PROTEINS: Structure, Function, and Genetics 27:290–308 (1997)

A Method for the Prediction of Surface "U"-Turns and Transglobular Connections in Small Proteins

Andrzej Kolinski,[1,2]* Jeffrey Skolnick,[1] Adam Godzik,[1] and Wei-Ping Hu[1]
[1]*The Scripps Research Institute, Department of Molecular Biology, La Jolla, California;*
[2]*Department of Chemistry, University of Warsaw, Warsaw, Poland*

ABSTRACT **A simple method for predicting the location of surface loops/turns that change the overall direction of the chain that is, "U" turns, and assigning the dominant secondary structure of the intervening transglobular blocks in small, single-domain globular proteins has been developed. Since the emphasis of the method is on the prediction of the major topological elements that comprise the global structure of the protein rather than on a detailed local secondary structure description, this approach is complementary to standard secondary structure prediction schemes. Consequently, it may be useful in the early stages of tertiary structure prediction when establishment of the structural class and possible folding topologies is of interest. Application to a set of small proteins of known structure indicates a high level of accuracy. The prediction of the approximate location of the surface turns/loops that are responsible for the change in overall chain direction is correct in more than 95% of the cases. The accuracy for the dominant secondary structure assignment for the linear blocks between such surface turns/loops is in the range of 82%. Proteins 27:290–308.** © 1997 Wiley-Liss, Inc.

Key words: protein folding; protein structure; supersecondary structure; structure prediction; turn prediction; statistical potentials

INTRODUCTION

At low resolution, the native conformation of a globular protein may be viewed as a series of linear blocks tied together by surface loops/turns that change the overall direction of the protein chain[1–5]; these regions we term "U" turns. Such a schematic model, where the local wiggles of the chain have been averaged out characterizes the topology or global fold of a protein. Within a given block, the backbone conformation often comprises a single dominant element of regular secondary structure; but sometimes it may contain two essentially collinear, regular secondary structural elements, such as helices or β strands. Other times, the secondary structure within the block may be so irregular that, by some definitions, the backbone would be classified as entirely unstructured. All such blocks run from one surface of the protein to another and are transglobular in nature. Thus, the emphasis is placed on aspects of the global as opposed to local characteristics of protein structure. In this paper, we describe a novel approach to predict the number and the positions of the surface U turns along the sequence, and concurrently, the dominant secondary structure within each transglobule block. Application of the methodology is then made to 38 proteins of known tertiary structure to assess the ability of the method to predict the location of the U turns and the identity of the secondary structure within the blocks.

Overview of the Method

Given this geometric model of protein structure, the goal is to predict the structural components of the model, that is, the locations of the U turns and the structural characteristics of the intervening blocks. In this approach, we maintain a global description of the interactions, that is, we recognize that the native conformation (of the whole protein) is in the global minimum energy state, which locally does *not* necessarily adopt the minimum energy conformation at each position along the chain. For example, the best solution for one fragment of the chain may be a helical hairpin; while in a second fragment, partially overlapping with the first, the best solution may be a β hairpin. However, the entire protein may have a lower energy if a triplet of helices is adopted. In other words, an approach that accounts for the entirety of interactions within the entire chain represents an important part of the nature of proteins. This will account for the fact that proteins locally may be energetically frustrated, that is, the best global solution for the chain conformation need not be the best local solution everywhere in the chain. To this end, we employ the observation that small single-domain, globular proteins form compact

Contract grant sponsor: NIH, contract grant number GM-48835; contract grant sponsor: University of Warsaw, contract grant number BST-502/34/95; contract grant sponsor: Howard Hughes Medical Institute, contract grant number 75195-543402.*Correspondence to: Dr. Andrzej Kolinski, The Scripps Research Institute, Department of Molecular Biology, 10666 North Torrey Pines Road, MB1, La Jolla, California. E-mail: andr@scripps.edu
Received 27 July 1996; accepted 6 August 1996.

structures whose radius of gyration, *S*, can be rather well estimated[6,7] [see Eq. (1)]. Depending on the dominant secondary structural type within a block, the length of the linear block between U turns spans a range of values which is a function of *S*. These geometric constraints preclude certain conformations such as the formation of a single 50-residue helix in a 50-residue protein. We also include the interplay of energetic terms such as intrinsic secondary structural preferences and the burial preferences of the amino acids. The energy of a given U turn reflects the tendency of its residues to be exposed as well their intrinsic preferences to adopt this particular type of secondary structure. Similarly, the energy of a given set of amino acids in a block will include the burial energy of those residues located on the face of the block pointing toward the core of the protein, and the solvation energy of those residues on the face of the block exposed to the solvent, and the intrinsic secondary structural preferences of the residues for the secondary structure adopted by the block. Thus, the preference for a particular division of the chain into N blocks connected by N-1 surface U turns is determined by the total energy of the chain. That is, both U turns and blocks are treated equivalently, and a self-consistent, global description of the energy is employed.

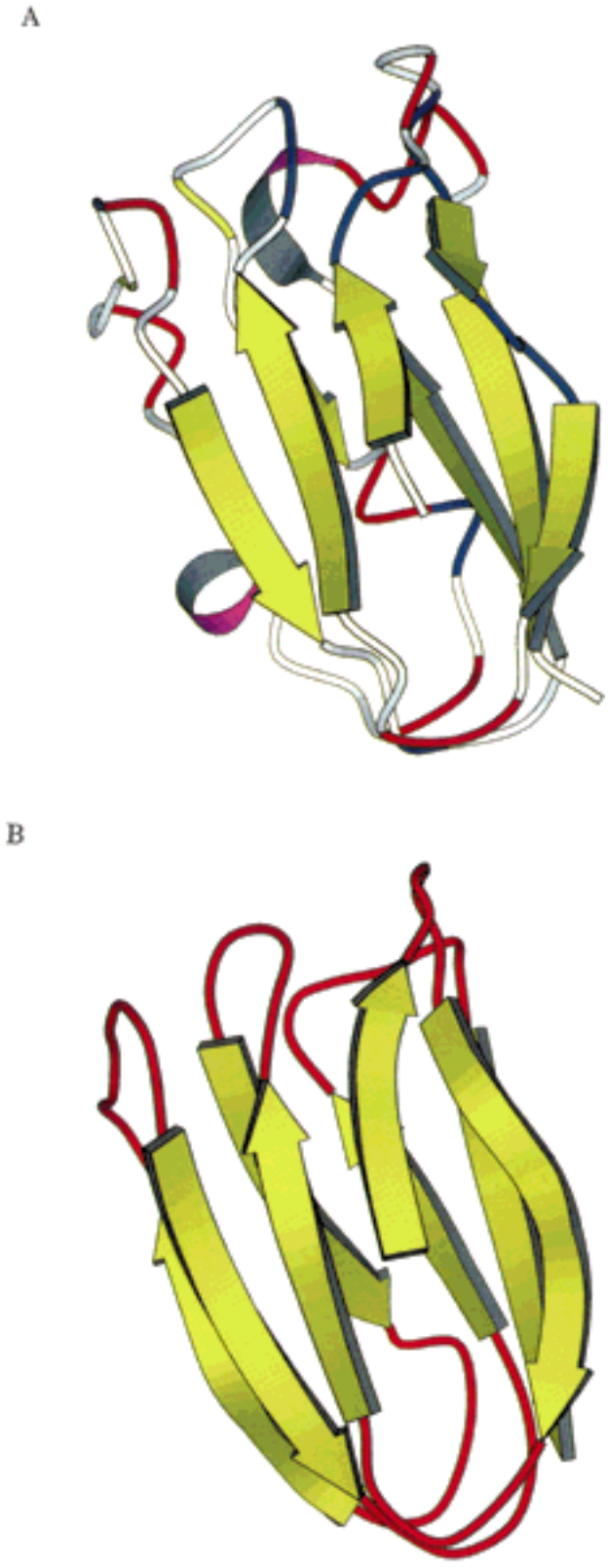
The goal of the calculation is to identify the optimum division of the chain into blocks and U turns with the lowest possible energy. We search for such a partitioning of the chain by means of a Monte Carlo (MC) algorithm. The elementary step is equivalent to randomly choosing a set of U turns and building blocks from a structural library extracted from known protein structures. The chain is then stitched together, and its energy is evaluated. The new conformation is accepted subject to a Metropolis criterion, and the procedure is repeated many times. Ultimately, the lowest energy states are collected and pooled, and their geometric properties yield a prediction of the blocks and U turns.

Comparison With Previous Studies

At this juncture, it is appropriate to compare and contrast our blocks and U turns model/predictive approach with existing work. The conceptual basis of our geometric model of protein structure bears a certain similarity to an idea originally proposed by Cohen and coworkers.[8] They, too, view a protein as being divided into turns and the secondary structure elements between them. However, there are important differences between that work and the approach described here. Cohen's group does not distinguish between turns that simply divide regions of secondary structure and those that reverse the overall direction of propagation of the chain; this distinction is crucial to our approach. We ignore the former and explicitly focus on the latter. Furthermore, the approach of Cohen and coworkers[8] is based on amino acid pattern rather than energy. Using purely local pattern information, their algorithm first attempts to find the turns, and then, given the location of the turns, they try to identify the secondary structure between predefined turns on the basis of amino acid sequence patterns. In our approach, turns and secondary structure fragments (blocks), are treated simultaneously; all are identified based on their energy, which reflects an interplay between local and tertiary interactions, as described by intrinsic secondary structural preferences and burial preferences, respectively.

Some aspects of the block and U turn geometric model of a protein have been captured by the Richardson[5] ribbon diagrams, which are widely used to depict and classify protein structure. However, she defines a ribbon that encompasses individual secondary elements as defined by more standard second structure assignments (as in the DSSP approach[9]). Thus, as shown in Figure 1A, a classical Richardson diagram of plastocyanin[10] prepared by the program MOLSCRIPT[11] contains breaks in the second β strand because of the presence of a bulge; whereas, as seen in Figure 1B, because they do not change the direction of the chain, the blocks and U turns model fuses these two strands into one. Hence our model is equivalent to a Richardson[5] protein diagram whose secondary structural elements have been smoothed to eliminate local irregularities that do not change the overall direction of chain propagation. The comparison between Figure 1A,B also highlights the difficulty in translating classical secondary structural classification schemes into three-dimensional models. For example, is a residue, classified by a classical secondary structure assignment method as a turn, simply a β bulge, which does not change the overall chain direction, or is it a U turn? Indeed, sometimes structures that are classified as bulges occur in what we call U turns, while DSSP[9]-type turns can occur in the middle of β strands. A representative example that illustrates this ambiguity is shown by a Richardson diagram of plastocyanin in Figure 1A where the bulges and turns as classified by the DSSP program[9] are indicated in blue and red, respectively. By contrast, given a set of blocks and U turns as in Figure 1B, it is much more straightforward to translate this information into an approximate model of the global topology of a protein.

