Determinants of secondary structure of polypeptide chains: Interplay between short range and burial interactions

Andrzej Kolinski[a)]
Department of Chemistry, University of Warsaw, Pasteura 1, 02-093 Warsaw, Poland and Department of Molecular Biology, The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, California 92037[b)]

Jeffrey Skolnick[c)]
Department of Molecular Biology, The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, California 92037

(Received 17 December 1996; accepted 8 April 1997)

The effect of tertiary interactions on the observed secondary structure found in the native conformation of globular proteins was examined in the context of a reduced protein model. Short-range interactions are controlled by knowledge based statistical potentials that reflect local conformational regularities seen in a database of three-dimensional protein structures. Long-range interactions are approximated by mean field, single residue based, centrosymmetric hydrophobic burial potentials. Even when pairwise specific long-range interactions are ignored, the inclusion of such burial preferences noticeably modifies the equilibrium chain conformations, and the observed secondary structure is closer to that seen in the folded state. For a test set of 10 proteins (belonging to various structural classes), the accuracy of secondary structure prediction is about 66% and increases by 9% with respect to a related model based on short-range interactions alone [Kolinski *et al.*, J. Chem. Phys. **103**, 4312 (1995)]. The increased accuracy is due to the interplay between the short-range conformational propensities and the burial and compactness requirements built into the present model. While the absolute level of accuracy assessed on a per residue basis is comparable to more standard techniques, in contrast to these approaches, the conformation of the chain now has a better defined geometric context. For example, the assumed spherical domain protein model that simulates the segregation of residues between the hydrophobic core and the hydrophilic surface allows for the prediction of surface loops/turns where the polypeptide chain changes its direction. The implications of having such self-consistent secondary structure predictions for the prediction of protein tertiary structure are briefly discussed. *© 1997 American Institute of Physics.* [S0021-9606(97)01527-4]

I. INTRODUCTION

It is frequently assumed that a key to prediction of the native conformation of a protein lies in the prior prediction of its secondary structure.[1,2] Having such information in hand, one could then assemble the native fold from its constituent secondary structural elements followed by fine-tuning the atomic details.[2–4] Assuming three classes of secondary structure [a helix (H), expanded β-type conformation (E), and everything else, i.e., coil/turn (−)], classical prediction methods achieve an accuracy ranging from 55% to 65%.[1,5] Even using the most elaborate methods that employ multiple sequence alignment information, the resultant level of accuracy is about 70%–75%.[6] A likely origin of the limitations in accuracy is the fact that all classical methods of protein secondary structure prediction are inherently local in nature. In reality, the secondary structure seen in the native conformation of globular proteins may reflect an energetic compromise between the local conformational propensities and global restraints emerging from close packing of globular proteins, specific patterns of side chain interactions, hydrogen bond restraints, etc. Indeed, some short sequence fragments adopt a helical conformation in one protein, while in another protein, the same fragments is part of a β sheet.[7] Consequently, exact prediction of secondary structure is equivalent to the prediction of tertiary structure, an as yet unsolved problem. While the idea that tertiary interactions modify secondary structure is widely believed to be true, this effect has not been explicitly investigated in any protein model. Thus, in the context of a reduced protein model, we explicitly examine whether incorporation of some tertiary information enhances the accuracy of secondary structure prediction and explore what additional information can be provided by such an analysis.

Recently we described a reduced model of protein structure and dynamics and proposed a factorization of short-range interactions that reproduced the secondary structure of globular proteins with an accuracy of about 60% (50%–75%, depending on the sequence) for three structural classes (helix, H, extended, E, and - everything else).[8] This model of short-range interactions was then implemented in a reduced protein model that allowed Monte Carlo (MC) folding of a number of small proteins.[9]

[a)] Author to whom correspondence should be addressed. Electronic mail: Kolinski@chem.uw.edu.pl.

[b)] Address for correspondence.

[c)] Electronic mail: Skolnick@scripps.edu; fax: (619) 784-8895.

In this work, we present a related model that incorporates some aspects of long-range interactions typical of single domain, globular proteins. These long-range interactions are limited to one body, residue specific mean-field potentials that reflect preferences for the location of various amino acids within a globular structure, that is, they play the role of a burial energy. In the absence of any pairwise (and higher order) long-range interactions, the Metropolis Monte Carlo (MMC) sampling[10] of the model is very fast. Furthermore, the assumption of a close to spherical shape also permits the imposition of some global restraints that can moderate the protein secondary structure. For example, because globular proteins are compact, regular secondary structural elements cannot be too long or too short.[11] Thus, the aim of the present study is to analyze the interplay between these global restraints: secondary structure propensities, protein compactness, and hydrophobic–hydrophilic phase separation as embodied by a one body approximation to the hydrophobic burial potential. We expect that the tertiary ‘‘perturbation’’ will moderate the local conformations of the model polypeptide, thereby allowing for more accurate secondary structure predictions based on the statistics of the chain geometry at low temperature. Due to the approximate treatment of the long-range interactions, the method is applicable to all single domain globular proteins or to well defined domains of multidomain proteins. In the latter, the division of the protein into domains must be done by a different method.

The assumption of a spherical domain globular protein model was recently employed by us in a very similar context.[12] There, the goal was to predict the most probable set of ‘‘hairpins’’ defined as a regular fragment of secondary structure followed by a surface loop or turn where the chain reverses global direction and then another regular fragment of secondary structure. This prediction was done for the sequence of interest by threading randomly selected fragments of protein structure through a hypothetical, spherical globule. That is, the protein consists of a set of hairpins that are in essence stitched together. The resulting model exhibited very high accuracy in the prediction of loop regions and the dominant secondary structure of regular (transglobular) fragments.[12] However, the accuracy of prediction of the secondary structure assignments on a per residue basis is moderate due to the ‘‘overregularization’’ of the structures and frequent errors near the loop regions. The latter are due to the very approximate way that the hairpins were constructed. In the present study, we explore a similar set of interactions, but which are now applied to a continuous chain, thereby enforcing a more self-consistent manifold of local conformations that define the secondary structure assignment.

The outline of this article follows. In Sec. II, we briefly describe the lattice model, the MC sampling technique, and the interaction scheme. Short-range interactions are exactly the same as those described previously,[8] and therefore only a short summary is provided for the reader's convenience. The approximations of the long-range (one body) interactions are discussed in more detail. Next, the method is applied to a set of 10 representative test proteins, and an analysis of the information provided by this approach is presented. We conclude with a discussion of our results and possible directions of future research.

II. METHODS

The method of secondary (and to some extent supersecondary) structure prediction presented here is based upon a high coordination number lattice model of protein structure and dynamics developed over the last few years by our group.[3,4,8,9,13,14] This model has been used for studies of polypeptide dynamics,[8] protein folding thermodynamics,[14] structure prediction,[4] and other aspects of protein biophysics.[4] Recently we undertook an effort to refine the entire force field of the model and to carefully reexamine the contribution of various interactions and their effects on model protein properties,[8,15,16] the overall goal being to develop better, more sensitive potentials. This article represents another step in that direction.

