Discretized model of proteins. I. Monte Carlo study of cooperativity in homopolypeptides

Andrzej Kolinski
Department of Molecular Biology, The Scripps Research Institute, 10666 N. Torrey Pines Road, La Jolla, California 92037 and Department of Chemistry, University of Warsaw, 02-093 Warsaw, Poland

Jeffrey Skolnick[a)]
Department of Molecular Biology, The Scripps Research Institute, 10666 N. Torrey Pines Road, La Jolla, California 92037

(Received 26 May 1992; accepted 4 September 1992)

A discretized model of globular proteins is employed in a Monte Carlo study of the helix–coil transition of polyalanine and the collapse transition of polyvaline. The present lattice realization permits real protein crystal structures to be represented at the level of 1 Å resolution. Furthermore, the Monte Carlo dynamic scheme is capable of moving elements of assembled secondary and supersecondary structure. The potentials of mean force for the interactions are constructed from the statistics of a set of high resolution x-ray structures of nonhomologous proteins. The cooperativity of formation of ordered structures is found to be larger when the major contributions to the conformational energy of the low temperature states come from hydrogen bonds and short range conformational propensities. The secondary structure seen in the folded state is the result of an interplay between the short and long range interactions. Compactness itself, driven by long range, nonspecific interactions, seems to be insufficient to generate any appreciable secondary structure. A detailed examination of the dynamics of highly helical model proteins demonstrates that all elements of secondary structure are mobile in the present algorithm, and thus the folding pathways do not depend on the use of a lattice approximation. Possible applications of the present model to the prediction of protein 3D structures are briefly discussed.

I. INTRODUCTION

For quite some time, discretized models of proteins and polypeptides have proven to be very useful in the elucidation of some general rules of protein folding.[1–5] These models are very convenient for formulating statistical thermodynamic models[6] as well as for computer simulations of proteins[7–10] and polypeptides. Over the years, this class of models has been employed to examine the static and dynamic properties of proteins. Here we present an extended, high resolution lattice model that affords all the computational advantages of being on lattice, and eliminates the disadvantages, the most salient of which are the presence of symmetry axes which limit the allowed orientations of secondary structure. The objective of the present paper is to examine the interplay of local vs tertiary interactions in determining the nature of the conformational transitions from the random coil to an ordered state in model helical and β-proteins which comprise a very simple realization of the folding of model globular proteins.

The early studies[2–4,11–13] of lattice models of idealized proteins examined the interplay of short vs long range interactions and their effect on the stability of the native state and on the folding kinetics. In spite of sometimes drastic simplifications, several studies provided new insights into the cooperativity of coil–globule transitions,[3,4,14] the origin of secondary structure, and some factors controlling the uniqueness of the globular state.[15–17] However, the ultimate goal of these studies, the prediction of tertiary structure from amino acid sequence alone and the folding of the structure on the computer, starting from an arbitrary, random coil state, has not yet been achieved. In the past, several simulations of the folding of small proteins have been reported,[17–19] although the accuracy and reproducibility of the predicted structures are very low.

In previous studies on the 210 lattice, we were able to simulate the folding pathway of relatively large proteins (plastocyanin,[8] and two TIM-type barrels[10]); however, it was necessary to use a small bias towards the local geometry of polypeptide chains consistent with the final folded conformation. In the case of TIM barrels for which some experimental data on folding intermediates are known,[20] the predicted intermediates were remarkably accurate and the folding process was very fast. Moreover, for several small and relatively simple idealized structural motifs, restricted to a high coordination lattice, the unique "native state" was generated when the only *a priori* bias towards the target structure was in the form of a marginal preference for nativelike turns.[9] The folding of these idealized motifs was driven mostly by tertiary interactions. These simulations showed that the local conformational propensities could even contradict locally the secondary structure found in the final, unique, 3D structure. Simulations on related, off-lattice models[21,22] show that the folding pathways are qualitatively the same for both types of representations of conformational space. Thus the results depend only on the general properties of the model of the protein and not the particular realization of the model.

[a)]To whom correspondence should be addressed.

 0021-9606/92/249412-15$06.00

The model for the polypeptide conformation and dynamics proposed here allows for a much more accurate representation of the polypeptide geometry than has been achieved in all the earlier lattice models, and yet still retains all the computational advantages of a discretized approach. The present paper provides the basic description of the model and examines its cooperative properties, which are discussed in the context of the helix–coil transition, the β-sheet–coil transition and some aspects of globular protein folding. A forthcoming paper will report on the folding of simple proteins from the amino acid sequence alone (characterized by a rather regular hydrophilic/hydrophobic pattern of amino acids). These simulations lead to well defined folds, without any bias or target potential built-in into the model. Possible applications of the model for studies of protein dynamics, protein design, and prediction of tertiary structure of globular proteins will be discussed.

The proposed model employs a lattice discretization of the conformation of polypeptide chains. The previously used high coordination lattice[23] (coordination number $z=24$) is replaced by its superset lattice where the virtual bonds connecting successive α-carbons can be selected from 56 possible orientations. This way potentially severe problems related to the local anisotropy of lattice models are avoided; e.g., helices can readily move between different orientations. Then, purely local lattice dynamics, which is still capable of moving assembled elements of secondary structure, is proposed for this lattice model. The residue specific potentials accounting for short range and long range interactions between side chains have been derived from a statistical analysis of a set of high resolution 3D structures present in the Brookhaven Protein Data Bank. Appropriate for this lattice model, a representation of hydrogen bonds between main chain atoms is also introduced.

For the purpose of clarity, the simulation experiments discussed below were performed for two homopolypeptides: poly-*l*-alanine, and poly-*l*-valine, at various, sometimes extreme, conditions. The major question addressed in these simulation experiments concerns the sources of cooperativity in protein folding and the interplay between secondary structure and packing of the polypeptide chain in the globular state. Alanine is known to be a strongly helix-forming amino acid, while valine tends to appear in β-sheets.[24] This allows us to examine for some extreme cases a number of general differences between the folding of helical and β-type proteins. Both amino acids have a well defined sidegroup conformation (also for valine, the spherical symmetry of the sidegroup is assumed), which depends only on the conformation of the main chain. The treatment of more complicated cases of the side chains having internal conformational degrees of freedom will be discussed in forthcoming work. Of course, we realize that the potentials of mean force extracted from a database of 3D structures of globular proteins when applied to the extended states of homopolypeptide have only a qualitative meaning. However, if the principle of the minimal frustration (of the native state) is on average applicable,[25,26] then the short range interactions, or rather the conformational propensities, should be the same as in globular states. Indeed, there is some experimental evidence that this is the case.[27] The principle of minimal frustration means that the long and short range interactions (conformational propensities) essentially do not contradict each other in the native state. In fact, the sufficient condition for applicability of statistically derived potentials of mean force is somewhat weaker. What is indeed required is a Boltzmann distribution of nativelike states. On the other hand, the high accuracy of the potential of the long range interactions for this particular case seems not to be so important. That is because there is no long range specificity for the homopolypeptides considered here. Consequently, the observed elements of supersecondary structure are perhaps generic for all proteins or at the very least, for some particular class of polypeptides.

The paper is organized as follows: First, the description of the polypeptide representation, the interaction scheme, and the model of Monte Carlo dynamics is given. Next, the results of simulations of the helix–coil and the coil–globule transitions are analyzed in the context of the interplay between the long range vs the short range interactions. We remind the reader that the term long range refers to the interactions between residues which are far away from each other down the polypeptide chain, but are spatially close. The discussion of the results focuses on comparison of the two classes of polypeptides, and a comparison of the model properties with analytical theories and experiments is made. Finally, the major conclusions are summarized, and some future applications are suggested.

II. POLYPEPTIDE MODEL AND ITS DYNAMICS

The model allows for a lattice representation of polypeptide conformations while simultaneously avoiding the anisotropy of space that is characteristic of regular lattices. As a matter of fact, quite regular helices may change their orientations almost continuously, and they can bend and twist just as long real helices do. These model chains can be fitted to the real structures with an accuracy measured by the coordinate root-mean-square (rms) deviation on the level of 1–1.5 Å for the α-carbons and the centers of mass of the sidegroups. The model of dynamics employs local random rearrangements which span the entire conformational space of the model. While it is difficult to give a proof of ergodicity of the dynamic Monte Carlo scheme, the comparison with other lattice models of polymer chain dynamics strongly suggests that the present model is very unlikely to suffer from any ergodicity problems.

A. Geometry of the model chain

The main chain α-carbon's backbone is constructed from the set of vectors type [2,1,0], [2,1,1], and [1,1,1]. Two consecutive backbone segments can join with an angle which is not smaller than 78.5° (e.g., the sequence [2,1,0], [0,−1,2]), and not larger than 143.1° (e.g., the sequence [2,1,0], [2,0,1]). Thus some acute angles, as well as the most open ones, including collinear sequences, are prohib-

ited. This is in accord with the observed polypeptide geometry. The high temperature distribution of planar angles defined by two consecutive vectors of the model chain semiquantitatively reproduces the distribution seen in x-ray structures of globular proteins. Similar agreement is seen when three consecutive vectors of the chain backbone are considered. Since some sequences of the basis vectors are excluded, the average length of the backbone vector is very close (within 1%–2%) to that of a [2,1,0]-type vector. Thus the distance $5^{1/2}$ for the model corresponds to 3.8 Å in real proteins. We should point out that some α-carbon–α-carbon distances in real proteins, specifically those involving *cis*-prolines are smaller; thus the present model dramatically improves the quality of fit to these real structures.

The major advantage of the fluctuating bond method emerges from fits to structures having several helices at various orientations. Taking into consideration that the resolution of most of the structures from the Protein Data Bank[28] (PDB) is not better than 1.5 Å, the adjustable bond length allows for a reasonable representation of packing of the main chain backbone and the discrete set of sidegroups. Furthermore, the lattice dynamics may employ smaller local rearrangements, therefore rendering the dynamics of the model very similar to off-lattice Monte Carlo dynamics or Brownian dynamics of related models.

