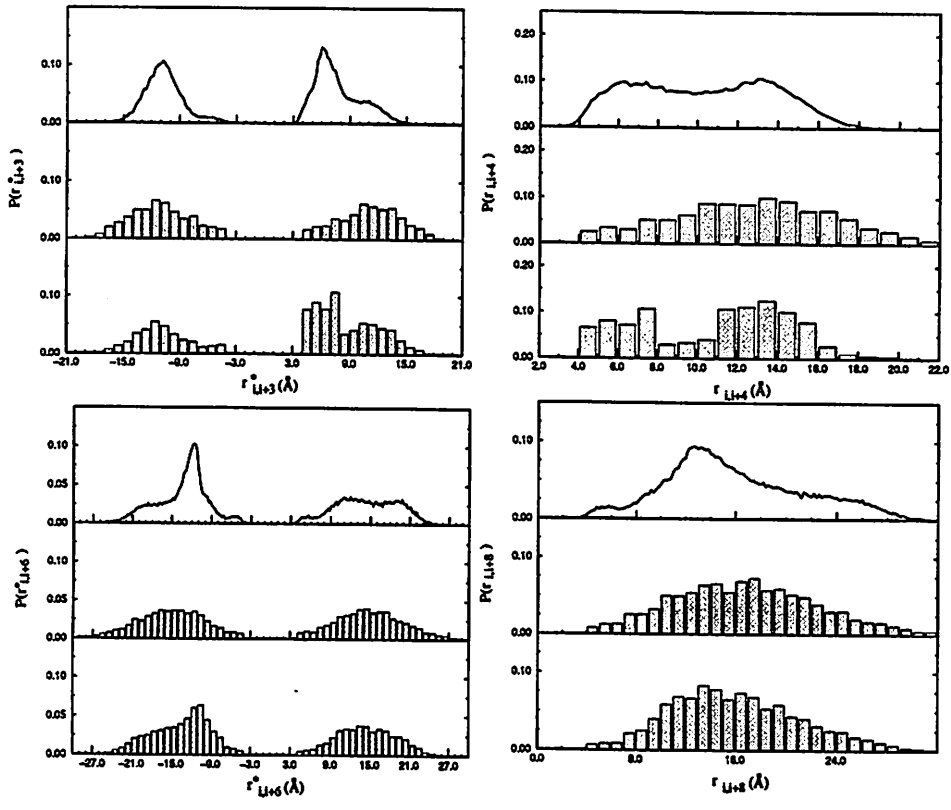

Figure 5: Comparison of distributions of distances between side chain centers of mass for known protein structures (extracted distributions neglected amino acid sequence identity, the top sections of each panel, smooth line) with the corresponding distributions for unrestricted model chains (the middle sections) and with that for the model chain controlled by the generic potential (extracted from long isothermal, T=1, simulation for the model chain consisting of N=100 units, in the bottom panels of each section). The plots are prepared for $r^*_{i,i+3}$, $r_{i,i+4}$, $r^*_{i,i+6}$ and $r_{i,i+8}$ distances, where (*) denotes a chiral value, i.e., the value multiplied by a sign depending on the fragment handness. In the $r^*_{i,i+6}$ case the handness is defined for a "superchain" constructed by connecting every second original chain bead.

6 Modeling the Secondary Preferences and the Long-Range Interactions.

The generic model of polypeptide chains described in the previous section could be made sequence specific in various ways. One of the simplest possible methods for designing the sequence specific secondary propensities employs statistics of pairwise dependent short range distances. The method is very similar to that used in the context of reduced models based on alpha carbon representation. The geometry employed in the definition of short-range interactions in the present studies is explained in Fig. 6.