Next, there is the question of the accuracy of classical secondary structure prediction schemes. After years of investigation,[12–17] the best secondary structure prediction methods[18] currently have an accuracy of 50–75% for the prediction of three structural classes (helix, β, or coil). The margin of error is, therefore, quite substantial; what is even worse, the predictions often miss entire elements of secondary structure. The reason for this high error rate is that these methods only address the local aspect of chain interactions and not the role of tertiary interactions. Of course, every sequence fragment has some intrin



sic secondary structural propensities.[19] In some regions of the chain, these may be quite strong, while in others they may be quite weak. In folded proteins, these intrinsic secondary structure propensities are always to some extent moderated, and perhaps are often overridden by tertiary interactions. Thus, secondary structure prediction methods attempt to solve what is perhaps a nonseparable subset of the protein folding problem.[20,21] Finally, because of the ambiguity between secondary structure assignment and the actual direction of the chain, even if they were to be successful, as discussed above, there would remain the problem of building a three-dimensional model from a classical secondary structure assignment.

Another class of methods attempts to attack the folding problem head on.[21,22] Starting from a random coil state, the global minimum of free energy is sought by means of a Monte Carlo method or another efficient search procedure.[23,24] To make the conformational search tractable, the model of the protein is often simplified.[21] Such an approach clearly demands much from the potential function. While much can be learned from such methods,[21] as yet, they have only been successful on very simple folds[6,25] or exaggerated model protein sequences.[24,26,27]

A final class of structure prediction methods is based on an inverse folding paradigm,[28–33] where one attempts to associate a sequence with a particular structure in a library of known protein structures. The advantage of such methods is that only a limited part of conformational space, that corresponding to already known protein structures, is searched. Therefore, the procedure is relatively rapid, and the demands on the potential function are far less. However, the inherent weakness of this method is the assumption that the structure of a protein similar to the one of interest already exists in the library of known structures. If this is not true, then the method will fail. By contrast, the present method uses only small pieces (of the size of a hairpin) of known structures as generic templates for which the sequence-structure fitness will be computed. Thus, an example of the overall topology of the protein need not have been previously determined.

Objectives of the Blocks and ∪ Turns Model

What is clearly required is some technique that combines the virtues, but not the shortcomings, of all these extant methods. It should incorporate global

Fig. 1. **A:** Richardson diagram of plastocyanin where the ribbons are defined on the basis of classical secondary structure (DSSP) assignments. Turns and bulges are indicated by red and blue, respectively. **B:** Richardson diagram of plastocyanin in the blocks and U turns model, where U-turns (red) occur when the direction of propagation of the chain changes global direction, and blocks (yellow) join successive U turns by running from one surface of the protein to another on the opposite side of the protein.

information into the prediction scheme to avoid the pitfalls of secondary structure prediction methods. It should limit the search to "proteinlike" regions of conformational space and yet not be limited by it. The work described in this paper is a step in this direction and lies midway between one-dimensional, local secondary structure prediction schemes and a full treatment of the folding problem, with all the myriad of interactions that constitute a real globular protein. As will be shown below, in most cases, the method is able to locate surface U turns which delineate the end of the transglobule blocks. It also provides a structural assignment for the blocks as helices or β-strand/expanded-long-loops. This information is sufficient to suggest a small number of possible folds; these could be further refined with the help of the constraints provided by the present model. Application of such a technique will be described in forthcoming work; here, we focus on the development and validation of the method.

The outline of the remainder of the paper is as follows. In the Results section, we describe the application of the method to 38 globular proteins of known structure. Most of these structures were made public in early 1995 and were unknown to us at the time the potential was built. All test proteins were excluded from the database used in the derivation of the statistical potentials. For nine representative examples, we present the results in graphic form and discuss the results in detail. For all proteins, we provide the statistics of the accuracy of the prediction of the blocks and U turns. Finally, we conclude with a discussion of the degree of success and limitations of the present approach and the outlook for future progress.

METHODS

The basic idea of the methodology described below is not tied to any particular protein model, nor to any particular force field. Thus, we separate the description of the basic spherical domain model from the details of implementation (lattice model and specific interaction scheme) employed in the present work.

Spherical Domain Model

The basic idea of the model is as follows: The protein sequence of interest is randomly divided into several partially overlapping sequence fragments. For a given sequence fragment, a structural template is assigned by random selection from a library of such structural templates, constructed using a database of known protein structures. In principle, one could employ an exhaustive search over all combinations of sequence fragments and structural templates; however, due to the large number of degrees of freedom involved, this would be impractical. Each structural template is comprised of two successive protein building blocks which may be viewed as all α, all β, or mixed motif hairpins. These structural templates lack sequence information and are used to provide a library of proteinlike structures to which the sequence of interest can be assigned. After division of the protein sequence into fragments, each fragment, now with an assigned structural template, is oriented with respect to the center of a hypothetical sphere that approximates the single-domain protein. Next, the burial energy and short-range interactions of the structural template are assessed. Hydrophobic residues, when placed in the inner part of the sphere, would decrease the "energy" of the fragments, while exposed hydrophilic residues will contribute accordingly. Similarly, the secondary structure preferences indicate whether or not, based on local considerations, the sequence favors the structural template. The division into sequence fragments and structural templates is repeated many times, and the top scoring results are used to make structural predictions.

Outline of Algorithm

The algorithm can be outlined as follows (see the flow chart given in Fig. 2):

1. For a given sequence, estimate the radius of gyration of the globule, S_0 (a single domain with a single hydrophobic core is assumed) from the number of residues, m, in the protein by

$$S_0 = 2.2\ m^{0.38} \quad \text{(in Å)} \qquad (1)$$

   The above formula has been derived from the statistics of single-domain proteins[7,27] with the mean square radius of gyration computed assuming the same mass for all the structural units. Note that if the proteins were long compact homopolymeric chains,[34] then the exponent would be equal to ⅓. The small deviation from this theoretical value is associated with a finite length effect; single-domain proteins being relatively short polymers.[35]

2. Estimate the range of the number of secondary structure elements that pass through the entire globule, N_{min}, N_{max}. This is done only to speed up the computations, since, for a given small protein, this range is fairly narrow. N_{min} and N_{max} are obtained from analysis of the database of structures and the use of a chain smoothing algorithm that automatically assigns the location of the blocks and surface U turns in proteins.[2]
3. Generate an initial division of the test sequence into N fragments, with $N_{min} \leq N \leq N_{max}$. The length of the fragments is limited by the size of the shortest expanded fragment and the longest helical fragment that can fit into the expected limits of the hydrophobic core (range: $1.8S_0$), and the entire globule (range: $2.5S_0$), respectively. This superimposes some obvious limitations on

the N_{min} and N_{max} values. For example, the lower limit for N_{min} could be estimated as N_{min} is the nearest integer to $(m/n_{max} + 1)$, where n_{max} is equal to 2.5 S_0/l_h, with l_h = 1.5 Å, the length of a helix per residue. In a similar fashion, the upper limit for N_{max} is associated with the repeat period of β-type structures and the estimated size of the globule (here n_{min} = 1.8 S_0/l_β, with l_β = 3.4 Å the length per residue of an expanded strand).

4. Select by lottery, N-1 structural templates whose lengths are appropriate to the current division of the protein chain. Each structural fragment has to be a hairpin, which goes across the globule and passes through three "checkpoints" (the beginning of the hairpin, the top of the hairpin, and its end) near its surface. One may just cut the templates from randomly selected fragments of a protein structural database; here, for simplicity, a lattice representation of high-resolution folds is used.
5. Modify by a random shift (by one residue at a single division point) the original division to produce N new sequence fragments.
6. Select the lowest energy [computed according to Eq. (7)] set of templates from many (on the order of 10^4) cycles consisting of operations 4–5. The hairpins partially overlap along the sequence; however, structurally they are constrained only by the requirements of the surface positioning of the top of the hairpin and its two ends.
7. Store the lowest energy set and use the corresponding division as the starting point in step 4.
8. Repeat the process of steps 4–6 many times. Each time, take the set of the lowest energy building blocks into the final sample of possible chain conformations.
9. Perform a clustering of the sample to provide the most probable number of secondary structure elements for the sequence of interest, make the secondary structure assignment, and compute the most probable locations of the surface turns and their distribution (range of uncertainty). Simple geometrical (and local) criteria are applied for the secondary structure assignment. Additional details are provided below in the sections describing the interaction scheme used by the Monte Carlo procedure.

The proposed procedure is depicted in Figure 3A–C. The hairpin composed of two blocks is used for computation of the short-range interactions (contributions defined in Eqs. (2)–(3); see below) (Fig. 3B); however, the two blocks are "decoupled," and each is treated separately in the burial energy calculations (these contributions are defined in Eqs. (4)–(6) below) (Fig. 3C). This is done for several reasons. A hairpin provides a more physical environment for

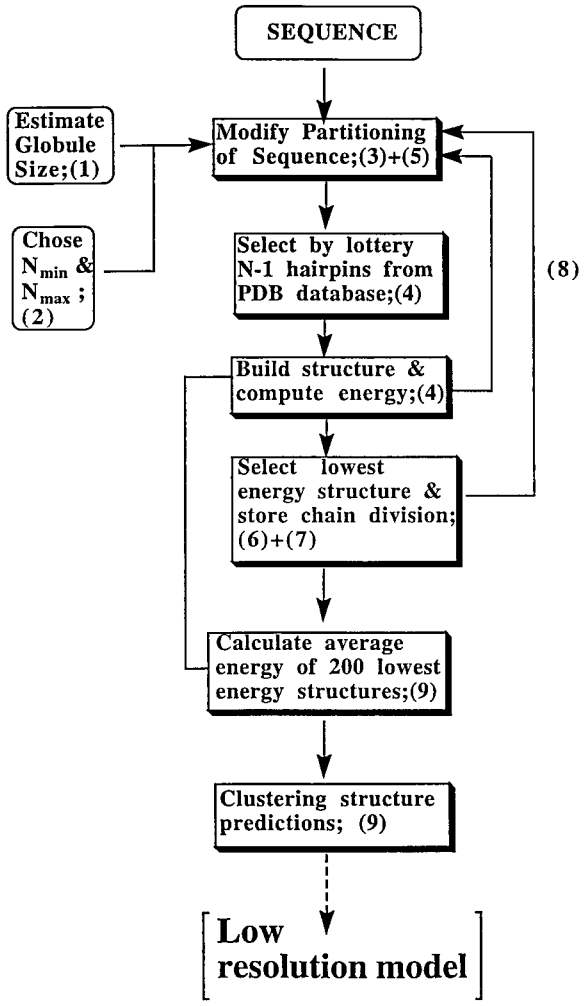


Fig. 2. Flow chart describing the iterative procedure for prediction of secondary structure information from the sequence of amino acids. The numbers in parentheses correspond to those in the outline of the method given in the text.

the estimation of the secondary structure preferences. A single block will leave too many dangling loop residues which behave like free ends. To enhance sampling for the estimation of the burial energy, it is much easier to place a single block (helical or expanded) within the protein sphere rather than the entire hairpin. Since most hairpins come from different size structures, in general, they will not fit within a given sphere. This technique enhances the sampling efficiency and maintains the surface location of the U-turn region and burial of the hydrophobic faces of the strands. Consequently, the burial energy is more reasonably estimated when the two branches of the "hairpin" are treated separately. This will make the loop location somewhat more diffuse, and while one can identify the secondary structures belonging to transglobule connectors, the detailed geometry of the hairpin is lost.