A. Lattice representation of polypeptides

The Cα trace is modeled as a lattice chain that consists of a sequence of vectors belonging to the following 90 basis vector set $\{(3,1,1),\ldots\ (3,1,0),\ldots\ (3,0,0),\ldots\ (2,2,1),\ldots\ (2,2,0),\ldots\}$. The best fit of such a lattice chain to high resolution protein structures in the Brookhaven Protein Data Bank (PDB)[17,18] is obtained when the mesh size of the underlying simple cubic lattice is assumed to be equal to 1.22 Å. As a result, the average length of a Cα–Cα segment on the lattice is equal to 3.8 Å, and the fluctuations of the Cα–Cα distance do not exceed ± 0.3 Å. In contrast to low coordination lattice models of proteins, the accuracy of protein representation is essentially independent of the orientation of the fitted structures with respect to the lattice principal axis.[19,20]

From the fit of a set of high resolution, nonhomologous proteins to the lattice, one can derive statistics of the occurrence of particular triplets of consecutive backbone vectors (say, $\mathbf{v}_{i-1},\mathbf{v}_i\,.\mathbf{v}_{i+1}$).[8] Many triplets never occur, while others are extremely rare, and perhaps result from database errors, structure inaccuracy, or fitting errors. Whatever their origin, it is assumed that such conformations are very unlikely, and they are prohibited in the model. Interestingly, the set of allowed three-vector conformations derived from the straightforward statistics of the lattice projection of protein three-dimensional structures almost exactly overlaps with the set resulting from restrictions superimposed on virtual bond angles (between two subsequent Cα vectors) and distance restraints for the ith and $i+3$th α-carbons.

Previously we have shown that the sequence of three α-carbon vectors defines the orientation of the central (for the fragment) planar (trans) peptide bond unit.[21] Two consecutive α-carbon virtual bonds provide a reference frame for the definition of the side chain position. In this work, we employed a single rotamer representation of the side chains (corresponding to the center of mass of the most probable rotameric isomeric state).[22] For Gly residues, the side chains coincide with the α-carbon positions. Side chain positions are employed in an approximation of the hydrophobic burial potential. Figure 1 shows a representative conformation of a

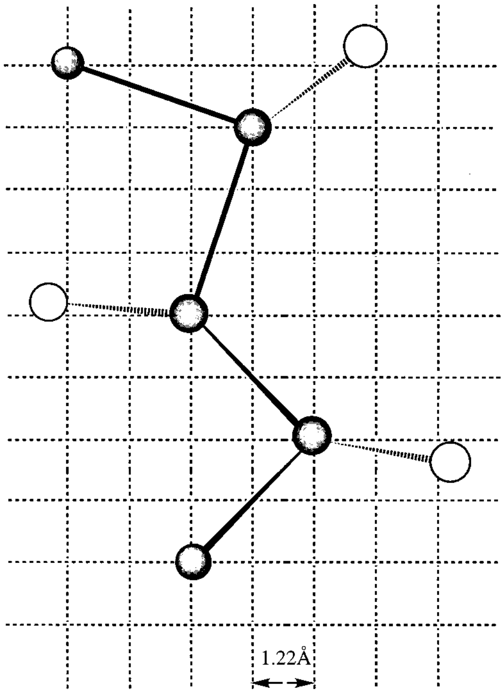

FIG. 1. Illustration of the lattice model of protein chain. The closed circles correspond to the α-carbon backbone, the open circles are side chains. A single rotamer (the most probable position of the side chain center of mass) approximation was used.

short fragment of the model polypeptide restricted to the ''310'' lattice.

B. Sampling procedure

Conformational space is sampled according to the standard asymmetric Metropolis[10] scheme with the transition probability from an ''old'' state to a ''new'' state equal to $P(\text{new/old}) = \exp[-(E_{\text{new}} - E_{\text{old}})/k_B T]$, with k_B Boltzmann's constant and T the absolute temperature. The conformational transitions attempted in a single MC cycle consist of two-bond end moves, two- and three-bond moves, larger fragment moves generated by long distance (up to 30 bonds along the chain according to a random selection of the distance) permutations of two chain vectors and similar longer distance moves employing permutations of two pairs of vectors. The longer distance moves facilitate faster rearrangements of more rigid secondary structure elements (e.g., helices). These are quite important due to global restraints superimposed onto the model chains. For a chain of length N, such medium distance moves are attempted with a frequency equal to $1/N$ with respect to the frequency of the local micromodifications. Examples of the short-range and medium-range moves are schematically shown in Fig. 2.

The sampling algorithm is ''local,'' i.e., the cost of attempting a single micromodification does not depend on the chain length. This was achieved by using a lattice occupancy list to detect self-overlaps of the chain units, as described later. Thus, the expense of a single sampling step (which corresponds to unit time in the model), consisting of N attempts at both types of local moves and single attempts at medium-range moves, is proportional to the chain length, N. Since the longest relaxation time of polymeric chains scales roughly as N^2, the total cost of the simulation scales as N^3. One experiment (chain collapsing upon the simulated thermal annealing followed by the isothermal sampling) for a 50-residue protein requires about 10–15 min CPU on a HP-735 workstation running at 125 MHz, and grows to a few hours for a 150-residue protein. Thus, the method is clearly unsuitable for massive screening of protein sequences, but it is relatively inexpensive when applied to selected cases.

To obtain reasonable estimates of the relevant conformational properties in the interesting low temperature range, the model chains were slowly ''cooled'' from random expanded states and were then subjected to an isothermal sampling run. This procedure was repeated several times, and the trajectories from the runs with the lowest average conformational energy were taken for the final analysis.

C. Short-range interactions

The short-range interaction scheme was described and examined previously.[8] Here, for the convenience of the reader, a brief overview of the various terms is given. Short-range potentials consist of four-contributions. Three are generic and do not depend on amino acid sequence. The role of the generic terms is to provide a strong bias toward a ''proteinlike'' distribution of main chain conformations. The first generic term comes from the statistics of the three-vector fragments of the PDB lattice replicas of globular proteins and is equivalent to an effective Ramachandran torsional potential[23] for these reduced models

$$\varepsilon_g = f(\mathbf{v}_{i-1}, \mathbf{v}_i, \mathbf{v}_{i+1}). \tag{1}$$

The potential is encoded in the form of a histogram, defined in terms of six bins of the ''chiral'' value of the square end-to-end distance for three-vector fragments, $r^2*_{i-1,i+2}$, defined as follows:

$$r^2*_{i-1,i+2} = r^2_{i-1,i+2} \ \text{sign}((\mathbf{v}_{i-1} \otimes \mathbf{v}_i) \cdot \mathbf{v}_{i+1}). \tag{2}$$

The binning definition and numerical values of this potential are given in Table I. The fourth bin corresponds to the right-handed helical conformations, while bin Nos. 1 and 6 correspond to the expanded, β-type conformations.