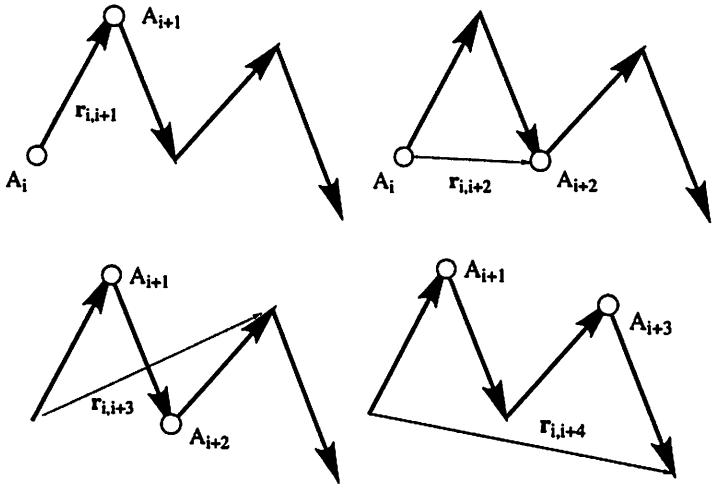

Figure 6: Intrachain distances employed in the definition of pair-dependent (amino acids marked in the figure) short-range potentials. In the case of $\mathbf{r}^*_{i,i+3}$, and $\mathbf{r}_{i,i+4}$ distances, the statistics also have been determined for the alternative (flanking) amino acids and taken into simulations with the same weigh factor as these shown in the figure.

The obtained potentials have several bin histograms, where the numerical values are obtained by comparison with the "generic" distributions:

$$\nu\left(A,B,\mathbf{r}_{i,i+k,m}\right)=\frac{-ln\left(f\left(A,B,\mathbf{r}_{i,i+k,m}\right)\right)}{f\left(\mathbf{r}^o_{i,i+k,m}\right)} \tag{12}$$

where $f(\mathbf{r}_{i,i+k,m})$ is the frequency of observation of the i-th bin of the $\mathbf{r}_{i,i+k,m}$ distribution, $f(\mathbf{r}^o_{i,i+k,m})$, denotes the database averaged, sequence-independent distribution. A and B denote identity of proper residues as marked in Fig. 6.

Similarly as for short-range interactions, a proteinlike generic bias could be introduced on the level of the tertiary (long-range) interactions. The generic terms consist of a cooperative "hydrogen-bond" scheme and a bias toward proteinlike local patterns of the side chain packing. The details are found elsewhere[21].

7 Assembly of Protein Structures from Known Secondary Structure and a Small Number of Tertiary Contacts

To check the performance of the proposed representation of protein structure and corresponding model interactions, we simulated the process of tertiary structure assembly given protein secondary structure (in the three-letter code) and a small number of known tertiary contacts.[21] This kind of experimental data are frequently available in the early stages of protein structure determination based on NMR experiments. The range of the number of tertiary restraints considered here is substantially smaller than that required for successful model building via more standard procedures. Recently, we developed a method of protein assembly from such sparse experimental data. The method employed alpha carbon plus side chain reduced representation of protein structure. A much simpler, present model was tested on a superset of protein sequences used previously.

A small number of tertiary restraints and loosely defined secondary structure is not sufficient for structure assembly. These restraints have to be superimposed on the top of the more general force field of the model. The force field consists of the short-range interactions described in the previous section, generic long-range interactions (a simple model of "hydrogen bonds", translated onto side group based model chain geometry, and some packing regularizing generic interactions) and sequence specific long-range interactions. These consist of a model of hydrophobic burial interactions and a pairwise interaction scale, which has been derived as a potential of mean force based on the statistics of side chain-side chain contact preferences in known protein structures.

The model of hydrophobic interactions consisted of several terms. First one regularizes model protein density under the assumption that we deal only with singe domain globular proteins. These are approximately spherical in shape and the radius of such spheres could be easily correlated with the number of residues in a folded polypeptide chain. Let us define protein radius of gyration:

$$S = \left(N^{-1} \sum (r_{CM} - r_i)^2 \right)^{\frac{1}{2}} \tag{13}$$

where r_{CM} indicated the center of mass of the chain, and r_i indicated coordinates of the center of mass of the i-th side chain. The following scaling relation could be derived from a proper statistical analysis of single domain proteins:

$$S = 1.52 N^{0.38} \text{ in lattice units} \tag{14}$$

The equivalent sphere could be derived on several shells, and a target number of residues in a particular shell could be prescribed to "an average" protein of a given size. Consequently, a density regularizing potential could be defined as follows:

