An Efficient Monte Carlo Model of Protein Chains. Modeling the Short-Range Correlations between Side Group Centers of Mass

Andrzej Kolinski, Lukasz Jaroszewski, Piotr Rotkiewicz, and Jeffrey Skolnick

Department of Chemistry, University of Warsaw, ul. Pasteura 1, 02-093 Warsaw, Poland, and Department of Molecular Biology, The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, California 92037

The Journal of Physical Chemistry B®

Reprinted from
Volume 102, Number 23, Pages 4628–4637

An Efficient Monte Carlo Model of Protein Chains. Modeling the Short-Range Correlations between Side Group Centers of Mass

Andrzej Kolinski,*,†,‡ Lukasz Jaroszewski,† Piotr Rotkiewicz,† and Jeffrey Skolnick‡

Department of Chemistry, University of Warsaw, ul. Pasteura 1, 02-093 Warsaw, Poland, and Department of Molecular Biology, The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, California 92037

Received: October 16, 1997; In Final Form: January 20, 1998

A new high-coordination lattice model of polypeptide chains has been designed and tested. The model employs a single united atom representation of amino acid residues. These atoms are centered on protein side groups. Characteristic short-range distance correlations have been built into the model, thereby providing a rather accurate description of proteinlike conformational stiffness. Sequence-specific interaction schemes have been derived from sequence similarity and sequence-structure compatibility criteria. The conformations of the model chain observed in isothermal Monte Carlo simulations reproduce protein secondary structure with high fidelity. Implications for structural studies of protein systems are briefly discussed.

1. Introduction

Proteins are relatively complex molecular objects.[1] Computer studies of large-scale conformational transitions in proteins (including the protein folding process) on an atomic level of resolution are not practical;[2] therefore, reduced models of protein chains are necessary.[3−6] In such models, groups of atoms are replaced by united atoms to decrease the number of explicitly treated degrees of freedom.[6,7] United atoms can replace entire residues, segments of the main chain and amino acid side groups, or smaller groups of atoms. A lattice-based[8,5,9−14,3,15−21] or continuous[22−28,7,29,30] representation of conformational space of a polypeptide could be assumed.

There is always some tradeoff between simplicity (and computational efficiency) and the geometric fidelity of a reduced model.[31] Models that employ a low-coordination lattice representation and a single united atom per residue are the simplest and could be studied in great detail.[32,5,9,33−43] On the other hand, more complex models can account for specific aspects of polypeptide chain geometry and, consequently, for some important properties that are characteristic of proteins.[10−14,3] Here, we propose a model that has the virtues of both approaches. Single united atoms placed at the centers of mass of protein side chains provide a simplicity similar to the so-called "simple exact" models of proteins. At the same time, the excluded volume and local geometry of these virtual chains are designed to reproduce local protein geometry and some details of protein packing with an accuracy comparable to earlier, intermediate resolution protein models.

The rigid structure of the peptide bond, the fixed chirality of Cα, and the restrictions of the main chain rotational degrees of freedom (resulting from the excluded volume effect of the side groups and some other local interactions) lead to a relatively high conformational stiffness of polypeptides and their specific geometry.[1,44] Generally, there are a few types of local chain geometries that correspond to the local minima of the above-mentioned set of interactions. Most typical are helices and almost fully expanded conformations that build β-sheets and expanded loops in globular proteins. The majority of the other rotational-isomeric states of a polypeptide are extremely unlikely.[45] Due to the absence of fine atomic details, these "generic" characteristics of the polypeptide chains have to be built into any nontrivial reduced model. Of course, the tendencies toward formation of a particular chain geometry are determined by the specific amino acid sequence. Having properly designed generic biases toward proteinlike chain geometry, it is possible to construct a sequence-specific potential that triggers formation of helices, turns, or expanded fragments. For several reasons, these rather weak "secondary-structure propensities" cannot be factorized into single amino acid properties. The most important reason is that the local geometry of protein chains results from a complex interplay of the short-, medium-, and long-range tertiary interactions,[46] probably with nonfactorizable multibody components as well.

Some of the conformational and energetic attributes of polypeptide chains are explored in this work by means of the Monte Carlo method. We compare the average equilibrium properties of three types of chains. The simplest is a designed lattice model chain whose behavior is controlled only by the excluded volume interactions. This model system provides a reference system for a chain with the generic "proteinlike" biases and for model chains that simulate specific sequences of amino acids. The last system requires a definition of the potentials that reproduce specific secondary-structure propensities encoded in sequences of amino acids. These potentials are derived from a statistical analysis of the geometry of locally homologous fragments of known protein structures. This kind of "knowledge-based" approach has been successfully used in many applications,[47−52] the most extreme being the homology modeling of protein tertiary structure.[53,54] Here, we employ only the homology (usually low) of small fragments of protein sequences, thereby allowing for the construction of a potential for sequences having no globally homologous counterparts in the structural database.

The purpose of this work is to analyze the role of the generic proteinlike regularities seen in protein chains, the role of sequence-specific short-range correlations of the side chain

* Author to whom correspondence should be addressed.
† University of Warsaw.
‡ The Scripps Research Institute.

S1089-5647(97)03371-3 CCC: $15.00
Published on Web 05/13/1998

positions, and this interplay. We will demonstrate that the proposed model provides a very efficient tool for modeling protein secondary structure. In future work, the model would be generalized by incorporating an appropriate set of tertiary interactions.

2. Model and Methods

2.1. Lattice Chain of the Centers of Mass of Side Chains. Consider a given conformation of a polypeptide chain. For each side group, it is easy to define its center of mass for a given rotational isomeric state of the side chain. For simplicity, let us assume the same mass for all heavy atoms. The error of this approximation is small, well-below the overall geometric accuracy of the model. For Gly residues, the center of mass is arbitrarily placed on the backbone Cα atoms. Having such a defined set of reduced residues (or chain beads), the geometry of the model lattice chain composed of N amino acids (residues) could be defined as a string of vectors $\{\mathbf{v}_i\}$, connecting successive residues (i.e., vector $\mathbf{v}_i$ denotes displacement between the ith and $(i+1)$st side chains of the model polypeptide) with $i = 0, 1, ..., N - 1$. Two dummy residues are added for a convenient definition of the "conformation" of the N- and C-terminal residues. Thus, the chain consists of $N + 2$ united atoms ("beads") restricted to the underlying simple cubic lattice. (As will become evident later, the following assumption allows for a straightforward definition of the chain's excluded volume.) The distance between two successive units (beads) of such a chain depends on the identity of the corresponding residues and on the actual rotameric state of their side chains. The shortest possible distance would be observed for a pair of consecutive Gly residues, and the longest distance would be almost 3 times greater for a pair of residues with long side chains in an extremely expanded conformation. To cover this rather wide distribution, a proper set of distances between the beads of the model lattice chains should be allowed. It has been arbitrarily assumed that a set of virtual bonds $\{\mathbf{v}_i\}$ could be defined as $\{\mathbf{v}_i\} = \{a\mathbf{q}_i\}$ where a is a constant equal to the lattice spacing of the underlying simple cubic lattice and the vectors $\mathbf{q}_i$ belong to the following set of lattice vectors:

$$\{\mathbf{q}_i\} = \{[\pm k, \pm l, \pm m]\} \qquad (1)$$

with

$$k,l,m = 0, 1, 2, 3, 4, \text{ or } 5 \qquad \text{and} \qquad 11 \le |\mathbf{q}_i|^2 \le 30$$

The above implies that the number of the lattice basis vectors is equal to 592. Assuming the lattice-spacing parameter a = 1.45 Å, the resulting distance between an arbitrary pair of side chains can change from 4.81 to 7.94 Å. This nicely covers the main portion of the distance distribution seen in real proteins that have an average value of about 6.6 Å and a standard deviation of about 1.3 Å; however, the wings of the distribution have been arbitrarily cut off. The resulting positional error of about 1 Å (for a small fraction of extreme cases) is still below the assumed inherent resolution of the model defined by the lattice spacing, a. On the other hand, restricting the range of virtual bond fluctuations allows for a simpler (and computationally more effective) handling of excluded volume and the definition of various interactions. For technical reasons, successive pairs of identical vectors are a priori excluded. Such sequences are also unlikely in real proteins.