The second generic short-range interaction term provides a longer distance bias toward a proteinlike distribution of states. This favors ''regular'' elements of secondary structure, i.e., helices and β-type expanded states.

$$\eta_i = f(r_{i-2,i+2}) \tag{3}$$

and the function η_i is of the following form, where

$$\begin{aligned} \eta_i &= -1, \quad \text{for } (r^2_{i-2,i+2})^{1/2} < 6.2 \text{ Å}, \\ \eta_i &= -1, \quad \text{for } (r^2_{i-2,i+2})^{1/2} > 10.6 \text{ Å}, \\ \eta_i &= 0 \quad \text{otherwise.} \end{aligned} \tag{4}$$

The third generic term is somewhat more complicated and reflects the stiffness of protein chains. The idea is based

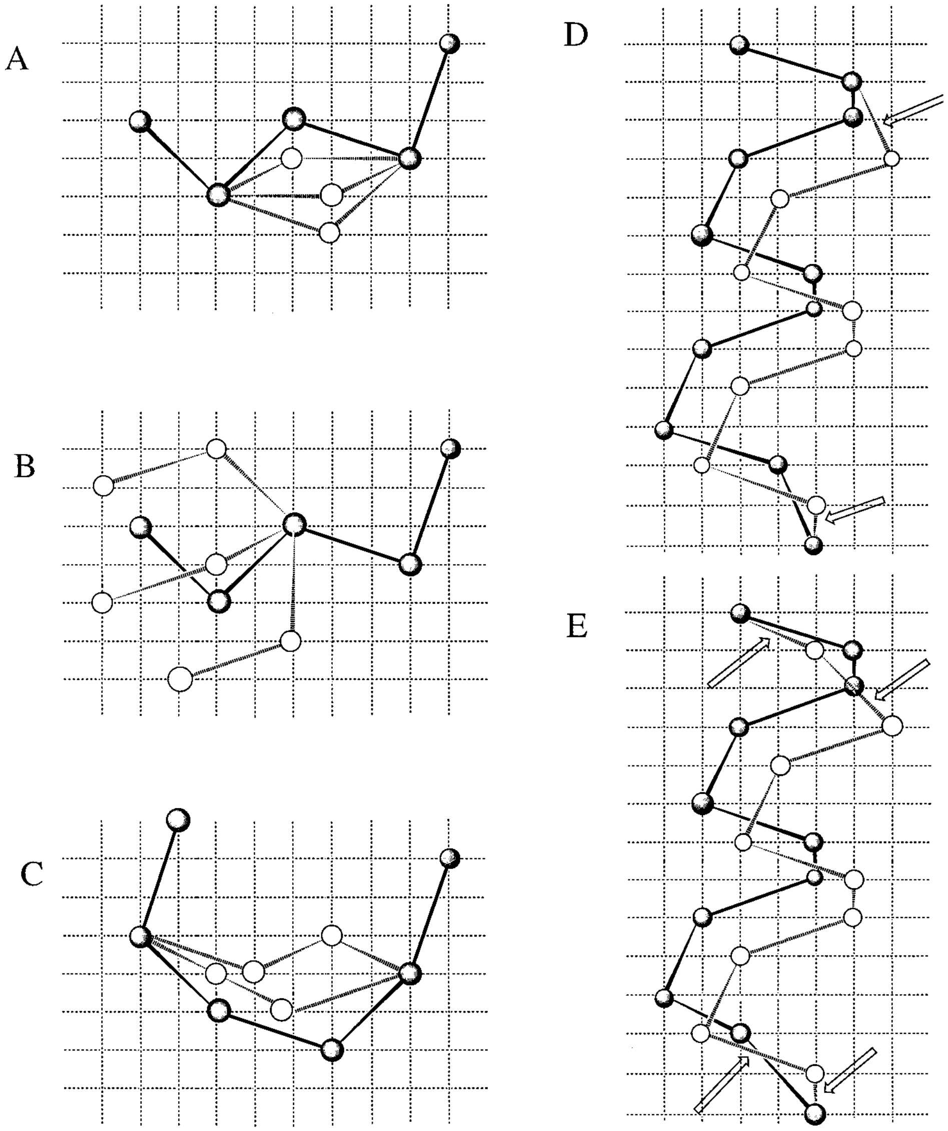

FIG. 2. Elementary moves employed in the Monte Carlo sampling algorithm. (A) examples of single residue, two-vector moves. The solid line marks an ‘‘old’’ conformation, while the dashed lines and open symbols of Cα united atoms symbolize a subset of possible new conformations. The number of new conformations depends on the old conformation and the maximum number is equal to 11. The specific new conformation is selected by a pseudorandom mechanism. Only moves that lead to ‘‘proteinlike’’ conformations of all involved three-vector fragments (i.e., conformations that occur in known protein structures) could be accepted by the algorithm. (B) examples of end move. Here the number of allowed new conformations is bigger due to the larger conformational flexibility of chain ends. (C) several examples of three-bond moves, (D) a longer distance, two bond permutation move. The virtual Cα bonds indicated by arrows are the only ones affected by this kind of move. (E) a longer distance four-bond permutation move. The bonds indicated by the arrows in the top of the figure are permuted with the two bonds on the bottom of the figure, and the intervening portion of the chain translates in a ‘‘rigid bodylike’’ fashion.

TABLE I. Sequence independent torsional potential.

Bin No.	Description of conformation	Range of $r^{2}*_{i-i,i+2}$ (in lattice units)	ε_g (in k_BT)
1	Expanded, beta	−89, −57	−0.052
2	Coil/turn	−56, −26	0.105
3	Left-handed helix	−25, 0	2.474
4	Right-handed helix	0, 25	−0.987
5	Coil/turn	26, 55	0.075
6	Expanded, beta	56, 91	1.043

on the observations that the mutual orientations of certain pairs of peptide bond plates are highly correlated in the elements of secondary structure found in folded proteins.

$$\varepsilon_p = \cos(\mathbf{h}_i, \mathbf{h}_{i+2}) + \cos(\mathbf{h}_i, \mathbf{h}_{i+4}), \tag{5}$$

where $\cos(\mathbf{h}_i, \mathbf{h}_j)$ denotes the cosine of the angle between the ith and jth vectors defining the orientation of the peptide bond plates (the vectors from amide hydrogen to the nitrogen and carbonyl oxygen). These peptide vectors are parallel along the helical fragments. In expanded states, every pair of second (and forth) peptide bond vectors is parallel. The idea

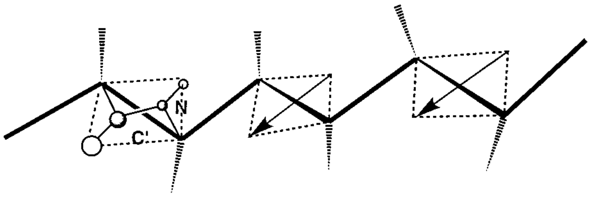

FIG. 3. Illustration of the geometry employed in the definition of the generic stiffness of the model polypeptide; some pairs of peptide bond vectors are almost always close to parallel orientations in real protein structures.

behind the reconstruction of all backbone atomic coordinates and the correlation between the peptide plates is depicted in Fig. 3.

The sequence specific part of the short-range interactions was defined in a similar way as the generic potential [Eqs. (1) and (2)]; however, this term depends on the identity of the two consecutive amino acids:

$$\varepsilon_s = f(A_i, A_{i+1}, r^{2*}{}_{i-1,i+2}). \tag{6}$$

The potential is used in the form of six bin histograms that are amino acid pairwise dependent depending on the identity of residue A_i and A_{i+1}, where A_k is the identity of the kth residue in the sequence. The potential is available upon request or can be downloaded from an anonymous ftp site.[24]

The total conformational energy associated with the short-range interactions has been computed as follows:

$$E_{\text{short}} = \sum (4\varepsilon_s + 1.5\varepsilon_g + \varepsilon_\eta + \varepsilon_p), \tag{7}$$

where the summation is performed along the peptide chain. The scaling of the sequence specific interactions relative to the generic terms is to some extent arbitrary. These scaling factors were previously adjusted by trial and error methods for a few representative proteins belonging to various structural classes.[14] In the presence of long-range interactions (mostly due to a surface effect associated with the segregation of polar and nonpolar residues on the protein surface; see Sec. II D), the short-range interactions have to compete with the tertiary preferences. Thus, in order to compensate for this effect, relative to our early work, the contribution of one of the generic potentials (ε_g) was increased from 1.0 to 1.5. Due to our approximate account of long-range interactions, the secondary structure is more regular than was seen previously in their absence. Consequently a more precise definition of the secondary structure in the scoring procedure (see Sec. II E) for the predictions may be used. This is another reason for applying stronger short-range terms than had been used in the simulations that ignored all long-range interactions. It has to be mentioned, however, that the method works quite well with the original scaling (with the predictions of secondary structure for some α/β-type proteins being poorer by 2%–3% using the original scaling) as well as other scale factors ranging over quite a broad range of values.