$$E_b = \varepsilon_b \sum \mid m_{0,i} - m_i \mid \tag{15}$$

where $m_{0,i}$ is the target number of amino acids in shell i from the center of the globule. Three inner shell, of equal thickness covering the volume of the sphere of radius S contain slightly more than half of the protein residues. The entire protein

is essentially contained in a sphere with a radius equal to $\frac{5}{3}$S. The value of the parameter ε_b was assumed to be in the range of 1.0 k_BT. A conservative estimation of the radius of the protein hydrophobic core is about $\frac{2}{3}$S. Faster collapse of the model protein is facilitated by the implementation of potential energy equal to $-\varepsilon_{KD}(i)/16$ for all residues within the hydrophobic core. Here $\varepsilon_{KD}(i)$ denotes the Kyte-Doolittle hydrophobicity parameter of the i-th residue. [22,23]

Protein residues have a different size and different shape. The fraction of the covered surface of a given amino acid being in contact with another amino acid depends on the identity of both partners. The effect is not additive. Appropriate parameters could be derived from the statistics of the known protein structures and translated into a number of occupied points in a coordination sphere. The corresponding potential describes a local burial effect.

$$E_{surface} = \varepsilon_{surf} \sum E_b\left(A_i, a_i\right) \tag{16}$$

where a_i is the covered fraction of the side chain of amino acid A_i and the resulting burial energies E_b could be derived from the proper statistics of the structural database. The scaling factor ε_{surf} for this term has been assumed to be equal to 0.25 in all reported simulations. The above approach to the hydrophobic interactions enables the omission of the previously defined centrosymmetric one-body potentials, and thereby opens up the possibility of extending the present approach to multidomain and multimeric proteins. In the present simulations, both terms have been used in parallel.

The pairwise contact potentials for the side chains are the same as used in the previous work. The restraints were also implemented in a similar fashion. The short-range restraints have been superimposed only with respect to the helical and extended (β-type assignment) conformations. They were implemented as a bias toward the loosely defined appropriate local geometry defined by four consecutive model chain vectors (virtual "bonds" between side chain centers of mass). Additionally, the proper "mixing rules" have been superimposed onto the model hydrogen bond network, i.e., hydrogen bonding between known helical residues and known β-residues has been ignored during the simulations, thereby penalizing hydrogen bonding inconsistent with known secondary structure.[21] The long-range restraints were defined as follows:

$$E_{ij,restrained} = \begin{cases} E^{rep} & \text{for } 3 \le r_{ij} < R_{i,j}^{rep} \\ \varepsilon_{ij} - 0.5 & \text{for } Ri, j^{rep} \le r_{ij} < R_{i,j} \\ \varepsilon_{res}\left(R_{i,j}^2 - R_{i,j}^{2\ rep}\right) & \text{for } R_{i,j} < r_{ij} < 10 \\ \varepsilon_{res}\frac{\left(100-\left(r_{ij}^2-100\right)\right)}{3} & \text{for } 10 < r_{ij} \end{cases} \tag{17}$$

where, E^{rep} denotes a finite (5 k_BT) repulsive interactions for larger amino acids, and $R_{i,j}^{rep}$ is the cut-off distance. The value of parameter ε_{res} in structure assembly runs was set equal to $\frac{1}{8}$, while during the low temperature refinement run needed for the proper identification of the lowest energy structures, it was set equal to $\frac{1}{2}$. Ten proteins of various sizes and various secondary structure classes have been tested in the present work. The secondary structures have been taken from the Kabsch-Sander assignment [24] (reduced to a three-letter code) and the pairs of

tertiary restraints have been generated randomly from the native contact maps. For proteins that were the subject of our previous related work, we used the same sets of long-range restraints. In the cases when the number of restraints is smaller than in the previous work, we took a subset of previously used restraints. [18]

In all cases, the folding simulations lead in a prevailing fraction of experiments to a proper fold of a moderate resolution. All misfolded structures (one per five to ten simulations) were identified as a topological mirror image folds. These could be easily dismissed based on the conformational energy of resulting structures, i.e., the model force field properly identifies nativelike structures. Table I contains a concise comparison of the results obtained via the present model with the results of our previous study employing a more complex model.

Table 1: Comparison of results for the present simulation of protein structure assembly with the results obtained by the MONSSTER method.