In Fig. 1(a), an example of a backbone drawing of a short lattice chain is given. The backbone of the main chain has an excluded volume envelope which is constructed as follows. The lattice point representing the α-carbons are strictly excluded for all the remaining elements of the model chain. Then, there are the six points closest to each α-carbon which are associated with the underlying cubic lattice (spacings of the type [1,0,0] in the model metric, sc envelope); these are also strictly mutually excluded. Additionally, 12 points (fcc envelope) at positions of the type [1,1,0] from the central vertex serve to introduce a long range excluded volume. They are strictly excluded for all residues for which $|i-j|>4$ (with i,j the residue index down the chain). For $|i-j|\leqslant 4$, these 12-point envelopes may partially overlap, and for $|i-j|\leqslant 2$, the fcc envelope may even overlap with the sc envelope. This way, the long range excluded volume is somewhat exaggerated with respect to the hard core of the main chain of real polypeptides.

The sidechains have their own excluded volume, where each heavy atom is projected onto the underlying sc lattice. The result of the projection is the sc lattice point which is closer to the central vertex of the residue than all other lattice points separated from the real coordinates of the atom by distances smaller than the sc-lattice spacing. Some projected atoms within the same residue may coincide (occupy the same lattice point). Moreover, some points representing sidegroup atoms can coincide with the main chain excluded volume envelope. If the last case holds, that part of excluded volume of the sidegroup is treated in the same way as the main chain. The excluded volume of any part of a sidegroup which extends beyond the main chain envelope is rigorously preserved. Of course, the orientation

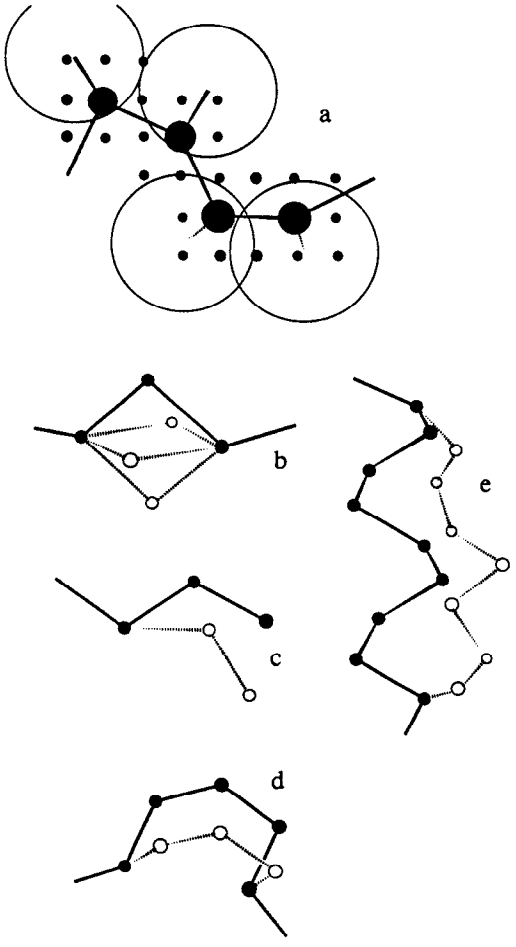

FIG. 1. Schematic drawing of the hard core excluded envelope of the model chain (a) and the elemental moves used in the Monte Carlo scheme (b)–(e). The large open spheres in (a) show the extension of the long range pairwise hydrophobic interactions for the case of polyvaline. The radius of the sphere depends on the particular pair of interacting residues. The location of the center of this interaction depends on residue type and the conformation of the main chain backbone. Possible modifications of the chain local conformations by the elemental moves are marked by dotted lines. (b) Single α-carbon spike move. (c) Chain end move. (d) Four bond rearrangement. (e) Eight bond rearrangement.

of the sidegroup depends on the conformation of the main chain, i.e., the particular sequence of two consecutive backbone vectors. For all amino acids, and for each backbone conformation, the particular projection of the sidechain excluded volume is extracted from the mean position of the atoms of the sidegroup, after the best possible fit of the two backbone vectors to the main chain coordinates of three residue fragments of PDB structures. The equivalent set of relative positions of the sidechain's center of mass with respect to the backbone (off-lattice coordinates) is also generated. The center of mass of each sidegroup serves as a center of the long range sidechain–sidechain interaction.

Let us note here that for relatively small sidegroups such as alanine and valine, the sidegroup excluded volume envelope is a subset of the main chain envelope, while for larger sidechains it can extend beyond the backbone envelope. Moreover, these larger amino acids have several con-

TABLE I. Angle dependent part of energy of interactions (in kT units) between sidegroups. Numerical values of $E\phi_{i,i+k}$.

$\cos(\mathbf{l}_i\mathbf{l}_{i+k})$ (From	To)	Short range interactions ($k=2$)	($k=3$)	($k=4$)	Long range ($k>4$)
		Alanine–Alanine			
−1.0	−0.8	0.875	0.405	0.484	0.305
−0.8	−0.6	0.721	−0.223	−0.027	−0.234
−0.6	−0.4	0.470	0.539	0.602	0.305
−0.4	−0.2	−0.223	−0.080	−0.027	0.054
−0.2	0.0	−0.166	−0.223	0.283	0.054
0.0	0.2	0.028	0.405	0.378	0.865
0.2	0.4	0.269	−0.223	0.283	−0.147
0.4	0.6	0.028	−0.080	−0.364	−0.388
0.6	0.8	0.269	−0.080	0.410	0.460
0.8	1.0	−0.944	−0.080	−0.497	−0.521
		Valine–Valine			
−1.0	−0.8	1.163	−0.105	0.456	0.570
−0.8	−0.6	0.000	−0.223	0.707	−0.002
−0.6	−0.4	0.134	0.269	1.044	−0.059
−0.4	−0.2	0.981	0.028	0.351	0.374
−0.2	0.0	0.693	0.721	0.168	0.464
0.0	0.2	0.693	0.102	0.088	0.375
0.2	0.4	0.288	−0.166	0.014	0.598
0.4	0.6	0.065	−0.629	0.014	0.464
0.6	0.8	−0.595	−0.105	−0.605	−0.040
0.8	1.0	−1.078	1.058	−0.812	−0.665

formers for each fixed conformation of main chain fragment. These aspects will be further explored in future work.

What are the reasons for such a choice of the excluded volume representation in a polypeptide? First, various amino acids should have representations reflecting their different volumes and shapes. On the other hand, within the framework of a simplified lattice model, too many geometrical details would be difficult to control and could lead to artificial steric effects when the chain is very compact. The total volume of the backbone plus sidechains should lead to a representation of the native conformation which is almost completely space filled. The present model achieves this objective. One additional feature should be noted. The exaggerated main chain excluded volume, which can also be viewed as an overall excluded volume of a polypeptide composed only from small amino acids, serves additionally to account for some of the anisotropy of sidechain–sidechain, long range, interactions. This "fat" main chain backbone prevents some nonphysical mutual orientations of the interacting sidegroups. Consequently, the conformational entropy of the system is further reduced.

B. Interaction scheme

The athermal version of the present model mimics remarkably well the local geometry of polypeptide chains. That is, the distribution of distances between the ith and $i+2$, $i+3$, and $i+4$th α-carbons in real proteins is closely reproduced. However, one wishes to reproduce sequence specific, local conformational propensities of real proteins. Surprisingly, we have found that the direct correlations of the backbone vectors do not contain such specificity. Much more information can be extracted from the angular correlations of the sidegroup vectors, i.e., the vectors pointing from an α-carbon to the center of mass of the sidegroup of the same amino acid. Such correlations can be extracted from the Protein Data Bank. The following entries of the PDB[28] have been used for all the statistics: 1*bp*2, 1*alc*, 1*ccr*, 1*cse*, 1*eco*, 1*fd*2, 1*fx*1, 1*gcr*, 1*gd*1, 1*hmq*, 1*hne*, 1l01, 1lz1, 1*mba*, 1*mbd*, 1*p*01, 1*paz*, 1*pcy*, 1*sgt*, 2*act*, 2*app*, 2*aza*, 2*ca*2, 2*ccy*, 2*cdv*, 2*cna*, 2*cpp*, 2*cyp*, 2*fb*4, 2l*h*2, 2l*hb*, 2*prk*, 2*rhe*, 2*sga*, 2*sns*, 2*sod*, 2*wrp*, 3*apr*, 3*c*2*c*, 3*est*, 3*grs*, 3*rnt*, 3*rp*2, 3*tln*, 3*tpi*, 4*fxn*, 4*hhb*, 4*i*1*b*, 5*cpa*, 5*tnc*, 6l*dh*, 7*rsa*, 8*dfr*, 9*pap*, and 9*wga*. In the case when there is more than one chain in the crystal structure, only one has been included in the statistics. As a result, there is no pair of chains with more than 30% sequence homology.

Angular correlations of the sidegroup vectors are sequence specific in the sense that the observed angular distribution depends on the pair of amino acids of interest. Inclusion of higher order orientational terms is impossible given the size of the PDB database. Table I gives the numerical values of the potential of mean force for the short range and long range interactions of the sidegroup pairs, which depends on the angle between the sidegroup vectors for ala–ala and val–val pairs. The numerical values are given in reduced units of kT for ten discreet intervals ranging from -1 to 1 for the cosine of the angle between pairs of sidegroup vectors.

The long range interaction potentials are also derived from statistics of the PDB. A pair of sidegroups is considered to have a nonzero interaction energy when the distance between their centers of interaction is smaller than a critical distance, which is pair specific (i.e., it depends on the identity of the interacting amino acids). These critical

TABLE II. Parameters of long range pairwise interactions between sidegroups of valine and alanine in polypeptides.