TABLE II. Distribution of centers of mass of protein side chains within the globule.

B in of S	Distance from the center of mass (fraction of S)	Percent of residues
1	0–1/3	2.43
2	1/3–2/3	16.51
3	2/3–1	36.47
4	1–4/3	34.29
5	4/3–5/3	9.61
6	5/3–2	0.56

D. Spherical protein model and long-range interactions

Let us consider the mean square radius of gyration, S.

$$S = \left[N^{-1} \sum (r_{\text{CM}} - r_i)^2 \right]^{1/2}, \tag{8}$$

where r_{CM} is the position of the center of mass of the globule, and r_i is the position of the center of mass of the ith side chain. In their native state, single domain globular proteins exhibit closely packed conformations with a very small number (and size) of cavities. Thus, based on a statistical analysis of known protein structures, the mean square radius of gyration S scales with the number of residues N according to

$$S = 2.2N^{0.38} \text{ in angstroms.} \tag{9}$$

The exponent 0.38 is very close to the value of 1/3 expected for a collapsed long polymer chain. This arises because the vast majority of monomeric, single domain globular proteins adopt a close to spherical shape, with hydrophobic residues predominantly buried inside the globule and polar residues exposed to the solvent. These observations constitute the basis of the one body burial potentials employed in this work. Again, there are generic components of these potentials and sequence specific potentials. The first potential is based on the statistics of the distribution of amino acids found at a given distance from the center of mass in a library of native protein structures. This distribution, in the form of a histogram, is given in Table II, and the model system is driven to adopt this distribution. The corresponding potential has the following form:

$$E_b = \varepsilon_b \sum |m_{o,i} - m_i|, \tag{10}$$

where $m_{o,i}$ is the target number of amino acids at a given distance from a fixed point (the center of the MC working box) that is also assumed to be the center of mass of the model chain. Of course, at the beginning of the simulation run, the random chain is always placed at the center of the MC box. Note that the sphere of radius S contains somewhat less than half amino acids (see Table II), and in part defines the hydrophobic core of globular proteins.

In order to achieve a more uniform distribution of protein fragments within the globule, an approximate excluded volume was introduced. A 3×3×3 cluster of underlying cu-

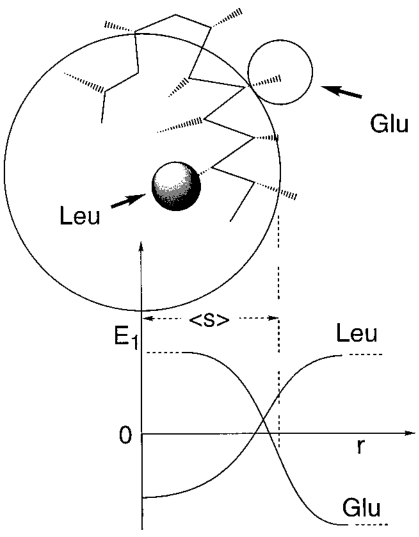

FIG. 4. Illustration of the idea of a centrosymmetric, one body burial potential. See the text for more detail.

bic lattice points is associated with each side chain. A MC working box serves as an occupancy array for the model polypeptide side chains. The excluded volume per residue is significantly underestimated, and therefore the cubic shape of the side chains does not distort the model chain geometry. Each time the two side chains overlap, the system energy increases by $\varepsilon_{\mathrm{rep}}$. The short-range excluded volume (up to the sixth neighbors down the chain) is treated more explicitly, prohibiting side chain–side chain distances below the values typical for proteins; again, the penalty for too close a distance is $\varepsilon_{\mathrm{rep}}$.

The situation when the polypeptide chain changes its direction inside the globule instead of reversing on the surface is extremely rare; thus, an additional penalty $\varepsilon_{\mathrm{rep}}$ is superimposed when the sharp turn is buried below the surface limited by the sphere of radius S. E_{rep} is the sum of the excluded volume and premature turn penalties over the entire chain.

The sequence specific burial potential consists of two terms. The first is a surface term defined with the help of the Kyte–Doolittle[25] hydrophobic scale by

$$\varepsilon_{i,\mathrm{KD}} = \begin{cases} \varepsilon_{Ai,\mathrm{KD}}, & \text{when } r_i > S, \\ 0, & \text{when } r_i < S. \end{cases} \tag{11}$$

The $\varepsilon_{Ai,\mathrm{KD}}$ are the Kyte–Doolittle (KD) hydrophobicity parameters.[25] The total contribution E_{KD} is the sum of this term over all residues.

The second sequence specific term, E_r, is derived from the straightforward statistics of particular amino acid occurrences at a given distance from the center of mass of the globule and was discussed previously. The idea is given in Fig. 4. The numerical values of the potential are given in the form of a histogram. For larger values of r (above 1.5 S), the potential for all amino acids was extrapolated by a monotonically increasing function. The data for $r < 0.3S$ (three first bins) were omitted due to the dispersion associated with the small volume of this region of the molecules and, consequently, the poor statistics in this bin. In the simulations, the statistical potential (numerical data could be found in PAPS supplementary material[26]) was smoothed by replacing it by a weighted average over three consecutive bins. The three first bins have values equal to the values in the fourth bin after smoothing. The scaling of the contributions to the burial potential is as follows:

$$E_{\mathrm{burial}} = E_b + E_{\mathrm{rep}} + 0.25{*}E_{\mathrm{KD}} + 4.0E_r. \tag{12}$$

Using this scaling, all the components are in the range of ±1–2 k_BT per residue, and the burial energy is of comparable magnitude as the short-range interactions. Nevertheless, the scaling is arbitrary, and it is possible that with a different scaling factor the performance of the secondary structure prediction method described here could be somewhat better. The total energy of the model system is the sum of long-range burial and short-range interactions.