The excluded volume of the chain units could be defined in the form of a cluster of 19 points of the underlying cubic lattice that are closest to a given chain bead. The cluster includes the central point, 6 simple cubic lattice points, and 12 face-centered cubic lattice points. If each pair of two clusters is not allowed to overlap, the resulting hard core is characterized by a distance of closest approach equal to 3 lattice units (4.35 Å). This is a reasonable, but somewhat underestimated, representation of the excluded volume of real polypeptide chains. Let us note that it is easy to introduce a soft core excluded volume envelope having different radii for different amino acid types. Here, however, we employ only the simplest, hard core representation of chain excluded volume. The geometrical properties of the model chain are illustrated in Figure 1.

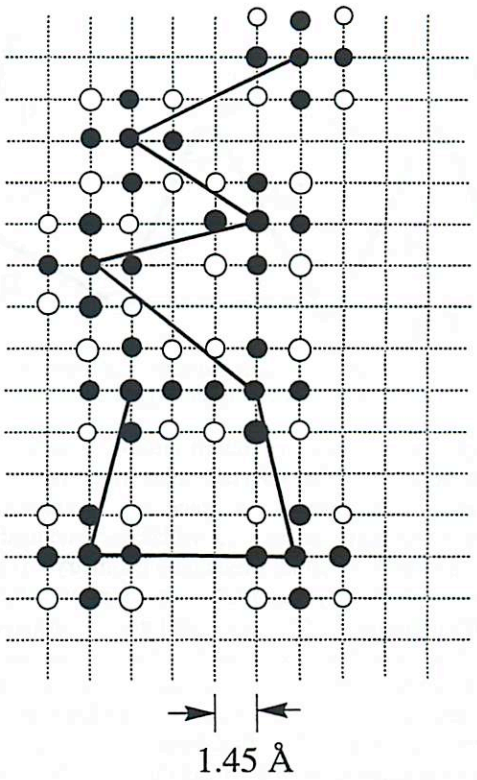

Figure 1. Lattice representation of polypeptide chains. The solid line connects the side chain centers of mass (at their actual rotational isomeric states). The solid dots indicate three points of excluded volume along the direction perpendicular to the drawing plane (one below the plane, one in the plane, and one above the plane). The open circles represent single excluded volume points in the drawing plane. Thus, for each side group, the excluded volume cluster consists of 19 points on the underlying simple cubic lattice. The spacing of the lattice is equal to 1.45 Å. For the sake of simplicity, a planar fragment of the model chain has been shown.

For a given chain bead, the number of points at which a second bead could be found at the distance of closest approach is 24. At slightly longer distances, the number of possible bead positions grows rapidly. Consequently, any effects of the lattice anisotropy are practically nonexistent in such a defined model. The lattice representation provides a very convenient way for modeling excluded volume and detecting nearest neighbors and the stochastic simulation of the chain dynamics. Monte Carlo simulations are 1−2 orders of magnitude faster (as measured by computer time per characteristic longest relaxation time) than those for a high-coordination lattice model based on Cα plus side chain representation. We will show that, at the same time, there is no loss of accuracy.

2.2. Model of Stochastic Dynamics. The simplicity of the polypeptide representation described in the previous section enables the design of a very straightforward description of chain dynamics. A single cycle of the algorithm consists of several

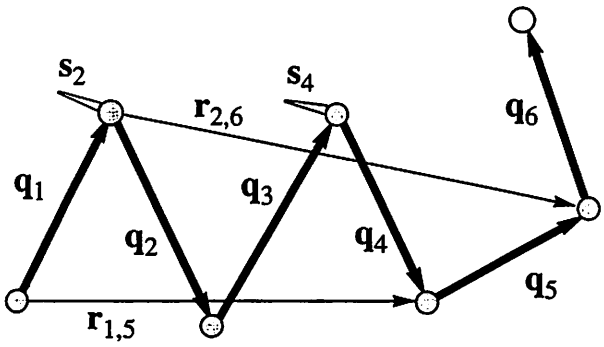

Figure 2. Explanation of the local geometry and annotations used in definitions of the generic part of the short-range potential.

simple operations. First, a random number generator selects the chain bead. For the selected chain fragment, a random conformational transition is attempted and accepted provided that the new conformation does not violate the excluded volume or assumed boundaries of allowed chain geometry. Transitions in models with energy parameters are further subject to a Metropolis criterion.[55] The set of local moves consists of one- and two-bead motions, and rigid-bodylike, small translations of somewhat longer fragments of the chains (of randomly selected lengths of up to 20 residues). The last type of move is attempted less frequently. One-bead motions are generated by the random translation of the bead direction up to a distance of $3^{1/2}$ lattice units. Consequently, the number of new possibilities for this type of move is equal to 26. Other moves are similarly constructed. A conventional time unit corresponds to fn attempts at the various conformational transitions, where n is the chain length and f is a constant (about 2) associated with the number of types of conformational transitions built into the algorithm.

Due to the large "coordination number" of the model chain, the proposed Monte Carlo dynamics are free of various ergodicity problems typical of low-coordination lattice chains. The present model could be considered to be a relatively fine discretization of the continuous space Rouse model of polymer dynamics,[56] which is known to be ergodic.

2.3. Modeling Generic "Proteinlike" Conformational Stiffness. Since proteins are relatively rigid copolymers, the number of accessible states of the model chain defined in the previous section is considerably larger than that in a real polypeptide chain (when projected onto a lattice). For example, an unrestricted chain has a close-to-Gaussian distribution of the distance between the ith and $(i+4)$th beads. The same distribution for proteins is bimodal, reflecting the existence of helices and expanded states. These elements of secondary structure tend to propagate over several residues. Modeling proteinlike stiffness could be essentially reduced to designing a switch between the two dominant types of secondary structure. Moreover, polypeptide chains have chiral preferences. The vast majority of helices are right-handed, while expanded conformations have a slight tendency to adopt a left-handed supertwist. The proposed set of energetic biases described below is consistent with helical and expanded structures, and also with some types of commonly occurring turns in globular proteins.

Let us consider a small fragment of the model chains consisting of beads numbered from 1 to 7 and connecting vectors numbered from 1 to 6 (see Figure 2 for set reference). Then, let us define a set of additional vectors according to the formula given below:

$$\mathbf{s}_i = (\mathbf{q}_{i-1} \otimes \mathbf{q}_i - \mathbf{q}_{i-1} - \mathbf{q}_i)/(|\mathbf{q}_{i-1}|\cdot|\mathbf{q}_i|) \qquad (2)$$

Per design, the vectors $\mathbf{s}_i$ are exactly parallel to a canonical helix (when the chain vectors are of the same length) axis, and consequently, they are parallel to each other for the residues within a helix, except for those residues that terminate a helix. For regular expanded states, every second vector is parallel. Of course, in a real case, the chain vectors $\mathbf{q}_i$ (depending on a sequence of amino acids) could assume various lengths as defined in eq 1. Consequently, for the appropriate pairs of $\mathbf{s}_i$, typically small deviations from exactly parallel orientations would be observed even for helical or β-type conformations. The $\mathbf{s}_i$ vectors are normalized such that their length is the longest for the values of the planar angles typical of helical and β-type states and have a magnitude roughly equal to unity (in the lattice units). Such defined vectors $\mathbf{s}_i$ provide a very convenient way to address intermediate range angular correlations in protein chains.

