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

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. 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 and full native structure prediction could be obtained with relatively high accuracy.

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 division points (or rather distribution peaks of the division) indicate surface loops/turns, where the polypeptide chain changes its direction. Prediction of the number of transglobular connections and their secondary structure assignment is obtained at the same time. The force field used here is a subset of the force field employed in our previous study of the lattice models of proteins [2-5]. Here we also use a high coordination lattice representation of the structural templates, however the method is rather general and different representations can be easily implemented. The short range interactions have two components [5]. The first one controls local geometry of the $C\alpha$ backbone and depends on identity of two subsequent amino acids. In similar way preferred mutual orientations of close along the chain (up to the forth neighbor) side chains are encoded. The burial energy is approximated by amino acid specific centrosymmetric force [2,5], pattern of hydrophilic and hydrophobic residues and their orientation in respect to the center of the hydrophobic core. These potentials of mean force were derived from statistics of the high resolution PDB structures.

For a given sequence, the Monte Carlo algorithm generates structure assignments in the form (sec_1-loop-sec_2-loop....sec_N) where -sec_i- denotes the i-th transglobular connection of a specific secondary structure and -loop- denotes the intervening surface loop or turn. The number of transglobular connections N is determined during the course of the optimization procedure. For 10 small test proteins (B1 domain of Streptococcus protein, B domain of protein A, pou-specific domain, telokin, ribosomal protein S6, wheat lipid transfer protein, fibronectin repeat of tenascin, thermolysin fragment 255-316, and major cold shock protein 7.4) 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.

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.

1. J. S. Richardson, *Nature* **268**, 495-500 (1977)
2. A. Kolinski, J. Skolnick, *Proteins* **18**, 338-352 (1994)
3. A. Kolinski, J. Skolnick, *Proteins* **18**, 353-366 (1994)
4. J. Skolnick, A. Kolinski, C. Brooks III, A. Godzik, A. Rey, *Curr. Biol.* **3**,
5. A. Kolinski, W. Galazka, J. Skolnick, *J. Chem. Phys.* in press (1995)