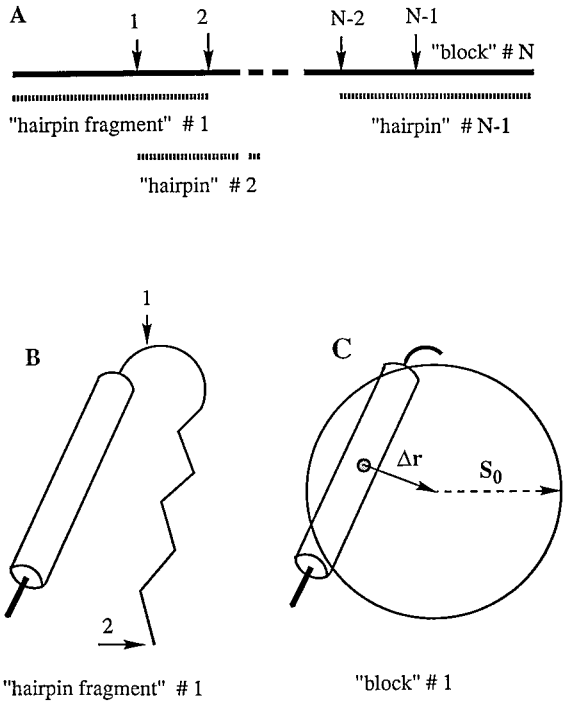


Fig. 3. Schematic representation of the method employed in the present work. **A:** The test protein sequence is divided onto *N* fragments. *N* is a variable that changes over a narrow range of "reasonable" numbers of secondary structure elements for a protein of a given size. Two sequence fragments are then matched to a "hairpin building block" from the structural data base. A set of geometrical restrictions is superimposed that limits the size of particular blocks, and the distance (less than expected radius of gyration of the globule) between hairpin ends. **B:** A hairpin (and the matching test sequence fragment) is used for secondary structure propensity calculations. **C:** Single blocks are used for burial energy calculations. First, the orientation of the hydrophobic and hydrophilic side chains is used to define the direction to the center of mass of the hypothetical globule. The length of the $\Delta\mathbf{r}$ vector is assumed to be equal to $S_0/2$. The procedure is repeated many times to estimate an optimum distribution of the number and location of division points, and subsequently, the locations of loops and secondary structure assignment. See text for more details.

Geometric realization

For simplicity and speed of computation, we use a library of structures that are projected onto a high coordination lattice, which represents the Cα-reduced backbones of PDB structures with an RMS (coordinate root-mean-square deviation) in the range 0.6–0.7 Å.[36] Each side group is represented as a single point positioned at the center of mass of the most probable side-chain rotamer for a given local main-chain geometry,[6,25,27] defined by two consecutive Cα–Cα vectors.

Interaction Scheme

Overview

The force field for this lattice model was developed previously; here, the relevant subset of interactions is employed.[27,37] Short-range interactions are reflected in two kinds of amino acid pair specific terms designed to account for intrinsic secondary structural preferences. The first term depends on three consecutive virtual Cα bond vectors, as encoded by their end-to-end distance and associated chain chirality and depends on the identity of the central two residues. The second class of local terms describes the angular preferences of pairs of side chains for the first through fourth neighbors down the chain and is defined in terms of the angle between respective vectors from the Cα to the side-chain center of mass. Tertiary interactions are accounted for in terms of a centrosymmetric burial energy and a term that reflects the preferences for the hydrophobic (hydrophilic) face of a block to point toward (away from) the center of the molecule. Both the short-range interactions and the burial energy are based on the statistical correlations seen in a database of protein structures. The numerical values of the statistical potentials have been previously published;[37] they are available upon request from the authors or are more easily accessible via anonymous ftp,[38] as is the list of the protein structures employed to derive the statistical potentials. None of the test proteins are in the database used for the derivation of the parameters.

Intrinsic secondary structural preferences

The short-range interactions reflecting intrinsic secondary structural preferences are described by Cα backbone correlations and correlations[37] between the side chain vectors.[6,7,27] The idea is schematically depicted by Figure 4A,B. With respect to the former class of terms, for each set of consecutive Cα virtual bond vectors, $\mathbf{v}_{i-1}$, $\mathbf{v}_i$, and $\mathbf{v}_{i+1}$, where $\mathbf{v}_i$ is the vector from the ith to $i + 1$th Cα, the secondary structural propensities depend on the identity of the two amino acids of the central two residues (Fig. 4A). The total energy associated with such triplets of backbone vectors, E_S is given by

$$E_S = \Sigma\, \epsilon_S\,(A_i, A_{i+1}, r^{2*}_{i-1,i+2})$$

with

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

where ϵ_S is the virtual backbone conformational energy for a consecutive triplet of backbone vectors, A_i, is the identity of the residue at position i, and $r^2_{i-1,i+2}$ is the square of the distance from Cα $i - 1$ to Cα $i + 2$. $r^{2*}_{i-1,i+2}$ is the "chiral" square of distance between the corresponding chain vertices. "Chiral" means a negative sign is assigned for left-handed conformations and a positive sign for right handed ones, respectively. The potential depends on the r^{2*} parameter and is coarse grained into 6 bins. There is a decrease in the system's energy when the overlapping "arms" of consecutive hairpins have the same secondary structure; this decrease equals −0.5 kT

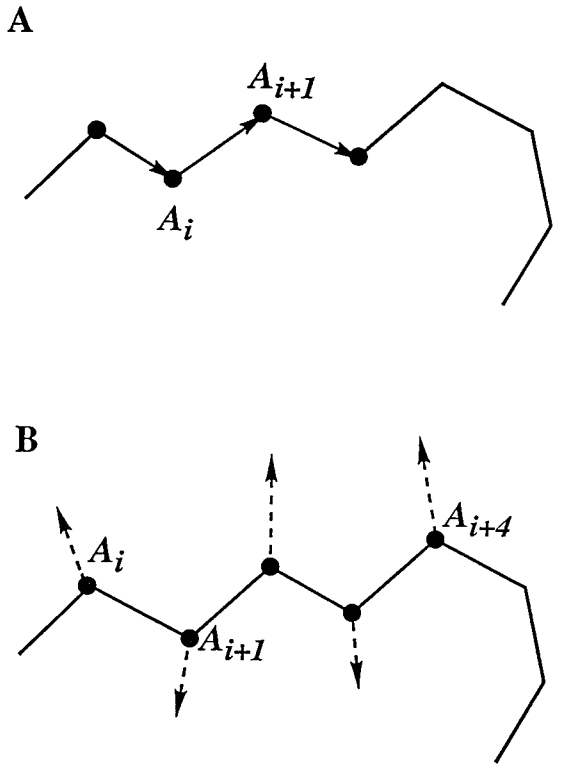


Fig. 4. Explanation of the method of factorization of the short-range interactions. **A:** The first contribution comes from the three vector (two amino acid specific) Cα backbone fragment energy. **B:** The second type of contribution is from orientational correlations between side groups *i*, and *k*, with $k = i + 1$, $i + 2$, $i + 3$, and $i + 4$. All four contributions are amino acid pair-specific. The most probable rotamer for a given residue and for a given backbone conformation is employed to define the vectors from the Cα to the center of mass of the side chain.

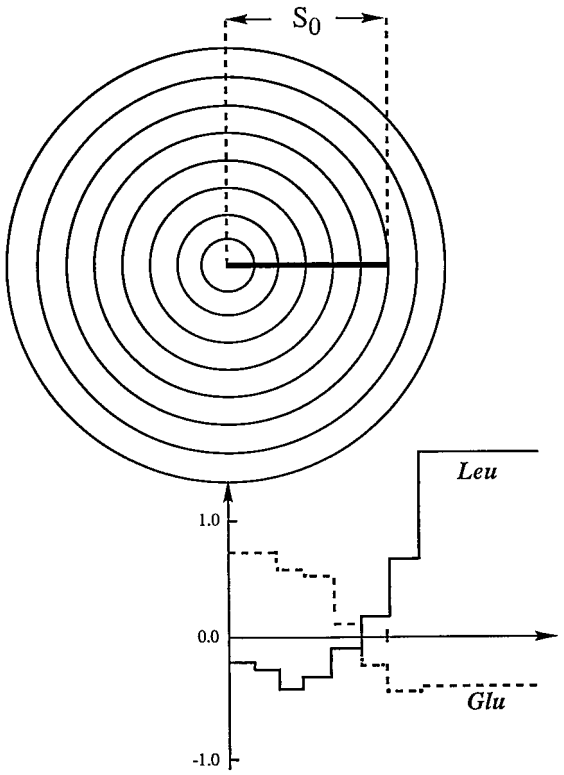


Fig. 5. Centrosymmetric burial energy definition. The expected radius of gyration of a globular protein, S_0, serves as the scaling factor for an 8-shell "onion" model of the protein packing. The resulting histograms of the burial energy are extracted from the statistics of the structural database of single domain proteins, assuming random packing of a protein of average composition as a reference state. In the lower part of the diagram, the resulting potential is shown for two amino acids, the hydrophobic residue, leucine, and the strongly hydrophilic residue, glutamic acid.

when a given position in each of the overlapping arms is in the same bin of r^2*.

Somewhat longer range conformational correlations are accounted for via angular correlations between side group vectors, that is, vectors from the Cα to center of mass of the current rotamer. This class of terms contributes for the first through fourth neighbors down the chain. These interactions are schematically depicted in Figure 4B. The most probable side chain rotamers are used for the computations. The total local energy associated with the angular orientational preferences of the side groups, $E_{\text{sg-local}}$, is defined as

$$E_{\text{sg}-\text{local}} = \Sigma\, \epsilon_k\, (A_i, A_{i+k}, \cos\,(\Theta_{i,i+k}))) \quad k = 1, 2, 3, 4 \tag{3}$$

where ϵ_k is the energy associated with the orientational coupling of side groups located at residues *i* and $i + k$, $\Theta_{i,j}$ is the angle between the side group vectors of residues *i* and *j*. The potential is encoded in the form of a histogram with an angular bin of 36° and a range of 0° to 180°.