Protein PDB name	Number of residues	Type of fold	Number of long-range restraints	cRMSD in Å from the present model[a,b]	cRMSD in Å from MONSSTER[a]
1gb1	56	α/β	8	3.4	3.3
1ctf	68	α/β	10	3.2	4.2
1pcy	99	β	46	3.8	3.5
1pcy	99	β	25	4.9	5.4
1pcy	99	β	15	5.7	—
2trx	108	α/β	30	3.1	3.4
2trx	108	α/β	16	3.5	—
4fab	113	β	27	4.4	—
4fab	113	β	16	5.9	—
3fxn	138	α/β	35	4.1	3.9
3fxn	138	α/β	20	4.1	—
1mba	146	α	20	4.3	5.9
Atim	247	α/β	62	5.1	—
Atim	247	α/β	50	6.0	—
Atim	247	α/β	36	6.7	—

[a] Average cRMSD of the Cα over an isothermal stability run.

[b] The average coordinate root mean square deviation from known tertiary structures 25 is reported from structures obtained from the present model based on side chain only representation, and converted into approximate coordinates of the Cα trace (computed as a simple linear combination of the three consecutive positions of the side chains).

Inspection of the data given in Table 1 leads to several observations. First, the average accuracy, measured by cRMSD from the native alpha carbon trace for the model presented in this work is no worse (in most cases, better) than that for the more complex MONSSTER model having two interaction centers per residue.[18] Second, proteins, especially the all-β type, could be assembled with a smaller number of restraints. One needs, at most, N/7 long-range restraints, where N is the number of residues; however, the fidelity of the assembled structures increases with the increasing number of long-range restraints. Third, much larger structures could

be assembled as has been shown for the 247 residue Atim protein. This is possibly due to the strictly lattice structure of the present model that makes simulations for these long proteins about two orders of magnitude faster. This may increase the possible applicability of the model in the early stages of protein structure determination from NMR data. Very few long-range restraints are necessary for building an approximate model. Then, the model can facilitate the identification of other NMR signals, providing additional restraints, and thereby the possibility of consecutive refinement of the structure of interest.

8 Local Sequence Similarity Based Potential for Short-Range Interactions

The statistical potential for short range interactions described in Section 6 can be replaced by a more elaborate potential of mean force employing sequence similarity and sequence-structure compatibility of short fragments of polypeptide chains.[20] The process of the derivation of such potentials for a given test sequence (in contrast to the simple statistical potential described before, this one has to be derived separately for each sequence of interest) could be outlined as follows:

(i) A multiple sequence alignment search is performed to find close homologues in the sequence database. When found, the homologous sequences were treated in exactly the same way as the test sequence, increasing the strength of the statistics. Only aligned fragments of the homologous sequences are taken into consideration.

(ii) The test sequence (and its existing close homologues) is divided into fragments by sliding a 19-residue window along the sequence. These are sequence fragments for which the potential is actually constructed.

(iii) The resulting set of N-18 sequence fragments (for the N residue test sequence) is then compared to all possible continuous fragments of the sequences from the structural database. The BLOSUM80 [26] sequence similarity criterion was employed for this purpose. For the central 9 residues of a 19-residue fragment, the value of the weight for the alignment scoring criterion was assumed to be equal to 1.0, while for the flanking residues, the weight decreased linearly up to a value of 0.1 for the first and 19th residue. The top 100 most sequentially similar fragments of protein structures are extracted for each 19-residues window of the test sequence. These fragments are then used for derivation of local residue-residue distance distributions for the test protein.

(iv) The distance distributions were collected for several intrachain distances, $r_{i,i+k}$, with k=1,2,3,4,6, and 8 for the central residues of the test window. The k=3 and k=6 distances were treated as "chiral", i.e., the distances were stored as negative numbers for left-handed conformations and as positive numbers otherwise. For the k=6 case, the chirality was defined using three consecutive vectors connecting every second side chain center of mass (vectors $r_{i,i+2}$, $r_{i+2,i+4}$ and $r_{i+4,i+6}$, respectively).

(v) The statistical potential of mean force was then generated by comparison of the observed distributions (in a form of histogram) of particular distances for the extracted top scoring structure fragments with the corresponding distribution for the entire structural database.

