AN ALGORITHM FOR PREDICTION OF STRUCTURAL ELEMENTS IN SMALL PROTEINS

Andrzej Kolinski[1,2], Jeffrey Skolnick[1] and Adam Godzik[1]

[1]The Scripps Research Institute
Department of Molecular Biology
10666 North Torrey Pines Road, MB1
La Jolla, California 92037 USA
(619) 554-8297

[2]Department of Chemistry
University of Warsaw
Pasteura 1, 02-093 Warsaw, Poland
Email: kolinski@chem.uw.edu.pl

Abstract

A method for predicting the location of surface loops/turns and assigning the intervening secondary structure of the transglobular linkers in small, single domain globular proteins has been developed. Application to a set of 10 proteins of known structure indicates a high level of accuracy. The secondary structure assignment in the center of transglobular connections is correct in more than 85% of the cases. A similar error rate is found for loops. Since more global information about the fold is provided, it is complementary to standard secondary structure prediction approaches. Consequently, it may be useful in early stages of tertiary structure prediction when establishment of the structural class and possible folding topologies is of interest.

1: Introduction

A simplified picture of a small single domain and monomeric globular protein could be summarized as follows. The polypeptide chain starts near the surface of a sphere confining the globule, passes several times throughout the interior of the globule and ends somewhere near its surface. The "transglobular linkers" almost always have a well-defined dominant secondary structure. It is either helical or expanded. In the last case, it would be most likely a part of β-sheet. This is very much in the spirit of Richardson [1] topological diagrams of the native structures.

Certainly, the above picture is in many cases oversimplified [2-5]. Nevertheless, it provides some important limitations for possible folds of small polypeptides. In this work, we use this model for construction of a very simple method for the prediction of surface loops (or turns), where the polypeptide chain changes its direction and the dominant secondary structure of the intervening transglobular linkers. Thus, information that is midway between the standard (one dimensional) secondary structure prediction [6-8] and full native structure prediction [9-17] could be obtained with relatively high accuracy.

2: Method

2.1: Monte Carlo scheme

In order to estimate the best location of the surface loops/turns, the protein sequence of interest is randomly divided into several partially overlapping sequence fragments. Then, for each sequence fragment, a structural template is assigned by random selection from a library of structural templates constructed using a database of known protein structures. Each structural template is comprised of two successive protein building blocks which may be viewed as generalized (all α, all β, or mixed motif) hairpins. These structural templates are devoid of any sequence information and are used to provide a library of "protein-like" structures onto which the sequence of interest is inserted. Having divided the protein into sequence fragments, each structural fragment, now with assigned sequence information, is oriented with respect to the center of the 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.

The algorithm is based on random Monte Carlo optimization of the division of the test sequence into structural fragments. The procedure could be outlined as follows. In the beginning, one needs to estimate the size of the globule, S_0 (a single domain with a single hydrophobic core is assumed), and the plausible range for the number of building blocks (linkers) N, with $N_{min} \le N \le N_{max}$ (if the algorithm selects one of the limit values of N, then the computation should be repeated with a different range of N). The number of blocks, N, superimposes limitations on the block size that corresponds to the shortest β-type (expanded) fragment that has the end-to end distance not shorter than 1.8 S_0 and the longest helical fragment of maximum size of 2.5 S_0.

The Monte Carlo iterative procedure consists of the following cycle:

1. Modification by a random shift (by one residue at a single division point) of the original division into N new sequence fragments.
2. Selection 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 "check-points" near its surface. The distance between the ends of hairpin has to be not larger than S_0. 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.
3. Selection of the lowest energy set of templates from many (on the order of 10^4) cycles consisting of operations 1-2. The hairpins partially overlap along the sequence; however, structurally they are bound only via the requirements of the surface positioning of the top of the hairpin and its two

ends. The corresponding division of the chain is stored for the next iteration.

After many iterations, a clustering of the sample is done in order to define the most probable number of secondary structure elements for the sequence of interest. At the same time, the secondary structure assignment of the linkers and the most probable locations of the surface turns and their distribution (range of uncertainty) could be established. Simple geometrical (and local) criteria are applied for secondary structure assignment; additional details are provided below in the sections describing the interaction scheme used by the Monte Carlo procedure.