E. Scoring procedure for the secondary structure prediction

In order to compare the properties predicted by the model to the structure of real proteins, it is necessary to define a method for assigning the secondary structure from low temperature, isothermal MC simulations. Of course, since long-range hydrogen bonds are not explicitly included into the potential, we cannot use a standard classical method[27] that starts from assignment of hydrogen bonds. We therefore opt for a classification based on backbone geometry.[28,29] There is, of course, a direct correspondence between the main chain conformations and the secondary structure of the polypeptide.[29] The method used here is the same (except for modifications resulting from use of more rigorous criteria for helical states) as in our previous work, and is based on a single distance and chirality parameter for a given residue. In particular, when

$$\begin{aligned} &r_{i-2,i+2} > 10.6\ \text{Å},\ \ \text{assign the } i\text{th} \\ &\qquad \text{residue as } \beta \text{ extended (E)}, \\ &r_{i-2,i+2} < 7.2\ \text{Å},\quad \text{and } r_{i-2,i+1}{}^{2*} \text{ and } r_{i-1,i+2}{}^{2*} \\ &\qquad \text{are right handed, assign as helix (H)}, \\ &\text{otherwise, assign as coil/turn } (-). \end{aligned} \tag{13}$$

This simple geometrical assignment correlates very well with the three-class reduced notation (commonly used to score various secondary structure prediction methods) of the Kabsch–Sander assignment.[27] It should be noted that the proposed method of secondary structure classification provides much more information due to the possibility of analyzing various geometrical properties. Actually, one can predict quite complex short-range conformational characteristics that are not available from standard methods. Thus comparison of the results from the scoring of secondary structure predictions proposed here with other methods should be understood as the most conservative estimate of the prediction accuracy and the utility of this approach.

The method also predicts the surface turns (or loops) where the polypeptide chains change their average direction (a U turn). The procedure for identifying a loop region is as follows. First, a constant simulation time interval, the chain is scanned, and the chain reversals are counted according to the criteria given below. The first scan detects a ''convex'' part of the chain; however, some end residues of various regular elements of secondary structure could be included

$$(\mathbf{r}_i-\mathbf{r}_{i-5})\cdot(\mathbf{r}_{i+5}-\mathbf{r}_i)<0, \quad \text{then} \quad lk(i)=1, \qquad \text{otherwise} \quad lk(i)=0. \tag{14}$$

Next, a second scan is performed to detect the ''straight'' regions of the chain, that presumably are the elements of regular secondary structure. This could be further used to remove false assignments of loop residues. Let all residues be assigned a structural index, $s(i)$, which is initially set equal to zero for $i=1,\dots N$. Then, the second scanning updates $s(i)$ according to the following criteria:

$$\text{when} \quad |(\mathbf{r}_{i+5}-\mathbf{r}_{i+1})-(\mathbf{r}_{i+4}-\mathbf{r}_i)|^2<13.4 \ \text{Å}^2,$$
$$\text{and} \quad |(\mathbf{r}_{i+5}-\mathbf{r}_{i+1})|>10.6 \ \text{Å}, \tag{15a}$$
$$\text{and} \quad |(\mathbf{r}_{i+4}-\mathbf{r}_i)|>10.6 \ \text{Å},$$

then the fragment is assinged to be expanded, and $s(i+k)=1$, with $k=1,4$.

$$|(\mathbf{r}_{i+5}-\mathbf{r}_{i+1})-(\mathbf{r}_{i+4}-\mathbf{r}_i)|^2<13.4 \ \text{Å}^2,$$
$$\text{and} \quad |(\mathbf{r}_{i+5}-\mathbf{r}_{i+1})|<7.2 \ \text{Å}, \tag{15b}$$
$$\text{and} \quad |(\mathbf{r}_{i+4}-\mathbf{r}_i)|<7.2 \ \text{Å},$$

then the fragment is helical, and $s(i)=1$, with $k=0.5$.

The third scan of the chain assigns loop residues combining the curvature index, $lk(i)$, and the secondary structure index, $s(i)$, according to

$$\text{loop}(i)=\text{loop}(i)+1, \quad \text{when} \quad s(i)=0 \quad \text{and} \quad lk(i)=1. \tag{16}$$

The idea of U turn (surface loops) detection is further clarified in Fig. 5. At the end of the simulations, one obtains a histogram of loop frequency loop(i); $i=1, N$; with values of loop(i) ranging from 0 to 20 (the number of scanning passes in a single run). For high values of loop(i) exceeding an assumed threshold value (six counts), the ith residue is assigned as part of a U turn. If two residues assigned as parts of a U turn are separated by less than four residues, the intervening residues are also assigned as being part of the same U turn. This filtering corrects for the false detection of very short regular elements of intervening secondary structure, i.e., it is assumed that a helix or beta strand (with possible flanking expanded coil fragments) cannot be shorter than four residues. Such short expanded fragments are usually parts of wide surface U turns.

At first glance, the frequency of collecting statistics for loop assignment may appear to be very low; however, the algorithm is very stable in this respect. While the overall chain conformation relaxes very quickly and changes many times during the single run, the loop signatures remain almost invariant. A test simulation with a 10 times higher sampling frequency of the loop geometry gave qualitatively the same results. At the end of each run, the secondary structure is assigned according to the criteria given in Eq. (13), based on the average (time average from the MC run) values of the corresponding intrachain distances and chirality factors. Residues that are detected as a member of U-turn regions are then reassigned as coil residues (−), regardless of the outcome of the initial assignment (the threshold value for the number of counts as a loop region is three). However, this reassignment very rarely changes the original one. Nevertheless, it contributes to a slightly more consistent final prediction.

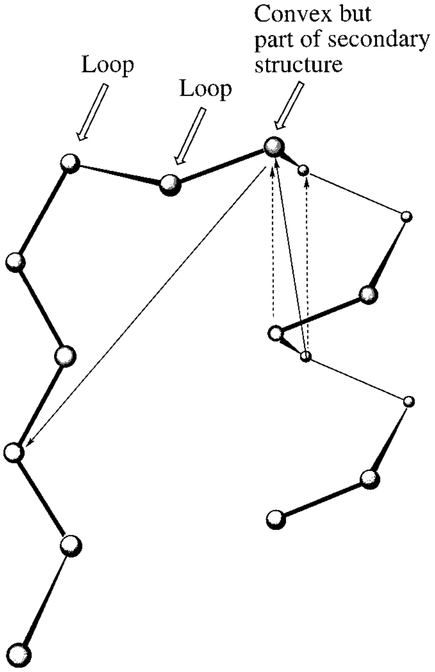

FIG. 5. The idea for detecting the loop fragments. The units indicated by the arrows are detected as belonging to the loop region by a convexity criterion (antiparallel orientation of the two solid vectors). One of the convex units is, however, part of a helix (parallel orientation of the two dashed-line vectors), therefore its loop assignment was disregarded.

III. RESULTS AND DISCUSSION

In the present study, we tested the same set of single domain globular protein sequences as was employed in our previous work without tertiary interaction. For each test sequence, at least three independent simulations were performed; each starts from a random coil state, and is subjected to simulated annealing and collapse of the chains. Then, an isothermal run at $T=1.0$, during which the final statistics were collected, is performed. The results of the simulations are very reproducible, and there is good correlation between the total energy and the accuracy of the secondary structure prediction.

TABLE III. Comparison of secondary structure predictions obtained in the present Monte Carlo simulations, in the simulations without burial interactions and by PHD method.

Proteins	Results of MC simulations(%)	Without burial term (%)	PHD (%)
1cd8	67.5, 68.4, 70.2	56.1	76.3
1crn	63.0, 67.4, 67.4	60.9	39.1
1ctf	66.2, 60.3, 61.8	58.8	60.3
1gb1	87.5, 82.1, 80.4	73.2	91.1
1mba	63.7, 58.9, 65.1	54.8	78.2
1pcy	65.7, 65.7, 65.7	60.6	75.8
351c	64.6, 67.1, 64.6	61.0	69.5
2pab.A	62.3, 61.4, 65.8	52.6	70.2
3fxn	70.3, 65.9, 66.7	60.1	73.9
2trx	59.3, 58.3, 63.9	50.5	63.0
Average	65.8	56.7	71.2

A. Burial energy and size restraints improve secondary structure prediction

The main question in this work is associated with the interplay of long-range (between residues that are at long distances along the chain) and short-range interactions in globular proteins. The secondary structure seen in the folded native state is a compromise between these two kinds of interactions. In Table III, we summarize the accuracy of secondary structure prediction for the 10 test proteins (additional details could be found in PAPS supplementary data[26]). The secondary structure for the test sequences was assigned according to the geometrical criteria described in Sec. II.

