Full Paper: A reduced model of polypeptide chains and protein stochastic dynamics is employed in Monte Carlo studies of the coil-globule transition. The model assumes a high-resolution lattice representation of protein conformational space. The interaction scheme is derived from a statistical analysis of structural regularities seen in known three-dimensional protein structures. It is shown that model polypeptides containing residues that have strong propensities towards locally expanded conformations collapse to β-like globular conformations, while polypeptides containing residues with helical propensities form globules of closely packed helices. A more cooperative transition is observed for β-type systems. It is also demonstrated that hydrogen bonding is an important factor for protein cooperativity, although, for systems with suppressed hydrogen bond interactions, a higher cooperativity of β-type proteins is also observed.

Helix-coil and beta sheet-coil transitions in a simplified, yet realistic protein model

Dedicated to Professor *Oskar Friedrich Olaj* on the occasion of his 65th birthday

Bartosz Ilkowski,[1] *Jeffrey Skolnick,*[2] *Andrzej Kolinski**[1, 2]

[1] Department of Chemistry, University of Warsaw, ul. Pasteura 1, 02-093 Warsaw, Poland
Fax: (+48)-22-8225996; E-mail: kolinski@chem.uw.edu.pl
[2] Donald Danforth Plant Science Center, 7425 Forsyth Blvd., St. Louis, 63130 MO, USA

1. Introduction

The conformational space of globular proteins is enormous.[1] A direct search for both the denatured state and folded state by means of traditional molecular mechanics (Molecular Dynamics or Monte Carlo method for all-atom detailed models) is, at present, not feasible.[2] Thus, simplified models are very helpful in understanding the general principles of protein dynamics and thermodynamics and may contribute to a partial solution of the protein folding problem – the theoretical prediction of the protein's three-dimensional structure from the sequence of amino acids in its polypeptide chain. There are various levels of simplification of protein-chain representations described in the literature; these range from simple lattice copolymers[3] to more complex lattice and off-lattice models.[4, 5] The simplest models could be treated in an exact (or almost exact) fashion, thereby providing valuable insight into the most fundamental aspects of protein-folding thermodynamics. On the other hand, meaningful studies of the effects of sequence on specific protein features may require more detailed models.[5, 6] For example, this is the case for the ab initio prediction of the approximate three-dimensional protein structure[7] and protein folding pathways.

Recently, we have developed a simplified lattice model. The simplifications were the result of a compromise between computational tractability and a meaningful, low resolution representation of specific polypeptide sequences. For some very simple, and relatively small proteins (in the range of 100 amino acids) this model enables the ab initio simulation of the entire folding trajectory from a random coil state to a well defined, low-resolution collapsed native state.[8] For more complex proteins, it was possible to assemble approximate structures with the aid of some experimentally or theoretically derived restraints.[9, 10] The model was also employed in a detailed analysis of the thermodynamics and dynamics[11] of the assembly of C-terminal β-hairpin of the B1 domain of protein G. The results of the simulations were in semi-quantitative agreement with recent experimental studies.[11]

The model employed here assumes a lattice representation of a virtual chain connecting the centers of mass of the side groups of a polypeptide chain. The force field of the model has been developed by means of statistical analysis of various structural regularities seen in known protein structures. Similar to our previous work, the Monte Carlo process with local conformational transitions simulates the stochastic dynamics of the model chains. In contrast to the previous applications, here we analyze simplified (realistic, but exaggerated) sequences and the coil-to-helix (or coil-to-beta sheet) transitions. Effects of the sequence-specific short- and long-range interactions and the strength of the model hydrogen bonds on the collapse transition is modeled and analyzed.

1022-1344/2000/0811–0523$17.50+.50/0

2. Protein representation and model of Monte Carlo dynamics

The majority of reduced protein models assume a C_α-representation of the polypeptide chain, i.e., the main-chain backbone is reduced to the alpha-carbon trace, i.e., the chain of virtual bonds connecting subsequent alpha carbons.[6, 12] In more detailed models, reduced representations of the side groups were sometimes introduced.[13] Here, we propose a slightly different philosophy. Namely, the chain modeled in an explicit way connects the centers of mass of the side groups. A number of arguments support such an approach. First, it is known that the sequence-specific interactions in polypeptides involve their side chains rather than the main-chain units. Side groups are closely packed. There is a well-defined peak from the first coordination sphere in the radial distribution function for side groups in folded proteins. This is not the case for alpha carbons; contacts between various elements of secondary structure lead to very broad distribution of distances between closest alpha carbons. The main-chain, long-range interactions in proteins are rather generic, and to a large extent, sequence independent. Second, having reasonable positions of the side chains in the folded structure, the rebuilding of a good approximation of the main-chain geometry is quite easy and straightforward. The opposite procedure, i.e., rebuilding of the side-chain positions from an alpha-carbon trace is significantly more difficult. In other words, when assuming a single, explicitly treated unit per amino acid, centers of side groups seems to be a better choice than main chain units (alpha carbons). The side chain option enables more straightforward simulations of closely packed protein-like structures. Moreover, due to well-defined packing of side groups, simple contact potentials are better justified and more accurate.

The model of a polypeptide chain composed of N amino acid residues consists of N + 1 vectors connecting the centers of mass of side chains in their actual rotational isomeric state. Two additional vectors on the chain ends define orientations of the N-terminal and C-terminal caps of the polypeptide. The centers of mass of the original side chains are computed for all heavy atoms (all atoms except hydrogen atoms) of the side chains + alpha carbons. For instance, the center of interaction for glycine coincides with C_α, for alanine it is located in the center of the C_α-C_β bond and for valine the interaction center is placed at the C_β position. The coordinates of the interaction centers are restricted to the nearest knots of the underlying simple cubic lattice with a lattice spacing corresponding to 1.45 Å (Angstroms) in real proteins. To cover a wide distribution of the allowed distances between the nearest (along the chain) centers of interaction, the chain vectors have the form of $\mathbf{r}_i = (\pm j, \pm k, \pm l)$, with j, k, l = 0, 1, 2, 3, 4 or 5 and $9 \le |\mathbf{r}_i| \le 30$, in lattice units. There are 646 possible bond vectors. The shortest vectors, type of (2, 2, 1) or (3, 0, 0) correspond to a distance of 4.35 Å. The longest vectors, type of (5, 2, 1), correspond to 7.94 Å. Thus, the wings of the distribution seen in real proteins have been cut off. This does not introduce any significant error, although it somewhat reduces the excess entropy of the model chain. A hard-core excluded volume is associated with each center of interaction. It has the form of a symmetric cluster of 19 lattice points; the central one, six nearest neighbors of the simple cubic lattice, and 12 second neighbors on distance $2^{1/2}$ lattice units from the central one (see Fig. 1). The distance of closest approach of two such clusters is equal to 3 lattice units and the number of relative orientations for the closest approach is equal to 30. For larger side groups, the hard core is supplemented by a soft repulsive spherical envelope which extends (depending on the pair of amino acids involved) up to $12^{1/2}$ lattice units. For a number of purposes, it is important to know the approximate