The idea of the proposed procedure is depicted in Fig. 1. The arrows in Fig. 1A indicate the Monte Carlo selected division of the test sequence into blocks that are successively threaded through structural templates. The hairpin structural template composed of two blocks (Fig. 1B) (randomly selected from the structural database) is used for computation of the short range interactions (Fig.1C); however, the two blocks are "decoupled", and each is treated separately in the burial energy calculations (Fig.1C). This is done for several reasons. A hairpin provides a more physical environment for 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 strand (helical or expanded) within the protein sphere rather than the entire hairpin. Since most hairpins come from different size structures, they will not fit within a given sphere (some proteins in the database are simply larger due to larger chain length). This technique maintains the surface location of the loop 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 transglobular connectors, the detailed geometry of the hairpin is lost.

2.2: Interaction scheme

For simplicity and speed of computation, we use a library of structures that are projected onto a high coordination lattice [16], which represents the Cα reduced backbones of PDB [18,19] structures with an RMS (coordinate root-mean-square deviation) in the range 0.6-0.7 Å. 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 [16], defined by two consecutive Cα-Cα vectors. The force field for this lattice model was developed previously [16]; here, the relevant subset of interactions is employed. Both the short range interactions and the burial energy are based on the statistical correlations seen in a database of protein structures. The numerical data for the statistical potentials have been previously published; they are available upon request from the authors or are easily accessible via anonymous ftp [20], 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, nor there is any noticeable sequence homology to any protein from the database.

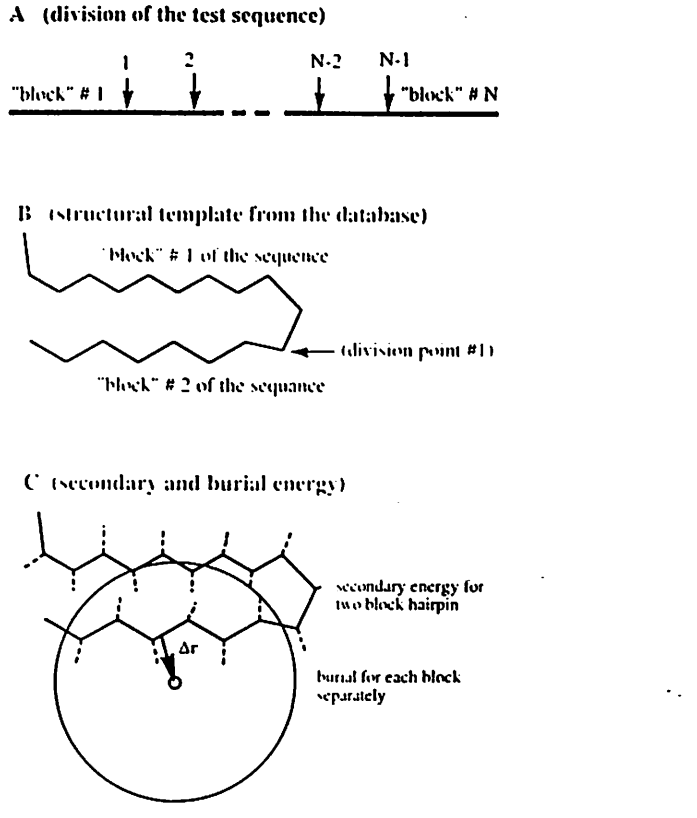

Figure 1. Schematic representation of the method employed in the present work. In (A), the test protein sequence is divided onto N fragments. N is a variable that changes over a narrow range of "reasonable" numbers of secondary structure elements for a protein of a given size. Two sequence fragments are then matched to a "hairpin building block" from the structural data base (B). 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. A hairpin (and the matching test sequence fragment) is used for secondary structure propensity calculations (C). Single blocks (C) are used for burial energy calculations. First the average 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 Δr vector is assumed to be equal $S_0/2$. 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.