As compared to simulations lacking the restraints of chain compactness and the contributions of chain burial, the accuracy of the secondary structure prediction increases substantially, on average by 9.1% (from 56.7% obtained in the previous work to 65.8% in this work, as an average weighted by the number of residues in each protein sequence); however, in some cases the improvement is of a qualitative nature while in others it is rather small, but always well above statistical error. This demonstrates that the burial potential and compactness restraints significantly influence the resulting secondary structure. First, there is a somewhat trivial effect that comes from globule size restrictions. A given secondary structural element simply cannot propagate for a substantial distance beyond the boundary of the globule. In contrast, such a situation occurs in most one-dimensional methods that overpredict the central helix of protein G in the direction of the N terminus. This effect was also observed in our previous MC studies, where the all tertiary interactions were neglected. Second, in single domain proteins, the hydrophobic side chains tend to be buried in the core of the globule, whereas the polar, hydrophilic side chains tend to be exposed to the solvent. This, of course, has to moderate the secondary structure. Some regularizing effect could also be due to the more (on average) hydrophilic loop regions. This also may regularize the secondary structure between the loops.

The results of particular runs for a given sequence differ due to the statistical character of the method. The fluctuations are larger for smaller proteins and become relatively smaller for larger structures. This tendency is demonstrated in Tables IV and Tables V where the results of three independent predictions for the sequence of the 56 residue $B1$ domain of protein G (1gb1) are compared with the results of three independent runs for the 138 residue protein flavodoxin (3fxn). Besides the secondary structure prediction, we also include the results of surface loop/turn assignments. Here, U means that the loop probability is very high $[\text{loop}(I)>6]$, while smaller values indicate the presence of more flexible and partly exposed (in the time averaged sense) residues. $\text{loop}(I)\geqslant 3$ overrides the secondary structure assignment to other states. Apparently, the magnitude of local fluctuations in the prediction accuracy (the extent of secondary structure

TABLE IV. Results of five independent simulations for protein G (1gb1).

```
87.5%, 82.1%, 80.4%, 80.4%, and 87.5% correctly predicted.
12345678901234567890123456789012345678901234567890123456
MTYKLILNGKTLKGETTTEAVDAATAEKVFKQYANDNGVDGEWTYTDDATKTFTVTE
-EEEEEE------EEEEEE---HHHHHHHHHHHHH------EEEEE----EEEEE-
-EEEEEEE------EEEE----HHHHHHHHHHHHHH-----EEEE---EEEEEEE-
-EEEEEEE-------EE-----HHHHHHHHHHHHHH------EEE---EEEEEEE-
-EEEEEEE------EEE-----HHHHHHHHHHHHHH-EEE-EEEE---EEEEEEE-
-EEEEEEE------EEE-----HHHHHHHHHHHHHH-EEE-EEEE---EEEEEEE-
-EEEEEEEE-----EEE-----HHHHHHHHHHHHHH-----EEEEE---EEEEEE-
--------5U21-----1UUU----------------15----114U4--------
-------21455UU4-13UUU------------------3UU111UU4--------
--------1-2UUU1-2UU2-----------------1--1----3UU1-------
--------UUUUU---1UU33----------------------21UU2--------
---------UUUU---2UUUU11--------------41------2UUU1------
```

Note: The first three lines describe the native structure. The first line of this panel gives the last digit of the residue number, the second line the one-letter codes of the protein G amino acid sequence, and the third line provides the three-letter code of the secondary structure assignment, according to DSSP method (H—helix, E–extended/beta, and ''-'' coil, or everything else). The next five lines are the secondary structure predictions from the five independent MC runs; the remaining lines provide the surface U-turn/loop predictions according to the procedure described in Sec. II. U denotes a strong prediction of the surface loop region (more than 5 per 20 counts during the simulations), the numbers from 1 to 5 denote weak loop predictions of various strengths, and ''-'' means that at a given position the loop conformation was never detected.

TABLE V. Results of three independent simulations for flavodoxin (3fxn). (Note: See footnote of Table IV; here, the results come from three independent Monte Carlo simulations.)

```
70.3%, 65.9%, and 66.7% correctly predicted.
12345678901234567890123456789012345678901234567890
MKIVYWSGTGNTEKMAELIAKGIIESGKDVNTINVSDVNIDELLNEDILI
EEEEEE----HHHHHHHHHHHHHHH-----EEEE-------------EEE
--EEEEEEEEE-HHHHHHHHHHHHH-----EEEEEE--EE-HHHHH-EEE
--EEEEEEEEE-HHHHHHHHHHHHH----EEEEEEE----HHHHH--EEE
-EEEEE------HHHHHHHHH-EEEE----EEEE-------HHHHH-EEE
---------11---------------13312-------------------
--------------------------243-------32----------1-
------4UU5--------------1-3U2-----UUU51-------4-2-
12345678901234567890123456789012345678901234567890
LGCSAMGDEVLEESEFEPFIEEISTKISGKKVALFGSYGWGDGKWMRDFE
EEEE-E---------HHHHHHHH-------EEEEEEEE-----HHHHHHH
EEE-HHHHHHHH---HHHHHHHHH-------EEEEEEEEE--HHHHHHHH
EEE----HHHHHH------HH-EEEE-----EEEEEEEEE----HHHHHH
E------HHHHHH-HHHHHHHHHH-EE---EEEEEEEEEEE--HHHH--H
-1---------1-----------224UUUUU2-------1U11-------
----2-1-------22---------23UUUU1-------1311-------
-1------------------------2UUU2-------------------
12345678901234567890123456789012345678
ERMNGYGCVVVETPLIVQNEPDEAEQDCIEFGKKIANI
HHHHH---EE----EEEE---HHHHHHHHHHHHHHHH-
HH---EEEEEE---EEEEE-HHHHHHHHHHHHHHHH--
HH---EEEEEE--EEEEE---HHHHHHHHHHH--HHH-
HH---EEEEEE--EEEEEE-HHHHH---HHHHHHHH--
----------1UU4----1-------------------
----------1UU2---1U3------------------
-----------U52----1-------------------
```

overprediction or underprediction) is on a very similar level for most proteins. Thus, larger relative fluctuations of global prediction accuracy are seen in smaller proteins, which usually consist of a smaller number of secondary structural elements. In general, however, the predictions are quite reasonable. For example, in the case of 1gb1, the main error that occurs in some runs is due to the overprediction of short extended fragments for residues 38–40. In the native state, these residues constitute a very broad surface loop/turn that has a rather extended conformation (the loop on the top of the native structure shown in Fig. 6); however, it is not a part of the β sheet. From visual inspection of the conformations generated by the MC algorithm, this could be deduced with rather high fidelity. Moreover, the automatic procedure of loop detections assigns these as loop residues (the black fragments of the MolScript[30] structure shown in Fig. 6). Another example of such an apparent overprediction is the helical fragment predicted in 3fxn for residues 41–46. If the flavodoxin fold were an ideal α/β barrel, these residues should be part of a helix. In the real 3fxn structure, these residues are a series of turns that indeed have a conformation that is close to helical, but which nevertheless is somewhat too expanded for the DSSP (Dictionary of Secondary Structure in Proteins) algorithm to assign the helical pattern of hydrogen bonds. Consequently, this overprediction of the MC algorithm paradoxically could even be helpful in three-dimensional model building.[4] Underprediction of one of the helical fragments in the second simulation for flavodoxin (residues 66–73 in the native state) is clearly an error of the algorithm.