Tertiary interactions

The form of the burial energy of a particular type of side chain is schematically depicted in Figure 5. The centrosymmetric potential, ϵ_l, is amino acid specific and depends only on the distance between the center of mass of the globule and the center of the side group of interest.[6,27] The total contribution of the centrosymmetric potential, E_l, is given by

$$E_1 = \Sigma\, \epsilon_1\, (R\, (A_i)/S_0) \tag{4}$$

where S_0 is the expected radius of gyration of a single domain protein consisting of m amino acids in their native conformation and is defined in Equation (1). $R(A_i)$ is the distance of the center of mass of the *i*th side group from the center of mass of the entire chain. The potential is derived from the statistics of single-domain proteins and is expressed in the form of a histogram.[6]

Some explanation is required for the procedure that positions the particular building blocks with respect to the center of the hypothetical protein. First, the orientation of the hydrophobic face of a strand is determined from a vector **f,** which is a burial energy-weighted sum of the side-chain center of mass vectors and is given by

$$\mathbf{f} = \Sigma\,(\epsilon(\mathrm{i})_{\mathrm{K-D}} \cdot \mathbf{g_i}/|\mathbf{g_i}|) \quad \text{for } \mathbf{r}_\mathrm{B} < S_0 \qquad (5)$$

where $\mathbf{g}_\mathrm{i}$ is the vector from the *i*th Cα to the *i*th side-chain center of mass (as in Fig. 3B), $\mathbf{r}_\mathrm{B}$ is the vector from the center of mass of the building block to the side chain of interest, and $\epsilon_{\mathrm{K\text{-}D}}$ are the Kyte-Doolittle hydrophobicity parameters.[39] Due to the different reference state used here, the Kyte-Doolittle parameters have been divided by a factor of 5. It should be pointed out that there is an implicit assumption that the face of each block can be defined and that the supertwist of the secondary structure is not too large. This is actually one more reason why the proposed method can only be applied to small, single domain globular proteins. After determining the direction of the hydrophobic face from Eq. 5, the center of the globule is placed at a distance $\Delta\mathbf{r} = (\mathbf{f}/(|\mathbf{f}|) \cdot S_0/2$ from the center of mass of the building block (Fig. 3C). At this point, the fragment is properly placed, and the centrosymmetric burial potential can be computed for all of the side groups. An additional and rather important contribution comes from the face separation term, and is given by $|\mathbf{f}|$, defined in Eq. (5).

Next, there is a correction (equal to $\epsilon(i)_{K-D}$) for the burial energy of loop residues, that is, those residues outside the radius S_0. Furthermore, the block residues are additionally energetically stabilized for a proper pattern of hydrophobic and hydrophilic residues associated with helical and β strands, respectively. This term is given by

$$\begin{aligned} E_{\text{pattern}} &= \Sigma\epsilon_{\mathrm{K-D}}(i) && \text{for } \mathbf{r}_\mathrm{B} > \mathrm{S}_0 \\ &= -\Sigma\epsilon_{\mathrm{K-D}}(i)\,\epsilon_{\mathrm{K-D}}\,(\mathrm{i}+2) && \text{for } \mathbf{r}_\mathrm{B} < \mathrm{S}_0, \text{ and } n < n^* \\ &= \Sigma\epsilon_{\mathrm{K-D}}(i)\,\epsilon_{\mathrm{K-D}}\,(\mathrm{i}+2) && \text{for } \mathbf{r}_\mathrm{B} < \mathrm{S}_0, \text{ and } n > n^* \end{aligned} \qquad (6)$$

Here, n is the length of the block, and n^* is the mean value between the largest possible block and the shortest possible block. These values are dictated by the total number of residues in the test sequence. More precisely, $n^* = (n_{\max} + n_{\min})/2$. The largest value of n ($n_{\max} = 2.5\ S_0/\mathrm{l_h}$) corresponds to the longest helix that fits into the globule, and the smallest value of n ($n_{\min} = 1.8\ S_0/\mathrm{l}_\beta$) corresponds to β strands that just cover the hydrophobic core diameter within the globule. The numbers $\mathrm{l_h} = 1.5$ Å and $\mathrm{l}_\beta = 3.4$ Å correspond to the approximate extension per residue (in Angstroms) of an α helix and β strand, respectively.

Total hairpin conformational energy

The total energy of a hairpin fragment can then be expressed as the sum of single contributions from the short-range interactions (comprising the hairpin) and two sets of contributions for the long-range interactions in each of the two blocks (1 and 2) in the hairpin; the latter having been independently positioned with respect to the center of mass of the globule.

$$E_{\text{hairpin}} = E_\mathrm{S} + E_{\text{sg-local}} + (E_1 + |\mathbf{f}| + E_{\text{pattern}})_1 + (E_1 + |\mathbf{f}| + E_{\text{pattern}})_2 \qquad (6)$$

The $|\mathbf{f}|$ contribution reflects the strength of the orientational separation of the hydrophilic and hydrophobic side groups, and E_{pattern} is defined by Eq. (6). The method is insensitive (over quite a broad range of parameters) to the specific weighting of the short- versus long-range interactions. The weighting of long- versus short-range interactions should be selected in a way that both contributions in the lowest energy assemblies are of the same magnitude. While the method is not too sensitive to the particular scaling of the long- versus short-range interactions, the above balancing of both types of interactions somewhat improves performance. This requires a scaling of about 1:2 for the ratio of the long- to short-range interactions (i.e., long-range potentials/short-range potentials = ½).

Analysis protocol

Classification of structures

Each simulation provides a set of 200 "lowest energy structures," which, based on the energy described in Eq. (7), are well suited for the test sequence. Each of the resulting structures consists of a series of overlapping hairpins. Those structures with energies below 1.05 times the average energy provide information about the location of the division points and the secondary structure of the blocks attached to the U turns; see below. The population of these low-energy structures varies and is dependent on the width of the energy distribution. In most cases, the algorithm selects a single set of division points (all the lowest energy structures have the same number of transglobular blocks). Thus, it provides the expected number of secondary structure elements and the corresponding number of surface U turns in the protein fold. In the remaining cases, there is a leading division that is taken for future analysis. However, the presence of a competing division of the chain may serve as a hint in more ambiguous cases where the turn distribution becomes flat.

Determination of "U" turn positions

The location of the division points between blocks usually exhibits a fairly narrow distribution. The

peaks of this distribution indicate the external loop/turn regions, and these U turns are assigned for all nonzero occurrences of the location of the most probable number of division points. The number of counts in the histogram representing these U-turn distributions (shown for example by the thick line in Fig. 6A) depends on the number of the lowest energy states selected in a particular run, and therefore, it reflects the energetic selectivity or the width of the energy distribution of the implicitly "assembled" chains. A more diffuse distribution of division points usually indicates a broad surface loop. In contrast, narrow, β-type turns exhibit a sharp distribution of division points.

The actual position of the U turns in the structure are assigned by a chain smoothing algorithm where the coordinate of the ith Cα is replaced by a weighted arithmetic average with a window of five neighbors on each side.

Let $\mathbf{X} = \{\mathbf{x_i}\}$ denote the actual coordinates of the α carbons. To account for end effects, in a chain containing N residues labeled 1 to N, we set

$$\mathbf{x_k} = \mathbf{x_1} \quad \text{and } k = -4, -3, -2, -1, 0 \tag{8a}$$

and

$$\mathbf{x_{N+k}} = \mathbf{x_N} \quad \text{and } k = 1, 2, 3, 4, 5 \tag{8b}$$

Then,

$$\mathbf{Y} = \mathbf{MX} \tag{8c}$$

with $\mathbf{M}$ an $N + 10$ by $N + 10$ matrix:

$$\begin{aligned} M(i, i) &= 51/243 \\ M(i, i \pm 1) &= 45/243 \\ M(i, i \pm 2) &= 30/243 \\ M(i, i \pm 3) &= 15/243 \\ M(i, i \pm 4) &= 5/243 \\ M(i, i \pm 5) &= 1/243 \\ M(k, \mathrm{l}) &= 0 \text{ otherwise.} \end{aligned} \tag{8d}$$

Next, for this smoothed Cα trace, an approximate value of the reciprocal of the radius of curvature at each position i is calculated from

$$R_c^{-1}(i) = |\{\mathbf{Y}(i + 2) + \mathbf{Y}(i - 2) - 2\mathbf{Y}(i)\}/4| \tag{9}$$

and the regions where the reciprocal of the radius of curvature is at a maximum are identified and values above a threshold of 0.11 are defined as the U turns. This defines the preliminary parsing of the chain into blocks and U turns. Next, the identity of the dominant secondary structure within the block is assigned [see Eq. (10)]. If the length of a extended/β block exceeds 10 residues or if the length of a helical block exceeds 15 residues, then the chain is rescanned to identify those positions, if any, where $R_c^{-1}(i)$ exceeds 0.06. There are generally regions that correspond to 90° kinks of the chain. This defines the location of all U turns. In practice, this definition sometimes extends the location of a β turn by one residue on each side, but this level of descriptor is compatible with the accuracy.

Determination of secondary structure of transglobular blocks

For each transglobule block, its secondary structure is extracted from the average local geometry of the backbones corresponding to the set of lowest energy structures which are less than 1.05 times the average energy. The criteria employed for assignment are based on the values of $r^{2*}_{i-1,i+2}$ [see Eq. (2)] of the particular fragments of structural building blocks, and they are as follows:

$$\begin{aligned} &\text{(H) helix} && 0 < r^{2*}_{i-1,i+2} < 37 \text{ Å}^2 \\ &\text{(T) turn} && r^2_{i-1,i+2} < 60 \text{ Å}^2 \quad \text{and not a helix} \\ &\text{(C) coil} && 59 \text{ Å}^2 < r^2_{i-1,i+2} < 75 \text{ Å}^2 \\ &\text{(E) extended} && r^2_{i-1,i+2} > 74 \text{ Å}^2 \end{aligned} \tag{10}$$

Alternatively, one could use the straightforward Kabsch-Sander method[9] to assign secondary structure. However, because our algorithm is driven by local backbone geometry and not the long-distance pattern of the hydrogen bonds, this would partially defeat the purpose of the current procedure. Consequently, assignments of β structure would be very inaccurate. Thus, we report the results according to geometry-based assignments.