The short-range, proteinlike correlations could be defined using the original chain vectors $\mathbf{q}_i$. More regular elements of secondary structure could be initiated as follows via certain biases superimposed onto the mutual orientation (conveniently measured by the appropriate dot products) of the model chain vectors:

$$E_{H1} = -\epsilon_s \quad \text{for} \quad \mathbf{q}_1\cdot\mathbf{q}_3 < -5 \qquad (3a)$$

and when the local conformation is "helical-compact";

$$E_{H2} = -\epsilon_s \quad \text{for} \quad \mathbf{q}_2\cdot\mathbf{q}_4 < -5 \qquad (3b)$$

and when the local conformation is "helical-compact";

$$E_{E1} = -\epsilon_s \quad \text{for} \quad \mathbf{q}_1\cdot\mathbf{q}_3 > 5 \qquad (3c)$$

and when the local conformation is "expanded";

$$E_{E2} = -\epsilon_s \quad \text{for} \quad \mathbf{q}_2\cdot\mathbf{q}_4 > 5 \qquad (3d)$$

and when the local conformation is "expanded" where "helical-compact" means that $\mathbf{r}_{1,5}^2 < 32$ and $\mathbf{q}_1\cdot\mathbf{q}_4 > 0$, and "expanded" means that $60 < \mathbf{r}_{1,5}^2 < 125$ (in lattice units). The above provides a bias toward either helical or expanded conformations of the model chain. The cutoff parameters -5 and 5 are arbitrary (however, they are geometrically quite permissive) bounds for specific regular conformations of the model polypeptide.

Now, we can use previously defined vectors $\mathbf{s}_i$ to build a propagation mechanism for proteinlike conformational stiffness. The following potential provides an additional local bias toward "regular" secondary structure, propagates the structure, and increases the chain persistence length:

$$E_{2,4} = -\epsilon_s \quad \text{for} \quad \mathbf{s}_2\cdot\mathbf{s}_4 > 0.25 \qquad (4)$$

$$= 0 \text{ otherwise}$$

Of course, the same holds for beads 4 and 6 (compare Figure 2). Additionally,

$$E_{2,6} = -\epsilon_s \quad \text{for} \quad \mathbf{s}_2\cdot\mathbf{s}_6 > 0.25 \qquad (5)$$

$$= 0 \text{ otherwise}$$

Again, the value of the cutoff parameter (0.25) is chosen in such a way that the conditions given in eqs 2 and 4 are satisfied for all expanded (β-type) and helical conformations of the model chain, regardless of the above-discussed fluctuations of the length of the chain vectors $\mathbf{q}_i$.

The next contribution (eq 6) to the generic potential propagates stretches of regular secondary structure and also acts to increase the chain's persistence length for unstructured fragments. In this case, the model system pays a penalty for large changes in the locally averaged direction of chain propagation. At moderate temperatures, it compensates to some extent for entropic effects that favor very irregular random-coiled conformations.

$$0 \text{ for } |\mathbf{r}_{2,6} - \mathbf{r}_{1,5}|^2 \le 15 \quad \text{(in lattice units)}$$

$$E_p = \epsilon_s |\mathbf{r}_{2,6} - \mathbf{r}_{1,5}|^2/40 \text{ for } 15 < |\mathbf{r}_{2,6} - \mathbf{r}_{1,5}|^2 \le 40 \quad (6)$$

$$\epsilon_s \text{ for } 40 < |\mathbf{r}_{2,6} - \mathbf{r}_{1,5}|^2$$

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

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

$$E_R = -\epsilon_s \quad \text{for} \quad \mathbf{r}^2_{i-1,i+2} < 32 \quad \text{and} \quad (\mathbf{q}_{i-1} \otimes \mathbf{q}_i)\cdot\mathbf{q}_{i+1} > 0 \quad (7)$$

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

$$E_{gen} = \sum(0.25E_{2,4} + 0.25E_{2,6} + E_{H1} + E_{H2} + E_{E1} + E_{E2} + 0.5E_p + E_R) \quad (8)$$

where $\sum$ means the summation along the chain. Note that the strength of the contribution of all generic components to the model interaction scheme is controlled by the single energy parameter $\epsilon_s = 1$ (in dimensionless kT units). Details of the above-defined model of proteinlike conformational stiffness are, to a large extent, arbitrary. Hypothetically, one can employ quite a different set of angular and distance correlations that would lead to a similar effect, provided that a set of general regularities of protein conformations is detected and properly implemented in the form of energetic biases. The advantages of the present design are simplicity and small sensitivity to the fluctuations of the length of the chain vectors. The last is an important feature of this model, allowing reasonable modeling of the dense packing of protein structures.

To maintain a reasonable proteinlike density (all the long-range interactions have been neglected, except the hard core repulsion of the side chains), the chains have been confined by a spherical density model based on the relatively well-preserved distributions of amino acids between particular shells, as defined by the expected radius of gyration of a protein. This weak, sequence-independent potential provides a bias against elements of generated secondary structure that are too long. Details can be found elsewhere.[57]

2.4. Sequence-Specific Short-Range Potentials. The potentials employed in this work are based on the regularities seen in protein fragments of known structure that exhibit certain levels of sequence similarity to the corresponding fragments of the test sequence. It should be noted that globally homologous proteins (more than 25% sequence identity) are always excluded from the structural database used in the derivation of the potentials. Also, known structural homologues have been eliminated in these test simulations. The procedure consists of several steps. First, for the test sequence, a multiple-sequence alignment search[58] is performed to find close homologues in the sequence database. When found, these enhance the statistics. In the entire procedure of potential derivation, the homologous sequences were treated as the test sequence. If close homologues are not found, only the test sequence is used. Then, the test sequence (and close homologues) is divided into fragments by sliding a 19-residue window along the sequence. The obtained $n - 18$ fragments are then compared to all possible continuous fragments from the structural database. The fragments of the highest sequence similarity are then used for the derivation of the statistical potential. The top 100 best-scoring fragments were taken for further consideration. The BLOSUM80[59] sequence similarity criterion was used with a trapezoid weight function for the alignment. For the central 9 residues of a 19-residue fragment, the value of the weight for the alignment was assumed to be equal to 1.0; for the flanking residues, the weight decreased linearly to a value of 0.1 for the 1st and 19th residues. The geometric characteristics were collected for various intrachain distances $\mathbf{r}_{i,i+k}$, with $k = 1$, 2, 3, 4, 6, and 8 for the central residues of the test window. The $k = 3$ and $k = 6$ contributions were assumed to be "chiral" (i.e., the distances were stored as negative numbers for left-handed conformations and as positive numbers otherwise). For the $k = 6$ case, the chirality was defined using three consecutive two-bond vectors (vectors $\mathbf{r}_{i,i+2}$, $\mathbf{r}_{i+2,i+4}$, and $\mathbf{r}_{i+4,i+6}$, respectively).

The potential was obtained by comparison of the observed distributions (in a form of histogram) of particular distances for homologous fragments (i.e., the top 100 fragments) with the corresponding distribution for the entire structural database.

$$\nu(i,\mathbf{r}_{i,i+k,m}) = -\ln(P(i,\mathbf{r}_{i,i+k,m})/P(\mathbf{r}^0_{i,i+k,m})) \quad (9)$$

where $P(i,\mathbf{r}_{i,i+k,m})$ is the weighted (by the sequence similarity matrix for the 100 top-scoring fragments) probability of observation of the ith bin of the $\mathbf{r}_{i,i+k,m}$ distribution; $P(\mathbf{r}^0_{i,i+k,m})$ stands for the database averaged distribution; i denotes the position along a given chain; m denotes the bin number of the distribution. There are 9, 7, 7, 10, 10, and 4 bins for $\mathbf{r}_{i,i+1}$, $\mathbf{r}_{i,i+2}$, $\mathbf{r}_{i,i+3}$, $\mathbf{r}_{i,i+4}$, $\mathbf{r}_{i,i+6}$, and $\mathbf{r}_{i,i+8}$, respectively. The potential for a given residue in a polypeptide chain depends on a 19-residue sequence window; consequently, the potential could be qualitatively different for the same central amino acid in two different sequences.

The local chain geometry could also be approximately translated into the DSSP[60] (three-letter code) secondary-structure assignment. For this purpose, three-dimensional histogram statistics of the structural database have been calculated first. For each possible set of bin numbers of the $\mathbf{r}_{i,i+k}$ values for $k = 3$, $k = 4$, and $k = 6$, the most probable secondary structure has been assigned (helix, extended, or coil). Analyzing the structural database for each combination of the three bins' numbers, one can count the number of DSSP assignments to the various structural classes. Combining these statistics (the highest population type of secondary structure is always assigned), the three-letter secondary-structure assignment for a representative set of proteins could be obtained with 85% accuracy (we remind the reader that this is structural translation accuracy, not predictive accuracy). The accuracy is somewhat lower than that

of a similar procedure based on the Cα trace of protein backbone. Assignment errors result from the various different size side groups and the rather irregular shape of the model chain. As a result, in some cases, very similar local chain geometry could correspond to different secondary-structure motifs, depending on the sequence of amino acids. Consequently, the secondary structure read from simulations is a lower bound for the actual geometric fidelity of the protein representation.