The short range interactions are factorized in the form of Cα backbone correlations and correlations between the side chain vectors. The detailed description of this factorization of the secondary propensities could be found elsewhere [21]; here, a brief description is given for reader convenience. The Cα based, three-vector secondary structural propensities depend on the identity of the two amino acids at the appropriate positions along the protein sequence. The local polypeptide conformation is defined by three consecutive reduced backbone vectors: $\mathbf{v}_{i-1}$, $\mathbf{v}_i$, and $\mathbf{v}_{i+1}$, where $\mathbf{v}_i$ is the vector from the i-th to i+1th Cα.

$$E_S = \Sigma\, \varepsilon(A_i, A_{i+1}, r^{2*}_{i-1,i+2}) \qquad (1)$$
$$\text{with:} \quad r^{2*} = \text{sign}((\mathbf{v}_{i-1} \otimes \mathbf{v}_i)\, \mathbf{v}_{i+1})\, r^2_{i-1,i+2}$$

where: 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 $C\alpha_{i-1}$ to $C\alpha_{i+2}$. $r^{2*}_{i-1,i+2}$ is the "chiral" square of distance between the corresponding chain vertices. "Chiral" means a negative sign for the left handed conformations and a positive sign for the right handed ones, respectively. The potential is used in the form of histogram, with the r^{2*} parameter divided into 6 bins. There is an additional decrease in the system's energy when the overlapping "arms" of consecutive hairpins have the same secondary structure (-0.5kT for each occasion when r^{2*} bin is the same). Somewhat longer range conformational correlations are accounted for via angular correlations between side group vectors (up to the fourth neighbor down the chain). There are four contributions to this part of the short range interactions [16]; each one is specific to the appropriate pair of amino acids. The most probable (over the structural database) rotamers are used for the computations.

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

where $\Theta_{i,j}$ is the angle between the side group vectors (vectors from the Cα to center of mass of the current rotamer) of residues i and j. The potential is also in the form of a histogram with an angular bin of 36 degrees and a range of 0 to 180 degrees.

The burial energy approximation employed here is consistent with the assumed spherical model of a single domain protein. The centrosymmetric potential [16] 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.

$$E_1 = \Sigma \varepsilon_1(R(A_i)/S_0) \qquad (3)$$

$$\text{with:} \quad S_0 = 2.2\, m^{0.38} \qquad \text{(in Angstroms)} \qquad (4)$$

where S_0 is the expected radius of gyration of a single domain protein consisting of m amino acids in their native conformation. $R(A_i)$ 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. The numerical values could be found elsewhere and are available via anonymous ftp [20].

Some explanation is required for the procedure that positions the center of the hypothetical domain with respect to the particular building blocks. First, the orientation of the hydrophobic face of a strand is determined from:

$$\mathbf{f} = \Sigma(\varepsilon_{K\text{-}D}(i) \cdot \mathbf{g}/|\mathbf{g}|) \qquad \text{for } r_B < S_0 \qquad (5)$$

where $\mathbf{g}$ is the side chain vector, $\mathbf{r}_B$ is the vector from the center of mass of the building block to the side chain of interest, and $\varepsilon_{K\text{-}D}$ are the Kyte-Doolittle [22] hydrophobicity parameters. Due to the different reference state used here, the K-D parameters have been divided by factor of 5. Next, there is a correction for the burial energy of loop residues, i.e., 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:

$$E_{pattern} = \begin{cases} \Sigma(\varepsilon_{K\text{-}D}(i)) & \text{for } r_B > S_0 \\ -\Sigma(\varepsilon_{K\text{-}D}(i))\,(\varepsilon_{K\text{-}D}(i+2)) & \text{for } r_B < S_0, \text{ and } n<n^* \\ \Sigma(\varepsilon_{K\text{-}D}(i))\,(\varepsilon_{K\text{-}D}(i+2)) & \text{for } r_B < S_0, \text{ and } n>n^* \end{cases} \qquad (6)$$

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 ($n_{max}=2.5\ S_0/1.5$) corresponds to the longest helix that fits into the globule, and the smallest value of n ($n_{min}=1.8\ S_0/3.4$) corresponds to β-strands that just cover the hydrophobic core diameter within the globule. The numbers 1.5 and 3.4 correspond to the approximate extension per residue (in Angstroms) of an α helix and β-strand, respectively.

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 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 globule is located at a distance $\Delta r = (\mathbf{f}/(|\mathbf{f}|) \cdot S_0/2$ from the center of mass of the building block (see the Fig. 1C). At this point, the fragment is properly placed, and the centrosymmetric burial potential can be computed for all side groups. An additional and rather important contribution comes from the face separation term, and is given by $|\mathbf{f}|$, defined in eq. 5; see below.