The obtained potentials could be symbolically written in the following way:

$$\nu\left(i, \mathbf{r}_{i,i+k,m}\right) = -ln\left(\frac{P\left(i, \mathbf{r}_{i,i+k,m}\right)}{P\left(\mathbf{r}^0_{i,i+k,m}\right)}\right) \tag{18}$$

where $P(i, \mathbf{r}_{i,i+k,m})$ is the weighted (by the sequence similarity criterion for the 100 top scoring fragments) probability of observation of the i-th bin of the $\mathbf{r}_{i,i+k,m}$ distribution, $P(\mathbf{r}^0_{i,i+k,m})$, and denotes the corresponding probability averaged over the entire structural database; i is the position along a given chain; and m denotes the bin number of the distribution histogram. Particular histograms consisted of 9, 7, 7, 10, 10 and 4 bins for $\mathbf{r}_{i,i+1}$, $\mathbf{r}_{i,i+2}$, $\mathbf{r}_{i,i+3}$, $\mathbf{r}_{i,i+4}$, $\mathbf{r}_{i,i+6}$ and $\mathbf{r}_{i,i+8}$, respectively. As mentioned before the potential for a given residue in the test sequence depends on a 19-residue window; consequently, the potential could be qualitatively different for the same central amino acid in two different protein sequences.

This potential, when applied together with the short-range generic structure regularizers discussed in previous sections, reproduces the secondary propensities of polypeptide chains with good fidelity. This could be observed in simulations of a protein chain in the absence of tertiary interactions.[20] When translated into a three-letter code, the average accuracy for the ten proteins listed in Table 1 is 72.2% when all homologous proteins (and proteins very similar to the test protein folds) are removed from the structural database. When the whole Fischers [27] database is employed in the derivation of the potential the secondary structure fidelity increases to 77.8%. It is worthwhile to note that when the potential is based only on homologous structures, then the secondary propensities are almost exact. Also, when the tertiary interactions are included, the secondary structure fidelity always increases; for some (but not all) structurally very simple proteins that fold to a nativelike compact state, it is almost exact. However, the outlined force field alone is insufficient to fold the majority of single domain proteins.

9 Protein Structure Assembly Based on Few Known Tertiary Contacts

Since the sequence similarity short-range potential described in the previous section reproduces secondary propensities of protein chains with good accuracy, it seems be natural to explore the possibility of structure assembly based on small number tertiary restraints only. The experiment is similar to that described in Section 7; however, here, knowledge of secondary structure is not assumed. Experimental methods exist that allow determination of some contacts between side chains. Thus, the model proposed here may be a valuable tool for prediction of low resolution protein structures. The folding experiments were done for a subset of proteins studied in section 7. We selected four proteins for the test simulations: 1gb1, 1ctf, 1pcy and 1mba. These proteins represent all structural classes of single domain protein folds. The sets of long-range restraints employed here were the same as those used in previous studies. We examined the possibility of folding these proteins with even smaller numbers of known tertiary contacts. In these simulations, we employed two times smaller number of restraints (every second have been extracted from the original sets). The results of experiments are summarized in Table 2.

Table 2: The results of folding experiments with a small number of long range restraints

PDB	Number of restraints	Number of simulations	range of cRMSD (in Å)	average cRMSD
1gb1	8	16/21*	2.6-4.2	3.2
1gb1	4	4/11	4.2-5.8	5.1
1ctf	10	7/8	3.9-5.0	4.5
1ctf	5	12/18	3.9-5.3	4.5
1pcy	25	7/8	4.6-5.5	5.0
1pcy	15	4/9	6.4-7.6	7.1
1mba	20	9/9	3.8-5.1	4.4

* n1/n2 - the number of successful (n1) folding experiments (all restraints satisfied and the topology of the fold correct per number of all experiments (n2) with satisfied restraints (misfolds have predominantly mirror image topology)

Inspection of the data given in Table 2 shows that the average accuracy of obtained structures is very similar to the accuracy of the structures obtained with assumed knowledge of secondary structure. The only exemption is the case of 1pcy, where the structures obtained with 15 restraints are of worse quality than those reported in Table 1. For the three remaining proteins, we attempted folding with an even smaller number of tertiary restraints: 4 and 5 for 1gb1 and 1ctf, respectively. 1gb1 and 1ctf fold under these conditions with reproducibility allowing easy identification of the nativelike structures. For 1mba with 10 restraints, the many folding experiments led to misfolded structures that violated some of the superimposed tertiary restraints. In the few successful experiments (all ten restraints satisfied, i.e., appropriate pairs of amino acids at contact distances), the resulting folds were correct. In all cases the lowest energy has been observed for one of the correct structures. Figures 7-10 show representative folded structures compared to the native structures. To prepare the ribbon diagrams[28], the Cα coordinates were determined using the approximate procedure mentioned before.