For the purpose of conventional visualization, we designed a simple method for the approximate reconstruction of the Cα trace given the positions of the side chain centers of mass of the original chain. Let

$$\mathbf{r}_i = \mathbf{R}_i + \Delta\mathbf{r}_i \tag{10}$$

where $\mathbf{R}_i$ is the position of the ith side group and $\mathbf{r}_i$ is the position of the ith Cα:

$$\Delta\mathbf{r}_i = (\mathbf{q}_i - \mathbf{q}_{i-1})/b \tag{11}$$

with normalization factor b calculated from

$$b = 6 + b_{i-1} + b_i - d_i \tag{12}$$

where

$$b_i = 1 \text{ for } \mathbf{q}_i^2 > 15 \text{ (and 0 otherwise)}$$

$$d_i = (\mathbf{r}_{i-1,i+1}^2 - 50)/10 \text{ for } \mathbf{r}_{i-1,i+1}^2 > 50 \text{ (and 0 otherwise)}$$

The "correction" factors b_{i-1}, b_i, and d_i account for the various distances between side groups; the accuracy of the above Cα coordinate estimation is better than 1 Å. Of course, a more exact procedure could be designed for the main chain reconstruction from the coordinates of the centers of mass of the side groups. We opt here for computational simplicity. This seems to be appropriate due to the aforementioned limits of resolution of the present model. Figure 3 shows a short fragment of the side group-based chain and the reconstructed Cα trace using eqs 11 and 12.

Protein structures from 301 proteins in Fischer's database[61] and the HSSP[62] sequence database were employed in this work. Always, for a given test case, all similar sequences and similar folds were removed from the structural database during the evaluation of the sequence-specific potentials.

3. Results

3.1. Conformational Properties of a Generic Sequence Chain. First, let us compare the distributions of the local distances seen in protein structures with those obtained from the Monte Carlo dynamics trajectories of the model chains that lack any sequence information (i.e., generic chains). Such distributions are given in Figures 4–8. The plots are arranged in three panels. In all figures, the topmost panel shows the distribution extracted from the structural database, the second panel shows the distribution for the athermal simulated chain ($N = 99$) with no short-range interactions (except for a small effect due to excluded volume), and the bottommost panel displays the corresponding distribution for the chain containing only the generic (sequence-independent) part of the short-range potential. It is clear that, for all short-range distances, there is a qualitative difference between the distribution seen for real proteins and for the athermal model chain. For example, the distribution of the distances between the ith and $(i+4)$th side chains in proteins is bimodal, reflecting the tendency to form helices (the first peak) or expanded conformations (the second peak), while the distribution for an athermal chain is unimodal, and close to a Gaussian distribution.

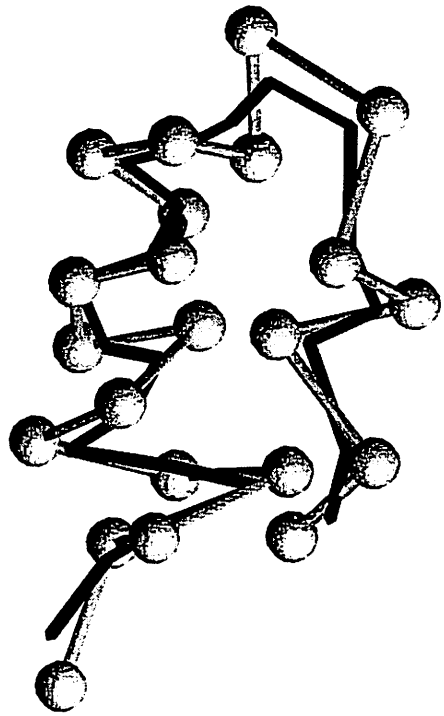

Figure 3. Short fragment of the model chain (excised from a snapshot of an isothermal Monte Carlo trajectory for protein G; see text for detail) constructed from united atoms located at the side chain centers of mass (in gray) and the approximate reconstruction of the Cα trace (black). The original chain could be fitted to known protein structure with an average accuracy of 0.8 Å cRMSD (coordinate root-mean-square deviation). Due to the favorable compensation of errors, the accuracy of the reconstructed Cα trace is slightly better.

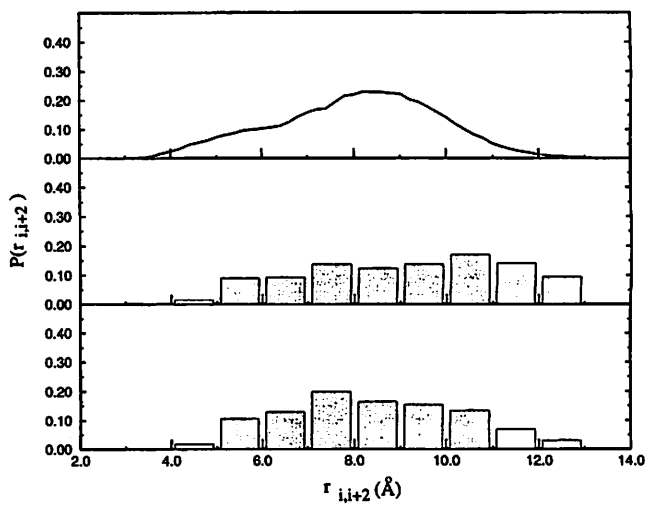

Figure 4. Distributions of the $r_{i,i+2}$ distance in real protein structures (upper panel), for an athermal chain with excluded volume (middle panel), and in the chain with the generic, sequence-independent potentials (bottom panel). The higher tick marks on the horizontal axes indicate the division into bins employed in derivation of the sequence-specific potential.

As confirmed by these figures, the generic potential mimics some of the short-range correlations seen in proteins. Indeed, the local geometry of these chains is closer to the "average" geometry of proteins. It reproduces the main features of proteinlike geometry. This is evident from an inspection of the

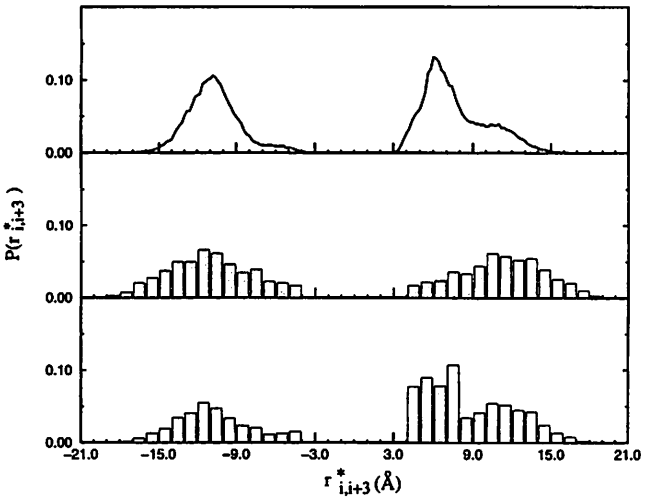

Figure 5. Distributions of the $r^*_{i,i+3}$ distance in real protein structures (upper panel), for an athermal chain with excluded volume (middle panel), and in the chain with the generic, sequence-independent potentials (bottom panel). The negative (positive) values correspond to the left-handed (right-handed) conformations. The larger tick marks on the horizontal axes indicate the division into bins employed in the derivation of the sequence-specific potential.