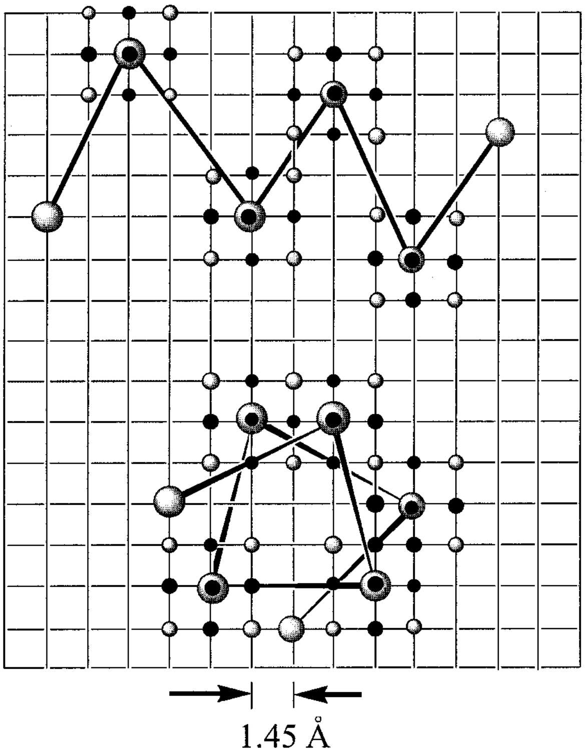

Fig. 1. Illustration of the model chain design. Two fragments (expanded, β-type in the upper part of the picture and helical in the lower part) are shown in a projection along the Z-axis. The expanded fragment is placed in the XY plane, the helix axis is along the Z-axis of the lattice. The larger gray spheres correspond to the side chain interaction centers. Smaller dots illustrate the excluded volume clusters associated with each residue. The black dots indicate three occupied lattice points along the Z-axis, the gray ones indicate single points in the XY plane of the interaction center. Each excluded volume cluster consists of 19 lattice points that are subject to single-occupancy tests in the Monte Carlo algorithm. The spacing of the underlying cubic lattice is equal to 1.45 Å.

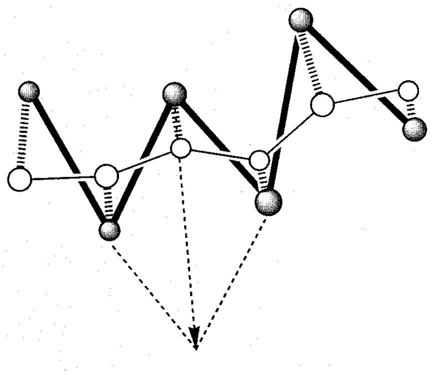

Fig. 2. Illustration of the procedure used for approximation of the alpha carbon positions. Given the side group chain (gray spheres and thick solid virtual bonds of the model chain) Cα positions are placed on the bisector of the chain fragment, and are not restricted to the lattice nodes. The distance from the appropriate side chain is residue dependent and could be extracted from the statistical analysis of the database of protein structures. For alanine such distance is equal to half of the length of a C—C bond (about 1/2 of a lattice unit), for valine it is two times larger, etc.

positions of the alpha-carbon atoms. It could be shown that for a given amino acid side chain, the position of the C_αs could be approximated from the positions of the three consecutive units of the model chain. The systematic and random errors of such a procedure are below the resolution of the model. The idea is explained in Fig. 2.

Monte Carlo sampling employs a random sequence of single-residue (interaction center) moves, moves of two and three subsequent chain units and small (rigid body like) shifts of larger randomly selected chain fragments. All moves are subject to bond length restrictions. Long sequences of such moves constitute the numerical solution of a stochastic equation of motion and thereby simulate the coarse-grained dynamics of model polypeptides.

The model outlined above somewhat resembles the "fluctuating bond" model[14] that is widely employed in studies of polymer dynamics.[15] It is assumed that the fluctuating-bond model significantly reduces the effects of lattice anisotropy (when compared to simple lattice models) and is ergodic in proper Monte Carlo sampling schemes. The present model is of higher resolution with some important geometrical properties of polypeptides encoded in the chain representation and in the force field of the model. Thus, by analogy to the fluctuating bond model, it is rather unlikely that this model exhibits any noticeable lattice anisotropy or experiences ergodicity problems.[15]

3. Interaction scheme

When taken at the athermal limit, the model outlined above has all of the properties of a random coil chain with excluded volume. However, proteins are quite special polymers; their chains are relatively stiff with specific short-range conformational correlations. Some properties of proteins are rather generic, while others depend on the sequence of amino acids in the polypeptide chain. Also, long-range interactions need to be designed in such a way that specific, protein-like packing in a globular state could be achieved at low temperatures. All of these potentials are derived from a proper statistical analysis of known folded structures of globular proteins. The underlying assumption is that such knowledge-based statistical potentials could be also applied to the denatured states of proteins. This supposition is difficult to prove in a rigorous way; however, this procedure is commonly accepted in studies of proteins, and there are strong reasons to believe that at least qualitatively this assumption is correct. Various components of the model force-field are briefly outlined below.

3.1 Modeling protein-like chain stiffness

Generic (sequence independent) short-range interaction potentials were designed to mimic protein-like chain stiffness. Details could be found elsewhere[9–11]. Here, a short summary is provided for the reader's convenience. Short fragments of polypeptide chains in proteins tend to adopt two types of conformations: compact ones, characteristic of helices and turns, and expanded ones, characteristic of β-sheets and expanded loop regions. As a result, the distribution of the distance between the ends of five residue fragments (connected by four virtual bonds) in proteins is bimodal. This should be contrasted with the single-maximum, Gauss-type distributions seen in flexible polymer models. Thus, a bias towards either of these conformations is introduced into the models. The energy gain ($-\varepsilon_g$) associated with protein-like conformations is the same for the right-handed helical turns and for expanded conformations, with the proper ranges of specific secondary structure geometry extracted from a statistical analysis of known protein structures. Additionally, the system is awarded the same negative energy increment, $-\varepsilon_g$, when such geometry extends by the next chain unit. For instance, this is the case when, for a given four-vector helical turn, the next vector also forms a helical turn with the three previous vectors. Such biases induce a bimodal distribution of the distance between the i-th and $i+4^{th}$ residues, and, in absence of the sequence-specific interactions, they enforce protein-like distributions of other short-range geometrical properties (bond angles, dihedral angles, and distances).