The total energy of the hairpin fragment can then be expressed as the sum of single contributions from the short range interactions (comprising the hairpin) and the 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.

$$E_{hairpin} = E_S + E_{sg\text{-}local} + (E_1 + |\mathbf{f}| + E_{pattern})_1 + (E_1 + |\mathbf{f}| + E_{pattern})_2 \qquad (7)$$

The |f| contribution reflects the strength of the orientational separation of the hydrophilic and hydrophobic side groups. 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 (as defined above) is selected in the way that both contributions in the lowest energy assemblies are of the same range. This requires a scaling of about 1:2 to 1:4 depending on sequence.

3: Results and discussion

3:1: Structure of the results

Each simulation provides a set of 200 "structures" that based on the energy described in eq. 7 are well suited for the test sequence. These structures consist of sets of overlapping (along the sequence of amino acids) hairpins. The lowest energy structures are selected from this set (those with energies less than 1.05 times the average energy), providing relevant information about the fold and the secondary structure of the blocks connecting the loops. In most cases, the algorithm selects a single most probable set of division points, and consequently, it provides the expected number of secondary structure elements, and the corresponding number of surface loops/turns in the protein fold. The location of the division points between blocks usually exhibits a very narrow distribution. The peaks of this distribution indicate the external loop/turn regions. The number of counts in the histogram representing these loop distributions depends on the number of the lowest energy states selected in a particular run and therefore on the energetical selectivity or the width of the energy distribution of the implicitly "assembled" chains. Thus, we assumed that the loop/turn length corresponds to the width of the peaks at the level of half of their average height. For each transglobule strand, its secondary structure could easily be extracted from the average local geometry of the backbone of the building blocks. The criterion employed for assignment is based on the values of the $r^2_{i-1,i+2}$ vectors (see eq. 1) of particular fragments of structural building blocks, and they are as follows:

(H) helix	$0 < r^{2*}_{i-1,i+2} < 37$		
(T) turn	$r^2_{i-1,i+2} < 60$	and not a helix	(8)
(E) extended	$r^2_{i-1,i+2} > 74$		
(-) coil	otherwise.		

All distances are in $Å^2$.

Alternatively, one could use the straightforward Kabsch-Sander [23] method to assign secondary structure. However, because the 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. Thus, we report the results of geometry based assignments. Moreover, due to dual character of the model with respect to both short range interactions and burial energy, the secondary structure assignments in the predicted loop regions are very ambiguous. Due to the structure of secondary propensities' factorization, the short range

interactions at the ends of hairpins are poorly defined. The loop structures in the hairpins are just used to extrapolate the secondary structural propensities of each single transglobule connection. This does not mean that the method does not provide dependable secondary structure information. In fact, the predicted geometrical information is quite rich. The secondary structural assignment in the middle portion of the particular "transglobular" building block is of highest accuracy (due to above mentioned structure of the model); indeed, such regions are perhaps the most important from the point of view of model building. Thus, we will use this output to estimate the accuracy of our method. More precisely, the leading secondary structure assignment for the three central residues (assigned according to rules from eq. 8) in each transglobular linker is used as the assignment for the entire linker.

In addition to predicting the secondary structure of the central region of the transglobule blocks, the algorithm also provides a substantial amount of intermediate distance information. First of all, the algorithm predicts the number of transglobule linkers, or secondary structure building blocks, and for all the secondary structure elements, their end-to-end distances are predicted to be close to $2S_0$. Furthermore, the "important" surface loops/turns, where the chain changes its direction, are predicted with good accuracy. 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. 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. Then, the algorithm provides two alternative structural assignments. Thus, a broad diversity of information is provided that can aid in subsequent tertiary structure prediction.

Finally, it should be pointed out that the present method converges very fast. For example, a reasonable first estimate could be obtained from a run that explores only a very small sample of building blocks (in the range of 10^3) per single connection for each of 200 structures generated. Subsequent 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 the seed number for the Monte Carlo process. Moreover, the use of half of the structural template database (instead of the full 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 (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 (any helical hairpin that fits the size and helical secondary propensities of a test protein will score well in the algorithm, regardless of specific realization), 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.

3.2: Application to known protein structures

The method is applied to ten test proteins. All are rather small proteins that are expected to have a single domain structure. With the exception of the B1 domain of protein G and the B domain of protein A [24], all the test proteins were chosen randomly from a set of coordinate entries made available by the Brookhaven Protein Data Bank, PDB [19], in early 95. Thus, we expect that they are representative of small globular, single domain proteins. None is a member of the structural database used for derivation of the statistical potentials employed here.