Because of the manner in which the hairpins are constructed, the secondary structure assignments in the predicted U-turn regions are very ambiguous. The U-turn fragments in the hairpins are used to extrapolate the secondary structural propensities of each single transglobule connection. The secondary structural assignment in the middle portion of the particular "transglobular" building block is of highest accuracy; such regions are perhaps the most important from the point of view of model building. Thus, we will use this output to assess the accuracy of our method. More precisely, the leading secondary structure assignment for the three central residues between the centers of the U-turn distribution peaks [assigned according to rules from Eq. (10)] in each

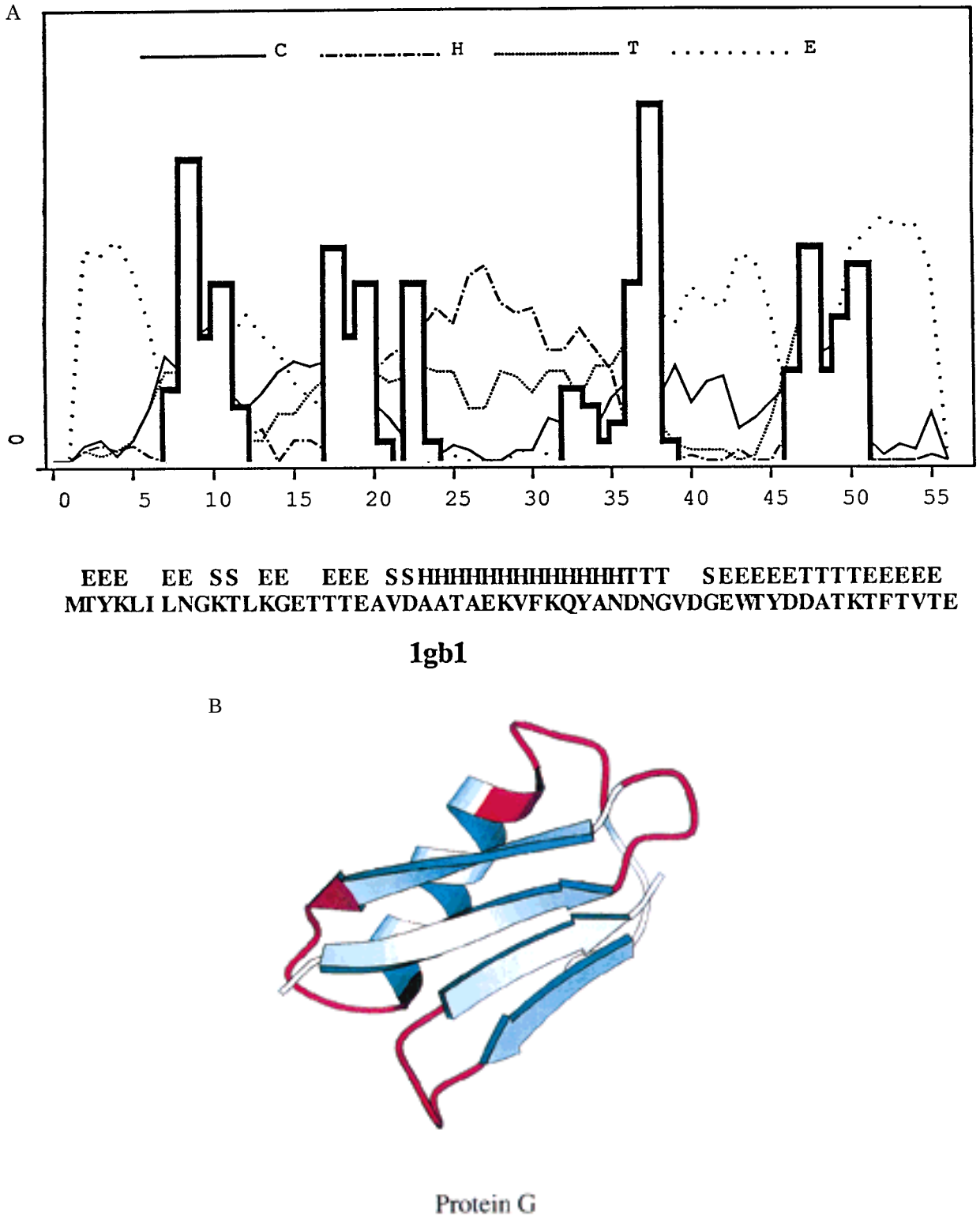


Fig. 6. **A:** The predicted U-turn distribution for the sample of the lowest energy structures assigned to the sequence of domain B1 of Streptococcus protein is indicated by the thick solid line. The secondary structure prediction (the most frequent structures seen in the building block, based on geometrical criteria for the Cα backbone fragments) is given by the thin solid line (coil), dot-dashed line (helix), dotted line (turns), and sparse dotted line (β), respectively. Below the figure, for purposes of comparison, the Kabsch-Sander assignment for the native state is presented. **B:** Ribbon drawing of the native conformation the B domain of 1gb1. Red indicates the location of the predicted U turns, blue indicates predicted expanded states, classified by the MC algorithm as possibly being of β-type.

transglobular block is used as the assignment for the entire block.

Summary of predicted information

The algorithm predicts the number of transglobule blocks, or secondary structure building blocks, and the secondary structure of the central region of the transglobule blocks. For all the secondary structure elements, their end-to-end distances are predicted to be close to $2S_0$. Furthermore, the location of surface U turns, where the chain changes its direction, are predicted. In a very rare number of runs, the lowest energy "structures" could be grouped into two clusters that correspond to two competing answers. In such a case, the secondary structure assignments for these two clusters could differ in some fragments, indicating weakly and strongly predicted regions. In such cases, the algorithm provides two alternative structural assignments. Thus, a broad diversity of information is provided that can aid in subsequent prediction of tertiary structure.

Convergence

Finally, it should be pointed out that the present method converges very fast. For example, a reasonable first estimate can be obtained from a run that explores only a very small sample of test block conformations (range: 10^3) per block location for each of 200 structures generated. Subsequent structural modifications after 10 times longer sampling essentially fine tune the predictions. Except for very short runs, the results do not depend on the starting division or on the seed number for the Monte Carlo process. Moreover, the use of half of the structural template database (instead of the entire set of possible fragments) has very little, if any, effect on the predictions. These are very important observations. First, let us note that even in a very long run, the algorithm is very far (by many orders of magnitude) from the limit of an exhaustive search through the database of the structural templates. Consequently, there are many structural fragments (and many combinations) that work equally well in the framework of the Monte Carlo search algorithm. This implies that the method is not sensitive to structural details, to the details of secondary structure assignments for the selected fragments, nor to other details of the model implementation, and is suggestive that an important physical effect (the interplay between the short- and the long-range interactions) has been correctly accounted for.

RESULTS

The list of examined proteins is given in Table I. All are rather small proteins that have a single-domain structure. With a few exceptions, which include the B1 domain of protein G[40] and the B domain of protein A,[41] the test proteins were chosen randomly from a set of coordinate entries made available by the Brookhaven Protein Data Bank, PDB, in early 1995.[42] Thus, we expect that they are representative of small globular, single-domain proteins. None is a member of the structural database[27] used in the derivation of the statistical potentials employed here.

Table II summarizes the statistics for the accuracy of the predictions as applied to the 38 test sequences. The third column of the table gives the number of correctly predicted surface U-turn regions and the actual number seen. In all structures, there are a total of 180 U turns. The location of each U turn is read from histograms (as in the one given for 1gb1 in Fig. 6). The prediction of turns is very permissive in the sense that, if there is at least one structure showing a U turn at a given position, then it is considered predicted. Thus, in general, the turn's span could be overpredicted. Except for those U turns that are assigned to the most exposed irregular terminal fragments of the test structures, the error in loop detection in most cases is generally on the order of a two-three residue shift in position along the chain.

As described in the Methods section, the secondary structure as read from the chain geometry of the building blocks is rather poorly defined in the U-turn regions, and becomes more and more precise as the centers of the building blocks are approached. Thus, the secondary structure assignments of the three central residues of each block between U turns are taken as the basis for the structural classification of the blocks given in Table II, column 5. It is easy to verify that the accuracy of the method depends very weakly on the details of the above definition and is rather high. As shown in column 5, most of the errors in the secondary structure assignment of the "transglobular" blocks occur in the terminal irregular fragments (mostly a series of expanded turns) that are wrongly classified as expanded states. Out of 195 assignments of the secondary structure in the inner sections of the transglobule blocks, 35 are in error. Parenthetically, let us note that the Rost-Sander method[18] (perhaps the most powerful one-dimensional secondary structure prediction algorithm) also fails for most of these fragments.

Another important question is the level of accuracy to be expected if the predictions were random. To address this point, we turned off the energy terms and simply selected chain division points that define the U turns and secondary structural fragments at random. The location of all the U turns is predicted correctly in 10% of the randomly chosen structures for the three-helix bundle protein A, to less than 1%

TABLE I. List of Test Proteins

No.	Symbol	Name	No. of Residues	Global fold
1	1gb1	*Streptococcus* protein G domain B1	56	β sheet + helix
2	proA	B domain of protein A	46	3-helix bundle
3	1fas	Fasciculin	61	5-stranded irregular β
4	1pou	POU-specific domain	71	Irregular 4α bundle
5	1tlk	Telokin	103	Immunoglobulin fold
6	1ris	Ribosomal protein S6	97	Up-down α/β protein
7	1lpt	Wheat lipid transfer protein	90	Irregular 4α bundle
8	1ten	Fibronectin repeat of tenascin	89	Immunoglobulin fold
9	1mjc	Major cold shock protein 7.4	69	6-stranded β protein
10	1gps	Wheat γ-1-P thionin	47	β sheet + helix
11	1tfi	Transcriptional elongation factor SII	50	4-stranded β protein
12	1tpm	Human plasminogen activator	50	5-stranded β protein
13	Alcc	*E. coli* LAC repressor	51	3α bundle
14	1pra	Bacteriophage 434 repressor protein	63	5α bundle
15	1c5a	Pig des-arg═74═complement C5A	66	Irregular 4α bundle
16	1trf	Turkey troponin C	76	2β strands + 4α
17	1lea	*E. coli* lexa repressor DNA binding domain	72	2β strands + 3α
18	2ptl	Peptostreptococcus magnus protein L (B1 domain)	78	β sheet + helix
19	1hdn	*E. coli* phosphotransferase	85	β sheet + 3α
20	1bta	*E. coli* ribonuclease inhibitor	89	β sheet + 4α
21	1ego	*E. coli* glutaredoxin	85	β sheet + 3α
22	1svq	Dictyostelium discoideum severin	94	β sheet + 2α
23	1ubq	Human ubiquitin	76	β sheet + helix
24	2utg	Rabbit uteroglobin	70	Irregular 4α bundle
25	1ctf	*E. coli* ribosomal protein L7/L12	68	β sheet + 3α
26	1crn	Cabbage seed crambin	46	2β strands + 2α
27	1msh	Human cytokine	72	β sheet + helix
28	1ftz	Fruit fly DNA-binding protein	70	3α bundle
29	1cis	Hybrid CI-2 protein	66	β sheet + helix
30	1tin	Pumpkin seed trypsin inhibitor V	69	β sheet + helix
31	1cvo	Taiwan cobra cytotoxin	62	5-stranded β protein
32	1adr	Salmonella bacteriophage P22 C2 repressor	76	5α bundle
33	1hme	Rat DNA-binding protein	77	3α bundle
34	1vna	Centruroides sculpturatus Ewing neurotoxin	65	β sheet + helix
35	2ait	Streptomyces tendae α-amylase inhibitor	74	6-stranded β protein
36	1cod	Taiwan cobra cobrotoxin II	62	5-stranded β protein
37	1cb1	Porcine calcium-binding protein	78	Irregular 4α bundle
38	1aca	Bovine acyl-coenzyme A binding protein	86	Irregular 4α bundle

in 1ten and 1tlk. In other words, our results (>90% accuracy) for the turns are substantially better than random. We further explored how well our method performs in comparison to more standard β-turn prediction algorithms such as that of Wilmot and Thornton.[43] As evidenced by Table III, the Wilmot-Thornton method predicts 71% of the actual turns; whereas the present approach is almost 100% accurate on the same set of proteins. While the present method is applicable to all types of protein motifs, since the Wilmot-Thornton method is restricted to the prediction of β turns, in the interest of fairness, here we limited the comparison to all beta structures.