3.2 Sequence dependent short-range interactions

Short-range, sequence dependent, conformational preferences of model polypeptides were encoded in four poten-

Tab. 1. Short-range potential for the most specific interactions between i-th and i + 3rd residues. Negative values of $R^{*}_{i,i+3}(A_i, A_{i+3})$ are for left-handed conformation of three-vector fragments with appropriate residues on the ends of the fragment.

	Range of $R^{*}_{i,i+3}(A_i,A_{i+3})$ (in Angstroms)													
	−12	-12, -10,	-10, -8,	-8, -6	-6, -4	-4, -2	-2, 0	0, 2	2, 4	4, 6	6, 8	8, 1 0	10, 12	12
VAL THR	1.75	−1.60	−0.91	1.01	1.86	2.00	2.00	2.00	2.00	−0.80	−0.53	0.25	−0.10	2.00
THR VAL	1.47	−1.68	−0.75	0.99	2.00	2.00	2.00	2.00	2.00	−0.81	−0.58	0.34	0.10	1.95
ALA SER	2.00	−0.80	−0.97	0.87	2.00	2.00	2.00	2.00	2.00	−1.62	−0.01	−0.49	0.13	2.00
LEU ALA	1.38	−0.72	−0.66	1.46	2.00	2.00	2.00	2.00	2.00	−1.22	−1.57	0.51	0.57	2.00
SER ALA	2.00	−0.75	−0.73	1.24	1.89	2.00	2.00	2.00	2.00	−1.70	−0.60	0.01	0.07	2.00
SER SER	2.00	−1.05	−0.81	0.82	1.67	2.00	2.00	2.00	1.85	−1.33	−0.37	−0.48	−0.12	2.00
ALA ALA	2.00	−0.35	−0.61	1.25	2.00	2.00	2.00	2.00	2.00	−2.08	0.04	0.06	0.95	2.00
ALA LEU	1.96	−0.46	−0.61	1.38	1.73	2.00	2.00	2.00	2.00	−1.95	−0.64	0.48	0.76	2.00

tials controlling short-range distances between chain units. Distances between i-th and i + k (k = 1, 2, 3, 4) residues were considered for all possible combinations of end residues in corresponding fragments. The distances between the i-th and i + 3rd residues were assumed to be "chiral" i.e., the handedness sign of the three-vector fragment defined the distance as being negative or positive, respectively. The statistics from the database of protein structures for all pairs of amino acids at appropriate positions were collected in the form of histograms (three bins for k = 1 and k = 2, 14 bins for k = 3 and 7 bins for k = 4) and related to flat reference-state distributions. The resulting potentials reflect the secondary-structure propensities encoded in protein sequences. For instance, with valine (Val) residues at position i and i + 4, the $r_{i,i+4}$ (Val, Val) has negative values for larger values of the corresponding distance, typical for β-sheets, while $r_{i,i+4}$ (Ala, Ala) for the helix-forming residue alanine (Ala) has a minimum for smaller distances corresponding to helical conformations. Numerical data for all potentials can be extracted from our web site.[16] Example data for the most specific short-range potentials are given in Tab. 1.

3.3 *Long-range pairwise interactions*

Long-range pairwise interactions are distance dependent (see Fig. 1). Distances up to 3 lattice units are prohibited. Then there is a finite-strength repulsive core, applicable to pairs of larger amino acids. Next, there is a square well of the pair-dependent interactions. The cut-off distance depends on the pair of amino acids involved. The values of the sequence specific part of the pairwise potential depend on the mutual orientation of the interacting units. There are separate values of these potentials for parallel contacts, acute/orthogonal contacts, and antiparallel contacts. For some pairs of amino acids this effect of orientation is large. There are several reasons for this effect. First, our model is reduced to a single interaction center per side chain. Side groups, especially larger, are not spherical. When they are closely packed, some mutual orientations could be energetically favorable. Second, in folded structures, polar and charged amino acids are located on the protein surface, with most side chains interacting in a parallel fashion. Since our model needs to discriminate not only against incorrectly folded structures but also against unfolded random structures, such a preference towards parallel packing needs to be accounted for in the interaction scheme. Indeed, statistical potentials for pairs of amino acids having opposite charges are strongly attractive for parallel orientations but repulsive for antiparallel contacts. The antiparallel packing of charged residues in proteins is extremely rare. The orientation dependent contact potential can also account for some more complex effect in protein packing like the acute orientation of aromatic rings, etc. Of course, the statistics for the derivation of such potentials is about three times worse than for simple orientation-independent pairwise potential. Fortunately, present databases of experimentally determined protein structures are large enough to provide (on average) hundreds of examples for every pair of amino acids and every mode of mutual packing. The numerical data are enclosed in Tab. 2. To increase the strength of tertiary interactions, all values of

Tab. 2. Values of the pair-specific potential of the long-range interactions for the residues included in the designed sequences.

Parallel contacts					
	ALA	SER	VAL	THR	LEU
ALA	−0.1	−0.2	−0.4	−0.1	−0.4
SER		−0.5	−0.2	−0.5	0.0
VAL			−1.0	−0.3	−1.0
THR				−0.5	−0.3
LEU					−1.2
Acute/orthogonal contacts					
	ALA	SER	VAL	THR	LEU
ALA	0.1	0.3	−0.2	0.2	−0.2
SER		0.1	0.4	0.2	0.4
VAL			−0.2	0.3	−0.4
THR				0.3	0.4
LEU					−0.5
Antiparallel contacts					
	ALA	SER	VAL	THR	LEU
ALA	−0.1	0.3	−0.5	0.2	−0.4
SER		0.3	0.3	0.5	0.3
VAL			−0.6	0.3	−0.4
THR				0.6	0.3
LEU					−0.4

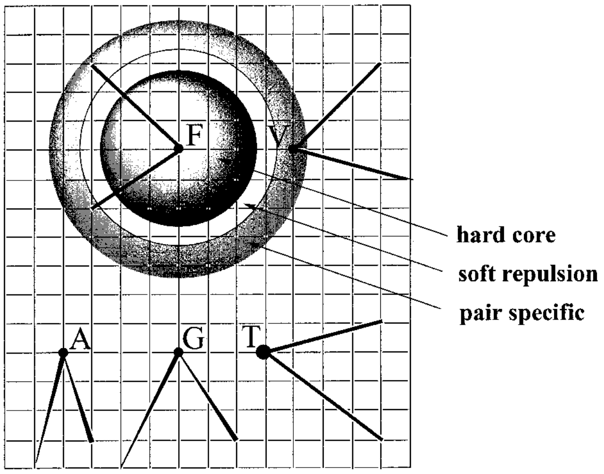

Fig. 3. Long-range interactions between side-chains. The central gray sphere corresponds to the hard-core excluded volume, the white envelope corresponds to the finite-strength repulsive interactions (for larger residues) and the gray envelope illustrates the range for pair-specific square-well potential. Examples of three types of contacts are shown in the graph: antiparallel for F and V residues, acute/orthogonal for G and T residues and parallel between A and G residues, respectively.