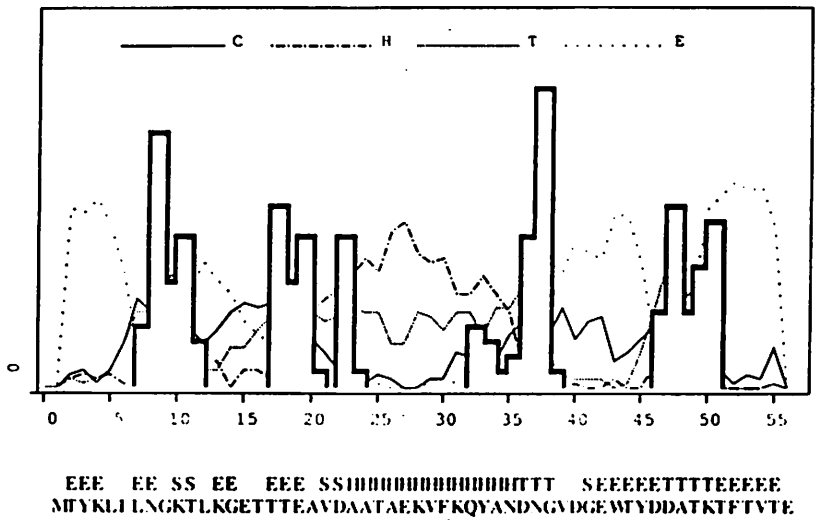

Figure 2. The loop distribution for the sample of 200 lowest energy structures assigned to the sequence of domain B1 of *Streptococcus* protein G. The secondary structure prediction (the most frequent structures seen in the building block, based on geometrical criteria for the Cα backbone fragments) is given for comparison.

Typical results from the calculation for domain B1 of *Streptococcus* protein G [25] are depicted in Fig. 2, where the distribution of division points is given in the form of a histogram. For easy reference, Fig. 3 contains a schematic drawing of the native backbone of the protein with the predicted surface turns/loops indicated in black. As one may see, the prediction is very good indeed. For this 56 residue sequence, the leading division selected by the algorithm is N=5. All of the 200 lowest energy "structures" consists of 5 fragments, i.e., it is built from 4 overlapping hairpins. Fig. 2 shows the distribution of division points that correspond to the locations of the external loops/turns. We remind the reader that the term loop/turn refers not only to local conformations, but also means that the polypeptide chain changes direction. This is clearly demonstrated in the schematic

drawing of the native conformation. Thus, the prediction clearly indicates the existence of five structural elements, which are connected via well-localized loops.

Fig. 2 also presents the secondary assignment (according to geometrical criterion defined by eq. 8) given by the algorithm, based on the most frequently observed structure of the selected building blocks for the lowest energy sets of building blocks (below 1.05 times the average score for the 200 lowest energy samples), and the secondary structure assignment for the native state is indicated below the plots. The agreement is rather good, especially when one compares the predictions of the loop/turn distribution histogram. Furthermore, some filtration of the results could easily be done. The assigned structure of the protein is ββαββ, and the biggest errors involve the shift in the location of two turns by a couple of residues along the sequence. The slight overprediction of the E (expanded) states actually takes into account the rather broad loops located next to the central helix. These discrepancies could be easily removed at the stage of three-dimensional model building [26].

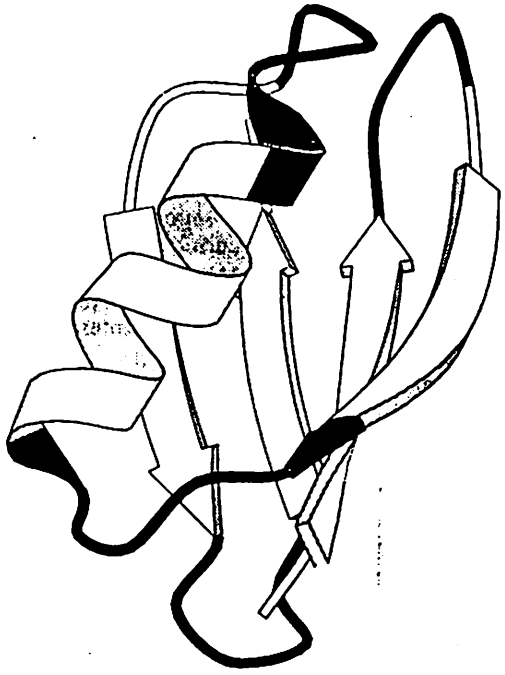

Figure 3. The predicted locations of the turns are shown on schematic drawings of the PDB 1gb1 structure.