Typical results from the calculation for domain B1 of Streptococcus protein G,[40] 1gb1, a 56-residue sequence, are depicted in Figure 6A, where the distribution of division points (i.e., the location of the U turns) is given in the form of a histogram. For easy reference, Figure 6B contains a schematic drawing of the native backbones of 1gb1 with the predicted surface U turns indicated in red. Yellow would indicate that the secondary structure of a block has been incorrectly predicted. Its absence in Figure 6B indicates that all blocks are correctly assigned in this case.

As one may see, the prediction is very good. For example, the leading division selected by the algorithm corresponds to five secondary structural blocks. All of the lowest energy "structures" consist of 5 fragments, that is, they are built from 4 overlapping hairpins. Figure 6A shows the distribution of division points that correspond to the locations of the "U"-turns. We remind the reader that the term U turn refers not only to local chain conformation, but also means that the polypeptide chain changes direc-

TABLE II. Summary of Prediction Statistics

Sequence no.	Protein name*	Surface U-turn prediction accuracy†	Errors of U-turn locations‡	Secondary structure block prediction accuracy§	Comments on wrong assignment
1	1gbl	4/4	0-2/2-3-0	5/5	—
2	proA	2/2	2-3	3/3	—
3	1fas	5/5	2-1-0-2-0	5/5	Terminal coiled assigned β; inserted β without a turn
4	1pou	4/3	4-3-2-0	4/4	Extended coil inserted
5	1tlk	8/7	5-3-2/1-1-2-4-7	7/8	Turn inserted into the C-terminal β strand
6	1ris	5/5	3-5-6-3/4-4	5/6	Second β strand predicted helical
7	1lpt	4/4	0-5-2-0	4/4	Shifted turns; hairpinlike C-terminus predicted as β
8	1ten	6/7	1/1-3-0-3-2-2/1	7/8	Shifted turns; one β strand missed
9	1mjc	5/5	1-2-0-2-0	6/6	Long central coil added as β
10	1gps	4/3	0-1-2-1	4/4	Long extended loop is predicted as another β
11	1tfi	4/3	0-0-0-1	4/4	Long extended loop is predicted as another β
12	1tpm	4/4	0-0-4-1	5/5	One shifted turn
13	Alcc	3/2	2-0-8	2/3	One short helix predicted as β; one β strand inserted
14	1pra	4/4	7-3-0-2	3/5	Two helices predicted as β; turn positions shifted at N-terminal
15	1c5a	3/3	4-0-5	3/4	One helix predicted as β; turn positions shifted
16	1trf	4/3	2-3-3-0	3/4	One helix predicted as β
17	1lea	4/4	2-2-4-0	4/5	N-terminal loop predicted as helical; shifted turns
18	2ptl	6/5	1-3-3-4-3-2	5/5	N-terminal loop predicted as β; one β strand inserted
19	1hdn	5/6	0-3-4-1-2/2	6/7	One short β strand missing
20	1bta	6/6	2-0-2-3-0-1	7/7	—
21	1ego	8/7	0-7-2-2-0-2-0-7	6/8	Two helices predicted as β; shifted turns
22	1svq	6/6	1-0-1-0-3-0	6/7	C-terminal helix predicted as β
23	1ubq	5/6	3-0-3-1-4	4/7	One β strand predicted as helical; one helix predicted as β; one β strand missing
24	2utg	3/3	3-2-7	4/4	Shifted turns
25	1ctf	4/5	0-3-0-1/1	5/6	One β strand missing
26	1crn	4/4	7-0-2-0	3/5	Two helices predicted as β strands
27	1msh	4/5	0-0-4-0	2/4	N-terminal loop and one β strand predicted as helical; one β strand missing
28	1ftz	4/4	4-3-3-0	3/3	Both terminal loops predicted as β
29	1cis	4/5	4/1-1/2-5-3/1	3/5	One β predicted as a helix; one β missing
30	1tin	6/5	0-4-3-0-3-0	5/5	Two β strands inserted in the loop regions
31	1cvo	5/5	0-3-1/2-4-0	4/5	C-terminal loop region predicted as another β strand; one β strand predicted as helical; one β strand missing
32	1adr	4/4	4-3-0-4	3/5	Two helices predicted as β strands
33	1hme	2/3	2-4	3/3	Terminal loops predicted within helices
34	1vna	4/4	4-2-3-0	2/5	One short helix predicted as β; one β strand predicted as helical; one β strand missing; C-terminal loop predicted as β
35	2ait	6/6	2/2-1/1-4-0-3-1	5/6	N-terminal loop predicted as β; C-terminal β strand predicted as helical
36	1cod	4/4	4-3-4-1	3/5	N-terminal loop predicted as β; one β strand predicted as helical
37	1cb1	3/3	3-0-0	4/4	Third helix predicted too long
38	1aca	4/4	1-1-0-4/4	3/4	One helix predicted β; one β strand inserted
Total % correct	—	173/180 = 96% +9 overpredicted U turns	—	160/195 = 82%	

*PDB descriptor; see Table I for protein name.
†The ratio of the correctly predicted number of surface U-turns to the actual number in the protein. A turn is said to be correctly predicted if its boundaries at least partially overlap with the actual turn location. "Overpredicted" indicates that a U turn is predicted, which does not occur in the protein structure, or when it separates an extended loop from a regular secondary structure fragment.
‡*i*/*j* means that *i* residues of the preceding block and the *j* residues of the following block have been incorrectly assigned as a part of the surface loop/turn. Otherwise, the number of over assigned residues of one of the transglobular blocks is given.
§The ratio of the correctly predicted number of secondary structure blocks to the actual number of the protein. The secondary structure of a given block is said to be correctly predicted when the secondary structure of the three central residues as defined by Equation (10) agrees with the experimental structure.

TABLE III. Comparison of the Blocks and U Turns Algorithm With the Wilmot-Thornton Approach

Protein	Wilmot-Thornton turn accuracy*	Present method U turn accuracy*
1fas	4/5	5/5
1tlk	5/7	8/7
1ten	5/6	6/6
1mjc	5/5	5/5
1tfi	2/3	4/3
1tpm	2/4	4/4
1cvo	3/5	5/5
2ait	3/6	6/6
1cod	3/4	4/4
Average accuracy	32/45 = 71%	45/45 = 100% + two overpredicted

*The ratio of the number of correctly predicted U turns to total number of U turns is reported.

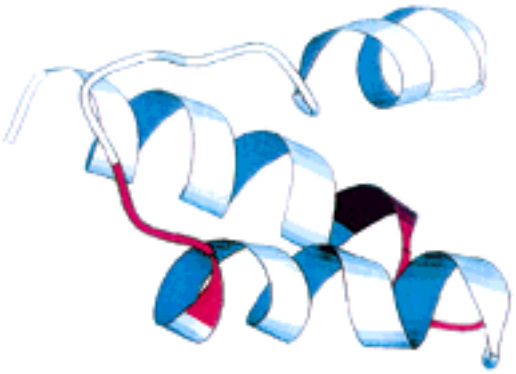
tion. This is clearly demonstrated in the schematic drawing of the native conformation. Thus, the prediction indicates the existence of five structural elements, which are connected via well-localized loops.

Figure 6A also presents the secondary assignment given by the algorithm, based on the most frequently observed structure of the selected building blocks for the lowest energy set of building blocks (below 1.05 times the average score for the 200 lowest energy samples); for comparison, the DSSP secondary structure assignment for the native state is indicated below the plot. The agreement is rather good. The assigned structure of the protein is ββαββ, and the biggest errors involve shifts in the location of two of the turns by a couple of residues along the sequence. The apparent slight overprediction of the E (expanded) states actually takes into account the rather broad loops located next to the central helix.

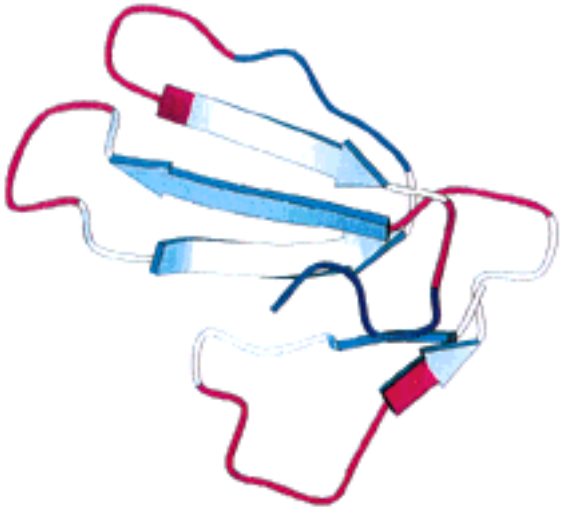
At this point, it seems worthwhile to examine the interplay of the short-range interactions versus burial energies. The first obvious question would be, do we really need the burial energy to predict the location of U turns and the secondary structure of the intervening blocks? Perhaps, just the requirement of "reentry" of the model chain into the globule when combined with secondary structure propensities would be enough to enforce a proper division of the chain into secondary structure elements. To check this possibility, we ran the algorithm without the burial energy. The result for the B1 domain of protein G is noticeably worse than the original prediction shown in Figure 6A,B. The number of predicted secondary elements is correct; however, the location of predicted surface U turns is more diffuse, and the secondary structure assignment is less accurate. Similarly, one may ask if the "reentry" condition and the burial energy together with the pattern of residues (but with the secondary structure contribution to the interactions, E_s and $E_{sg\text{-}local}$ [see Eqs. (2) and (3), deleted] could be used instead. The results of this experiment again show that the prediction is much weaker. We found that divisions containing 4, 5, and 6 blocks compete strongly; but the correct division is dominant. In conclusion, these studies strongly suggest that there is an interplay between the short-range interactions and the burial interactions that reflect the effects of both energetic competition and cooperation. Some fragments are driven by their intrinsic secondary structure propensities, while other fragments are mostly controlled by the pattern of hydrophilic and hydrophobic residues and the need to bury hydrophobic faces and expose hydrophilic faces.[44] Thus, these systems experience considerable energetic frustration as the tertiary and short-range interactions compete with each other to determine the lowest energy conformations of the chain.

In spite of the fact that the predictions of the present method are driven by both secondary structural preferences and burial interactions, the particular scaling (over quite a wide range) of the two sets of terms has very little influence on the results. This would suggest that very rarely is there a very strong contradiction between the secondary structure propensities and the burial energy. However, when such a contradiction of a moderate magnitude occurs, the balance of the two terms is important, and it is precisely this balance that enhances the accuracy of the present method. Consequently, other realizations of the method, employing a different representation of the building blocks and different factorizations of the short-range interactions, should be similar in accuracy.