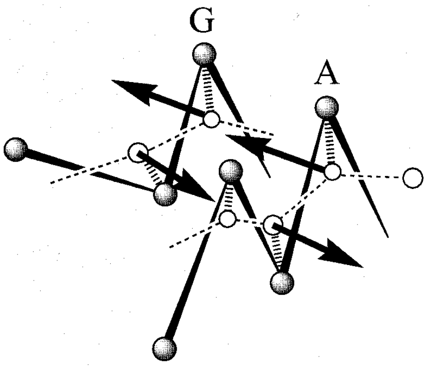

Fig. 4. Geometry of the model hydrogen bonds. Residues A and G are "hydrogen bonded" because the following geometrical criteria are satisfied. First, the alpha carbons of these residues are close to each other (with a distance less than 4.5 lattice units). Second, their corresponding chain fragments interact in a "parallel" fashion. Namely, the vector from the G's alpha carbon (solid arrow) length of three lattice units, orthogonal to the local chain plane, points into close vicinity (into a sphere of radius $2^{1/2}$ lattice units around the alpha carbon of residue A) of the A's alpha carbon. One such "hydrogen bond" vector originates from each Cα.

pairwise potential have been shifted by a constant value of −0.5. This accounts in a very approximate way for the average, chain-collapsing, hydrophobic effect of the solvent. A specific type of the contact is defined by the value of the angle between the bisectors of two virtual chain covalent angles corresponding to the residues in contact. Three types of side-chain contacts are illustrated in Fig. 3.

As mentioned before, positions of side groups allow an estimation of the alpha carbon coordinates. Finite-strength, the same as for pairs of side chains, repulsive interactions (ε_{rep}) are associated with side group-alpha carbon and alpha carbon-alpha carbon overlaps.

The separation of the hard-core and the soft-core excluded volume was done for technical reasons. The same size and shape of the hard core clusters for all amino acids enables fast (and chain-length independent), and a look-up type of detection of closely situated chain units.

3.4 Model of hydrogen bonds

Strongly directional, main-chain hydrogen bonds play an important structure-regularizing role in proteins. Instead of calculating hydrogen bond electrostatics, we opt for modeling the resulting structural regularities on the level of Cα and side group packing. Hydrogen bonding between two chain units implies a close spatial proximity of the corresponding alpha carbons and an almost parallel orientation of the planes defined by two-bond fragments of the chain. The idea is explained in Fig. 4. The system gains in energy ($-\varepsilon_H$) for each such interaction. The "hydrogen bond" network was designed to be weakly cooperative in an explicit way. Namely, an additional energy gain ($-0.5\ \varepsilon_H$) is assigned to each case when two fragments of the chain, each composed of three consecutive units, form a net of three parallel contacts. Such a pattern is characteristic of helices as well as of β-sheets.[17]

3.5 Total conformational energy

Total conformational energy of the model system is a sum of the above-outlined contributions. The scaling coefficients for particular interactions were adjusted by trial-and-error procedures in ab initio folding of a number of small globular proteins. The total energy reads as follows:

$$E = 1.0\ E_g + 0.375\ E_s + 4.0\ E_{rep} + 2.0\ E_{pair}$$
$$+ 1.25\ E_H + 0.125\ \delta(S) \qquad (1)$$

Where: E_g is the generic chain stiffness energy, E_s is the contribution from the short-range sequence-specific interactions, E_{rep} denotes the effect of soft (generic) repulsive interactions, E_{pair} is the sequence-dependent energy of pairwise interactions of the side groups, E_H is the energy of the hydrogen bond network, and δ(S) is a small energy contribution from a centrosymmetric potential which depends only on the radius of gyration (S) of the model chain. The centrosymmetric potential has a shallow minimum for compact globular states. All generic parameters (ε) have the same values equal to 1.0. The centrosymmetric potential contributes very little to the total energy; however, it accelerates the folding process, thereby penalizing extremely expanded conformations. In some simulations, the strength of hydrogen bonds ($-\varepsilon_H$) was modified as discussed later.

4. Results and discussion

The model outlined in previous sections can be used for modeling natural and artificial proteins. To some extent, each natural protein is different, and the differences among them are sometimes difficult to quantify. Thereby, studies of designed sequences may be a useful means for better understanding of general aspects of protein dynamics and thermodynamics. In designed sequences, one may use a smaller number of amino acids, exaggerated sequence patterns, and design otherwise similar "proteins" containing a different number of residues, etc. Consequently, computer modeling experiments for these polypeptide sequences could be better controlled. Some, more general but certainly not all, findings for such simplified co-polymeric systems may also apply to natural proteins.

4.1 Model polypeptide sequences

Two artificial polypeptide sequences have been designed for the purpose of this study. The first (Seq1) could be written as (-Val-Thr-)$_n$ and constitutes an example of an exaggerated β-forming sequence pattern. Valine and threonine residues have strong propensities to appear in β-sheet structures. Valine is hydrophobic while threonione is polar. Thus, there is a possibility of a phase segregation into β-sheet-like compact globular conformations. The second designed sequence (Seq2) could be expressed as (-Ala-Leu-Ser-Ser-Ala-Ala-Ser-)$_m$. It has the characteristics of helical structures: a 7-residue repeat period, with well defined hydrophobic (alanine and leucine) and polar (serine residues) faces in helical conformations. Alanine and serine residues have a strong tendency to appear in helical fragments of globular proteins. Short range potentials for the most specific interactions between i-th and i+3rd residues are compared in Tab. 1. Numerical data for the remaining potentials can be extracted from our home page.[16] The leucine residues in Seq2 provide a similar strength of tertiary interactions as valine in the first sequence. Tab. 2 contains the values of the pair-specific potential of long-range interactions for the residues included in the designed sequences. All numerical data are for the potentials before scaling given in Eq. (1).

Both sequence patterns were used to build polypeptides of lengths varying from 56 residues to 224 residues. Monte Carlo experiments consisted of a number (10 or more) of independent simulations at various temperatures. In each case, simulations started from a relatively high temperature, corresponding to the random coil regime. In subsequent runs, the temperature was lowered up to a temperature characteristic of high-density globular states.