At this point, it seems worthwhile to note that the interplay of the short and long interactions is responsible for the relatively high accuracy of the proposed method. The prediction is much worse in either case of just secondary interactions (with the burial contribution neglected) or using the burial energy alone.

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 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 TABLE I, the predictions of the external loops/turns are compiled for the test sequences. In general, more regular structures, especially with narrow surface turns (or loops) are predicted with higher accuracy. That is certainly the effect of the assumed simplified view of a globular protein. Let us comment here on some interesting cases. First, let us note that in majority of all cases, the secondary structure of the central fragments of the building blocks (extended or helical, roughly speaking) is correctly assigned except for the clear qualitative error in the third fragment of the 1ris sequence. This fragment 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 propensities 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 it could be perhaps "pulled-out" by the neighboring well-defined fragments into an extended state. On other hand, even in this case, the number of secondary structure elements and location of the surface turns/loops have been correctly predicted. Perhaps for such a strongly energetically frustrated sequence fragment, there is no way to predict correctly its secondary structure without invoking more detailed tertiary interaction related effects, such as hydrogen bonding and pair interactions. Clearly, this problem requires further investigation.

Another interesting example is the 1pou sequence. The Kabsch-Sander [23] 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. Of course, there is no monomeric protein with a single β-strand, and therefore, it has to be interpreted as an 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 native structure shows that this is exactly the case.

Probably, the largest errors in loop prediction occur in less regular helical proteins. Here the algorithm sometimes selects exposed fragments of helices near the helix end as a surface loop/turn. That is the situation seen in 1pou and more dramatically in 1lpt. It is interesting to notice that the "overpredicted" beta hairpin at the C-terminus of 1lpt reflects geometrical shape of this "irregular" fragment of the native structure.

TABLE I. Comparison of the native (N) secondary structure with predicted (P) structure.*)

1gb1 56 residues (Streptococcus protein G domain B1)

N:beta-(8-13)-beta-(22-22)-helix-(36-41)-beta-(47-50)-beta
P:beta-(9-11)-beta-(18-23)-helix-(37-38)-beta-(48-51)-beta

proA 46 res. (fragment domain B of protein A, residues 9-54)

N:helix- (9-15)-helix-(29-32)-helix
P:helix-(16-17)-helix-(33-34)-helix

1fas 61 residues (Fasciculin)

N:beta-(5-13)-beta-(17-21)-beta-(28-33)-beta-(40-47)-beta-(54-61)
P:beta-(8-13)-beta-(20-22)-beta-(28-32)-beta-(47-49)-beta-(55-56)-
P:beta

1pou 71 residues (pou-specific domain, residues 5-75)

N:helix-(20-24) -helix-(34-40)-helix-(48-55)-helix
P:helix-(16-17)-beta-(24-24) -helix-(36-37)-helix-(51-53)-helix

1tlk 103 residues (telokin, residues 33-135)

N:beta-(15-18)-beta-(23-26)-beta-(37-39)-beta-(46-47)-beta-(50-55)
P:beta-(13-16)-beta-(24-28)-beta-(35-39)- beta- (52-55)
N:beta-(61-64)-beta-(71-74)-helix-(78-78)-beta-(88-89)-beta
P:beta-(63-66)-beta- (77-80) beta-(92-92)-beta

1ris 97 residues (ribosomal protein S6)

N:beta-(11-15)-helix-(33-35)- beta-(53-54)-beta-(68-68)-helix-
P:beta-(13-17)-helix-(35-38)-helix-(56-59)-beta-(67-70)-helix-
N:(80-84)-beta
P:(83-88)-beta

1lpt 90 residues (wheat lipid transfer protein)

N: helix-(18-29)-helix-(38-40)-helix-(48-49)-helix-(58-62)-helix-
P: helix-(19-22)-helix- (44-46)-helix-
N: (67-90) turns-coil
P: (68-69)-beta-(77-81)-beta