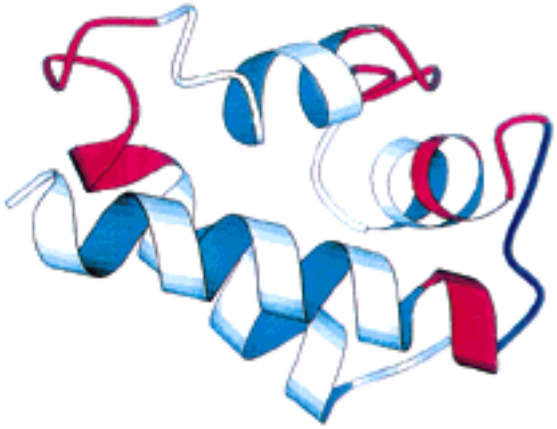
In Figure 7, the predictions for test proteins 2–9 of Table I are illustrated in terms of their location in the native structure, with the same convention as was used in Figure 6B. Red indicates the location of the predicted U turns; yellow indicates that the secondary structure of a block has been incorrectly predicted, and its absence indicates that the block is correctly assigned. Dark blue indicates predicted expanded states that have been classified by the algorithm as possibly being of the β type. Let us comment here on some interesting cases from this subset. First, let us note that in all cases displayed in Figure 7, the secondary structure of the central fragments of the building blocks (extended or helical, roughly speaking) is correctly assigned except for the clear qualitative errors in the assignment of a β strand as a part of the turn in tenascin (1ten) and in the third block of the ribosomal protein (1ris) sequence. This 1ris block has been predicted to be helical, in direct contradiction to the PDB structure where it is in a beta conformation. This is a rare example when both our secondary structure propen-



Protein A

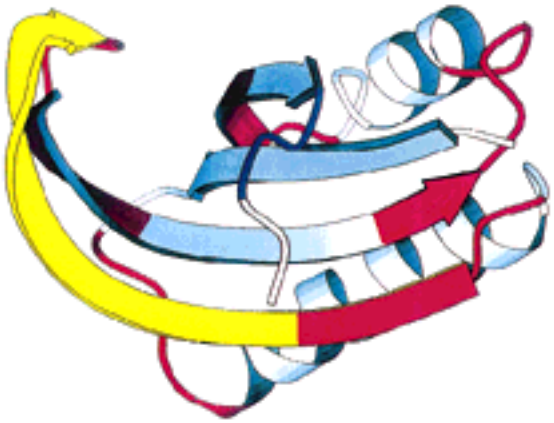


Fasciculin

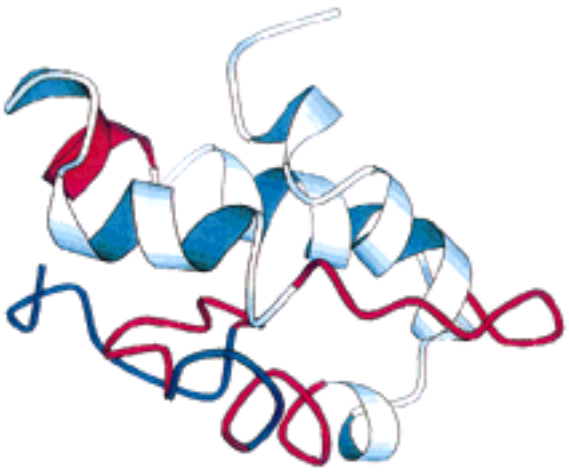


DNA binding domain

Telokin



Ribosomal protein



Lipid Transfer Protein

Fig. 7.

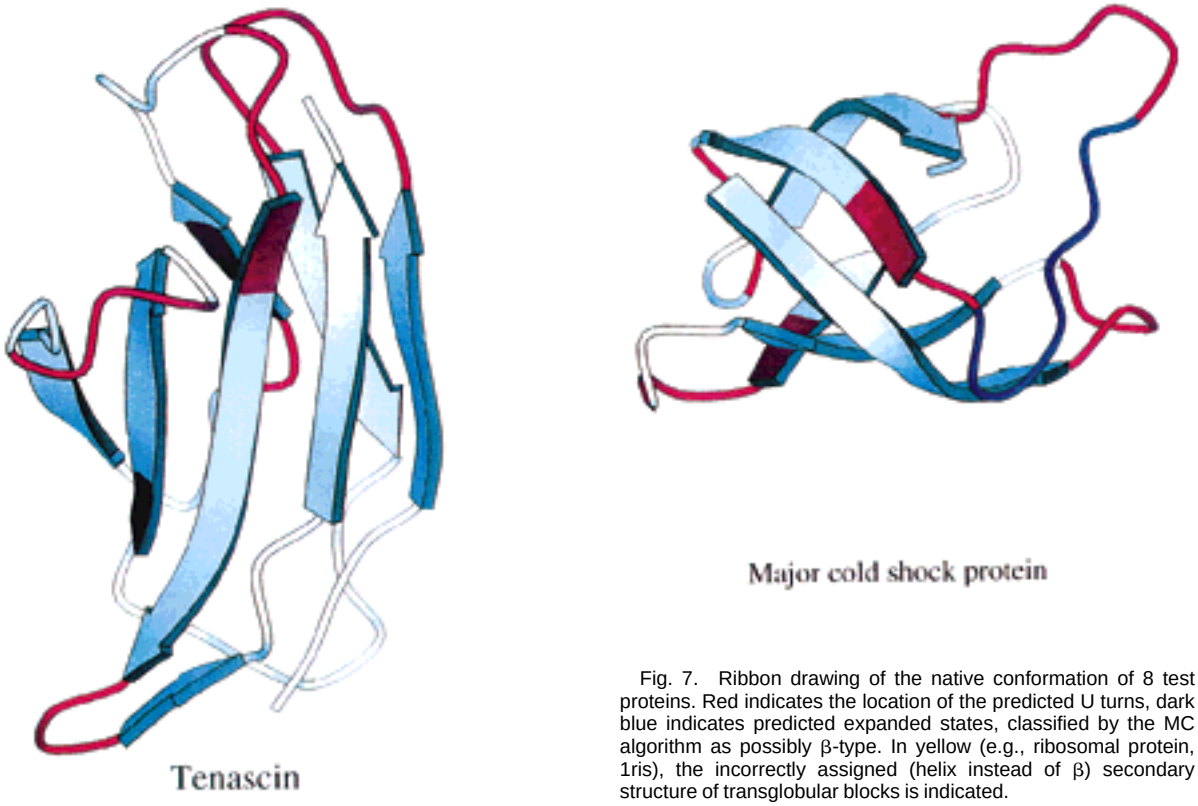


Fig. 7. Ribbon drawing of the native conformation of 8 test proteins. Red indicates the location of the predicted U turns, dark blue indicates predicted expanded states, classified by the MC algorithm as possibly β-type. In yellow (e.g., ribosomal protein, 1ris), the incorrectly assigned (helix instead of β) secondary structure of transglobular blocks is indicated.

sities and burial terms (including the pattern of hydrophobic and hydrophilic residues, where the leading repeat for the hydrophobic residues is 3, certainly more acceptable for helices than β-type structures) strongly favor the wrong (helical) assignment. If this β strand were shorter (in the native structure, it is the longest one), then in this algorithm, it could be perhaps induced by the neighboring, well-defined fragments to adopt an extended state. On the other hand, even in this case, the number of secondary structure elements and location of the surface turns/loops, U turns, have been correctly predicted. Perhaps, for such a strongly energetically frustrated sequence fragment, there is no way to correctly predict its secondary structure without invoking more detailed tertiary interaction related effects, such as hydrogen bonding and pair interactions. Clearly, this problem requires further investigation.

An interesting example is the 1pou sequence (DNA binding domain). The Kabsch-Sander[9] assignment of secondary structure indicates a fold that is built from four helices, and the first connection between helices appears to be the narrowest. Our method predicts four helices; however, it also indicates one more turn near the end of the first helical fragment defined according to the Kabsch-Sander assignment. This way, relative to the Kabsch-Sander method of secondary structure assignment, the present method indicates that there is an additional extended fragment (shown in blue in Fig. 7). Of course, monomeric proteins do not have single β strands, and therefore, it has to be interpreted as a expanded coil structure. Consequently, the fold could be safely predicted as being of the αααα type, with a rather broad loop between the two first helices. Indeed, the inspection of the schematic drawing of the native structure (see Fig. 7) shows that this is exactly the case. Probably, the largest errors in U-turn predictions occur in the less regular helical proteins. Here, the algorithm sometimes selects exposed fragments of helices near the helix end as a surface U turn. This effect is seen in 1pou and in 1lpt.

Another interesting case is the major cold shock protein, 1mjc, where residues 42–46 are predicted to be in an extended (possibly β-type) state, while according to the Kabsch-Sander assignment, it is a coil fragment. Inspection of the 3D native structure shows that this fragment is very expanded, with a β-type conformation (shown in blue), except for the lack of hydrogen-bonded partners. This is another illustration of the kind of structural information the present method provides. Similarly, a long expanded coil fragment has been assigned as β type in fasciculin, 1fas. This kind of assignment accounts for about half of the cases of wrongly predicted blocks (as

listed in Table I) in the entire test set. Consequently, the "overpredicted" turns actually do reflect changes of chain direction. Since the fragment of 1fas (and of other similar cases) is in reality very expanded, this is a geometrically correct assignment. However, the algorithm missed the change of the chain direction between the preceding β strand and the expanded coil fragment. In many cases, some ambiguities could be easily resolved and the type of fold could be precisely defined, while in other situations (especially for less regular β proteins), one might have to consider several alternative topologies.

We have also applied the proposed method to a set of 10 proteins that were members of the structural data base used for derivation of potentials and in the pool of structural templates (building blocks). What is very interesting is that for these proteins the prediction quality was the same as that obtained for members of the test set. This suggests that the proposed method captures the general features of protein folds, while the local details have little influence on the results. This is not surprising, since the tertiary interactions are accounted for in a very approximate way, reflecting essentially only the burial status and some aspects of specific hydrophobic and hydrophilic residue pattern along protein sequences.

In general, the predictions for proteins that have more regular folds (where a larger fraction of residues that could be assigned as helical or extended) are very good, while less regular structures account for most of the qualitative errors. For about one half of the tested proteins, the predictions are accurate enough (i.e., they yield the correct number of U turns and transglobular connections, the correct assignment of leading secondary structure, and a small error in all turn locations) that they can be used to propose plausible low-resolution structures. This way of addressing the protein structure prediction problem is now being investigated.