Comparison of several independent runs may help in building a consensus prediction. For some proteins, the algorithm tends to converge very quickly with a small dispersion of the final results. For other proteins, the dispersion of the results is greater and seems to correlate with the overall prediction quality. The more reproducible the results of the simulation are, the better is the accuracy of the secondary structure prediction. In this respect, 1cd8 and 3fxn are examples of very well behaved proteins, while the algorithm is less stable for 1ctf or 1crn.

B. Simulations provide medium-range geometrical characteristics

The present method of the study of protein chains that are gently restrained to occupy the proper volume of a globular state provides a wealth of geometrical information that is not available from standard secondary structure prediction methods. To illustrate, we compile the comparisons of predicted secondary structure elements with the native state of the protein G domain. The comparison is given in Table VI. The examples show that the elements of secondary structure are not only correctly predicted with respect to their structural classes (helix, extended, coil), but also that their geometry is quite accurate. This result is not surprising, since the model has a quite accurate description of short-range inter-

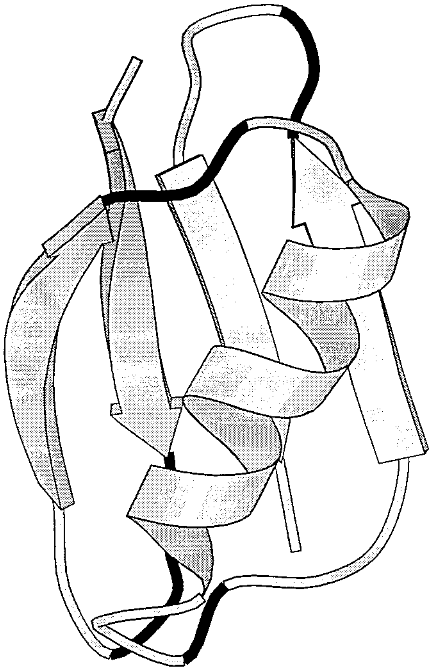

FIG. 6. MOLSCRIPT (Ref. 30) drawing of the native structure of 1gb1. The black fragments of the diagram indicate the surface loops/turn detected by the algorithm.

actions. Noticeably, those fragments of secondary structure that are reproducibly predicted and never ''contaminated'' by sparse loop predictions also have a much better geometrical fidelity. For the protein G sequence, this is the case for the central helical fragment and for two terminal β strands. These strands are located in the center of the four-stranded β sheet. Usually such strands (in contrast to the edge strands) have a better defined pattern of hydrophobic/hydrophilic residues. This, perhaps, increases the geometrical accuracy of the prediction.

Of course, due to the lack of specific tertiary interactions (pairwise interactions of the side chains, hydrogen bonds, etc.), the topology of the global fold is not defined by the present method. An example of a conformation generated by the algorithm is shown in Fig. 7. While the individual β strands, surface loops, and helical fragments are present, the overall topology is wrong. Nevertheless, for a substantial fraction of proteins, the accuracy of the secondary structure and the loop predictions allows one to propose a small number of possible global folds that could be subsequently refined and tested by other methods.

TABLE VI. Comparison of the predicted geometry for protein G (1gb1) domain with the native state.

	Average Cα rms from native (minimum rms) Å				
Run	Protein fragments				
No.	1–9	12–20	23–35	39–47	48–56
1	1.64	2.99	1.26	3.15	1.87
	(0.88)	(2.27)	(0.71)	(1.58)	(1.20)
2	1.80	2.73	1.22	2.44	1.79
	(1.22)	(2.12)	(0.76)	(1.83)	(1.08)
3	1.60	2.72	1.43	2.57	1.53
	(0.82)	(2.45)	(0.91)	(1.39)	(1.02)

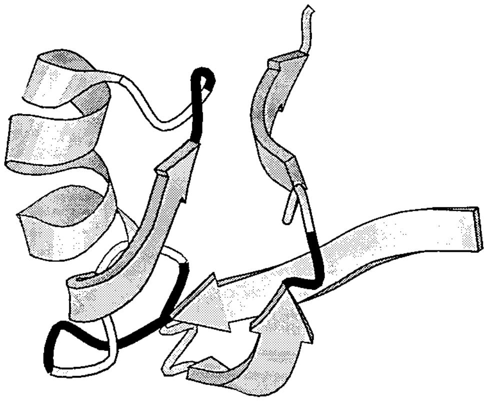

FIG. 7. An example of the tertiary ''structure'' of the B1 domain of protein G generated by the algorithm. All secondary structure elements and loops are correctly assigned; the fold topology is not defined, however.

C. Comparison with other surface U-turns and secondary structure prediction methods

With respect to the spherical domain model of protein structure, the method proposed and examined in this work is somewhat similar to the recently published method for predicting surface U turns and transglobular connections.[12] However, the sampling method and the chain model employed here are more complex and more realistic. While in the previous work[12] we achieved a high accuracy of U-turn prediction (about 95% of surface turn/loops correctly predicted for 38 test sequences) and prediction of the leading secondary structure of the transglobular connections (82% correct prediction), the secondary structure is better defined here and is more accurate with respect to a residue-by-residue comparison. However, the previous method[12] led to a large scattering of secondary structure assignment near the loop fragments. Consequently, the overall accuracy on the residue-by-residue level was low, in the range of 55%. Here, the secondary structure assignment was more accurate but, somewhat surprisingly, the surface loop predictions were less accurate. For the set of 10 proteins tested here the previous method gives 74% correct assignments, while the present method correctly predicted 69% of the surface U turns. Given that the test set used here contains proteins that are on average larger and represent a more diverse collection of topologies, this level of accuracy is probably acceptable. Note that the most known method of turn prediction (applicable only for β proteins) by Wilmot and Thornton has an accuracy of about 71%.[12,31] The accuracy of our previous method[12] for β proteins was close to 100%, and here the accuracy for β proteins is also higher (96%). It should be added that the present model carries the geometry of the entire chain. Therefore, it could be expanded very easily to include, for example, some long-range distance restraints or more specific packing restraints.

IV. SUMMARY AND CONCLUSION

In this work, our previously developed models of short-range interactions for the study of lattice protein dynamics are supplemented by approximate excluded volume interactions and a hydrophobic burial potential. The burial potential was implemented in the form of one body functions that are suitable for single domain proteins. The generic part of the potential drives the model system into conformations whose residue density in a hypothetical core is comparable to the average density of folded proteins. The sequence specific burial potential simulates the distribution of various amino acids with respect to the center of mass of the globular protein. Inclusion of this approximate burial potential leads to a better definition of the secondary structure seen during the MC simulations. The accuracy of the predictions of secondary structure, as defined by the α-carbon chain geometry, increases by 9%. This is because the global restraint of the collapsed structure to realistic dimensions moderates the surface segregation of the hydrophilic and hydrophobic residues, and perhaps to a lesser extent some finer burial preferences of various residues. Together with this prediction of secondary structure in a three-letter code (helix, extended, and coil, which are predicted with an accuracy of 66%), the method allows for the prediction of surface loops/turns where the polypeptide chain changes its direction. This enhances the overall prediction accuracy and its potential value for protein structure prediction.