10 Conclusions

In this work, we have proposed a new reduced model of protein structure and dynamics. The model is based on the lattice representation of the protein side chain centers of mass. The main chain is treated in an implicit way. Due to the high effective coordination numbers of such model chains, the model has the advantage of both simple on-lattice and off-lattice representations. The cost of Monte Carlo simulations of protein dynamics or the folding process for this model could be estimated at about two orders of magnitude smaller than that for our previous lattice models of a similar resolution that treated the main chain and side groups in an explicit way.

The new model was tested in three special cases. First, we used this model for protein structure assembly given the knowledge of protein secondary structure (as a three letter code) and small number tertiary contacts. The new method [21] performed better than our MONSSTER method.[18] However, it should be men-

tioned that the MONSSTER method seems be more exact and more versatile than other previously published algorithms for structure assembly [29,30] from sparse experimental data; thus, the present development appears to be important. Next, we examined the possibility of implementing the homology-based statistical potential of short-range interactions within the framework of the new model. The possibility of quite accurate modeling of secondary propensities has been demonstrated. Subsequently, the model was employed for protein structure assembly given only a few long-range (tertiary) native contacts between protein side chains. The preliminary results reported in this work are very encouraging.

Acknowledgements

This work was partially supported by University of Warsaw grant BST 34/97. A.K is an International Scholar of the Howard Hughes Medical Institute.

References

1. Kolinski, A., Skolnick, J. Lattice models of protein folding, dynamics and thermodynamics. Austin, TX.: R. G. Landes, 1996.
2. Creighton, T. E. Proteins: structures and molecular properties. New York: W.H.Freeman and Company, 1993.
3. Brooks, C. L. I., Karplus, M., Pettitt, B. M. Protein: A theoretical perspective of dynamics, structure, and thermodynamics. *Adv. Chem. Phys.* **71**, 1-259 (1988).
4. Levitt, M. Protein folding. *Curr. Opinion Struct. Biol.* **1**, 224-229 (1991).
5. Dill, K. A., Bromberg, S., Yue, K., Fiebig, K. M., Yee, D. P., Thomas, P. D., Chan, H. S. Principles of protein folding - A perspective from simple exact models. *Prot. Sci* **4**, 561-602 (1995).
6. Dinner, A. R., Sali, A., Karplus, M. The folding mechanism of larger proteins: Role of native structure. *Proc. Natl. Acad. Sci. USA* **93**, 8356-8361 (1996).
7. Karplus, M., Sali, A. Theoretical studies of protein folding and unfolding. *Curr. Opinion Struct. Biol.* **5**, 58-73 (1995).
8. Skolnick, J., Kolinski, A. Computer simulations of globular protein folding and tertiary structure. *Annu. Rev. Phys. Chem.* **40**, 207-235 (1989).
9. Go, N. Theoretical studies of protein folding. *Ann. Rev. Biophys. Bioeng.* **12**, 183-210 (1983).
10. Skolnick, J., Kolinski, A. Simulations of the folding of a globular protein. *Science* **250**, 1121-1125 (1990).
11. Godzik, A., Skolnick, J., Kolinski, A. Simulations of the folding pathway of TIM type α/β barrel proteins. *Proc. Natl. Acad. Sci. USA* **89**, 2629-2633 (1992).
12. Kolinski, A., Godzik, A., Skolnick, J. 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 (1993).

