

4-7 Structure from Sequence: Profile-Based Threading and “Rosetta”

Profile-based threading tries to predict the structure of a sequence even if no sequence homologs are known

The most important method that has been developed so far for the identification of a protein fold from sequence information alone in the absence of any apparent sequence identity to any other protein, is the method of “profile-based threading”. In this method, a computer program forces the sequence to adopt every known protein fold in turn, and in each case a scoring function is calculated that measures the suitability of the sequence for that particular fold (Figure 4-25).

The function provides a quantitative measure of how well the sequence fits the fold. The method is based on the assumption that three-dimensional structures of proteins have characteristics that are at least semi-quantitatively predictable and that reflect the physical-chemical properties of strings of amino acids in sequences as well as limitations on the types of interactions allowed within a folded polypeptide chain. Does, for example, forcing the sequence to adopt particular secondary structures and intra-protein interactions place hydrophobic residues on the inside and helix-forming residues in helical segments? If so, the score will be relatively high.

Experience with profile-based threading has shown that a high score, indicating a good fit to a particular fold, can always be trusted. On the other hand, a low score only indicates that a fit was not found; it does not necessarily indicate that the sequence cannot adopt that fold. Thus, if the method fails to find any fold with a significantly high score, nothing has been learned about the sequence. Despite this limitation, profile-based threading is a powerful method that has been able to identify the general fold for many sequences. It cannot provide fine details of the structure, however, because at such low levels of sequence identity to the reference fold the local interactions and side-chain conformations will not necessarily be the same.

The Rosetta method attempts to predict protein structure from sequence without the aid of a homologous sequence or structure

Ideally, one would like to be able to compute the correct structure for any protein from sequence information alone, even in the absence of homology. Ongoing efforts to achieve this “holy grail” of structure prediction have met with mixed success. Periodically these methods are tested against proteins of known but unpublished structures in a formal competition called CASP (critical assessment of techniques for protein structure prediction). Perhaps the most promising at the moment is the Rosetta method. One of the fundamental assumptions underlying Rosetta is that the distribution of conformations sampled for a given short segment of the sequence is reasonably well approximated by the distribution of structures adopted by that sequence and closely related sequences in known protein structures. Fragment libraries for short segments of the chain are extracted from the protein structure database. At no point is knowledge of the overall native structure used to select fragments or fix segments of the structure. The conformational space defined by these fragments is then searched using a Monte Carlo procedure with an energy function that favors compact structures with paired strands and buried hydrophobic residues. A total of 1,000 independent simulations are carried out for each query sequence, and the resulting structures are clustered. One selection method was simply to choose the centers of the largest clusters as the highest-confidence models. These cluster centers are then rank-ordered according to the size of the clusters they represent, with the cluster centers representing the largest clusters being designated as the highest-confidence models. Before clustering, most structures produced by Rosetta are incorrect (that is, good