Pair	Distance cutoff[a] (Angstroms)	Energy ($r<r_{crit}$) (in k_BT units)
Ala–Ala	5.55	−0.20
Ala–Val[b]	5.63	−0.50
Val–Val	6.48	−0.80

[a]The average distance of the center of interaction from the α-carbon is 1.5 for alanine and about 2.0 for valine; the distances slightly vary with changes of conformation of the main chain.
[b]Not used in this work, given for comparison.

distances are obtained as follows: Two sidegroups are considered to be in contact when at least one pair of heavy atoms are closer than 4.75 Å. For a given pair of residue types, the mean distance and its standard deviation between all such pairs in the PDB data base have been calculated; the critical cutoff corresponds to the mean distance plus one standard deviation. The long range interactions between residues are allowed for third nearest neighbor down the sequence and beyond. Table II summarizes the parameters of this pairwise long range potential. An additional contribution to the long range interactions comes from the angular correlations of interacting sidegroups and is constructed in a similar fashion to the short range angular correlations.

A hydrogen bond is simulated as follows. First, due to other simplifications of the model, all hydrogen bonds, in which every segment of the main chain can participate, start or end in the central vertices of the main chain backbone. Two residues i and j ($|i-j|\geqslant 4$) are considered to be hydrogen bonded when their distance is smaller than $20^{1/2}$ in model units (7.6 Å), and the following two geometrical requirements are satisfied:

$$|(\mathbf{b}_i-\mathbf{b}_{i+1})\cdot\mathbf{r}_{ij}|\leqslant 6, \quad |\mathbf{r}_{ij}|<\sqrt{20}, \tag{1}$$

$$|(\mathbf{b}_j-\mathbf{b}_{j+1})\cdot\mathbf{r}_{ij}|\leqslant 6, \quad |\mathbf{r}_{ij}|<\sqrt{20}, \tag{2}$$

where the $\mathbf{b}_k$ are the backbone vectors, and $\mathbf{r}_{ij}$ is the vector between α-carbons of the two model residues under consideration. The 7.6 Å distance cutoff is somewhat larger than the corresponding average distance 6.2 Å between ith and $i+4$th α-carbons in the α-helices of real proteins. The origin of this difference is twofold. First, there is a distribution of these distances in real systems, and second, the lattice approximation has an accuracy of about 1 Å, which has to be taken into account. Only main chain–main chain hydrogen bonds are considered, and consequently, every residue can participate in at most two such interactions. Within the main chain, unsaturated hydrogen bonds are considered to participate in interactions with the solvent, which is not explicitly simulated.

This definition allows for the formation of a proper pattern of hydrogen bonds down helices or across β-sheets. When the above definition is translated into an off lattice representation and subsequently applied to the C_α coordinates of real proteins, more than 90% of the hydrogen bonds defined by the Kabsch and Sander[29] method are correctly reproduced for α-helices and β-sheets. Our geometrical definition misses 1–3 and 1–4 (mostly in turns or 3_{10} helices) hydrogen bonds. This omission can be corrected when the folding of proteins with more realistic sequences is attempted. Here the simplified definition has been selected for the sake of clarity of presentation of the general features of the polypeptide model. The model hydrogen bond may be considered as a semiempirical potential of strength E^{H} which accounts not only for H bonds in real proteins, but also for other backbone–backbone interactions. These interactions are known to be rather cooperative, and therefore we introduce an additional cooperativity parameter. The system gains an additional energy E^{HH} in every case when residues i and j are H bonded at the same time as residues $i+1$ and $j+1$ ($i-1$ and $j+1$, or $i+1$ and $j-1$, or $i-1$ and $j-1$) are H bonded.

The total conformational energy of the model polypeptide may be symbolically written as follows:

$$E=\Sigma E\varphi_{i,i+2}+\Sigma E\varphi_{i,i+3}+\Sigma E\varphi_{i,i+4}+\Sigma\Sigma(e_{i,j}+E\varphi_{i,j}) +nE^{H}+mE^{HH}, \tag{3}$$

where the three first terms correspond to the energy of interactions defined by the angular correlations of the sidegroup vectors of the ith and $i+k$th residues, with $k=2$, 3, and 4, the fourth term corresponds to the long range interactions, which includes an anisotropic hydrophobic force and long range, angle dependent interactions of sidegroups (the double sum is performed over $r_{ij}<r_{crit,ij}$), and the last two terms account for the hydrogen bond network of the model system. For these terms, n is the number of H bonds, and m is the number of nearest neighbor pairs of "parallel" H bonds.

C. Model of Monte Carlo dynamics

The model of Monte Carlo dynamics uses a random sequence of various types of local micromodifications of the model chain conformations. The first kind of modification is a spike move of a single α-carbon. The number of possible choices depends on the actual conformation of the fragment of the chain and varies from 1 to 9. When the spike move is performed, the sidegroups of the two neighboring residues also change their orientations. An example of a one residue spike move is shown in Fig. 1(b).

The second type of micromodification involves chain end moves [see Fig. 1(c)]. In this case, one end residue moves to a new, randomly selected, position. At the same time, a dummy residue, attached to the end residue, also changes its orientation. These dummy residues serve as N-terminal and C-terminal caps; they define the conformation of the end residues, and consequently, the orientation of their sidegroups.

The third type of move involves the random rearrangement of the four bond vectors (three α-carbons). A trial conformation is generated by removing three residues and inserting them in a new conformation [Fig. 1(d)]. Here a slight bias towards the proper handedness of the model chains is introduced. Namely, in all the cases, when the distance between the ends of inserted fragment corre-

sponds to a compact, presumably helical state, the inserted fragment is right handed. When a more extended conformation has to be built, the handedness is random. Note that this bias is rather weak for a couple of reasons. The four bond moves, although they are attempted with the same average frequency as the spike moves, have a much lower acceptance ratio. Consequently, the handedness can be rapidly changed by successful spike moves. Moreover, a right-handed fragment insertion may produce left-handed conformations in the neighboring chain fragments. The lack of a specific average handedness in the high temperature, random coil, states shows that the bias is really marginal. This bias just increases the acceptance ratio for more helical proteins, in which a proper handedness is driven by the short range angular potential and not by the kinetic bias itself.

The final type of local motion is an eight-bond move [Figure 1(e)]. It is rather local in character in spite of the fact that it affects several residues. It is constructed as follows: First, a randomly selected, α-carbon vertex is moved to one of the 26 nearest points on the underlying sc-lattice. Then, starting from this vertex, two four-bond deletion–insertions are made on both sides of the central vertex [Fig. 1(d)]. This move, although rarely accepted, allows for the motion of larger pieces of assembled structure. Like the other moves, the eight-bond rearrangement changes the orientation of a number (9) of sidegroups.

A long series of such randomly selected local moves modifies the conformation of the entire molecule. Of course, the mobility of the model polypeptide depends on temperature and amino acid sequence; the latter dictates the intramolecular interactions. A single unit of time is defined as that when one attempt on average at every kind of local move per residue is made. More precisely, there are L-2 spike moves, two end moves, L-4 four-bond moves, and L-8 attempts at eight-bond moves for a chain composed of L residues per model unit of time. In order to speed up the simulation, almost all moves, except the end moves, are selected randomly from a large collection of already built blocks, which are generated only once by the full enumeration of all possibilities. Thus, the discretized nature of the model is fully exploited.

The sampling proceeds according to the following procedure: A single step of the Monte Carlo algorithm generates a micromodification of the system. Then, the new trial conformation is subjected to a geometrical test (excluded volume has to be preserved, and all other geometric restrictions cannot be violated). Finally, the change of conformational energy associated with the attempted move is computed and used to calculate the acceptance probability using an asymmetric Metropolis scheme. This single step is repeated many times invoking various local micromodifications.

III. RESULTS AND DISCUSSION

All simulations were done for homopolypeptides composed of $L=99$ amino acids. In order to elucidate the effect of short range interactions and hydrogen bonding on the cooperativity of the observed transitions from the random coil state to a more regular one (helix or β-globule), a number of simulations for both polypeptides (poly-*l*-alanine and poly-*l*-valine) have been performed assuming $e_{ij}=0$. That is, sidechain tertiary interactions are turned off, and the only long range interactions in these simulations arise from hydrogen bonds and from the angular correlations of the sidegroup vectors. For both polypeptides, the energy of a hydrogen bond is assumed to be equal to $-1\ kT$, which is in the appropriate range for the difference between an intrapolypeptide hydrogen bond and a hydrogen bond between main chain atoms and solvent molecules.[30] The cooperativity parameter for the hydrogen bond formation (E^{HH}) is also assumed to be equal to $-1\ kT$ in all simulations where E^{H} was non zero. In a complementary series of simulations, we exclude short range angular correlations and retain tertiary interactions, and finally, the system lacking hydrogen bonding and short range interactions is examined. Of course, in all simulations the long range excluded volume effect was accounted for, according to the description of the model given above. Table III summarizes the numerical values of the interaction parameters in the four models studied in this work.

TABLE III. Numerical values of the interaction parameters for four various models of polyalanine and polyvaline (99 residues).[a]

	Short range interactions			Long range interactions		Hydrogen bonds	
Model	$E\phi_{i,i+2}$	$E\phi_{i,i+3}$	$E\phi_{i,i+4}$	$e_{i,j}$	$E\phi_{i,j}$	E^{H}	E^{HH}
Model A	1	1	1	0	1	−1	−1
Model B	1	1	1	1	1	−1	−1
Model C	0	0	0	1	0	−1	−1
Model D	0	0	0	1	0	0	0

[a]Scaling factors for parameters given in Table I and Table II.