1ten 89 res (fibronectin repeat of tenascin, residues 803-891)

N:beta-(11-16)-beta-(23-29)-beta-(38-44)-beta-(51-54)-beta-(59-65)
P:beta-(12-16)-beta-(30-31)-beta-(38-40)-beta-(54-56)-beta-(65-67)
N:beta-(76-82)-beta
P:beta-(75-76)-beta

1trl 62 res. (thermolysin fragment, residues 255-316)

```
N:        (1-8)-helix-(21-26)-helix-(43-46)-helix
P:  beta- (8-9)-helix-(25-28)-helix-(43-45)-helix
```

1mjc 69 residues (major cold shock protein 7.4, residues 2-70)

```
N:beta-(13-16)-beta-(23-28)-beta- helix- coil-(37-48)-beta-(56-61)
P:beta-(13-16)-beta-(27-29)-beta-(35-40)-beta-(49-50)-beta-(56-57)
N: beta
P: beta
```

---

*)The numbers in brackets give the ranges of loops/turns fragments according to Kabsch-Sander assignment for (N) and the surface loops/turns predicted by the present algorithm. For (P) case the secondary structure of transglobular linkers is classified as beta for linkers with dominated expanded states, and helix for leading helical assignments.

Another case is the 1mjc sequence analysis, see TABLE I, where residues 41-48 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, except for the lack of hydrogen bonded partners. This is another illustration of the kind of structural information the present method provides. 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 would perhaps need to consider several alternative topologies.

Finally, let us note that for small, disulfide crosslinked proteins the structure could, to a large extent, be enforced by the pattern of S-S bridges. The information about the crosslinks may be employed by introduction of implicit mixing rules for the building blocks (hairpins). This could perhaps further increase the accuracy of the method. For the sake of clarity, however, we defer from analyzing this possibility in the present work.

4: Conclusion

In this work, using just sequence information, for small, single domain globular proteins, we have developed a method that allows for the quite accurate prediction of the location of surface loops/turns (loop) and the dominant type of secondary structures (sec) of the transglobule blocks that join such loops or turns. For a given sequence, the Monte Carlo algorithm generates structure assignments in the form (sec_1-loop-sec_2-loop....sec_N), with N determined during the course of the optimization procedure. For 10 small test proteins, there is just a single case where a secondary structure fragment is incorrectly classified. In all cases, the surface loops (or turns) that are characterized by a change of direction of the polypeptide chain are also quite accurately predicted.

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 the two kinds of terms. This is suggestive that proteins, on the average, need not necessarily satisfy the principle of minimal frustration for a given type of interaction. Thus, burial preferences which state that all hydrophobic side groups 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, intrinsic secondary preferences cannot always be satisfied. This is evidenced by the presence of pentapeptide fragments in more than one type of secondary structure. It should be noted that the accuracy of the present method could be perhaps further improved when combined with recent developments in prediction of protein structural classes [27].

The ultimate significance of the present method for protein modeling needs to be established; however, two points seem clear. First, the method quite accurately predicts the location of surface loops/turns, and therefore provides important complementary information for various 3D protein modeling procedures. Furthermore, for small proteins of rather regular secondary structure, the present method provides sufficient information to propose a few (sometimes just one) low resolution alternative folds that could be further refined by various techniques. Thus, it offers a new (albeit limited) path towards solving the protein folding problem. The method provides self-consistent global information about the character of the fold, and 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. This possibility is now being explored.

Acknowledgments

This work was supported in part by NIH Grant No. GM-37408 and University of Warsaw Grant BST-502/34/95.

References

1. J. S. Richardson, *Nature* **268**, 495-500 (1977)
2. C. B. Anfinsen, *Science* **181**, 223-230 (1973)
3. J. S. Richardson, D. C. Richardson, N. B. Tweedy, K. M. Gernert, T. B. Quinn,M.H. Hecht, B. W. Ericson, Y. Yan, R. D. McClain, M. E. Donlan. *Biophysical J.* **63**, 1185-1209 (1992)
4. O. B. Ptitsyn, *Journal of Protein Chemistry* **6**, 273-293 (1987)
5. R. L. Jernigan, *Current Opinion in Structural Biology* **2**, 248-256 (1992)
6. X. Zhang, J. P. Mesirov, D. L. Waltz, *J. Mol. Biol.* **225**, 1049-1063 (1992)
7. B. Rost, R. Schneider, C. Sander, *TIBS* **18**, (1993)
8. B. Rost, C. Sander, *Proteins* **19**, 55-72 (1994)
9. J. M. Thornton, T. P. Flores, D. T. Jones, M. B. Swindells, *Nature* **354** 105-106 (1991)
10. M. Levitt, *Current Opinion in Structural Biology* **1**, 224-229 (1991)
11. M. Karplus, E. Shakhnovich, Protein folding: *Theoretical studies of thermodynamics and dynamics, in Protein Folding*, T.E.Creighton, Ed., (W.H. Freeman, 1992)