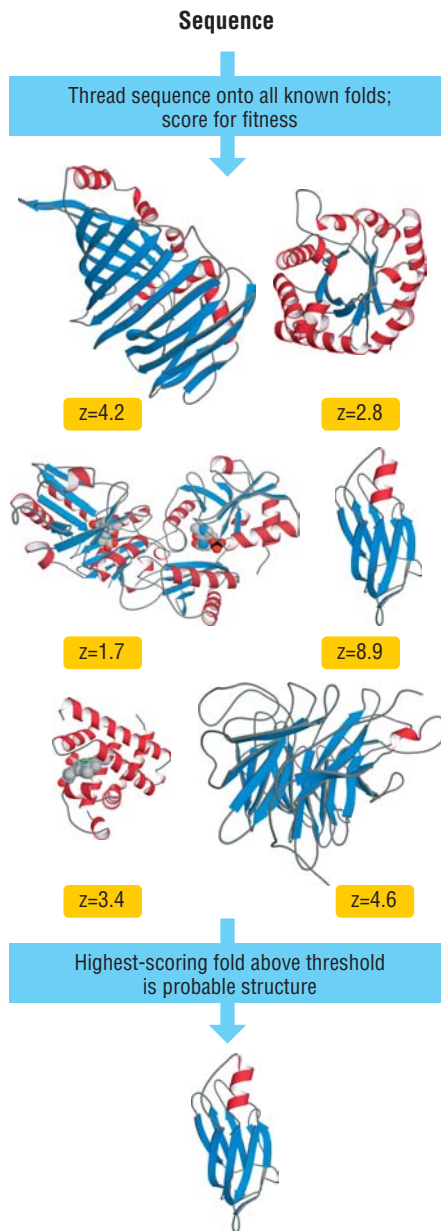


Figure 4-25 The method of profile-based threading A sequence of unknown structure is forced to adopt all known protein domain folds, and scored for its suitability for each fold. The z-value relates the score for the query sequence to the average score for a set of random sequences with the same amino-acid composition and sequence length. A very high z-score indicates that the sequence almost certainly adopts that fold. Sequences can be submitted online for threading by PSIPRED (<http://bioinf.cs.ucl.ac.uk/psipred/index.html>).

References

- Bonneau, R. *et al.*: **Rosetta in CASP4: Progress in *ab initio* protein structure prediction.** *Proteins* 2001, **45(S5)**:119–126.
- Bowie, J.U. *et al.*: **A method to identify protein sequences that fold into a known three-dimensional structure.** *Science* 1991, **253**:164–170.
- de la Cruz, X. and Thornton, J.M.: **Factors limiting the performance of prediction-based fold recognition methods.** *Protein Sci.* 1999, **8**:750–759.

Fischer, D. and Eisenberg, D.: **Protein fold recognition using sequence-derived predictions.** *Protein Sci.* 1996, **5**:947–955.

Miller, R.T. *et al.*: **Protein fold recognition by sequence threading: tools and assessment techniques.** *FASEB J.* 1996, **10**:171–178.

Simons, K.T. *et al.*: **Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions.** *J. Mol. Biol.* 1997, **268**:209–225.

structures account for less than 10% of the conformations produced); for this reason, most conformations generated by Rosetta are referred to as decoys (Figure 4-26). The problem of discriminating between good and bad decoys in Rosetta populations is still under investigation. Still, in some test calculations, the best cluster center has been shown to agree fairly well with the overall fold of the protein (Figure 4-27).

Both the Rosetta method and the method of profile-based threading suffer from some of the same limitations that beset homology modeling. The issue of false positives and negatives is significant, because the failure to generate a model does not mean one cannot be generated, nor that the structure is a novel one. And the generation of a model does not mean it is right, either overall or, more usually, in detail. At best one should look to these methods, at least for the present, for rough indication of fold class and secondary structure topology. And it is important to remember that all methods of model building based on a preexisting structure, whether found by sequence homology or by threading, suffer from massive feedback and bias. The structure obtained will always look like the input structure, because the computational tools for refining the model are unable to generate the kinds of shifts in secondary structure position and local tertiary structure conformations that are likely to exist between two proteins when their overall sequence identity is low (see Figure 4-19). *Ab initio* methods like Rosetta at least do not suffer from this problem, whatever their other limitations.

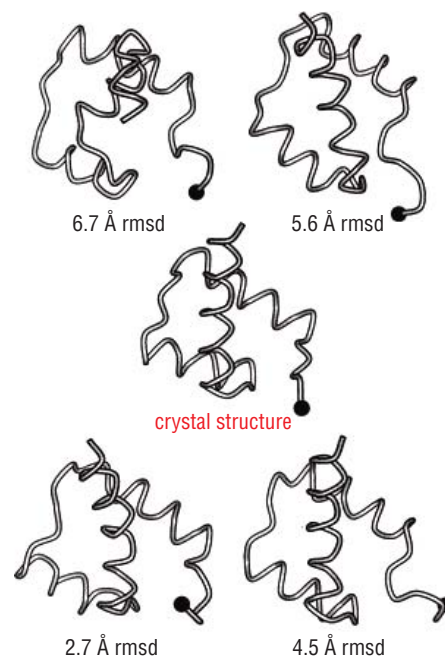


Figure 4-26 Some decoy structures produced by the Rosetta method. The structure at the center is the target, the experimentally determined structure of a homeodomain. The other structures are generated by the Monte Carlo approach in Rosetta, using only the sequence of the protein. Although some of the structures are quite far from the true structure, others are close enough for the fold to be recognizable. Rmsd is the root mean square deviation in α -carbon positions between the computed structure and the experimentally determined structure. (Taken from Simons, K.T. *et al.*: *J. Mol. Biol.* 1997, **268**:209–225.)