Each series of computations consisted of two series of long runs, one following a cooling route and the other a heating route. At each temperature, a short equilibration run was performed in order to allow the system to adjust to the new conditions. For the purpose of estimation of the statistical error of the simulations, several runs were performed at selected temperature points. Various parameters describing local and global dimensions of the model chains were calculated as time averages.

A. Systems without pairwise attraction of sidegroups

These model polypeptides have an interaction scheme incorporating hydrogen bonds and orientational correlations of sidegroup vectors (vectors from the α-carbon center to the center of mass of the sidegroup). Both short range and long range angular correlations of sidegroups are included (see Table I). Consequently, there is a slight long range sidegroup–sidegroup interaction. Additional long range interactions may emerge from hydrogen bonds formed between distant (down the chain) residues. Hydrogen bonds may also form between close residues, say between the ith and $i+4$th. Consequently, there is an interplay between short and long range interactions of various

types, including excluded volume interactions. The model studied in this section is designated as model A.

1. Helix–coil transition in model polyalanine

The simulations start at a relatively high temperature, which for this particular set of interaction parameters corresponds to $T=3.0$ (a reduced, dimensionless temperature scale is used). At this temperature, the model chain has the conformational properties of an expanded random coil. The average number of hydrogen bonds per residue equals 0.211, a value far from the saturation limit, of 1 (every residue may participate in at most two hydrogen bonds; if so, it is fully hydrogen bonded, with a saturation value of 1). The helix content is also rather small, below 0.1. The helix content for the model polypeptide is calculated as the fraction of residues i which have the following conformational properties: First, there is a hydrogen bond between the $i-2$nd and $i+2$nd residue, and second, the fragment of the main chain backbone is right handed. The averaging is done down the chain and along the Monte Carlo trajectory at the temperature of interest. The ratio of the mean square radius of gyration to the mean square end-to-end distance is close to 1/6, the proper value for an ideal random coil. The relaxation spectrum of the chain vectors is also typical for a random coil. The amplitudes of the acceptance ratios for various moves at this temperature are rather high: about 24% of single bond spike moves succeed, 50% of end moves, about 24% of the four bond moves, and about 12% of eight bond moves are accepted by the algorithm.

With decreasing temperature, the helix content increases, and the low temperature ($T=2.0$) equilibrium state is an extended helix. This occurs because the long range interactions are too weak to break the helical pattern of hydrogen bonds and induce collapse. The average number of hydrogen bonds per residue at this temperature equals 0.91, the value close to the saturation limit. An additional contribution to helix stabilization comes from the short range, angular potential. Approximately 50% of the stabilization energy comes from these interactions; the remaining 50% is almost entirely the contribution of hydrogen bonds and associated cooperative terms. The long range angular energy of interaction between sidegroups is very small.

Observe that the helix state is not static. While the longest relaxation time at $T=2.0$ is very large, there are substantial fluctuations of the structure, especially at the helix ends. The middle part of the helix moves very slowly, with small oscillations of bond angles and bond lengths. The global diffusive motion of this long helix, while nonzero, is many times slower than the global motion of the random coil structures. Since the model has no built in rigid body translation, the chains can only move by means of segmental diffusion. We will discuss the dynamic properties of the model later.

In Fig. 2(a), in the curves denoted by the solid diamonds, the helix content of the model system is plotted against reduced (dimensionless) temperature. As indicated by the solid line, the data are fit rather well by the homopolymer version of Zimm–Bragg theory,[31] with a helix

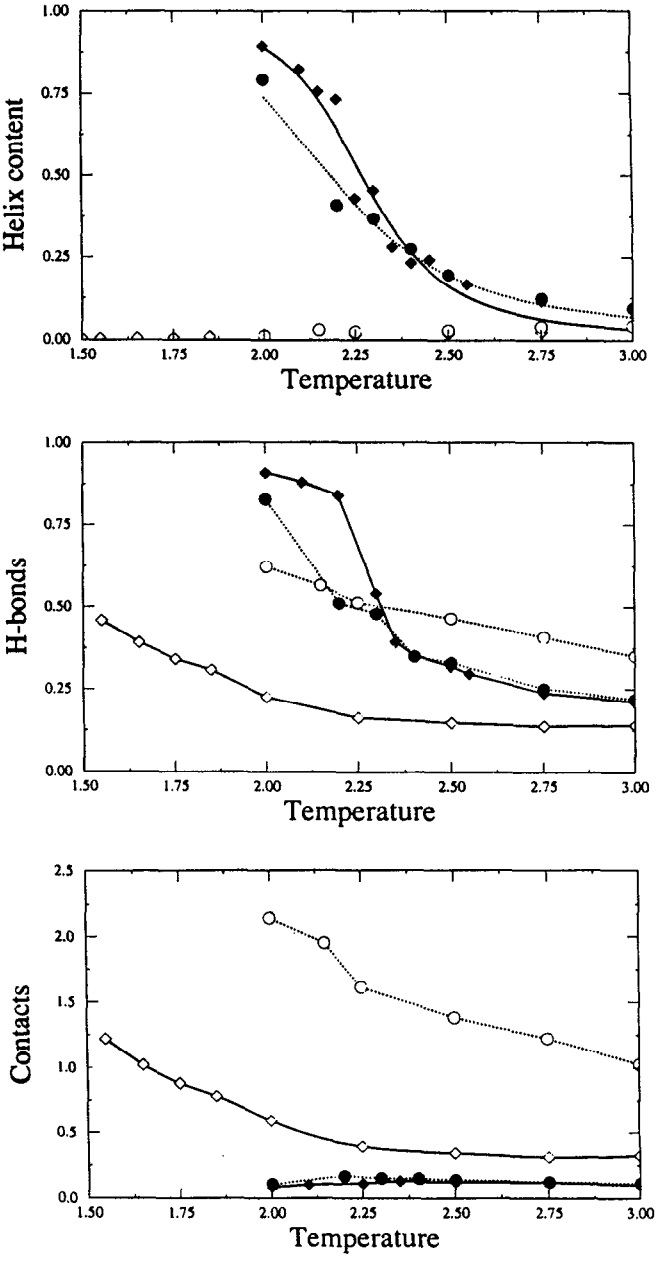

FIG. 2. Comparison of various properties of model A (diamonds) with model B (circles) as a function of reduced temperature (dimensionless). For each case, the solid (open) symbols refer to polyalanine (polyvaline). (a) Helix content. The lines correspond to the least squares fit of the Zimm–Bragg model; the points represent Monte Carlo results. The statistical error of simulations lies in the range of the symbol size. (b) The average number of hydrogen bonds per residue. The lines represent an arbitrary interpolation of the simulation data. (c) The average number of pairwise side group contacts per residue.

stability parameter s being temperature dependent according to $\ln(s)=A+B/T+C/T^2$. The fourth adjustable parameter is the helix initiation constant σ. The last parameter equals 0.087, indicating substantial cooperativity in the system. With $A=-6.4$, $B=16.00$, and $C=-3.28$, the helix stability changes from $s=2.18$ at $T=2.0$ to 0.24 at $T=3.0$. Taking into consideration that some interactions are omitted, these values of the Zimm–Bragg parameters

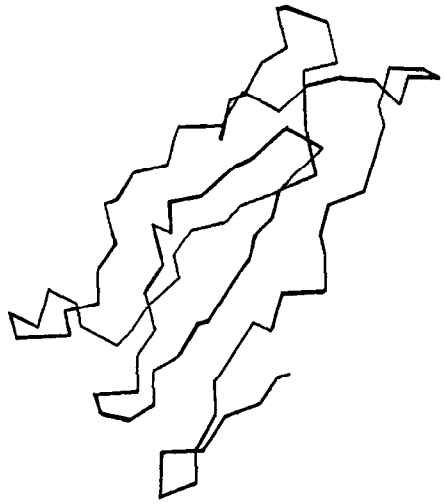

FIG. 3. Representative snapshot of model A polyvaline conformation at $T=1.75$

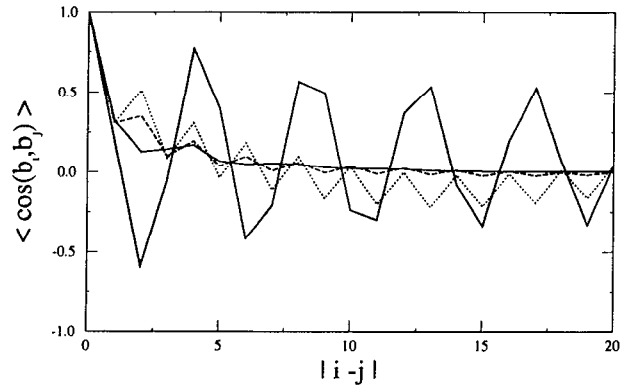

FIG. 4. Bond–bond correlations down the model A chain. Solid lines: polyalanine at $T=2.0$ (largest oscillations) and at $T=3.0$. Dashed line, polyvaline at $T=3.0$, and dotted line, polyvaline at $T=1.75$.

are within the physically plausible range. Note that a substantial contribution to the helix stability in this case comes from the down chain sidechain angular correlations. Also included in curves denoted by the solid diamonds in Figs. 2(b) and 2(c) are the number of hydrogen bonds and number of contacts vs reduced temperature.

Now let us compare the properties of model poly-*l*-valine with exactly the same interaction scheme, with of course, the angular coupling appropriate for valine.

2. "Folding" of model polyvaline

As can be seen from the short range angular correlations extracted from the database, valine exhibits a strong preference for rather extended conformations. Indeed, over the entire, broad range of temperatures in the curve indicated by the open diamonds in Fig. 2(a), the model polypeptide has rather marginal helix content, changing from 0.025 at $T=3.0$ to about 0.005 at $T=1.5$. At the same time, as the temperature decreases, the number of hydrogen bonds increases. However, it is always smaller than for polyalanine. The comparison is given in Fig. 2(b) for the solid and open diamonds for polyalanine and polyvaline, respectively, where the average number of H bonds per residue is plotted against reduced temperature. The difference in short range interactions in the two polypeptides causes the different pattern of hydrogen bonds. The contribution of the hydrogen bond energy to the total stabilization energy of the low temperature states is somewhat smaller in polyvaline with respect to polyalanine and is in the range of 1/3. One difference is clear—the transition for polyalanine is much sharper. The dependence of the configurational energy, and its fluctuation (heat capacity) on temperature also show a more narrow transition for polyalanine (see Figs. 6 and 7 below).