CONCLUSION

In this work, using just sequence information, for small, single-domain globular proteins, we have developed a method that predicts the location of the U turns and the dominant type of secondary structures of the transglobule blocks that join such loops or turns. Thus, the focus of our approach is on the prediction of the global topological elements and their identity, as opposed to a more locally precise, but globally far more ambiguous definition of secondary structure. In the prediction of low-resolution tertiary structure, it is the knowledge of the number and location of the U turns as well as the dominant secondary structure within the blocks that provides the most useful information. This knowledge would serve to effectively limit the number of possible topologies and thereby enhance the efficiency of a global conformational search algorithm, provided that the prediction is of sufficient accuracy. In practice, for 38 small test proteins, in almost all (96%) cases, the surface loops or turns, U turns, that are characterized by a change of overall direction of the polypeptide chain are predicted with errors in the range of 2–3 residues. For U-turn predictions, the requisite level of accuracy has been achieved. Furthermore, there are only 35 out of 195 cases where the secondary structure in the blocks is incorrectly classified; thus, the method is correct 82% of the time. This aspect of the algorithm does require some improvement.

The success of this method is predicated on the interplay of tertiary and secondary structure preferences. While at times the two tendencies may act in the same direction, in other cases, the resulting secondary structure reflects a compromise between these two kinds of terms. This is suggestive that proteins, on the average, need not satisfy the principle of minimal frustration[45] for each sequence fragment simultaneously. Thus, burial preferences that state that all hydrophobic residues should lie in the protein core are not completely satisfied; otherwise, there would be no unburied hydrophobic residues and no buried hydrophilic residues. While on average this is true, in general, there are many exceptions to this rule. Similarly, due to presence of long-range interactions, the intrinsic secondary preferences cannot always be satisfied. This is evidenced by the presence of pentapeptide fragments in more than one type of secondary structure.[46,47]

The ultimate significance of the present method for protein modeling needs to be established; however, two points seem clear at this point. First, the method accurately predicts the location of surface loop/turns, U turns, where the chain reverses its direction, and therefore, as discussed above, it provides important information for various three-dimensional protein modeling procedures. Second, for small proteins with regular structure, the present method provides sufficient information to propose a few (sometimes just one) low-resolution alternative (due to various possible handedness of the connections of known secondary structure elements[5,48]) folds that could be further refined by various techniques. Based on the conjuncture of the exact prediction of the number of loops and the secondary structure assignments of the transglobular blocks, this is the case in at least half of the tested sequences. The method provides self-consistent global information about the character of the fold. Thus, with some help from knowledge-based topological rules, this information may be sufficient for building low-resolution models of the native structure for

many monomeric globular proteins. Promising results along these lines are now being pursued. Preliminary results indicate that protein G can be successfully folded using information provided by this algorithm, and if the blocks and U-turn predictions are used to screen predicted tertiary contact information, then a variety of proteins can be folded.[49] These include 3icb, 6pbti, 1shg, ICIS, and Ipoh.

ACKNOWLEDGMENTS

We thank Professor Jan Hermans and the referee for their suggestions, which helped to clarify this paper. This work was supported in part by NIH grant GM-48835 and in part (A.K.) by University of Warsaw grant BST-502/34/95. The research of Andrzej Kolinski was supported in part by an International Research Scholars grant from the Howard Hughes Medical Institute (grant 75195-543402).

REFERENCES

1. Rose, G.D. Prediction of chain turns in globular proteins on a hydrophobic basis. Nature 272:586–590, 1978.
2. Rose, G.D., Seltzer, J.P. A new algorithm for finding the peptide chain turns in a globular protein. J. Mol. Biol. 113:153–164, 1977.
3. Lewis, P.N., Momany, F.A., Scheraga, H.A. Chain reversals in proteins. Biochem. Biophys. Acta 303:211–229, 1973.
4. Rooman, M.J., Wodak, S.J., Thornton, J.M. Amino acid sequence templates derived from recurrent turn motifs in proteins: Critical evaluation of their predictive power. Prot. Eng. 3:23–27, 1989.
5. Richardson, J. The anatomy and taxonomy of protein structure. Adv. Prot. Chem. 34:167–339, 1981.
6. Kolinski, A., Skolnick, J. Monte Carlo simulations of protein folding I. Lattice model and interaction scheme. Proteins 18:338–352, 1994.
7. Kolinski, A., Godzik, A., Skolnick, J. A General method for the prediction of the three dimensional structure and folding pathway of globular proteins: Application to designed helical proteins. J. Chem. Phys. 98:7420–7433, 1993.
8. Cohen, F.E., Abarbanel, R.M., Kuntz, I.D., Fletterick, R.J. Turn prediction in proteins using a pattern-matching approach. Biochemistry 25:266–275, 1986.
9. Kabsch, W., Sander, C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22:2577–2637, 1983.
10. Guss, J.M., Freeman, H.C. Structure of oxidized poplar plastocyanin at 1 · 6 Å resolution. J. Mol. Biol. 169:521–563, 1983.
11. Kraulis, P. MOLSCRIPT: A program to produce both detailed and schematic plots of protein structures. J. Appl. Crystallogr. 24:946–950, 1991.
12. Chou, P.Y., Fasman, G.D. Prediction of protein secondary structure. Adv. Enzymol. 47:45–148, 1978.
13. Garnier, J., Ousguthorpe, D.J., Robson, B. Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. J. Mol. Biol. 120:97–120, 1978.
14. Zhang, X., Mesirov, J.P., Waltz, D.L. Hybrid system for protein secondary structure prediction. J. Mol. Biol. 225: 1049–1063, 1992.
15. Rost, B., Sander, C. Progress of 1D protein structure prediction at last. Proteins 23:295–300, 1996.
16. Kneller, D.G., Cohen, F.E., Langridge, R. Improvements in protein secondary structure prediction by an enhanced neural network. J. Mol. Biol. 214:171–182, 1990.
17. Holley, L.H., Karplus, M. Protein secondary structure prediction with a neural network. Proc. Natl. Acad. Sci. USA 86:152–156, 1989.
18. Rost, B.S., Sander, C. Combining evolutionary information and neural networks to predict protein secondary structure of globular proteins. Proteins 19:55–72, 1994.
19. Chou, P.Y., Fasman, G.D. Conformational parameters for amino acids in helical, β-sheet and random coil regions calculated from proteins. Biochemistry 13:211–222, 1974.
20. Karplus, M., Shakhnovich, E. Protein folding: Theoretical studies of thermodynamics and dynamics. In: "Protein Folding." Creighton, T.E. (ed.). W.H. Freeman, 1992:127–196.
21. Karplus, M., Sali, A. Theoretical studies of protein folding and unfolding. Curr. Opin. Struct. Biol. 5:58–73, 1995.
22. Levitt, M. Protein folding. Curr. Opin. Struct. Biol. 1:224–229, 1991.
23. Hao, M.-H., Scheraga, H.A. Statistical thermodynamics of protein folding: Sequence dependence. J. Phys. Chem. 98:9882–9893, 1994.
24. Hao, M.-H., Scheraga, H.A. Monte Carlo simulations of a first-order transition for protein folding. J. Phys. Chem. 98:4940–4948, 1994.
25. Kolinski, A., Skolnick, J. Monte Carlo simulations of protein folding II. Application to protein A, ROP, and crambin. Proteins 18:353–366, 1994.
26. Sali, A., Shakhnovich, E., Karplus, M. How does a protein fold? Nature 369:248–251, 1994.
27. Kolinski, A., Galazka, W., Skolnick J. Computer design of idealized β-motifs. J. Chem. Phys. 103:10286–10297, 1995.
28. Wodak, S.J., Rooman, M.J. Generating and testing protein folds. Curr. Opin. Struct. Biol. 3:247–259, 1993.
29. Godzik, A., Skolnick, J., Kolinski, A. A topology fingerprint approach to the inverse folding problem. J. Mol. Biol. 227:227–238, 1992.
30. Jones, D.T., Taylor, W.R., Thornton, J.M. A new approach to protein fold recognition. Nature 358:86–89, 1992.
31. Flockner, H., Braxenthaler, M., Lackner, P., Jaritz, M., Ortner, M., Sippl, M.J. Progress in fold recognition. Proteins 23:376–386, 1996.
32. Madej, T., Gibrat, J.F., Bryant, S.H. Threading a database of protein scores. Proteins 23:356–369, 1995.
33. Bowie, J.U., Luethy, R., Eisenberg, D. A method to identify protein sequences that fold into a known three dimensional structure. Science 253:164–170, 1991.
34. de Gennes, P.G. Scaling concepts in polymer physics. Ithaca, New York: Cornell University Press, 1979:46–53.
35. Kolinski, A., Skolnick, J., Yaris, R. Dynamic Monte Carlo study of the conformational properties of long flexible polymers. Macromolecules 20:438–440, 1987.
36. Godzik, A., Kolinski, A., Skolnick, J. Lattice representation of globular proteins: How good are they? J. Comput. Chem. 14:1194–1202, 1993.
37. Kolinski, A., Milik, M., Rycombel, J., Skolnick, J. A reduced model of short range interactions in polypeptide chains. J. Chem. Phys. 103:4312–4323, 1995.
38. Kolinski, A., Skolnick, J. "Parameters of Statistical Potential." Available by anonymous ftp from public directory: scripps.edu (pub/andr/MCSP). 1995.
39. Kyte, J., Doolittle, R.F. A simple method for displaying the hydrophatic character of protein. J. Mol. Biol. 157:105–132, 1982.
40. Gronenborn, A., Filpula, D.R., Essig, N.Z., Achari, A., Whitlow, M., Wingfield, P.T., Clore, G.M. A novel, highly stable fold of the immunoglobulin binding domain of streptococcal protein G. Science 253:657–660, 1991.
41. Gouda, H., Torigoe, H., Saito, A., Sato, M., Arata, Y., Schimada, I. Three-dimensional solution structure of the B-domain of staphylococcal protein A: Comparisons of the solution and crystal structures. Biochemistry 40:9665–9672, 1992.
42. PDB Quarterly Newsletter, No. 71, January 1995.
43. Wilmot, C.M., Thornton, J.M. Analysis and prediction of

the different types of β-turn in proteins. J. Mol. Biol. 203:221–232, 1988.
44. Privalov, P.L., Gill, S.J. Stability of protein structure and hydrophobic interaction. Adv. Prot. Chem. 39:191–235, 1988.
45. Go, N., Abe, A. The consistency principle in protein structure and pathways of protein folding. Adv. Biophys. 18:149–184, 1984.
46. Argos, P. Analysis of sequence-similar pentapeptides in unrelated protein tertiary structures. J. Mol. Biol. 197:331–348, 1987.
47. Kabsch, W., Sander, C. On the use of sequence homologies to predict protein structure: Identical pentapeptides can have completely different conformations. Proc. Natl. Acad. Sci. USA 81:1075–1078, 1984.
48. Chothia, C., Finkelstein, A. The classification and origins of protein folding patterns. Annu. Rev. Biochem. 59:1007–1039, 1990.
49. Goebel, V., Sanders, C., Schneider, R., Valencia, A. Correlated mutations and residue contacts in proteins. Proteins 18:309–317, 1994.