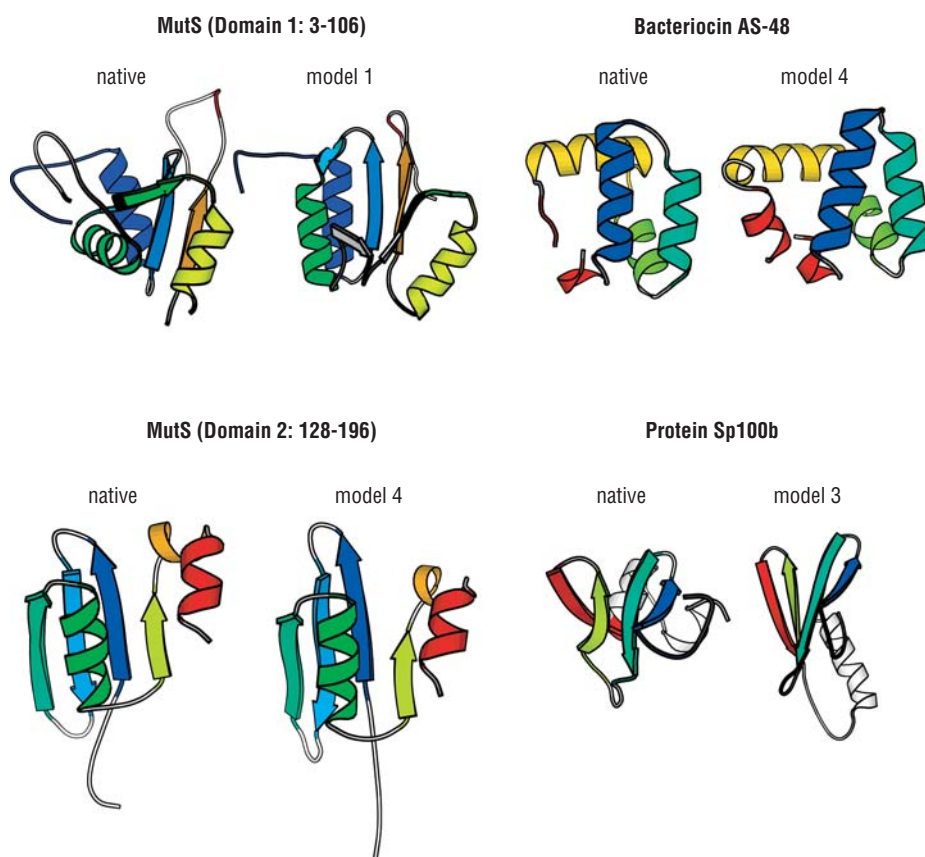


Figure 4-27 Examples of the best-center cluster found by Rosetta for a number of different test proteins. The level of agreement with the known native structure varies, but in many cases the overall fold is predicted well enough to be recognizable. Note, however, that the relative positions of the secondary structure elements are almost always shifted at least somewhat from their true values. Graphics kindly provided by Richard Bonneau and David Baker. (Adapted from Bonneau, R. *et al.*: *Proteins* 2001, **45**(S5):119–126.)

Simons, K.T. *et al.*: **Prospects for *ab initio* protein structural genomics.** *J. Mol. Biol.* 2001, **306**: 1191–1199.

URL for threading website:
<http://bioinf.cs.ucl.ac.uk/psipred/index.html>

URL for CASP:
<http://moult.carb.nist.gov/casp>