Chapter 1
Threading methods for protein structure prediction

David Jones and Caroline Hadley
Department of Biological Sciences, University of Warwick, Coventry CV4 7AL, UK.

1 Introduction

As the attempts to sequence entire genomes increases the number of protein sequences by a factor of two each year, the gap between sequence and structural information stored in public databases is growing rapidly. In stark contrast to sequencing techniques, experimental methods for structure determination are time-consuming, and limited in their application, and therefore will not be able to keep pace with the flood of newly characterized gene products. The development of practical methods for predicting protein structure from sequence is therefore of considerable importance in the field of biology.

Several different approaches have been used to predict protein structure from sequence, with varying degrees of success. *Ab initio* methods encompass any means of calculating co-ordinates for a protein sequence from first principles—that is, without reference to existing protein structures. Little success has been seen in this area, with more theory produced than actual useful methodology. Comparative (or homology) modelling, attempts to predict protein structure on the strength of a protein's sequence similarity to another protein of known structure (following the theory that similar sequence implies similar structure). Some success has been achieved, but several limitations to this method, not least of which are its dependence on alignment quality and the existence of a good sequence homologue, indicate it is not applicable to a large fraction of protein sequences. The third main category of protein structure prediction, falling somewhere between comparative modelling and *ab initio* prediction, is fold recognition, or threading.

2 Threading methods

The term 'threading' was first coined in 1992 by Jones *et al.* (1), but the field has grown considerably since then with many different methods being proposed: for example, Godzik and Skolnick (2); Ouzounis *et al.* (3); Abagyan *et al.* (4); Overington *et al.* (5); Matsuo *et al.* (6); Madej *et al.* (7); Lathrop and Smith (8);

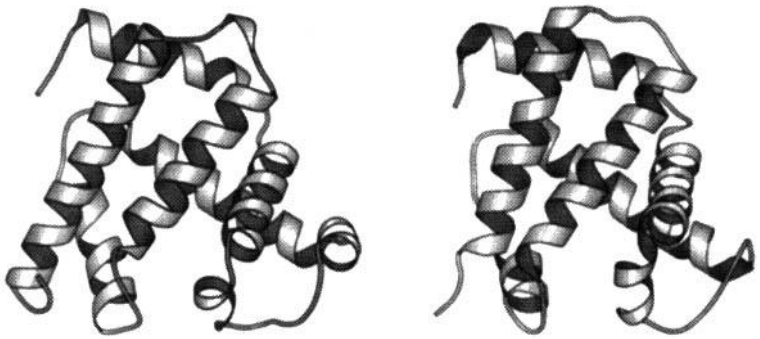

Figure 1 An example of a pair of protein structures in the same family. (a) Human myoglobin [2mm1], (b) pig haemoglobin, alpha chain [2pghA]. At the family level, proteins have higher sequences identity (in this case, 32%) and have highly similar structures. Figures created using Molscript (19).

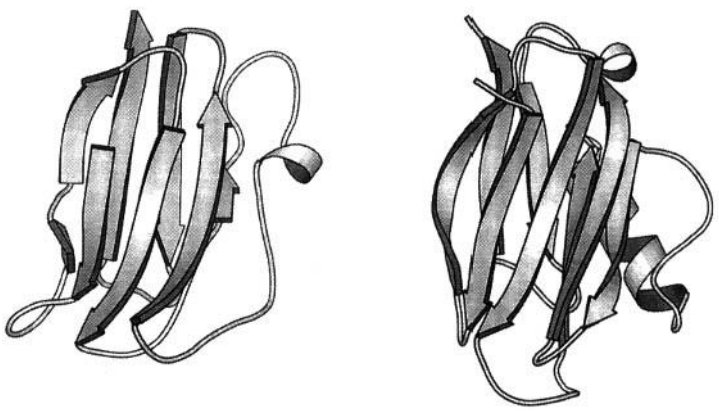

Figure 2 A pair of structures within the same superfamily. (a) *A. denitrificans* azurin [1azcA], (b) poplar plastocyanin [1plc]. Members of the same superfamily may have insignificant sequence identity (16% in this case), but still share most features of the protein fold, reflecting a common evolutionary origin.

Taylor (9) amongst others. The idea behind threading came about from the observation that a large percentage of proteins adopt one of a limited number of folds (*Figures 1–3*). In fact, just 10 different folds (the 'superfolds') account for 50% of the known structural similarities between protein superfamilies (18). Thus, rather than trying to find the correct structure for a protein from the huge number of all possible conformations available to a polypeptide chain, the correct (or close to correct) structure is likely to have already been observed and already stored in a structural database. Of course, in cases where the target protein shares significant sequence similarity to a protein of known 3-D structure, the 'fold recognition' problem is trivial—simple sequence comparison will identify the correct fold. The hope was, however, that threading might be able to detect structural similarities that are not accompanied by any detectable sequence similarity, and this has subsequently been proven to be the case.

Figure 3 A pair of analogous folds. (a) Chicken triosephosphate isomerase [1timA], (b) *E. coli* fructose bisphosphate aldolase [1dosA]. Members of the same fold family have the same major secondary structure elements with the same arrangement and connectivity. Very low sequence identity and large variations in the details of the structures reflects the lack of common ancestry between analogous folds. Although the aldolase structure contains additional helices, the TIM barrel fold is obviously present in both proteins. The TIM barrel is one of 10 'superfolds' identified by Orengo *et al.* (18).