13. Kolinski, A., Skolnick, J. Monte Carlo simulations of protein folding. I. Lattice model and interaction scheme. *Proteins* **18**, 338-352 (1994).
14. Kolinski, A., Skolnick, J. Monte Carlo simulations of protein folding. II. Application to protein A, ROP, and crambin. *Proteins* **18**, 353-366 (1994).
15. Godzik, A., Kolinski, A., Skolnick, J. Lattice representation of globular proteins: How good are they? *J. Comp. Chem.* **14**, 1194-1202 (1993).
16. Kolinski, A., Galazka, W., Skolnick, J. On the origin of the cooperativity of protein folding. Implications from model simulations. *Proteins* **26**, 271-287 (1996).
17. Ortiz, A. R., Kolinski, A., Skolnick, J. Nativelike topology assembly of small proteins using predicted restraints in Monte Carlo folding simulations. *Proc. Natl. Acad. Sci. USA* , (1997).
18. Skolnick, J., Kolinski, A., Ortiz, A. R. MONSSTER: A method for folding globular proteins with a small number of distance restraints. *J. Mol. Biol.* **265**, 217-241 (1997).
19. Skolnick, J., Kolinski, A. Dynamic Monte Carlo simulations of a new lattice model of globular protein folding, structure, and dynamics. *J. Mol. Biol.* **221**, 499-531 (1991).
20. Kolinski, A., Jaroszewski, L., Rotkiewicz, P., Skolnick, J. An efficient Monte Carlo model of protein chains. Modeling the short-range correlations between side group centers of mass. *J. Phys. Chem.* , (in press).
21. Kolinski, A., A., S. Assembly of protein structure from sparse experimental data: An efficient Monte Carlo Model. *Proteins* , (in press).
22. Kyte, J., Doolittle, R. F. A simple method for displaying the hydrophatic character of protein. *J. Mol. Biol.* **157**, 105-132 (1982).
23. Eisenberg, D., McLauchlan, A. D. Solvation energy in protein folding and binding. *Nature* **319**, 199-203 (1986).
24. Kabsch, W., Sander, C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**, 2577-2637 (1983).
25. Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer Jr, E. F., Brice, M. D., Rodgers, J. R., Kennard, O., Simanouchi, T., Tasumi, M. The protein data bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.* **112**, 535-542 (1977).
26. Henikoff, S., Henikoff, J. G. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. USA* **89**, 10915-10919 (1992).
27. Fischer, D., Eisenberg, D. Protein fold recognition using sequence-derived predictions. *Protein Sci.* **5**, 947-955 (1996).
28. Koradi, R. MOLMOL: a program for display and analysis of macromolecular structures. *J. Mol. Graph.* **14**, 51-55 (1996).
29. Smith-Brown, M. J., Kominos, D., Levy, R. M. Global folding of proteins using a limited number of distance restraints. *Prot. Engn.* **6**, 605-614 (1993).
30. Aszodi, A., Gradwell, M. J., Taylor, W. R. Global fold determination from a small number of distance restraints. *J. Mol. Biol.* **251**, 308-326 (1995).

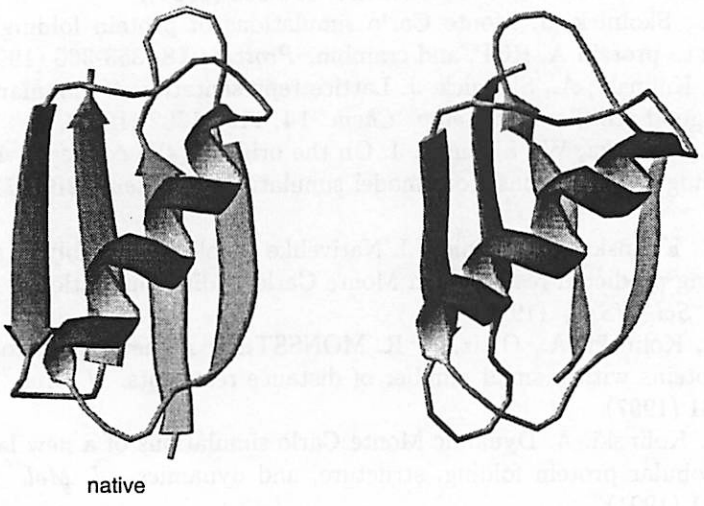

Figure 7: Comparison of the native structure of 1gb1 with a representative structure obtained from simulation with 4 long range restraints.