4.2 Effect of sequence

As expected, the two designed sequences exhibit qualitatively different behavior in the Monte Carlo experiments. At a sufficiently low temperature, both sequences collapse to a high-density globular state. In the collapsed states, the model chains exhibit a high level of local ordering. For Seq1, the packing is characteristic of β-globular proteins, while Seq2 adopts a packing pattern typical for helical proteins. Since both sequences have the repeating patterns of a few residues along the chain, the "folded" structures are not unique. There are no sequence signatures for turns or breaks in the secondary structure propensities of the model polypeptides. However, it should be noted that the average length of the secondary structure elements does not change very much from experiment to experiment. This is due to on-average attractive pair-specific interactions between side chains that act as a compacting force. Fig. 5 and Fig. 6 show representative snapshots of the alpha-carbon traces for Seq1 and Seq2 polypeptides at various temperatures (above the folding transition, at the transition, and below the transition, respectively).

Interestingly, the collapse transition for Seq1 polypeptides occurs at a higher temperature and is steeper, indicating a more cooperative character of the transition. This

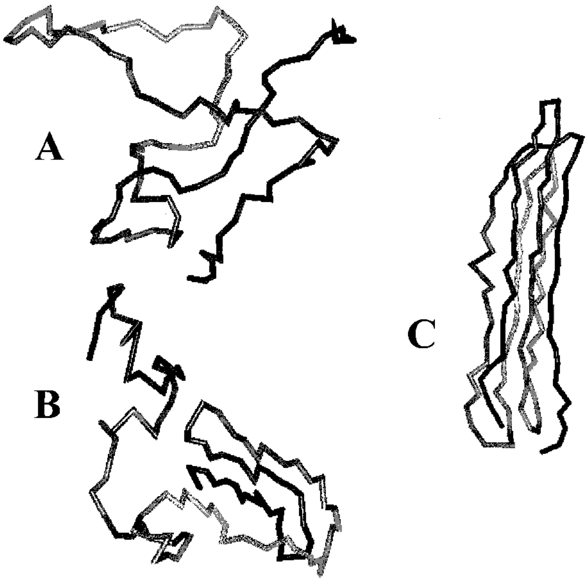

Fig. 5. Example conformations of the Seq1 polypeptide consisted of 112 residues: A- above the transition temperature, B- at the transition temperature and C- folded structure below the transition temperature. N = 112.

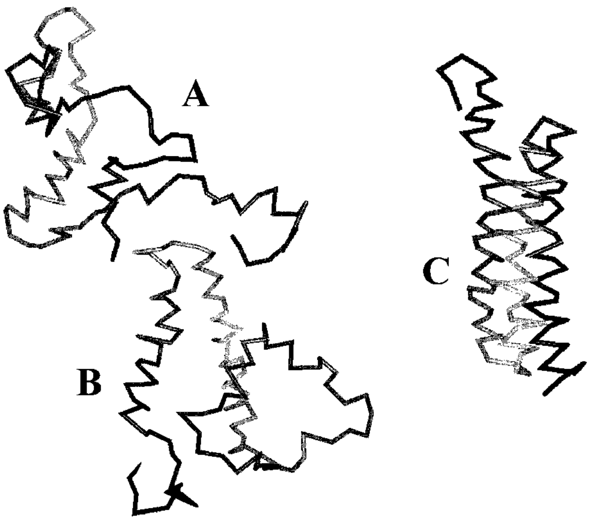

Fig. 6. Example conformations of the Seq2 polypeptide consisted of 112 residues: A- above the transition temperature, B- at the transition temperature, and C- folded structure below the transition temperature. N = 112.

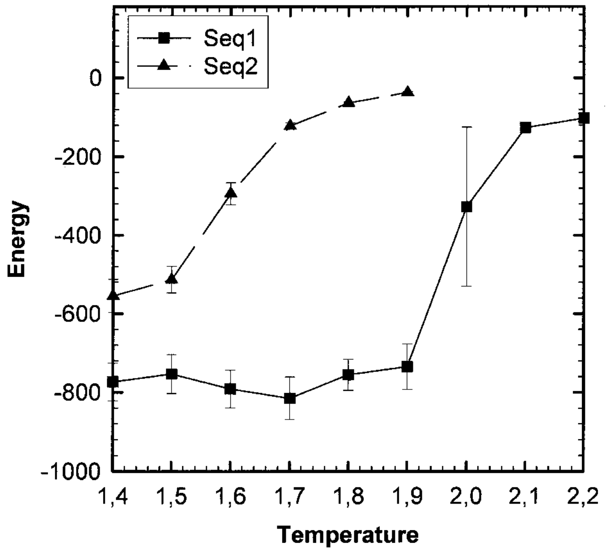

Fig. 7. Average conformational energy as a function of temperature for Seq1 (solid line) and Seq2 (dashed line), respectively. N = 112.

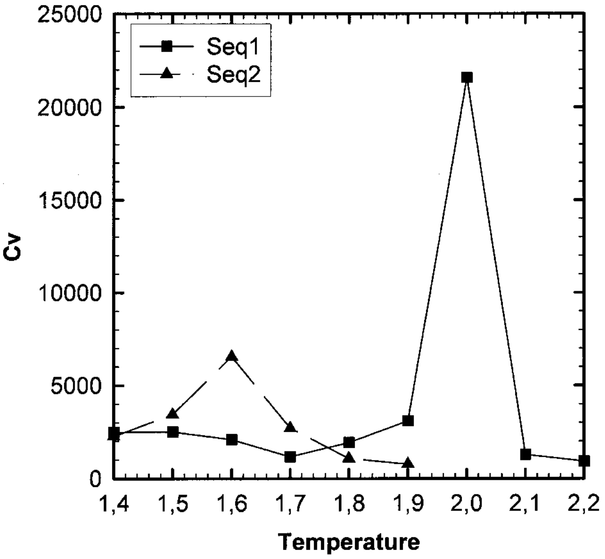

Fig. 8. Heat capacity of polypeptide chains computed from fluctuations of conformational energy as a function of temperature for Seq1 (solid line) and Seq2 (dashed line), respectively. Chain length N = 112.