As indicated by a representative snapshot in Fig. 3 at $T=1.75$, the low temperature state of polyvaline is a loosely defined β-structure, stabilized by short range interactions (angular coupling potential) and hydrogen bonds, which have (in contrast to polyalanine) the character of long range interactions. The conformation resembles the 3D structures of β-proteins. However, these folds are not at all unique. The structures are very mobile, resembling a protein liquid and rearrange many times during a single run. This is because these systems lack specificity in both the long range and short range interactions. Consequently, the observed folds are generic and are the result of the particular balance between the entropy of the system and the interplay of long and short range interactions. Polyalanine at low temperature has almost all of its hydrogen bonds saturated, the global mean conformation is very well defined, and (while fluctuating) the local conformation is also very well defined. In contrast, the polyvaline globular state consists of several expanded strands and turns (or loops) enforced by long range interactions. This results in some energetic frustration of the system, which is even more exaggerated since a β-sheet pattern of hydrogen bonding is much more difficult to achieve than an α-helical pattern, due to the large number of possibilities. Consequently, the entropy of this system is considerably larger than it is for the helical structure.

Quantitatively, the difference in generic secondary structure of both polypeptides can be measured by comparison of backbone vectors correlations down the chain. These correlations are measured as the average cosine between the ith and jth backbone vectors (we remind the reader that an α-carbon representation of the backbone is used). The averaging is done down the chain and over the trajectory. The resulting correlations for both systems are illustrated in Fig. 4. The solid (dashed) lines show two extreme cases of polyalanine at $T=2.0$, in the curve exhibiting the largest oscillations, and at $T=3.0$, respectively. In polyalanine at low temperature, there is a very slowly decaying helical pattern of angular correlations of the backbone. The average pattern suggests a repeat period of 4; however, this is the result of averaging, where the well defined helices with a more physical repeat period are mixed with quite extended, right-handed states. In con-

trast, the periodicity of high temperature polyalanine is hardly noticeable. There is just a weak signal from the first turn of a helixlike state. The correlations in polyvaline at $T=3.0$ are also weak (dashed line); only for the first three vectors is there a significant difference between the two polypeptides. Instead of forming turns, polyvaline tends to adapt a rather extended state. At low temperature ($T=1.75$), the first few vectors (dotted line) show a stronger preference for extended conformations. After five to six residues down the chain, the average correlations of the chain show a well defined tendency towards the antiparallel orientation of the backbone. The above reflects the average length of β strands in antiparallel configurations. Note that in spite of the lack of short and long range specificity, the average length of β strands in these generic folds ranges from between four and eight residues per strand. This nicely coincides with the distribution of β states in real proteins. This rather strongly suggests that the conformational properties of the present model, and consequently its conformational entropy per residue, mimic rather well the situation seen in real globular proteins.

There are several reasons to believe that this is not just a coincidence. First, the number of independent conformers per residue of the model chain is about ten, which is in agreement with real systems. Then, the particular conformations are weighted by the angular correlation potential, further driving the system towards a proper distribution of rotational isomers. Additional corrections to the conformational entropy of the model come from the excluded volume and the long range interactions, both of which provide some additional chain stiffness.

Concluding the discussion of the models with rather generic hydrogen bond interactions and amino acid specific angular correlations of the sidegroups, it should be emphasized that the striking, qualitative differences in properties of both models are the result of a different equilibrium between the long vs short range interactions. The qualitative change of the coil–helix into a coil–globule collapse transition is triggered by quantitative changes in the potential of short range interactions, which defines the conformational propensities of the model chains.

B. The effect of long range pairwise interactions of sidegroups

The parameters of the pairwise potential of sidegroup interactions for these model systems are given in Table II. It may be noticed that the interactions between valines are much stronger than the rather weak ones between alanine sidegroups. Moreover, the cutoff radius for valine is larger, which further makes the long range interactions more important. Certainly, this model (model B) should provide a more plausible correspondence to real proteins.

The helix–coil transition of polyalanine is only slightly affected by sidegroup interactions. The transition is somewhat broader, being slightly shifted towards lower temperatures and is probably a bit less cooperative. Fit of the Zimm–Bragg model to the Monte Carlo data [see Fig. 2(a), dashed line and solid circles, respectively] gives $\sigma=0.093$, a value slightly larger than that for model A, and $s=\exp(-3.62+7.70/T+0.52/T^2)$. An explanation of this effect (although it is small) emerges from the competition between short range interactions and weak long range interactions. As one may see from Fig. 2(c), solid circles (lower curve), the number of interactions per sidegroup in polyalanine is still small; however, there is a maximum at intermediate temperatures, just prior to the coil–helix transition. These long range interactions tend to make more compact structures than a single extended helix. Although at the transition temperature, one may observe short lived, rather unstable, helical hairpins or irregular bundles, at lower temperature, the short range interactions (conformational propensities and regular helical pattern of H bonds) dominate, and the final low temperature state is very much like that in model A.

In contrast to polyalanine (solid circles), polyvaline (open circles) exhibits a monotonic increase in the number of pairwise long range interactions with decreasing temperature [see Fig. 2(c)]. At all temperatures, this number is much larger than for polyalanine. These tertiary interactions do not make the collapse transition of polyvaline sharper than it was in the case of model A. This becomes readily apparent after examination of Fig. 2(b), where the number of H bonds per residue is plotted against reduced temperature for polyvaline and polyalanine in both models. The major effect of sidegroup interactions is that the collapse of polyvaline is shifted towards considerably higher temperature. The long range hydrophobic interactions contribute about 1/3 of the energy of stabilization of the globular state, somewhat less than 1/4 comes from short range conformational propensities, and the rest is due to hydrogen bonding. The globular state is a densely packed β bundle with rather tight turns. Elements of the characteristic Greek-key topology of β-barrels[32] can be identified in almost all folds generated by model B (and model A as well). It appears that the Greek-key topology is induced by the small bias towards the right handed conformation of more compact states, i.e., turns. It is difficult to conclude to what extent this mechanism reflects the basic physics of the real proteins. Perhaps other factors, such as the chirality of amino acids and specific long range interactions should also be considered. In Fig. 5, snapshots of low temperature states of (a) polyalanine and (b) polyvaline are given for representative experiments on model B.

Comparison of models A and B shows that short range conformational propensities and hydrogen bonds are responsible for the cooperativity of the coil–helix or coil–globule transitions. Strong tertiary interactions, i.e., hydrophobic interactions of sidegroups, lead to collapse at higher temperature; however, the transition is less cooperative. This is an apparent contrast to the Go *et al.*[3,4] and Krigbaum and Lin[2] simulations of highly simplified models. The reason is that these works employ highly specific masks that allow only nativelike tertiary interactions to occur. The transition temperature T_t may be approximated not only from inspection of the helix content data, but also from an analysis of the temperature dependence of the configurational energy and heat capacity. The results of

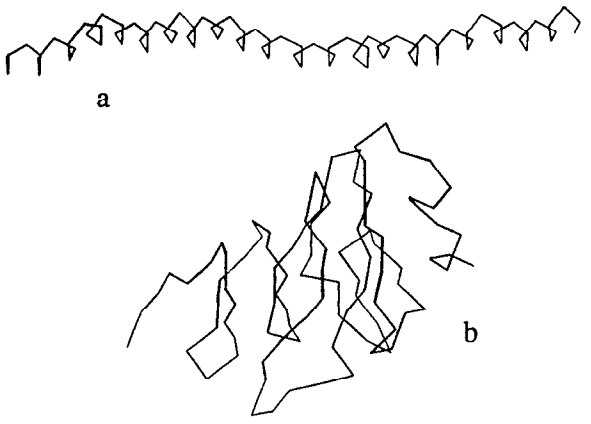

FIG. 5. Representative snapshots of backbone conformation of model (a) polyalanine at $T=2.0$, and (b) polyvaline at $T=2.25$. Both snapshots are from a simulation of model B.

simulations on various models are shown in Figs. 6(7) for polyalanine (valine). The energy increase at the transition is more gradual and the peaks of the heat capacity (computed from fluctuation of configurational energy) are more diffuse for both polypeptides in the case of model B when

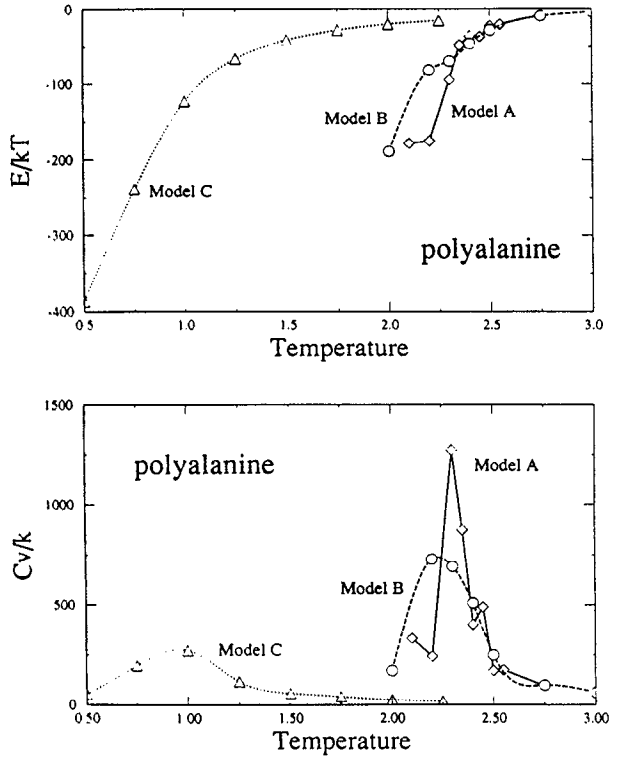

FIG. 6. Thermodynamic properties of various models of polyalanine. The results of simulations of models A to C are represented by diamonds, circles, and triangles, respectively. The statistical error for (a), the reduced (dimensionless) conformational energy, is on the range of the symbol size. For the heat capacity plots (b), the error is on the range of the symbol size except the points near the transition peaks, where the errors are about five times larger. The curves correspond to an arbitrary interpolation of the simulation data.