How do these predictions compare with the existing methods? We limit our comparison to perhaps the most powerful standard method of secondary structure prediction—the Rost–Sander PHD (profile fed neural network system from Heidelberg) neural network based method.[32] For the set of 10 test proteins, the PHD predictions were 5.4% better (71.2% vs 65.8% from present study); however, all the proteins considered by PHD are either in the training set or are closely homologous to members of the training set. The PHD method was used without multiple sequence alignment. Note that multiple sequence alignment information could be employed as well within the framework of the method presented here. The secondary structural propensities for the model chains could be combined to a form ''consensus'' sequence and the prediction of the secondary structural properties of such a composite could be readily implemented. This possibility will be explored in the future. However, in contrast to the standard secondary structure prediction methods, the method presented here gives a quite dependable (and consistent with regular fragments of secondary structure) prediction of the surface loop/turns fragments. Moreover, the present method gives the direct geometrical characteristics of the predicted fragments. This has to be contrasted with the nonphysically long helices predicted (for example, the 1 mba case) by PHD, or a helix changing directly into a stretch of extended states (as in the 1 gbl case). In addition, one-dimensional methods (such as PHD) do not distinguish between false turns (as β bulges) and real turns, where the chain reverses its global direction. Thus, the proposed method seems to be a useful tool for secondary structure prediction, as well as the prediction of protein structure in general.

Of course, the predictions are not 100% accurate. One technical reason for the limited accuracy of the method is that we translated the local geometry of the α-carbon chain to the secondary structure of protein. This was necessary because in this model the long-range hydrogen bonds are undefined. While the main chain geometry correlates very well with the secondary structure, some misalignments are certainly possible. However, the more fundamental reason for the inexact predictions is probably the lack of any sequence specific pairwise interactions. The results of the present work suggest that these interactions may have a significant effect on secondary structure. This is, of course, a somewhat trivial qualitative conclusion; however, on a quantitative level it is not. Our present studies, as well as those of our previous work,[14] show that reproduction of local chain geometry is possible. A further increase of the accuracy of secondary structure prediction, without invoking the computationally very expensive details of long-range interactions, could be achieved in some specific cases. For example, superimposing some (very few) long-range pairwise restraints (such as $S-S$ crosslinks, metal binding site, etc.) might further increase the fidelity and applicability of the present method.

ACKNOWLEDGMENTS

This work was supported by NIH Grant No. GM-37408. One author (A.K.) acknowledges partial support from the University of Warsaw, Grant No. BST-532-34/96, and the Howard Hughes Medical Institute (International Research Scholar Grant No. 75195-543402). Helpful discussions with Dr. Adam Godzik and the assistance of Leszek Rychlewski are gratefully acknowledged.

[1] B. Rost and C. Sander, Proteins **23**, 295 (1996).
[2] A. Monge, R. A. Friesner, and B. Honig, Proc. Natl. Acad. Sci. USA **91**, 5027 (1994).
[3] A. Kolinski and J. Skolnick, Proteins **18**, 353 (1994).
[4] J. Skolnick, A. Kolinski, and A. R. Ortiz, J. Mol. Biol. **265**, 217 (1997).
[5] K. C. Chou, Proteins **21**, 319 (1995).
[6] B. Rost and C. Sander, Proteins **19**, 55 (1994).
[7] P. Argos, J. Mol. Biol. **197**, 331 (1987).
[8] A. Kolinski, M. Milik, J. Rycombel, and J. Skolnick, J. Chem. Phys. **103**, 4312 (1995).
[9] J. Skolnick and A. Kolinski, in *Computer, Simulations of Biomolecular Systems. Theoretical and Experimental Studies*, edited by W. F. van Gunsteren, P. K. Weiner, and A. J. Wilkinson (ESCOM Science, Leiden, The Netherlands, 1996).
[10] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbulth, A. H. Teller, and E. Teller, J. Chem. Phys. **51**, 1087 (1953).
[11] C. Branden and J. Tooze, *Introduction to Protein Structure* (Garland, New York, 1991).
[12] A. Kolinski, J. Skolnick, A. Godzik, and W.-P. Hu, Proteins **27**, 290 (1997).
[13] A. Kolinski and J. Skolnick, Proteins **18**, 338 (1994).
[14] A. Kolinski, W. Galazka, and J. Skolnick, Proteins **26**, 271 (1996).
[15] K. A. Olszewski, A. Kolinski, and J. Skolnick, Protein Eng. **9**, 5 (1996).

[16] K. A. Olszewski, A. Kolinski, and J. Skolnick, Proteins **25**, 286 (1996).

[17] F. C. Bernstein, T. F. Koetzle, G. J. B. Williams, E. F. Meyer, Jr., M. D. Brice, J. R. Rodgers, O. Kennard, T. Simanouchi, and M. Tasumi, J. Mol. Biol. **112**, 535 (1977).

[18] PDB Q. Newsletter No. 71, January (1995).

[19] A. Godzik, A. Kolinski, and J. Skolnick, J. Comp. Chem. **14**, 1194 (1993).

[20] A. Kolinski and J. Skolnick, *Lattice Models of Protein Folding, Dynamics and Thermodynamics* (R. G. Landes Co., Austin, TX, 1996).

[21] M. Milik, A. Kolinski, and J. Skolnick, J. Comput. Chem. **18**, 80 (1997).

[22] J. Skolnick, A. Kolinski, C. Brooks III, A. Godzik, and A. Rey, Curr. Biol. **3**, 414 (1993).

[23] G. N. Ramachandran and V. Sassiekharan, Adv. Protein Chem. **28**, 283 (1968).

[24] A. Kolinski and J. Skolnick, *Parameters of Statistical Potential*. Available via ftp from public directory, scripps.edu (pub/MCSP) 1995.

[25] J. Kyte and R. F. Doolittle, J. Mol. Biol. **157**, 105 (1982).

[26] See AIP Document No. PAPS JCPSA6-107-953-6 for 6 pages of Tables I and II. Order by PAPS number and journal reference from American Institute of Physics, Physics Auxiliary Publication Service, Carolyn Gehlbach, 500 Sunnyside Boulevard, Woodbury, New York, 11797-2999. Fax: 516-576-2223, e-mail: paps@aip.org. The price is $1.50 for each microfiche (98 pages) or $5.00 for photocopies of up to 30 pages, and 0.15 for each additional page over 30 pages. Airmail additional. Make checks payable to the American Institute of Physics.

[27] W. Kabsch and C. Sander, Biopolymers **22**, 2577 (1983).

[28] S. Rackovsky, Proteins **7**, 378 (1990).

[29] T. J. Oldfield and R. E. Hubbard, Proteins **18**, 324 (1994).

[30] P. J. Kraulis, J. Appl. Crystallogr. **24**, 946 (1991).

[31] C. M. Wilmot and J. M. Thornton, J. Mol. Biol. **203**, 221 (1988).

[32] B. Rost and C. Sander, J. Mol. Biol. **232**, 584 (1993).