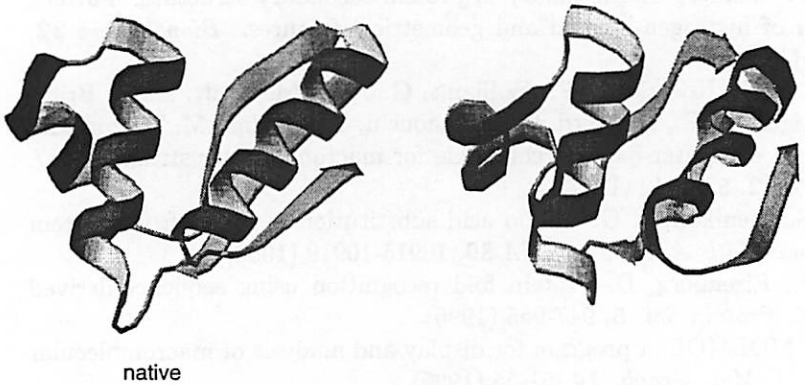

Figure 8: Comparison of the native structure of 1ctf with a representative structure obtained from simulation with 5 long range restraints.

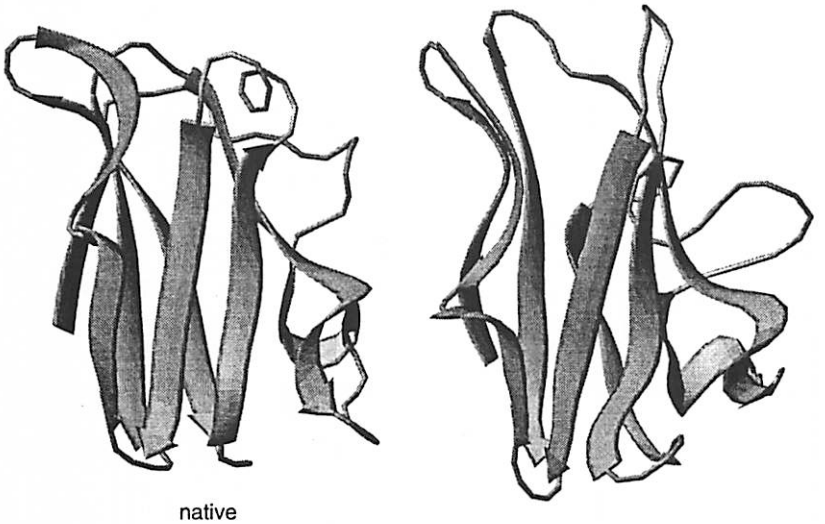

Figure 9: Comparison of the native structure of 1pcy with a representative structure obtained from simulation with 25 long range restraints.

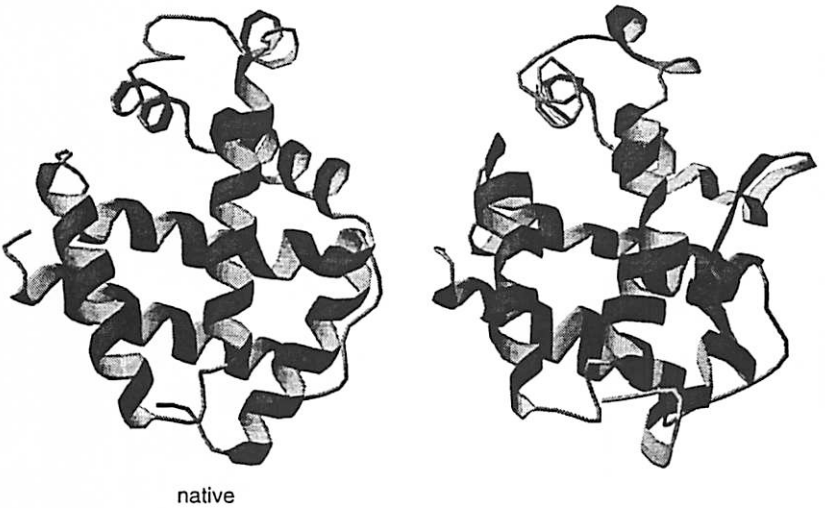

Figure 10: Comparison of the native structure of 1mba with a representative structure obtained from simulation with 20 long range restraints.