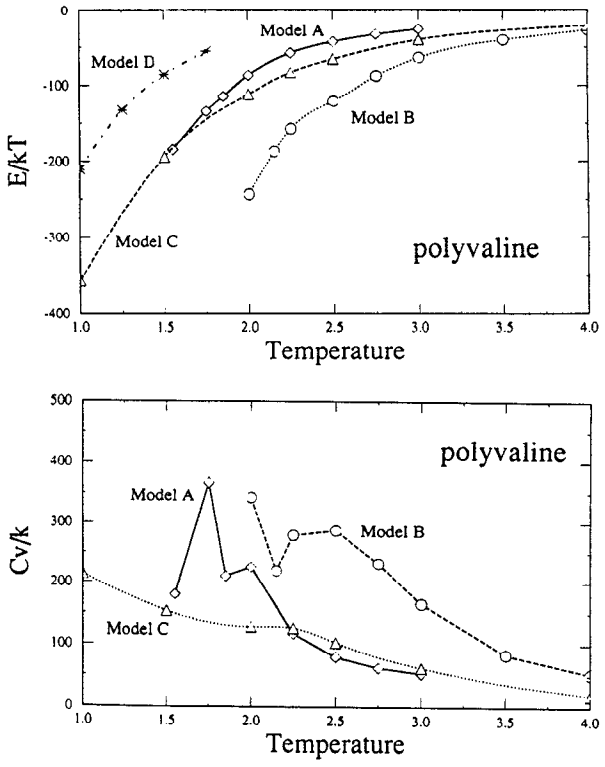

FIG. 7. Thermodynamic properties of various models of polyvaline. The results of simulations of models A to D are represented by diamonds, circles, triangles, and stars, respectively. The statistical error for (a), the reduced (dimensionless) conformational energy, is on the range of the symbol size. For the heat capacity plots (b), the error is on the range of the symbol size except the points near the transition peaks, where the errors are about ten times larger. The curves correspond to an arbitrary interpolation of the simulation data.

compared to model A. This observation is in agreement with earlier studies of even more simplified models of protein collapse.[7,33,34] The present work, however, having a quite accurate description of the conformational space of polypeptide chains, addresses this question on a more quantitative level.

Let us note that inspection of changes of the thermodynamic properties E/kT and C_V/k with temperature shows that in general the collapse (or folding) of polyvaline is much less cooperative than the coil–helix transition of polyalanine. As a matter of fact, the plots of the reduced heat capacity vs temperature do not indicate any well defined transition for all the studied models of polyvaline, with perhaps the possible exception of model A (see Fig. 7). This is because in the case of polyvaline, the low temperature state is highly degenerate, while the helical conformation of polyalanine is very well defined. Consequently, the entropy of the low temperature state of polyalanine is lower. The degeneracy of the folded state in model polyvaline results from lack of specificity of the long and short range interactions.

C. Systems without local conformational propensities

In the context of the results described above it is interesting to examine the behavior of the systems in which

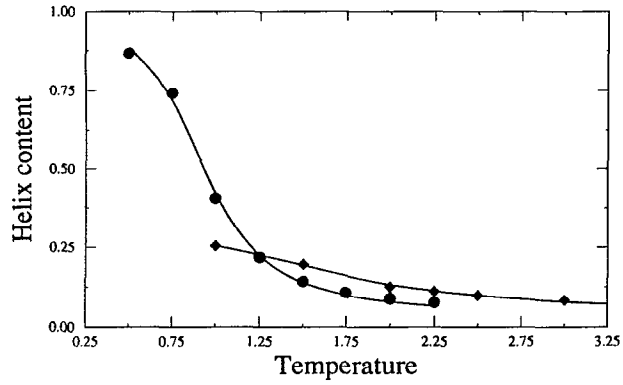

FIG. 8. Helix content versus temperature for polyalanine (solid circles) and polyvaline (diamonds), Model C. The lines correspond to least square fits of the Zimm–Bragg model to the Monte Carlo data.

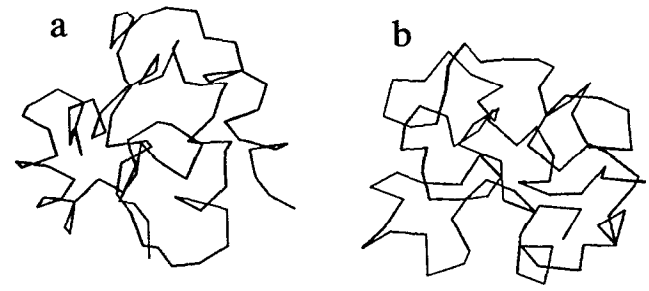

FIG. 9. Comparison of the effect of the hydrogen bonds on the structure of globular state. (a) Snapshot of a conformation of polyvaline at $T=1.5$, with parameters of model C. The average helix content at this temperature is above 0.25. (b) Dense globular state of model polyvaline at $T=1.0$. The collapse is driven only by pairwise long range hydrophobic interactions, model D.

all the short range interactions are neglected; i.e., the sidechain angular correlations are turned off. Thus, the system is controlled by hydrophobic interactions and hydrogen bonding. Again, as indicated by the solid circles of the helix content versus temperature curve of Fig 8, polyalanine folds to a long helical state. The two reasons for helix assembly in this case are: the marginal strength of the tertiary interactions, the locations of centers of mass of the sidegroups, and the regular pattern of hydrogen bonds in the low temperature state. The major difference between this model and two previous models is found in the dynamic properties. Model C without short range propensities is more mobile; the helices bend and move much faster. It is difficult to discern to what extent this is a physical effect and to what extent it is due to our particular realization of the hydrogen bond potential. Certainly, the long range, hydrophobic interactions in model polyalanine do not affect the system's mobility due to their rather moderate strength.

A quite different situation is seen for polyvaline. In this case, the hydrophobic interactions lead to collapse to a globular state, with quite substantial helical content within the densely packed globule, see Fig. 8, solid diamonds. The cooperativity parameter σ of the coil–helix transition for polyalanine is equal to 0.0087 and $\sigma=0.15$ for polyvaline. Consequently, these systems seem to be more cooperative in comparison with the models having nonzero short range conformational preferences. At least in the case of polyalanine, it may be partially explained by the considerably lower transition temperature, (the transition temperature is around 1.0 for this model of polyalanine, while model A has a transition temperature equal to 2.25 ± 0.05, and model B is about 0.05 lower than model A). The helix stability s at the transition temperature is similar for all models of polyalanine and is close to 1.0. However, with the same strength and cooperativity of hydrogen bonds in all models, the low transition temperature makes the contribution of these rather cooperative interactions substantially larger.

Observe that the present model of polyvaline adopts a predominantly helical instead of a β-sheet conformation. First of all, the lack of local conformational propensities makes all the systems with small sidegroups similar with respect to secondary structure preferences. In expanded states, the only type of secondary structure which satisfies the saturation of main chain hydrogen bonding is a helix. In collapsed states, the hydrogen bonds can be formed across the globule. However, even in this case, a mixture of helices and more expanded states allows for a larger number of intrachain hydrogen bonds. Considerations of local entropy also favor the zipping up of helical states, especially in the case when all the residues are identical. In the case of real proteins, a specific pattern of hydrophobs and hydrophils may qualitatively change this result. In such a situation, only hydrophobic interactions might be sufficient to select between α or β patterns of hydrogen bonding.

Effect of compactness alone on inducing secondary structure: The secondary structure seen in the collapsed state of polyvaline (model C) raises an interesting question. Perhaps, the secondary structure seen in this case is induced not only by hydrogen bonding but also just by dense packing itself, or maybe there is a strong bias in the model dynamics which intrinsically favors helical states? Indeed, the model employs dynamics which possess a marginal bias towards right handed twists in more compact local conformations. While this certainly has no effect on high temperature conformations, the low temperature dynamics may be potentially more sensitive to this bias.

In order to clarify these questions, a series of simulations were performed on a model which has only long range hydrophobic interactions (model D). In this case, both model polyvaline as well as polyalanine undergo a smooth transition to a very dense globular state (the collapse of polyalanine occurs at very low temperature). A representative snapshot of a polyvaline conformation at $T=1.0$ is given in Fig. 9(b). For both polypeptides, the globular state lacks any well defined secondary structure whatsoever. For example, the helix content is below 0.1 and is quite close to that seen in a *high temperature random coil state*, which means a random distribution of local conformations. In Fig. 10, we display the distribution of $r_{i,i+3}$

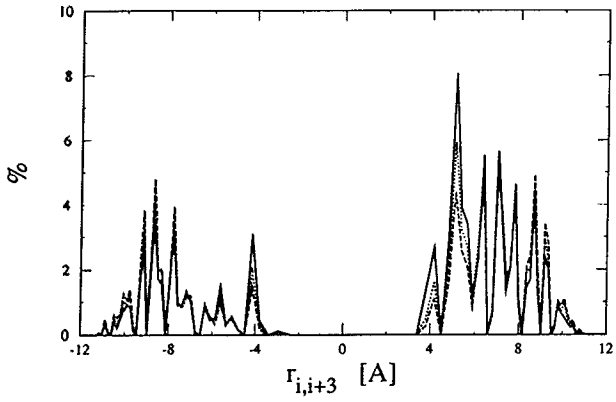

FIG. 10. Comparison of distributions of chiral $r_{i,i+3}$ for polyvaline (model D) in the collapsed globular state at $T=1.0$, (solid line), in the random coil state at $T=1.75$, (dashed line); and for a model where only residues $|k-j|<7$ can interact (dotted line).