12. A. Godzik, J. Skolnick ,A. Kolinski, *J. Mol. Biol* **227**, 227-238 (1992)
13. J. Skolnick, A. Kolinski, C. Brooks III, A. Godzik, A. Rey, *Curr. Biol.* **3**, 414-423 (1993)
14. A. Godzik, A. Kolinski, J. Skolnick, *J.Comp. Aided Mol. Des.* **7**, 397-438 (1993)
15. A. Kolinski, A. Godzik, J. Skolnick, *J. Chem. Phys.* **98**, 7420-7433 (1993)
16. A. Kolinski, J. Skolnick, *Proteins* **18**, 338-352 (1994)
17. A. Kolinski, J. Skolnick, *Proteins* **18**, 353-366 (1994)
18. F. C. Bernstein, T. F. Koetzle, G. J. B. Williams, E. F. Meyer Jr, M. D. Brice, J. R. Rodgers, O. Kennard, T. Simanouchi, M. Tasumi, *J.Mol.Biol.* **112**, 535-542 (1977)
19. PDB, Quaterly Newsletter, **17**, January 1995
20. A. Kolinski, J. Skolnick, *Parameters of statistical potential.* Avaiable by ftp from public directory: scripps.edu (pub/andr/MCSP), 1995
21. A. Kolinski, W. Galazka, J. Skolnick, *J. Chem. Phys.* in press (1995)
22. J. Kyte, R. Doolittle, *J. Mol. Biol.* **157**, 105-132 (1982)
23. W. Kabsch, C. Sander, *Biopolymers* **22**, 2577-2637 (1983)
24. H. Gouda, H. Torigoe, A. Saito, M. Sato, Y. Arata, I. Schimada, *Biochemistry* **40**, 9665-9672 (1992)
25. A. Gronenborn, D. R. Filpula, N. Z. Essig, A. Achari, M. Whitlow, P. T. Wingfield ,G. M. Clore, *Science* **253**, 657-660 (1991)
26. M. S. Johnson, N. Srinivasan, R. Sowdhamini, T. L. Blundell, *Crit. Rev. Biochem. & Mol.Biol.* **29**, 1-68 (1994)
27. K.-C. Chou, *Proteins* **21**, 319-344 (1995)

Sequence Sizes of Eukaryotic Enzymes

Eugene Kolker and Edward N. Trifonov
Department of Structural Biology, The Weizmann Institute of Science, Reh
76100, Israel

Abstract

We have shown in earlier studies that an appreciable fraction of proteins disp sequence size periodicity with periods of ≈123 aa and ≈152 aa for eukaryotes prokaryotes, respectively. For any firm conclusions to be made, the issue possible bias due to an overabundance of some protein families should addressed in more than one way. Here we present the size distributions for vari sequence ensembles of eukaryotic enzymes that differ by level of data b cleaning. The sequences were purged by applying several successive threshold relatedness irrespective of the sequence lengths. The previously obser preference to typical sizes is confirmed. Possible reasons for the observed exces the typical size sequences are discussed.

Key words: eukaryotic enzymes, sequence length, typical size, cleani segments.

Introduction

In 1929 Svedberg [1] suggested that proteins could have standard molecu masses, multiples of a certain unit mass. Since then, this simple and attractive i has been tested several times by different techniques on different data sets, al two main parallel lines: i) analysis of protein structures, and ii) analysis of prim sequences. When the structure of chymotrypsin was solved it was suggested th sufficiently long polypeptide chain might be "piled on itself" or "folded arou nuclei of highly stabilized local conformation" [2], thus making two distinct foldi domains of that protein [3]. According to Matthews *et al.* [4], long polypepti chains could be considered to be a combination of smaller independently folde