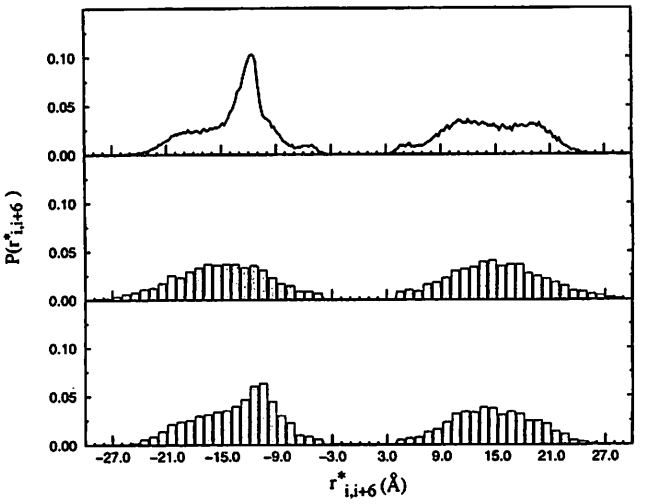

Figure 7. Distributions of the $r^*_{i,i+6}$ distance in real protein structures (upper panel), for an athermal chain with excluded volume (middle panel), and in the chain with the generic, sequence-independent potentials (bottom panel). The negative (positive) values correspond to the left-handed (right-handed) conformations. The larger tick marks on the horizontal axes indicate the division into bins employed in the derivation of the sequence-specific potential.

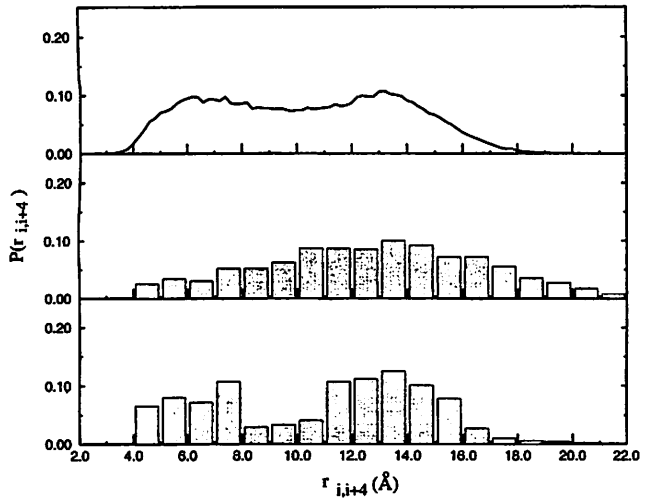

Figure 6. Distributions of the $r_{i,i+4}$ distance in real protein structures (upper panel), for an athermal chain with excluded volume (middle panel), and in the chain with the generic, sequence-independent potentials (bottom panel). The larger tick marks on the horizontal axes indicate the division into bins employed in the derivation of the sequence specific potential.

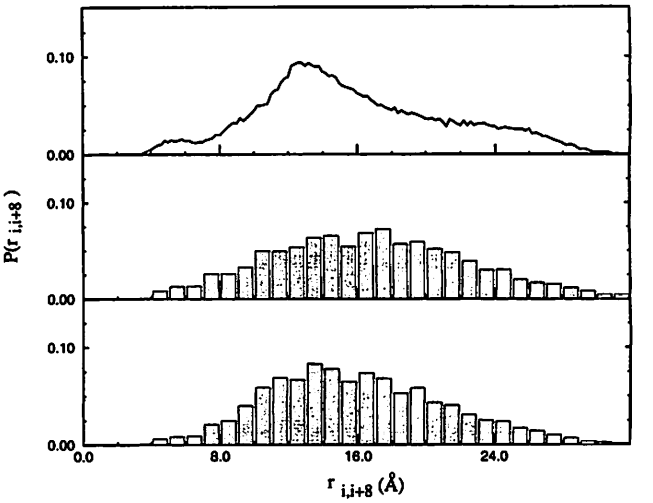

Figure 8. Distributions of the $r_{i,i+8}$ distance in real protein structures (upper panel), for an athermal chain with excluded volume (middle panel), and in the chain with the generic, sequence-independent potentials (bottom panel). The larger tick marks on the horizontal axes indicate the division into bins employed in the derivation of the sequence-specific potential.

histograms in the bottom panels of Figures 4−8 and after comparison with the distribution extracted from native proteins. For all internal distances, the distributions for the chains with the generic potentials are very similar to the distributions of real proteins, but they differ qualitatively from the unrestricted athermal chain distributions. For instance, after application of the above-described simple structural regularizers (all Monte Carlo simulations performed at $T = 1$), the model chain exhibits characteristic peaks for "chiral" distances $r^*_{i,i+3}$ and $r^*_{i,i+6}$. These peaks correspond to helical conformations. Also, the bimodal distribution of the $r_{i,i+4}$ distance is reproduced. It is very interesting to note that the distributions of those distances that were not explicitly regularized by generic potentials such as $r_{i,i+2}$ and $r_{i,i+8}$ become "proteinlike". This provides additional confirmation that the designed generic potentials rather nicely reproduce the average features of polypeptide chains. This potential also substantially reduces the contribution of very unlikely polypeptide conformations. When combined with sequence-specific potential, these generic terms should facilitate the formation of quite a regular "proteinlike" geometry of the model chains.

3.2. Simulation of Protein Secondary Structure. Monte Carlo simulations were done for a small but representative set of single-domain proteins. These proteins, abbreviated by their PDB[63] code, are listed in Table 1. All simulations were performed under isothermal conditions. Each run was very long compared to the longest relaxation time for all model systems. For each sequence, the simulations were done with two types of short-range interactions: with the generic part of the potential and without. In the former case, the relative weighting of the two contributions was set to 0.5:1.5 (generic:sequence-specific). This scaling is arbitrary. A lower weight factor for the generic potential leads to a system whose properties are *intermediate* between those of a system lacking the generic potential and

TABLE 1: Test Proteins Selected for Monte Carlo Simulations

PDB code	N	secondary-structure type	name
1gb1	56	$\alpha+\beta$	immunoglobin binding protein B1 domain
1ctf	68	$\alpha+\beta$	50s ribosomal protein (C-terminal domain)
1pcy	99	β	apo-plastocyanin
2trx	108	α/β	thioredoxin
4fab	111	β	immunoglobin FAB fragment
3fxn	138	α/β	flavodoxin (oxidized form)
1mba	146	α	myoglobin
1tim	247	α/β	triose phosphate isomerase

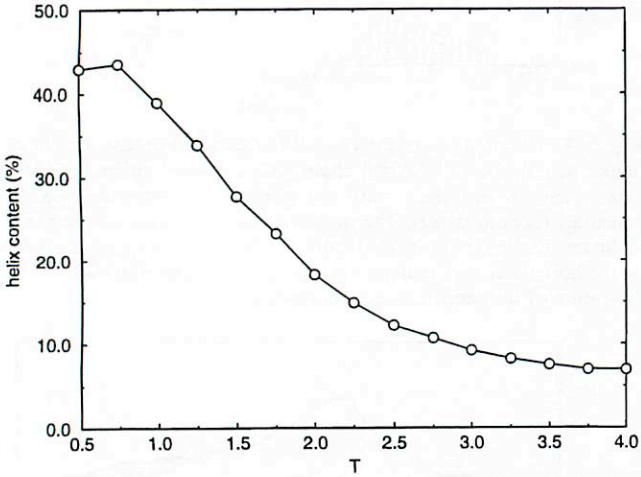

Figure 9. Average instantaneous helix content of the model chain of myoglobin as a function of temperature. See text for additional details.

one with this particular weight ratio. A stronger scaling of the generic term leads to a system in which the secondary structure is "overregularized" and ultimately to a system whose behavior becomes essentially sequence-independent.