is true for the entire range of chain length studied here. In Fig. 7, the conformational energy is plotted versus the temperature for two types of amino acid sequences and the same chain length. Indeed, the β-type polypeptide folds into a globular state at higher temperatures, the change of conformational energy is larger and the transition is narrower. Inspection of heat capacity curves (Fig. 8) also indicates a more cooperative transition for Seq1 chains. A possible explanation of this difference could be ascribed to the different characteristics of the hydrogen bond network in two types of proteins. Namely, the main-chain hydrogen bonds in helical proteins have characteristically short-range interactions, while in β-type proteins the hydrogen bonds involve residues that are farther apart along the chain. Consequently, in spite of having approximately the same number of hydrogen bonds in both structures and similar magnitudes of attractive interactions between the side groups, the larger fraction of long-range interactions stabilizing the globular conformations of β-type proteins lead to a more cooperative collapse transition. The drop of conformational energy is associated with a rapid decrease of the average chain dimensions. This is illustrated in Fig. 9A, where the mean-square radius of gyration for both sequences is plotted against temperature. Again, the effect is stronger for β-type proteins, since more interactions are strictly dependent on the spatial proximity of the polypeptide fragments. For the helical sequence, Seq2, the chain dimensions gradually decrease over a quite broad range of temperature. On the contrary, for Seq1 the decrease of the average chain dimensions is very rapid at a narrow range of temperature around the collapse transition. The small increase of the average dimensions for β-type globules well below the transition is associated with a small shift of the equilibrium between almost spherical globules and more elongated globules with a smaller number of longer secondary structure elements. As shown in Fig. 9B, at high temperatures both sequences exhibit very low (a few percent) helix content. The properties of the random coil state are very similar for both sequences. During the collapse transition, the helix content for the sequence containing helix-forming residues in a helix-like sequence pattern (Seq2) increases rapidly. The formation of the majority of the secondary structure coin-

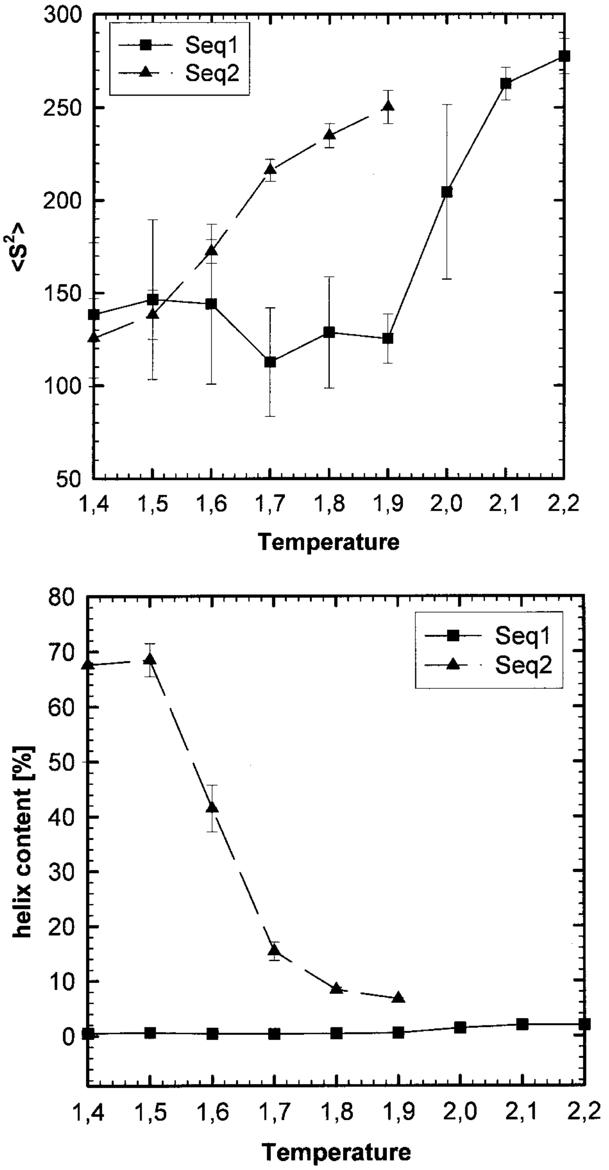

Fig. 9. Mean square radius of gyration (A) and average helix content (B) as a function of temperature for Seq1 (solid line) and Seq2 (dashed line). The data are for the same systems as presented in Fig. 7–8. Chain length N = 112.

cides with the folding transition. For β-forming Seq1, the folding transition leads to the decrease of the helix content from a small (but non-zero) value at the random coil state to essentially zero in globular state.

The mechanism of folding could be abbreviated by a combination of collision-diffusion[18] and sequential on-site assembly.[19] The second mechanism – where already assembled fragments provide a scaffold for the folding of the remaining portion of a model polypeptide – dominates, especially for the β-type sequence. Illustration of such a mechanism can be found in the intermediate temperature snapshots (B) in Fig. 5 and in Fig. 6. Indeed the snapshot (a typical one) for Seq1 shows an already assembled three-stranded sheet. The fourth strand is in the process of assembly, with conformation of the rest of the chain being more-or-less random. Comparison of various trajectories shows that the folding nucleus can form at any position along the chain. A similar mechanism of assembly can be observed for helical structures. Fig. 6B shows two loosely defined helices that zip-up to open a helical hairpin. Inspection of several trajectories indicates that the elements of sequential assembly and the diffusion collision mechanism could be seen in all simulations for this sequence.

Fig. 5C and Fig. 6C show snapshots of low-temperature folded structures. These belong to the lowest energy globular conformations for Seq1 and Seq2, N = 112 chains. The structures of the globules are not unique. For instance, Seq1 in a fraction of simulations formed two β-sheets that crossed at almost straight angles. In these structures, the average number of β-strands was larger than in the conformation shown in Fig. 5, allowing a saturation of the tertiary interactions in such orthogonal structures. For Seq2, three-member, four-member, and five-member helical bundles were observed. In some simulations (about 1/3 of the total), other types of packing were observed. Interestingly, a small fraction of folded structures from Seq2 showed intrusions of β-type packing of small chain fragments between helices.

4.3 *Effect of chain length*

Tab. 3 contains the numerical values of the folding transition temperature for various values of the chain length. Transition temperatures were read as positions of the maxima in the heat capacity curves. With an increasing number of residues in polypeptide chains, the transition temperature increases for β-type polypeptides (the helical sequence changes are below the resolution of our simulations), i.e., the globular structures become more stable. While the trend is rather clear over the entire range studied here, it cannot be extrapolated to longer chains. Formation of a hydrophobic core is an important aspect of protein folding. Above some critical chain length, formation of a single hydrophobic core could be impossible or not plausible due to competition of various interactions. One may expect a division of long chains into domains. This possibility will be addressed in future work.

Tab. 3. Transition temperatures for various values of the chain length.

Sequence		T
Seq1	56 residues	1.80
Seq1	112 residues	2.00
Seq1	224 residues	2.05
Seq2	56 residues	1.60
Seq2	112 residues	1.60
Seq2	224 residues	1.60

With increasing chain length, the transition becomes somewhat sharper for both sequences (larger jump of conformational energy in a narrower range of temperature and higher peak of specific heat). This indicates more cooperative folding.

