Beta-Sheet Structure Prediction Methods

Fang Fang Yin

The amino acid sequence rules that correspond to beta-sheet structures in proteins are still not well understood. Current protein structure prediction methods are more accurate for alpha-helical structures than for beta-sheet structures. One reason that beta-strand structure prediction is more difficult is because of its high prevalence of nonlocal interactions between regions of the protein chain that are not necessarily consecutive in the amino acid sequence. These long range interactions make it difficult to approximate structures from amino acid sequence information. However, several methods and approaches have been developed for predicting beta-sheet structures.

Typical protein structure prediction methods, including Hidden Markov models, sequence profile searches, and protein threading methods, assign structures to protein sequence using folds of known structures as templates [1]. These methods are limited by the number of known structures which can serve as templates, thus causing strong biases in the predicted results because there are relatively fewer known beta-sheet structures from higher eukaryotes than for prokaryotes. These methods do not work for many folds that have very low sequence homology with each other. For example, beta-helix is characterized by very little regular repeats in its sequence, thus preventing the use of structure templates for structure prediction.

Ab initio methods use physical and knowledge-based information to predict protein structures. Rosetta is a common ab initio protein structure prediction algorithm that is based on protein folding in which local sequence segments sample between different possible local structures, and folding occurs when the conformations and relative orientations of the segments satisfy burial of hydrophobic residues and pairing of beta-strands without steric clashes [2]. The distribution of the sampled structures are approximated by the distribution of the conformations adopted by the sequence segment and related sequence segments in the protein structure database.

However, recent experimental studies have contributed to the improvement of ab initio structure prediction, which is one of the most reliable method for predicting protein structure in the absence of homologue. Experimental studies showed that protein folding rates are correlated with the relative contact order (CO) of the native structure, which is the average sequence separation of residues that form contacts in the 3D structure divided by the length of the protein [3]. Proteins with more local contacts fold more rapidly than proteins with nonlocal contacts. This correlation

between folding rates and CO reflects the contribution of the entropic loss of the protein to the folding free energy barrier. Thus, proteins with low CO have lower folding free energy barriers and can make stabilizing interactions with less entropic loss [3].

Experimental relationships between CO and protein folding rates prompted examination of Rosetta ab initio folding simulations on the native state CO. Bonneau et al. showed that ab initio structure prediction parallels these experimental studies. Experimental studies showed that folding of small proteins is a single exponential process. The probability of folding to a native structure is independent of time for an individual polypeptide chain. Based on this observation, Bonneau et al. showed that many short simulations is a more effective method for structural prediction. Large numbers of independent short simulations can be used to generate structures that can then be clustered to identify the broadest minima in folding. This technique allows one to improve ab initio structure prediction by incorporating information from experimental studies on protein folding. Critical assessment of structure prediction experiments, CASP3 and CASP4, showed that Rosetta is currently one of the best methods for structure prediction in the absence of a homologue of known structure.

Although ab initio structural prediction is a very powerful tool for predicting structures in the absence of a homologue, it has several disadvantages for beta-strand prediction. Because beta-sheet are characterized by predominantly nonlocal interactions, ab initio method's reliance on sampling of local sequence segments makes it difficult to predict beta-strand structures. Also, beta-sheet structures have very low sequence homology, making sequence profile comparisons of sequence segments unamenable to beta-sheet prediction. Distribution of the conformations sampled in this method depends on the distribution of available structures in the current protein structure databases. Because relatively fewer beta-sheet structures are know, this bias in the database does not allow equal distribution of structures sampled in the method, introducing errors in the predicted structure. Another problem with ab initio structure prediction is that it can yield different final structures depending on the starting structure used to begin the simulation.

A recent advance in the method of parallel beta-helix prediction demonstrates the importance of incorporating structural information for protein structure prediction because structures are more conserved than amino acid sequences. This method, BETAWRAP, addresses some of the critical challenges faced in beta-sheet structure prediction [4]. The right-handed parallel beta-helix motif is characterized by a series of coils, each contributing to the three long beta-sheets that come together to form the fold. In this fold, hydrophobic amino acids are buried in

the cylindrical core. This fold is also characterized by stacks of hydrophobic side chains and ladders of hydrogen bonding side chains. The right-handed parallel beta-helix motif is not common, with only 12 known structures in the Protein Data Bank [4]. This fold does not have regular repeats in its amino acid sequence and has very low sequence homology with each other, making it very difficult to predict the fold from multiple sequence alignment methods such as PSI-BLAST and HMMER. Due to the difficulties in predicting beta-helix motif from sequence information, computational methods have been developed to utilize structural information for more accurate predictions of the fold.

Phil Bradley et al. developed a new method, BETAWRAP, for recognizing beta-helices. BETAWRAP utilizes structural information from solved beta-helix structures and mutants defective in the folding of beta-helices, which indicated that strand side chain interactions in the buried core are critical determinants of the fold. These analyses indicated strong statistical preferences for certain amino acids residues in the folded beta-structural motifs. BETAWRAP predicts beta-helix by dynamically assessing an amino acid segment into stacking beta-strands separated by variable and fixed length turns. First, the program identifies possible sequences of the well conserved B2-T2-B3 rung segment by using hydrophobic residue sequence patterns. Each segment is then screened for neighboring rungs that align well. Score for the best alignment incorporates beta-sheet pairwise correlations, turn lengths, stacking preferences, and other structural information. It makes the assumption that the core packing interactions within globular beta-sheets are similar to those of beta-helices. This assumption allows one to utilize alignment correlations from non-beta-helix proteins and avoids over-training on the very limited set of known beta-helix structures. Rung-rung alignment score is then calculated from the weighed sum of seven alignment scores for the aligned pairs in the beta-strands B2 and B3. This score is used to generate wraps of amino acid sequence into the beta-helical structure. The program then searches for strands of B1 sheet in the sequence gaps between B2-T2-B3 segments. The final score is the average of top ten wrap scores.

BETAWRAP is one method that addresses some of the difficulties in beta-sheet protein structure prediction. One of the strengths of this method is the incorporation of experimental data, such as previous studies of mutants defective in the folding of beta-helices. Theories applied in computational methods should be supplemented with experimental data. For example, much can be learned from studies of how amyloidogenic peptides with different amino acid sequences can form similar beta-sheet structures. These peptides serve as good models for understanding the

factors that contribute to beta-sheet formation. Systematic residue substitutions or mutants of wild-type peptides throughout the sequences of these peptides give insight to the residues that are critical for beta-sheet formation. Biochemical studies that trap or allow visualization of amyloid intermediates also contribute to the understanding of coordination and interplay between amino acids, kinetics, and thermodynamics of beta-sheet formation.

Protein folding is a very complex process. To date, most protein structure prediction methods rely on the extraction of information from known structures in databases and statistical methods to find the best alignments against a certain template. Because structures are often more conserved than amino acid sequences, sequence comparison and template-based prediction methods are often inaccurate. The methods are also limited by the number of known structures in the databases. Some structures are under-represented in the databases. Biochemistry and biophysical studies of the dynamics of protein folding are needed to rigorously test the parameters that guide protein folding. Systematic studies of known proteins and their folding dynamics should be incorporated with sequence and structural information in databases in order to improve beta-sheet protein structure prediction methods.

References:

1. Steward, R. and Thornton, J. *Proteins* (2002) 48:178-191.
2. Ruczinski, I. et al. *Proteins* (2002) 48:85-97.
3. Bonneau, R. et al. *Protein Science* (2002) 11:1937-1944.
4. Bradley, P. et al. *PNAS* (2001) 98:14819-14824.