To establish a reasonable temperature at which the simulation of secondary structure could be close to optimal, we first performed a series of isothermal simulations for the myoglobin model at various temperatures. We started from a very high temperature ($T = 4.0$) and then reduced the temperature to a value where the system mobility becomes very low; this is somewhere below $T = 1$. In Figure 9, the resulting average (instantaneous) helix content is plotted as a function of dimensionless temperature. At high temperatures, the chain essentially samples random coil conformations whose helix content is very low, about 5%. At lower temperatures, the helix content increases gradually to about 45%. This is still below the overall helix content of the native myoglobin, which is about 75%. Our simulations were performed over the range of temperatures where the chain is very mobile and the secondary structure is subject to significant fluctuations. Under such conditions, the instantaneous content of secondary structure is lower due to the nonnegligible entropic contribution that favors random coil type conformations. However, when an average from the trajectory is used with an appropriate threshold for assigning the fraction of particular observations, a higher helix content is obtained over a rather broad range of moderately low temperatures. Probably, by careful annealing to a very low temperature, one may obtain a more exact instantaneous representation of native secondary structure. Indeed, at $T = 0.9$, the helix content of the model myoglobin chain increases, but the dynamics are significantly slowed. Similar results were obtained for other sequences. We selected $T = 1$ for further Monte Carlo simulations of all test proteins. This choice, while somewhat arbitrary, is motivated by the aforementioned tradeoff between fast sampling (large-chain mobility) and the amount (and regularity) of emerging structure. When averaged, according to the procedure described below, the fidelity of the resulting secondary structure is not sensitive to small changes in the sampling temperature.

TABLE 2: Averaged Accuracy of Secondary Structure with Respect to the Native State Using a DSSP Assignment (Three-State: Helix, Extended and Coil Assignment)

protein	percent of correctly identified secondary structure (H, E, and coil): model with generic and sequence-specific terms	model with sequence-specific potential only	derived from the potential[a]
1gb1	75.0 (89.3)[b]	75.0	60.7
1ctf	70.6 (70.9)	63.2	69.1
1pcy	68.7 (79.8)	65.7	67.7
4fab	76.6 (79.3)	79.3	77.5
2trx	69.4 (85.2)	67.6	62.0
3fxn	78.3 (86.3)	79.0	65.9
1mba	73.3 (70.3)	68.5	70.6
1tim	68.4 (72.1)	67.9	65.2
weighted average	72.2 (77.8)	69.6	67.3

[a] Weighted statistics of the 100 top-scoring structural fragments (after elimination of homologous sequences) (i.e., assignment based on local sequence alignment and secondary structure "translation" from the side chain correlations). The most favorable energetic bins of $\mathbf{r}^*_{i,i+3}$, $\mathbf{r}_{i,i+4}$, and $\mathbf{r}^*_{i,i+6}$ distances were employed in the translation. [b] The case when the homologous sequences have not been excised from Fischer's structural database.

Since the model studied here lacks sequence-specific tertiary interactions (except for some medium-range interactions implicitly encoded in the short-range potential), the tertiary structure of the model chains is not well-defined. Thus, while possible in principle, it would not be very informative to compare the global chain geometry with the geometry of native proteins. Instead, we translate the observed local geometry into the three-letter code: helix (H), extended state (E), and everything else or "coil" (−) state. As mentioned above, for a given set of $\mathbf{r}^*_{i,i+3}$, $\mathbf{r}_{i,i+4}$, and $\mathbf{r}^*_{i,i+6}$ distances, the statistics of the structural database give three numbers, f_H, f_E, and f_{coil}, corresponding to the observed fraction of a particular structural class. Thus, the assignment for the entire trajectory (typically 200 snapshots taken at equal time intervals) could be done as follows:

$$g_{ss} = \sum f_{ss} \quad \text{with} \quad ss = \text{H, E, or "coil"} \tag{13}$$

where the summation is along the trajectory, and the secondary structure is assigned as "ss" when

$$g_{ss} = \max(1.1g_H, 0.75g_E, g_{coil}) \tag{14}$$

The weight factors correct for the diffuse character of the observed secondary structure.

Table 2 gives the appropriate statistics obtained at a selected temperature, $T = 1.0$, for the two models of interactions discussed above. For all test sequences, except for 1gb1, a multiple-sequence alignment was implemented. As mentioned above, highly homologous sequences, when found in the sequence database, were treated as the test sequence, thereby increasing somewhat the strength of the statistics (a larger number of 19-residue-sequence fragments have to be compared to the structural templates via the threading procedure). Of course, those sequences whose structures exist in the structural

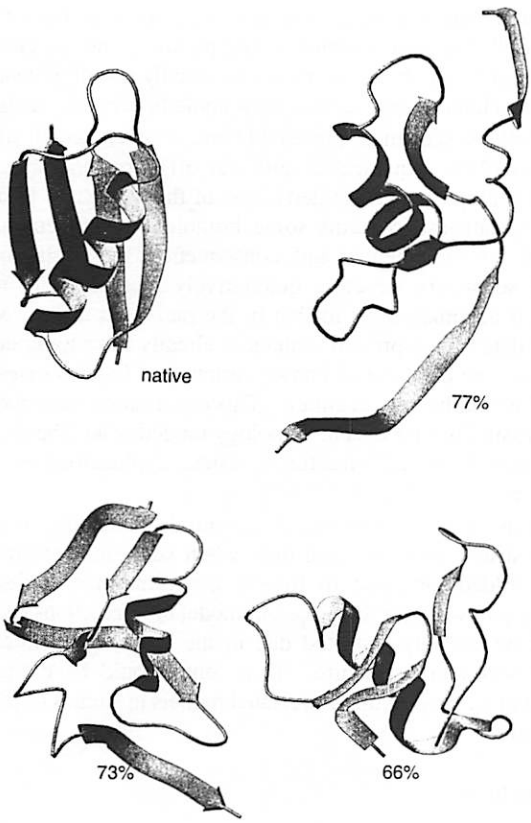

Figure 10. Native structure of 1gb1 and three representative snapshots from the isothermal (T = 1) simulation of the model protein. MOLMOL[64] drawings are based on Cα traces reconstructed from the side group coordinates. The helical ribbons and arrows correspond to the fragments assigned by the algorithm as helical and extended, respectively.

database employed in this work have been omitted. The instantaneous (for a single snapshot) assignment of secondary structure is often worse than the average numbers given in Table 2. However, the system is not frozen, and it is reasonable that the average propensities reproduced by this model of polypeptide dynamics are more relevant for future modeling of the protein folding process and the knowledge-based scheme of short-range interactions.

Some representative snapshots for the relatively short chain of 1gb1 are given in Figure 10. The native structure is given for comparison. The elements of secondary structure (helices and expanded fragments) are marked according to DSSP assignment for the native structure and according to the assignment procedure described above for the conformations obtained from the Monte Carlo dynamics. For convenience, we plot the estimated Cα trace instead of the original, side group-based chain.

Analysis of the data given in Table 2 leads to several interesting observations. First, the secondary-structure propensities of the model potential reproduce native secondary structure with good fidelity. The average accuracy for the three-letter code assignment equals 72.2%. This is on the same level as the most accurate methods of secondary-structure prediction.[49−51] But, of course, due to computational limitations, we consider a much smaller testing set here. The test predictions were obtained using potentials that were derived after removing all similar (25% threshold) sequences from the structural database. When the potential is built without this restriction (which could be the case for a "blind" prediction), the average accuracy is considerably improved and increases to 77.8% (see Table 2). In this case, the homologous sequences (and consequently similar structures) give rise to the fidelity of the potential, but do not control the resulting model. Otherwise, the secondary-structure accuracy would be much higher. Indeed, for different purposes, the short-range potential could be derived from only a structural library containing homologous sequences (when available). Then, the resulting secondary structure is almost exactly similar to classical homology modeling. These experiments show that the model (even in its preliminary form, without any tertiary interactions) could be used as a tool for secondary-structure prediction.

The fidelity of secondary structure seen in the model chain during Monte Carlo simulations is higher than the accuracy of the secondary structure recovered just from local sequence alignment via appropriate analysis of the resulting potential of mean force. The last column of Table 2 gives the result obtained from the statistical analysis. (We used the same procedure for secondary-structure translation as in the Monte Carlo simulations, except that all three secondary-structure states are equally weighted; see eq 14 (i.e., $g_{ss} = \max(g_H, g_E, g_{coil})$ of the geometry of the top-scoring fragments from Fischer's database.) Comparison with the other data from Table 2 clearly shows that the fidelity of the obtained secondary structure increases noticeably due to chain connectivity, excluded volume, and confinement of the chain into a loosely defined sphere. This is because the presence of a geometrical context provides for more self-consistent predictions. The information content about chain geometry is actually much richer than that provided by the three-letter code. For example, one may identify with the high-fidelity positions of the protein surface turns where the chain changes the direction of propagation.[57,66] This extension, however, lies beyond the scope of the paper.