Increase of the chain length from 56 to 224 residues slightly increases the average length of secondary structure elements. In all folded structures, the turns were located at the surface of the globule. This is also a common feature of natural globular proteins. With increasing chain length, the fraction of globules exhibiting the nonparallel packing of helices or β-sheets increased. However, for all studied systems, the parallel mode of packing dominates.

4.4 Role of hydrogen bonds

The model of hydrogen bonds employed here stabilizes both types of regular secondary structure. The formation of a particular secondary structure is triggered by short-range interactions that may locally favor either very expanded or more compact helical conformations. The interplay between conformational stiffness and long-range attractive interactions itself provides some cooperativity of the folding transition.[6] The increase of the scaling factor for short-range interactions leads to somewhat augmented cooperativity. However, a model with significantly stronger short-range interactions than employed here would be less physical. Namely, the level of secondary structure at temperatures above folding transition would be higher than observed in unfolded globular proteins. It is interesting to see to what extent the observed collapse transition is controlled by the interplay between the local conformational stiffness and the long-range interactions between the side chains, and to what extent the directional properties of hydrogen bonds are important. To address this problem, two additional series of simulations were performed; the first with the strength of the hydrogen bonds twice reduced, and the second without any hydrogen bond interactions. When the original series of simulations was done, we employed a certain scaling of the magnitude of hydrogen bond interactions that appears to be close to an optimal value for simulations of the folding process of natural protein sequences. Simulations with reduced-strength hydrogen bonds show that these interactions significantly contribute to the folding cooperativity. The collapse transition temperature decreases with the decreasing strength of hydrogen bonds. Since it leads to an overall decrease of the strength of the long-range interactions (at least for β-type proteins), it seems to be a rather trivial effect.

However, the situation is more complex. The lower scaling of hydrogen bond potential, or the lack of these interactions, makes the collapse transition more diffuse. This is further illustrated in Fig. 10, where the conformational energy curves are compared for various systems. The relative content of regular secondary structure (helices or β-sheets) noticeably decreases, especially at temperatures below the collapse transition. The effect is stronger for the helical sequence. This is illustrated in Fig. 11 where three representative snapshots of folded conformations of β-type proteins are presented for three values of the hydrogen bond scaling factor. A similar comparison for Seq2 is given in Fig. 12. Clearly, changes of hydrogen bond interactions influence the behavior of helical polypeptides in a qualitative fashion, while for β-type sequences the changes are significantly smaller. In

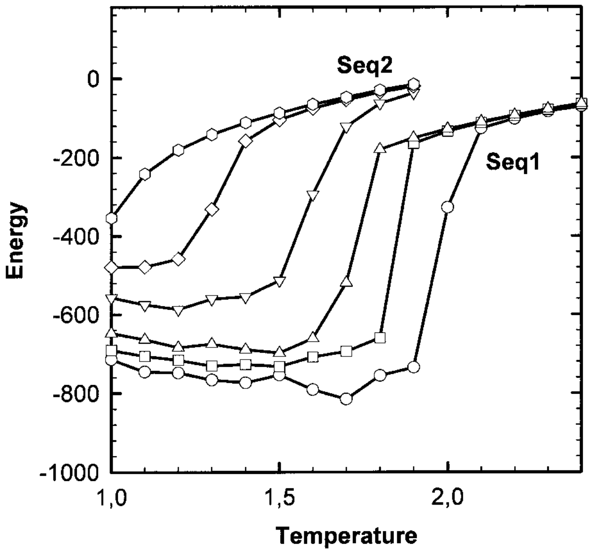

Fig. 10. Average conformational energy as a function of temperature for Seq1 (lower/right side curves, in triangles – no hydrogen bonds, squares – half-strength of hydrogen bonds, and in circles – full strength of hydrogen bonds) and Seq2 (upper curves, hexagons – no hydrogen bonds, diamonds – half strength hydrogen bonds, triangles – full strength of hydrogen bonds) for three values of the scaling factor of the hydrogen potential. Chain length N = 112.

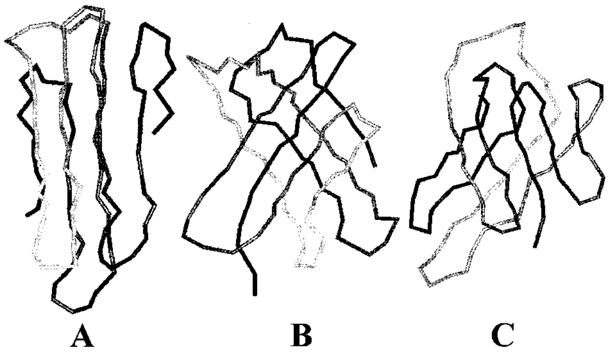

Fig. 11. Snapshots of globular conformations of β-type proteins for three values of hydrogen bond potential: A- without hydrogen bond interactions at T = 1.5, B- with 1/2 of the original strength of the hydrogen bond network at T = 1.6, and, C- with the full strength of hydrogen bond interactions, for temperature, T = 1.8 (the same conditions as for snapshot C in Fig. 5).

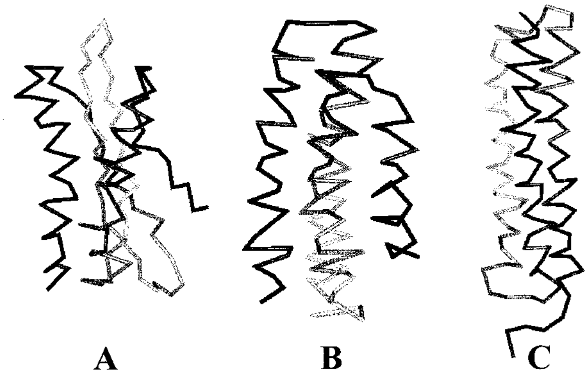

Fig. 12. Snapshots of globular conformations of β-type proteins for three values of hydrogen bond potential: A- without hydrogen bond interactions at T = 1.0, B- with 1/2 of the original strength of the hydrogen bond network at T = 1.2, and, C- with the full strength of hydrogen bond interactions, for temperature, T = 1.4 (the same conditions as for snapshot C in Fig. 6).

order to separate the effect of hydrogen bonds, all globules have approximately the same average packing density as measured by their radii of gyration. Consequently, the temperatures for particular structures are different – lower for the systems with reduced strength of the hydrogen bond interactions.