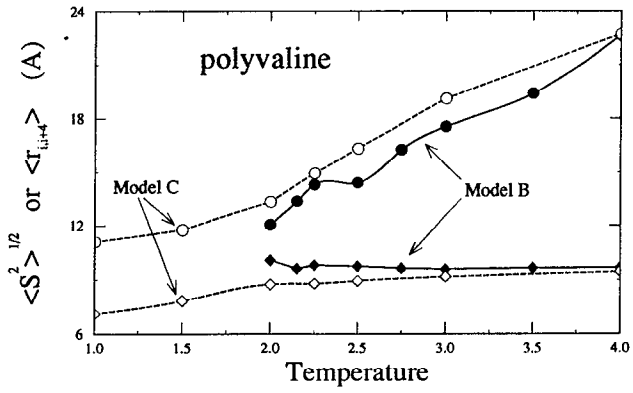

FIG. 11. Temperature dependence of the root-mean-square radius of gyration (upper curves) and the average distance between the ith and $i+4$th residue (lower curves), averaged over the chain and over the trajectory, for model B (solid symbols) and model C (open symbols) of polyvaline. The error is on the range of the symbol size; the lines correspond to arbitrary interpolations of the simulation data.

distances including chirality at $T=1.75$ and 1.0 in the dashed and solid lines, respectively. That is, we look at both the magnitude and chirality of distances of third nearest neighbors ($+1$ means right handed, and -1 indicates a left handed conformation). Both the distribution in a high temperature random coil state and a low temperature globular state are shown. The distribution of conformations is very similar, with the exception that compactness of the globule slightly favors more local compact conformations that are neither helical nor sheet. Moreover, the local handedness of the chain backbone is on average the same as for an expanded random coil. In fact, the marginally enhanced secondary structure (about 2% and 4%, respectively, for left and right handed turn type states) mainly comes from *local* tertiary interactions and not the constraint of compactness. In the dotted curve, we plot the corresponding distribution of $r_{i,i+3}$ at $T=1.0$ for a model in which only residues $|j-k|<7$ interact. What is important to note is that, in the last case, the volume is essentially that of a random coil, about ten times larger than in the case of globular state, and is about two times larger than in the case of the $T=1.75$ system with all sidegroups interacting. Consequently, the "local" tertiary interactions and not compactness give rise to the majority of enhanced local turnlike states. The number of possible hydrogen bonds (geometrical criterion applied with zero energy for H bond) is considerably smaller than in models A–C. Thus, in agreement with Gregoret and Cohen,[35] we find that the supposition that compactness alone induces substantial secondary structure is incorrect. This is not surprising, for we are basically examining the statistics of 3D Hamiltonian walks which exhibit local conformational correlations which are very similar to those in a random coil.[36]

While Dill and co-workers[37,38] have reported an apparent increase of secondary structure in small two dimensional and some three dimensional close packed cubic lattice polymers, as Gregoret and Cohen[35] have observed for α-carbon chain representations, this only occurs at densities which are about 30% higher than that seen in globular proteins. A possible counterargument to Gregoret and Cohen might have been to point out that sidechains effectively increase the packing density over an α-carbon model. But these calculations on very high coordination number lattice models with sidechains rule out that possibility as well. Thus, consistent with our previous conjectures,[5] the secondary structure seen in the folded state is predominantly the result of short range interactions, or conformational propensities, which are more or less in accord with packing requirements in the dense globular state.

To clarify this point further, let us compare the collapse transition in two models of polyvaline, models B and C. Figure 11 shows the dependence on temperature of the root-mean-square radius of gyration of the model chains (upper curves) and the average distance between the ith and $i+4$th α-carbons (lower curves). The first parameter measures the size of the coil or globule. The value of $\langle S^2 \rangle^{1/2}$ of about 10–12 Å which is observed at low temperatures for both models lies in the proper range for globular folded proteins having a similar number of residues. For example, plastocyanin (1pcy) composed of 99 amino acids has a radius of gyration of about 12.5 Å. Taking into consideration that the average size of the sidegroups for this protein somewhat exceeds the size of valine sidegroup, the agreement is very good indeed. $r_{i,i+4}$ measures the average local conformation and is larger for β structures than for helical states. Observe that the change of local conformation at the collapse transition to the globular state is rather small. Moreover, depending on the local conformational propensities, these small changes of local conformation, reflecting the presence of particular secondary structure elements, are of opposite sign for both models. In model B, short range interactions prefer extended β strands, whose average length increases only marginally with the chain collapse. Model C collapses to a globule of similar size with a considerably smaller average distance between ith and $i+4$th residue. This reflects the presence

of helical elements seen in the globular state. These helices are, however, induced by the hydrogen bonding, not by the compactness itself. Model D, without hydrogen bonds, collapses to a globule of the similar size, which is completely disordered.

Furthermore, the above provides proof that the model has no built-in "by hand" secondary structure. At the same time, the proper distribution of local conformations, consistent with that seen in real globular proteins, excludes a rather artificial bias "against" secondary structure. Comparison of properties of the low temperature states of models A–C with model D strongly suggests that requiring dense packing of the globular state does not by itself induce secondary structure. Secondary structure preferences encoded in sequence may be, however, moderated or/and magnified by various cooperative interactions (hydrogen bonds, the pattern of hydrophobic interactions, etc.) which stabilize the globular state.

D. Dynamic properties of the model polypeptides

In previous work,[23] we have shown that a similar high coordination lattice number model of polypeptides exhibit dynamic properties very close to the dynamics of Rouse chains, which is the physically correct model of polymer dynamics in the absence of hydrodynamic interactions. The present model includes additional geometrical details and has a finer resolution. The combination of a more flexible representation of the chain conformation and a larger set of elemental moves makes this model much closer to an off-lattice α-carbon (plus sidegroups) representation of real proteins. However, it is still a lattice model, and its dynamics should be examined. We limit the discussion of the results to the case of the most restricted model, i.e., model B of polyalanine. This model, having strong local helical preferences and which forms long helices at low temperatures, exhibits the lowest mobility in comparison with other models studied in this work.

At a temperature of 2.2, model B has a helix content of about 0.5, using a rather conservative definition (e.g., the residues on the ends of a helix are not counted). The model chain can move by segmental diffusion of helical fragments, which preserves the hydrogen bond pattern, and/or by dissociation of the existing helix and the formation of new helical fragments. Both mechanisms are observed. As a result, the model chain exhibits dynamic properties similar to Rouse chains, regardless of its highly helical structure. In Fig. 12, two autocorrelation functions which characterize chain diffusion are plotted on a log–log scale. The single residue autocorrelation function, $g(t)$, (the mean square displacement of an α-carbon vertex averaged over the chain and the trajectory) exhibits two regimes which are typical of a Rouse chain. There is a short time collective relaxation when $g(t)$ is proportional to $t^{1/2}$, and long time diffusion when $g(t)$ is proportional to t. The center of mass of the model chain moves as a single diffusing particle with a mean square displacement $g_{CM}(t) \sim t$. In the limit of infinite t, both autocorrelation functions should coincide.

The internal relaxation of the chain conformation can be measured by the autocorrelation functions of the chain

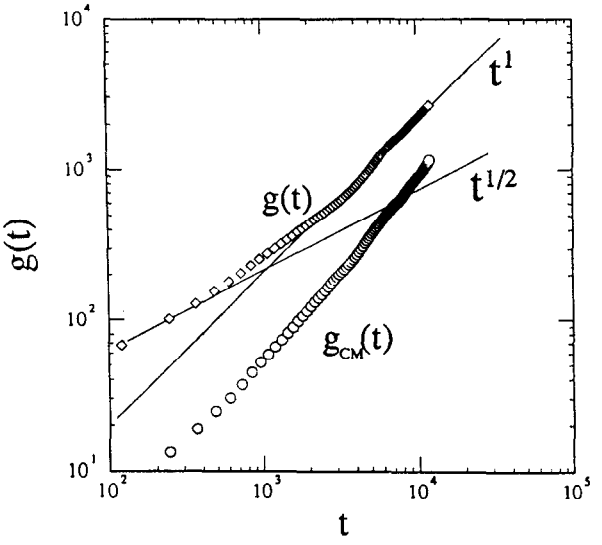

FIG. 12. Log–log plot of the single residue autocorrelation function $g(t)$ and the center of mass autocorrelation function $g_{c.m.}(t)$ vs simulation time (the number of MC cycles) for model B of polyalanine at $T=2.2$.

vectors $g_R(t)=\langle \mathbf{R}(t)\cdot\mathbf{R}(0)\rangle/\langle R^2\rangle$, with $\mathbf{R}$ being the intrachain vector of interest, at time t and at time 0, respectively, and $\langle R^2\rangle$ being the mean square length of the vector. Figure 13 shows semilog plots of some autocorrelation functions for polyalanine (model B) at $T=2.2$. It is clear that the orientation of subchains of various lengths decays in a fashion close to that expected for a Rouse chain. Some deviation from simple exponential decay of the chain end-to-end vector may be partially explained by rather poor statistics, and partially by the dual mode of chain relaxation (i.e., segmental diffusion of long helices and relaxation by dissolution–formation of helical fragments). The value of the longest relaxation time qualitatively agrees with the time regime when $g(t)$ crosses over from the collective diffusion exponent of 1/2 to the free diffusion limit.