Second, one may note that the generic contribution to the model potential (despite being completely sequence-independent) somewhat increases the average accuracy of the prediction. The difference is small for the prediction obtained from the weighted average over the Monte Carlo trajectory described above. We note that the prediction is much more exact due to the contribution of generic terms. These propagate secondary structure and regularize the chain geometry by interpolating between various "strongly predicted" fragments in a more "proteinlike" fashion. When generic terms are included, the average length of a secondary-structure element increases by 7.3%, and the number of longer, regular secondary-structure elements (longer than 8-residue helices and 5-residue expanded fragments) increases by 37.5% and becomes closer to that seen in real proteins. However, it should be pointed out that the purpose of this work was not to build yet another scheme for secondary-structure prediction. Analysis of the secondary-structure fidelity on a limited (but representative) set of proteins was done to evaluate the possibility of building a reasonable interaction scheme into a very simple lattice model in the form of generic proteinlike biases and a local sequence similarity based statistical potential for short-range interactions. Perhaps this could be used as a useful tool for medium-resolution computer studies of realistic protein models.

4. Discussion

In reduced modeling of protein structure and dynamics, one always has to take into consideration a tradeoff between computational simplicity and geometric fidelity. Simple lattice models could be studied in great detail; however, such an approach may distort some important aspects of protein physics. On the other hand, more detailed, especially continuous space,

models could be more difficult to control. One needs much more extensive simulations to estimate the resulting properties of these complex model systems. The protein model proposed in this work has some virtues of both approaches (i.e., it uses quite a simple lattice representation (single-interacting unit per residue)) and has relatively good geometric fidelity. Yet, the large number of lattice vectors representing hypothetical bonds between the centers of mass of the protein side chains makes the dynamic properties of such a model close to continuous space representations. Thus, the potential problems of lattice anisotropy could be safely dismissed. Such a fluctuating bond lattice chain itself is, perhaps, a good general model of a long flexible polymer. To build a reasonable protein model, it is necessary to introduce strong, proteinlike conformational biases.

The data presented in this work show that it is feasible to design a side chain model of a protein chain that, despite its extreme simplicity, is capable of reproducing polypeptide chain geometry with rather high fidelity. This was made possible by the careful design of model geometry and the inclusion of generic structure regularizing potentials. In the past, we have shown that an equivalent regularization of reduced protein models was necessary for somewhat more complex models comprised of two united atoms per residue (Cα + side group).[13] In that case, as here, we also attempted to build into the model the characteristic stiffness of protein chains and a generic tendency toward formation of helices and expanded conformations. Due to the more regular geometry of the Cα reduced backbone for the Cα-based models, this task was relatively easy. In this work, it has been shown that the same results are possible in a much simpler side chain-based representation. We also seem to have achieved a better average accuracy of secondary-structure representation in the present model of potential and representation.

Quantitative comparison with our previous work employing reduced lattice models[13,57] is rather difficult because here, for the first time, we applied a potential based on the straightforward usage of sequence-structure compatibility. In different contexts, such an approach is by no means new. Salamov and Solovyev[65] employed query sequence alignments with sequences whose structures are known as a method of secondary-structure prediction. The achieved accuracy (in the three-letter secondary-structure code) was about 71% and increased to 73.5% when the multiple-sequence alignment was employed, similar to the first step of the approach employed in this work. Common for all such approaches are the probable distributions of the local distance geometry that could be built for any sequence of interest. The idea of using these geometric characteristics for derivation of various local potentials of mean force is also not new and has been employed in many works.[10,11,26,66,67,68] A classical implementation has been proposed by Sippl.[67,68] The use of distance restraints is also typical for more advanced methods of homology modeling.[69] It should be noted that very good secondary-structure fidelity has been obtained without including any tertiary interactions. Consequently, application of this model for tertiary-structure predictions could be very promising due to the expected reduced competition between secondary-structure propensities and tertiary interactions. Marginally, let us note that trajectories obtained from simulations described in this work provide quite a bit of supersecondary-structural information. For example, even a casual inspection of the snapshots of the 1gb1 chains given in Figure 10 (forgetting for the moment the native conformation) strongly suggests that the protein consists of a helix and four-stranded β-sheets, as is indeed the case. In this respect, the application presented here resembles the approach used in the LINUS method.[70] For longer proteins, the picture is not as clean as for the 1gb1 case; however, most (but usually not all) secondary-structure elements and surface turns could be correctly assigned.

For all test sequences presented here, we excluded all similar proteins (25% sequence identity cut off) from the structural database employed in the derivation of the potential. Usually, when a database contains some homologous sequences, the accuracy of the potential and, consequently, the fidelity of the model secondary structure qualitatively increases. In many cases, it becomes close to that in the native structure. Many newly determined protein sequences already have weak homologues in the database of known structures. In such cases, the model would be very accurate. This observation may open up new possibilities for distant homology modeling and the de novo prediction of protein structures. These applications are now being investigated.

The use of homology-based "potentials" extracted from the native structures is justified only when structural information about folded, or close to folded, conformations is desired. Folding pathways for the proposed model of interactions perhaps would be strongly distorted due to the very large content of native secondary structure. Thus, one should be extremely careful in employing this and related models in studies of protein dynamics.

5. Conclusion

A new, very simple and efficient model of protein chains and polypeptide short-range interactions has been proposed. The model employs a single side chain-based united atom representation of amino acid residues located at the centers of protein side chains. The resulting chain has a complex geometry and is embedded into an underlying simple cubic lattice with a fine mesh spacing of 1.45 Å. Each side group occupies a cluster of lattice points, and the "virtual bond" between two successive chain units can assume a spectrum of lengths and orientations. These implicitly account for the correlated conformational transitions of the protein backbone and side chains (for those residues having internal degrees of conformational freedom). Due to its purely lattice-based structure, the resulting model is computationally extremely efficient (the cost of a local move, including energy computations, is independent of chain length), thereby allowing simulations of large-scale transitions in protein systems. The model's force field consists of generic terms that account for geometric regularities seen in all proteins and sequence-specific potentials that are different for each protein. Inclusion of characteristic proteinlike short-range distance correlations result in an accurate representation of protein conformational stiffness and secondary-structure propensities. Sequence-specific "potentials" are local homology-based and are rather accurate. The predicted secondary structure exhibits an accuracy comparable to the best secondary-structure prediction methods. The advantage of this model is that such secondary-structure information is now provided in a more complete geometric content. Thus, it has been shown that the proposed very simple protein representation enables meaningful modeling of secondary-structural propensities. As a result, this model may be a good candidate for other applications in structural studies of protein systems.

Acknowledgment. This work was partially supported by NIH Grant GM37408, by the University of Warsaw Grant BST-34/97, and by HHMI (A.K. is an International Scholar of the Howard Hughes Medical Institute) Grant 75195-543402. Helpful

discussions with Drs. Adam Godzik and Angel Ortiz are gratefully acknowledged.