Monte Carlo experiments with a reduced strength of hydrogen bond interactions indicate that these directional interactions are necessary to achieve the highly regular local packing in a globular state. Strong secondary structure short-range conformational propensities and a very regular pattern of hydrophobic and polar residues are not sufficient to induce protein-like packing. This has to be contrasted with findings for very simple lattice models. In such models, proper sequence patterns and/or chain stiffness led to highly ordered structures. The difference could probably be ascribed to the much larger number of possible conformations per chain unit in the present model. The effective number of degrees of conformational freedom per amino acid in the model studied here is close to that calculated for real polypeptides. Thus, it appears that the interplay between the short-range interactions and sequence-specific long-range interactions of the side groups and the main chain hydrogen bonds seen in the present model mimics, to a certain extent, the complex interactions determining real protein structures.

The effects of solvent and resulting hydrophobic interactions are partially encoded in the side-chain interactions. Nevertheless, hydrophobic interactions may need more explicit treatment. This aspect of reduced modeling of proteins will be addressed in the forthcoming work.

5. Conclusion

In this work we describe a reduced model of protein structure and stochastic dynamics. The resolution of the lattice representation of polypeptide conformations in this model (about 1.5 Å) allows one to model some details of protein structures. The interaction scheme of the model, derived from statistical analyses of the structural correlations seen in real proteins, mimics the local stiffness of polypeptides, the mean effects of side-chain interactions, and the main-chain hydrogen bonds. Here, two artificial sequences of amino acids were studied. Both were designed using idealized and exaggerated patterns of beta-forming and helix-forming residues. Monte Carlo simulations have shown that indeed the two sequences undergo collapse, or folding transitions that lead to highly ordered globular structures. The β-type model polypeptides exhibited sharper, apparently more cooperative folding transitions. For these polypeptides, their rapid collapse, as measured by a steep decrease of chain dimensions, was accompanied by the fast formation of regular secondary structures in the densely packed globular state. For helical proteins, the transition occurs more gradually, and the formation of secondary structure (helices) progresses over a broader range of temperatures. Moreover, the assembly of helices is somewhat less coupled to the onset of chain collapse than in the case of β-sheet formation.

Varying chain length (polypeptides composed of 56, 112 and 224 residues were studied) had a small influence on the folding transition. For longer chains, the folding temperature slightly increased and the collapse transition was sharper.

The folded structures were not unique; however, the distribution of the size of the globule was quite narrow. Few types of structures were observed – mostly parallel β-sheets and helical bundles. For longer chains, more complex folds, with orthogonal packing of secondary structure elements, were also sometimes observed.

While the sequence of amino acids used here, and the resulting short-range and long-range packing preferences dictated the type of folding pattern, it was also found that interactions mimicking the orientational effects of hydrogen bonds are necessary for a high level of packing order in the globular state. The effect is stronger for helical sequences.

Folding transitions observed in simulations occurred by a combination of two mechanisms. The β-type sequence folding was dominated by a sequential folding mechanism where already folded fragments served as a scaffold for "on site" assembly of the rest of the chain. For helical proteins, a "collision-diffusion" mechanism was somewhat more pronounced, where loosely defined helices, fluctuating (in time and space), sometimes collided and subsequently adjusted their geometry and packing patterns.

In the forthcoming work, we will attempt to modify the sequences to achieve more unique globular structures. In particular, the effect of turn/loop inducing sequence patterns will be investigated. Also, a more explicit treatment of the solvent effect and hydrophobic interactions in the framework of the reduced model will be introduced and its effect on cooperativity of the thermodynamics of folding transitions will be addressed. The purpose of these studies is to establish a minimal model of molecular interactions in proteins that is sufficient for semi-quantitative studies of the dynamics and folding thermodynamics of globular proteins, multidomain proteins, and multimeric protein assemblies.

Acknowledgements: This work was supported partly by the *Polish Committee for Scientific Research* (KBN) grant # 6 PO4A 014 13 and NIH grant No. GM-37408. Andrzej Kolinski is an International Scholar of *The Howard Hughes Medical Institute*.

Received: January 24, 2000
Revised: March 28, 2000

[1] T. E. Creighton, *"Proteins: structures and molecular properties"*, W. H. Freeman and Company, New York 1993.

[2] C. L. Brooks III, *Curr. Opin. Struct. Biol.* **1993**, *3*, 92.

[3] K. A. Dill, S. Bromberg, K. Yue, K. M. Fiebig, D. P. Yee, P. D. Thomas, H. S. Chan, *Protein Sci.* **1995**, *4*, 561.

[4] D. Hoffmann, E. W. Knapp, *Phys. Rev. E* **1996**, *53*, 4221.

[5] J. Skolnick, A. Kolinski, *"Protein Modeling"*, in *"Encyclopedia of Computational Chemistry"*, Vol. 3, P. Schleyer, T. Clark, J. Gasteiger, P. A. Kollman, H. F. Schaefer III, P. R. Shreiner, Eds., John Wiley & Sons, Chichester, UK 1997, pp. 2200–2211.

[6] A. Kolinski, J. Skolnick, *"Lattice models of protein folding, dynamics and thermodynamics"*, R. G. Landes, Austin, Texas 1996.

[7] A. R. Ortiz, A. Kolinski, P. Rotkiewicz, B. Ilkowski, J. Skolnick, *Protein Suppl.* **1999**, *3*, 117.

[8] A. Kolinski, P. Rotkiewicz, J. Skolnick, *"Application of High Coordination Lattice Model in Protein Structure Prediction"*, in *"Monte Carlo Approach to Biopolymers and Protein Folding"*, P. Grassberger, G. T. Barkema, W. Nadler, Eds., World Scientific, Singapore/London 1998, pp. 100–130.

[9] A. Kolinski, J. Skolnick, *Proteins* **1998**, *32*, 475.

[10] A. Kolinski, P. Rotkiewicz, B. Ilkowski, J. Skolnick, *Proteins* **1999**, *37*, 592.

[11] A. Kolinski, B. Ilkowski, J. Skolnick, *Biophys. J.* **1999**, *77*, 2942.

[12] M. Levitt, *Curr. Opin. Struct. Biol.* **1991**, *1*, 224.

[13] S. Sun, *Protein Sci.* **1993**, *2*, 762.

[14] I. Carmesin, K. Kremer, *Macromolecules* **1988**, *21*, 2819.

[15] K. Binder, *"Monte Carlo and Molecular Dynamics Simulations in Polymer Science"*, Oxford, New York 1995.

[16] A. Kolinski, *http://biocomp.chem.uw.edu.pl*.

[17] A. Godzik, J. Skolnick, A. Kolinski, *Protein Eng.* **1993**, *6*, 801.

[18] M. Karplus, E. Shakhnovich, *"Protein folding: Theoretical studies of thermodynamics and dynamics", in "Protein Folding"*, T. E. Creighton, Ed., W. H. Freeman, 1992, pp. 127–196.

[19] J. Skolnick, A. Kolinski, *Science* **1990**, *250*, 1121.