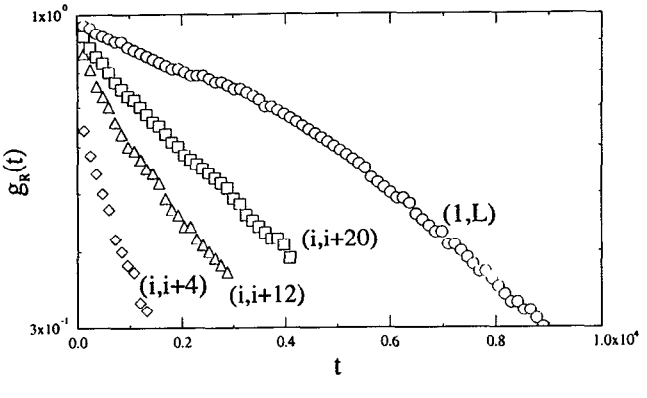

FIG. 13. Semilog plots of autocorrelation functions for several intrachain vectors versus time t (the number of MC cycles) for model B of polyalanine at $T=2.2$. The symbol (i,j) next to a curve indicates the autocorrelation function for the vector between the ith and jth residue. The upper curve corresponds to the case of the chain end-to-end vector.

At temperatures well below the coil–helix transition temperature, the model chains are still able to move as a long helix with fluctuating bond lengths and bond angles. However, the 99 residue helix moves very slowly; thus the dynamics of long isolated helices is rather artificial. Therefore, we performed a test of the mobility of a shorter, 40 residue helix, at $T=2.0$ with all other parameters of model B. This helix never completely dissolved during the 10^5 time units (cycles of MC algorithms, where every residue has a chance to be subject to each elemental move) run, moving at the same time by a distance many times larger than its size and changing its orientation. Consequently, the dynamics of the lattice model of polypeptides seems be quite physical and can be employed with confidence to study protein folding.

IV. SUMMARY

The discretized model of protein conformation presented in this work seems to possess several general features similar to those observed in real proteins. It allows for a quite accurate (on the range of 1 Å root-mean-square deviation of α-carbons from high resolution x-ray structures) representation of protein geometry. The proposed interaction scheme includes local conformational propensities and pairwise hydrophobic interactions extracted from the statistics of 3D structures of globular proteins. The cooperativity of the model is enhanced by a simplified, but realistic, model of hydrogen bonds between main chain atoms which is highly correlated with the Kabsch–Sander definition of hydrogen bonds.

The equilibrium and dynamic properties of the model are sampled by means of a dynamic Monte Carlo scheme. The dynamics is purely local and presumably has no bias to any particular scheme of protein folding. The flexibility of the model and its dynamics allows for the fast relaxation of assembled structural elements, including cooperative motion of helices or β structures. This way several limitations of the earlier lattice models[2–7] have been avoided.

In this work, the model was applied to studies of two extreme cases of homopolypeptides: polyalanine and polyvaline. Monte Carlo studies of both models show a cooperative transition from the random coil state to low temperature ordered states. In the case of polyalanine, the low temperature state is highly helical, while polyvaline forms fluctuating β-barrels. In both cases, the hydrogen bonds are responsible for the cooperativity of the transition, and the local short range conformational propensities select the type of ordered structures. The long range interactions themselves (hydrophobic force), when the hydrogen bonds and the short range interactions are artificially suppressed, lead to the collapse into a dense globular state without any appreciable enhancement of elements of secondary structure relative to that seen in an expanded random coil.

What is worth noticing is that exactly the same procedure was used for the elucidation of interaction parameters for both polypeptides. Geometrically, they differ only in the location of the centers of interaction and the range and angular dependence of interactions for the sidegroups. These parameters were also extracted from the statistics of x-ray structures of globular proteins. Thus considerable effort was directed to develop a reasonable representation of known properties of globular proteins. Of course, potentials derived on the basis of a rather limited statistical ensemble, and applied to relatively small systems (on the order of a hundred interacting units), have obvious limitations in accuracy. Due to the qualitative character of the present simulations, we have not attempted a more exact normalization of the interaction parameters. The applicability of this kind of approach requires the assumption that on average proteins obey a Boltzmann distribution of states. Otherwise, the statistical potentials will be useless. Applications of various statistically derived potentials in theoretical studies of protein systems are now commonly accepted, and in a number of cases, confirm this hypothesis.[39] What is apparently less commonly realized, is that the level of detail in models with such potentials cannot be too large. Otherwise, one is trying to extract nonexistent regularities. The above discussion leads us to the conclusion that the prediction of 3D protein structures from their amino acid sequences is feasible only if the rules of protein folding are sufficiently robust. Both earlier work and the present studies seem to indicate that this is the case.

In forthcoming work, the model will be used to study the folding process of heteropolypeptides. Preliminary results show that for a well defined amphipathic pattern of sequences of amino acids in proteins (model sequences and sequences of some designed proteins), the model presented here is able to predict a unique, stable 3D fold. Further refinements of the model provide the possibility for the prediction of tertiary structure of some natural proteins.

ACKNOWLEDGMENTS

This research was supported in part by Grant Number GM-37408 from the Division of General Medical Sciences of the National Institutes of Health. We thank Dr. Adam Godzik for preparing the tertiary interaction scale in this work, as well as Dr. Antonio Rey and Dr. Mariusz Milik for stimulating and valuable discussions.

[1] M. Levitt, Curr. Op. Struc. Biol. **1**, 224 (1991).
[2] W. R. Krigbaum and S. F. Lin, Macromolecules **15**, 1135 (1982).
[3] N. Go and H. Taketomi, Proc. Natl. Acad. Sci. USA **75**, 559 (1978).
[4] *Protein Folding*, edited by N. Go, H. Abe, H. Mizuno, and H. Taketomi (Elsevier/North Holland, Amsterdam, 1980).
[5] J. Skolnick and A. Kolinski, Annu. Rev. Phys. Chem. **40**, 207 (1989).
[6] K. A. Dill, Biochemistry **24**, 1501 (1985).
[7] J. Skolnick, A. Kolinski, and R. Yaris, Proc. Natl. Acad. Sci. USA **86**, 1229 (1989).
[8] J. Skolnick and A. Kolinski, Science **250**, 1121 (1990).
[9] J. Skolnick and A. Kolinski, J. Mol. Biol. **221**, 499 (1991).
[10] A. Godzik, J. Skolnick, and A. Kolinski, Proc. Natl. Acad. Sci. USA. **89**, 2629 (1992).
[11] V. G. Dashevskii, Mol. Biol. **14**, 105 (1980), (translation from).
[12] Y. Ueda, H. Taketomi, and N. Go, Biopolymers **17**, 1531 (1978).
[13] H. Abe, Biopolymers **20**, 1013 (1981).
[14] S. I. Segawa and T. Kawai, Biopolymers **25**, 1815 (1986).
[15] D. Covell and R. L. Jernigan, Biochemistry **29**, 3287 (1990).
[16] E. Shakhnovich, G. Farztdinov, and A. M. Gutin, Phys. Rev. Lett. **67**, 1665 (1991).

[17] D. A. Hinds and M. Levitt, Proc. Natl. Acad. Sci. USA **89**, 2539 (1992).
[18] I. D. Kuntz, G. M. Crippen, P. A. Kollman, and D. Kimelman, J. Mol. Biol. **106**, 983 (1976).
[19] M. Levitt, J. Mol. Biol. **104**, 59 (1976).
[20] C. R. Matthews, in *Protein Folding*, edited by L. M. Gierasch and J. King (AAAS, Washington, 1990), pp. 191–197.
[21] A. Rey and J. Skolnick, Chem. Phys. **158**, 199 (1991).
[22] J. D. Honeycutt and D. Thirumalai, Proc. Natl. Acad. Sci. USA **87**, 3526 (1990).
[23] A. Kolinski, M. Milik, and J. Skolnick, J. Chem. Phys. **94**, 3978 (1991).
[24] M. Levitt, Biochemistry **17**, 4277 (1978).
[25] J. D. Bryngelson, and P. G. Wolynes, Proc. Natl. Acad. Sci. USA **84**, 7524 (1987).
[26] J. D. Bryngelson, and P. G. Wolynes, Biopolymers **30**, 177 (1990).
[27] T. E. Creighton, Biochem. J. **270**, 1 (1990).
[28] F. C. Bernstein, T. F. Koetzle, G. J. B. Williams, E. F. Meyer, Jr., M. D. Brice, J. R. Rodgers, O. Kennard, T. Shimanouchi, and M. Tasumi, J. Mol. Biol. **112**, 535 (1977).
[29] W. Kabsch and C. Sander, Biopolymers **22**, 2577 (1983).
[30] E. N. Baker and R. E. Hubbard, Prog. Biophys. Mol. Biol. **44**, 97 (1984).
[31] D. Poland and H. A. Sheraga, *Theory of Helix–Coil Transitions in Biopolymers* (Academic, New York, 1970).
[32] J. S. Richardson, Adv. Protein Chem. **34**, 167 (1981).
[33] A. Sikorski and J. Skolnick, Biopolymers **28**, 1097 (1989).
[34] A. Sikorski and J. Skolnick, Proc. Natl. Acad. Sci. USA **86**, 2668 (1989).
[35] L. M. Gregoret and F. E. Cohen, J. Mol. Biol. **219**, 109 (1991).
[36] T. G. Schmalz, G. E. Hite, and D. J. Klejn, J. Phys. A **9**, 751 (1984).
[37] H. S. Chan and K. A. Dill, Proc. Natl. Acad. Sci. USA **87**, 6388 (1990).
[38] H. S. Chan and K. A. Dill, Macromolecules **22**, 4559 (1989).
[39] S. H. Bryant and C. E. Lawrence, Proteins **9**, 108 (1991).