References and Notes

(1) Branden, C.; Tooze, J. *Introduction to protein structure*; Garland Publishing: New York, London, 1991.
(2) Brooks, C. L., III.; Karplus, M.; Pettitt, B. M. *Adv. Chem. Phys.* **1988**, *71*, 1−259.
(3) Kolinski, A.; Skolnick, J. *Lattice Models of Protein Folding, Dynamics and Thermodynamics*; R. G. Landes: Austin, TX, 1996.
(4) Levitt, M. *Curr. Opin. Struct. Biol.* **1991**, *1*, 224−229.
(5) Dill, K. A.; Bromberg, S.; Yue, K.; Fiebig, K. M.; Yee, D. P.; Thomas, P. D.; Chan, H. S. *Protein Sci.* **1995**, *4*, 561−602.
(6) Skolnick, J.; Kolinski, A. Monte Carlo Lattice Dynamics and The Prediction of Protein Folds. In *Computer Simulations of Biomolecular Systems. Theoretical and Experimental Studies*; van Gunsteren, W. F., Weiner, P. K., Wilkinson, A. J., Eds.; ESCOM Science Publ.: Leiden, The Netherlands, 1997.
(7) Sun, S. *Protein Sci.* **1993**, *2*, 762−785.
(8) Dill, K., A. *Curr. Biol.* **1993**, *3*, 99−103.
(9) Dinner, A. R.; Sali, A.; Karplus, M. *Proc. Natl. Acad. Sci. U.S.A.* **1996**, *93*, 8356−8361.
(10) Kolinski, A.; Skolnick, J. *Proteins* **1994**, *18*, 353−366.
(11) Kolinski, A.; Skolnick, J. *Proteins* **1994**, *18*, 338−352.
(12) Kolinski, A.; Galazka, W.; Skolnick, J. *J. Chem. Phys.* **1995**, *103*, 10286−10297.
(13) Kolinski, A.; Milik, M.; Rycombel, J.; Skolnick, J. *J. Chem. Phys.* **1995**, *103*, 4312−4323.
(14) Kolinski, A.; Galazka, W.; Skolnick, J. *Proteins* **1996**, *26*, 271−287.
(15) Hao, M.-H.; Scheraga, H. A. *J. Phys. Chem.* **1994**, *98*, 4940−4948.
(16) Hao, M.-H.; Scheraga, H. A. *J. Phys. Chem.* **1994**, *98*, 9882−9893.
(17) Hao, M.-H.; Scheraga, H. A. *J. Chem. Phys* **1995**, *102*, 1334−1348.
(18) Hao, M.-H.; Scheraga, H. A. *Proc. Natl. Acad. Sci. U.S.A.* **1996**, *93*, 4984−4989.
(19) Hao, M.-H.; Scheraga, H. A. *J. Phys. Chem.* **1996**, *100*, 14540−14548.
(20) Godzik, A.; Kolinski, A.; Skolnick, J. *J. Comput. Chem.* **1994**, *14*, 1194−1202.
(21) Hinds, D. A.; Levitt, M. *Proc. Natl. Acad. Sci. U.S.A.* **1992**, *89*, 2536−2540.
(22) Rey, A.; Skolnick, J. *Chem. Phys.* **1991**, *158*, 199−219.
(23) Rey, A.; Skolnick, J. *Proteins* **1993**, *16*, 8−28.
(24) Hagler, A. T.; Honig, B. *Proc. Natl. Acad. Sci. U.S.A.* **1978**, *75*, 554−558.
(25) Levitt, M.; Warshel, A. *Nature* **1975**, *253*, 694−698.
(26) Levitt, M. *J. Mol. Biol.* **1975**, *104*, 59−107.
(27) Hoffmann, D.; Knapp, E. W. *Eur. Biophys. J.* **1996**, *24*, 387−403.
(28) Hoffmann, D.; Knapp, E. W. *Phys. Rev. E* **1996**, *53*, 4221−4224.
(29) Honeycutt, J. D.; Thirumalai, D. *Biopolymers* **1992**, *32*, 695−709.
(30) Knapp, E. W.; Irgens-Defregger, A. J. *Comput. Chem.* **1993**, *14*, 19−29.
(31) Park, B. H.; Levitt, M. *J. Mol. Biol.* **1995**, *249*, 493−507.
(32) Go, N. *Annu. Rev. Biophys. Bioeng.* **1983**, *12*, 183-210.
(33) Sali, A.; Shakhnovich, E.; Karplus, M. *J. Mol. Biol.* **1994**, *235*, 1614−1636.
(34) Sali, A.; Shakhnovich, E.; Karplus, M. *Nature* **1994**, *369*, 248−251.
(35) Shakhnovich, E.; Farztdinov, G.; Gutin, A. M. *Phys. Rev. Lett.* **1991**, *67*, 1665−1668.
(36) Shakhnovich, E. I. *Folding Des.* **1996**, *1*, R50−R54.
(37) Skolnick, J.; Kolinski, A. *Science* **1990**, *250*, 1121−1125.
(38) Skolnick, J.; Kolinski, A. *J. Mol. Biol.* **1991**, *221*, 499−531.
(39) Skolnick, J.; Kolinski, A.; Yaris, R. *Proc. Natl. Acad. Sci. U.S.A.* **1988**, *85*, 5057−5061.
(40) Skolnick, J.; Kolinski, A. *Annu. Rev. Phys. Chem.* **1989**, *40*, 207−235.
(41) Skolnick, J.; Kolinski, A.; Yaris, R. *Biopolymers* **1989**, *28*, 1059−1095.
(42) Skolnick, J.; Kolinski, A.; Yaris, R. *Proc. Natl. Acad. Sci. U.S.A.* **1989**, *86*, 1229−1233.
(43) Skolnick, J.; Kolinski, A. *J. Mol. Biol.* **1990**, *212*, 787−817.
(44) Richardson, J. *Adv. Protein Chem.* **1981**, *34*, 167−339.
(45) Ramachandran, G. N.; Sassiekharan, V. *Adv. Protein Chem.* **1968**, *28*, 283−437.
(46) Anfinsen, C. B.; Scheraga, H. A. *Adv. Protein Chem.* **1975**, *29*, 205−300.
(47) Bowie, J. U.; Reidhaar, O. J. F.; Lim, W. A.; Sauer, R. T. *Science* **1990**, *247*, 1306−1310.
(48) Thornton, J. M.; Flores, T. P.; Jones, D. T.; Swindells, M. B. *Nature* **1991**, *354*, 105−106.
(49) Rost, B.; Sander, C. *J. Mol. Biol.* **1993**, *232*, 584−599.
(50) Rost, B.; Sander, C. *Proteins* **1994**, *19*, 55−72.
(51) Rost, B.; Sander, C. *Proteins* **1996**, *23*, 295−300.
(52) Godzik, A.; Skolnick, J.; Kolinski, A. *J. Mol. Biol.* **1992**, *227*, 227−238.
(53) Ponder, J. W.; Richards, F. M. *J. Mol. Biol.* **1987**, *193*, 775−791.
(54) Sali, A.; Potterton, L.; Yuan, L.; van Vlijmen, H.; Karplus, M. *Proteins* **1995**, *23*, 318−326.
(55) Metropolis, N.; Rosenbluth, A. W.; Rosenbluth, M. N.; Teller, A. H.; Teller, E. *J. Chem. Phys.* **1953**, *51*, 1087−1092.
(56) Rouse, P. E. J. *J. Chem. Phys* **1953**, *21*, 1272−1278.
(57) Kolinski, A.; Skolnick, J. *J. Chem. Phys.* **1997**, *107*, 953−964.
(58) Tonges, U.; Perry, S. W.; Stoye, J.; Dress, A. W. *Gene* **1996**, *172*, GC33−GC41.
(59) Henikoff, S.; Henikoff, J. G. *Proc. Natl. Acad. Sci. U.S.A.* **1992**, *89*, 10915−10919.
(60) Kabsch, W.; Sander, C. *Biopolymers* **1983**, *22*, 2577−2637.
(61) Fischer, D.; Eisenberg, D. *Protein Sci.* **1996**, *5*, 947−955.
(62) Schneider, R.; de Daruvar, A.; Sander, C. *Nucleic Acids Res.* **1997**, *25*, 226−230.
(63) Bernstein, F. C.; Koetzle, T. F.; Williams, G. J. B.; Meyer, E. F., Jr.; Brice, M. D.; Rodgers, J. R.; Kennard, O.; Simanouchi, T.; Tasumi, M. *J. Mol. Biol.* **1977**, *112*, 535−542.
(64) Koradi, R. *J. Mol. Graph.* **1996**, *14*, 51−55.
(65) Salamov, A. A.; Solovyev, V. V. *J. Mol. Biol.* **1997**, *268*, 32−36.
(66) Kolinski, A.; Skolnick, J.; Godzik, A.; Hu, W.-P. *Proteins* **1997**, *27*, 290−308.
(67) Sippl, M. J. *J. Mol. Biol.* **1990**, *213*, 859−883.
(68) Sippl, M. J. *Curr. Biol.* **1995**, *5*, 229−235.
(69) Sali, A.; Blundell, T. L. *J. Mol. Biol.* **1993**, *234*, 779−815.
(70) Srinivasan, R.; Rose, G. D. *Proteins* **1995**, *22*, 81−99.