Figure 4 shows an outline of a generic fold recognition method. Firstly, a library of unique or representative protein structures needs to be derived from the database of all known protein structures. Different groups use different selection criteria for their fold libraries: in some cases, complete protein chains are used in the library, but in other cases, structural domains or even conserved proteins cores are used. Each fold from this library is then considered in turn and the target sequence optimally fitted (or aligned) to each library fold (allowing for relative insertions and deletions in loop regions). Many different algorithms have been proposed for finding this optimal sequence–structure alignment, with most groups using some form of dynamic programming algorithm (including the examples described below), but other algorithms such as Gibbs sampling (7) or branch-and-bound searching (8) have also been used with some success. Finally, some kind of objective function is needed to determine the goodness of fit between the sequence and the template structure. It is this objective function which is optimized during the sequence–structure alignment. Again opinions differ as to the form of this objective function. Most groups use some kind of 'pseudo energy' function based on a statistical analysis of observed protein structures, but other more abstract scoring functions have also been proposed (see ref. 20 for a recent review). The final result of a fold recognition method is a ranking of the fold library in descending order of 'goodness of fit', with the best fitting fold (typically the lowest energy fold) being taken as the most probable match.

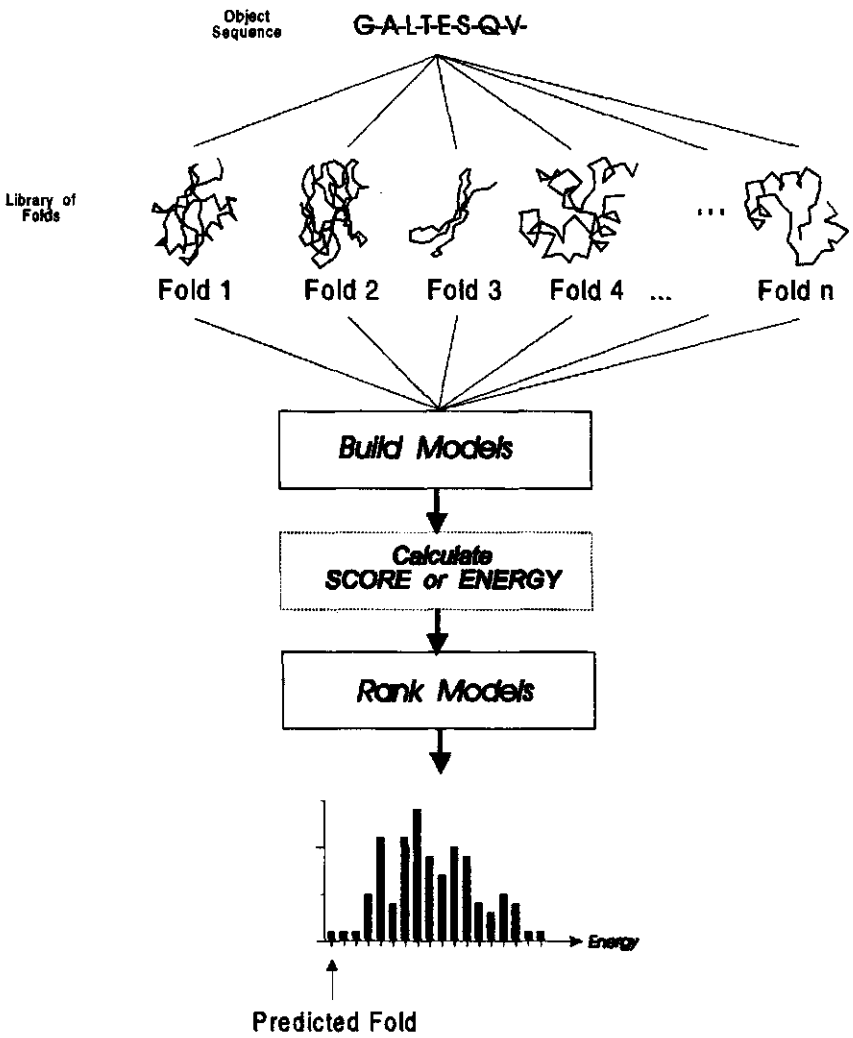

Figure 4 This is an outline of the fold recognition approach to protein structure prediction, and identifies three clear aspects of the problem that need consideration: a fold library, a method for modelling the object sequence on each fold, and a means for assessing the goodness-of-fit between the sequence and the structure.

No matter what algorithm or scoring function is used, fold recognition is not without its limitations, and some progress must be made before it can be considered a routine protein structure prediction tool. Several different aspects of this method are particularly open to improvement, namely the question of potential functions (i.e. the calculations used to determine the energy of a particular sequence once fitted onto a template fold), improvements in alignments (i.e. correctly aligning the sequence onto the template fold, to produce the best fit), and the need for progress in post-processing the results (i.e. from

the energy calculations, etc., choosing the best 'fit'). Significant progress may also arise from improvements in the threading library used (i.e. the templates upon which the sequences will be threaded).

To get some idea of the variety of methods which have been developed, four distinct approaches to the fold-recognition problem will be described. Virtually all fold-recognition methods are similar to at least one of these methods, and some newer methods incorporate concepts from more than one.

2.1 1-D–3-D profiles: Bowie *et al.* (1991)

The first true fold recognition method was by Bowie, Lüthy, and Eisenberg (10), where they attempted to match sequences to folds by describing the fold in terms of the *environment* of each residue in the structure. The environment was described in terms of local secondary structure (3 states: α, β, and coil), solvent accessibility (3 states: buried, partially buried, and exposed), and the degree of burial by polar rather than apolar atoms. The basic idea of the method is the assumption that the environment of a particular residue thus defined is expected to be more conserved than the actual residue itself, and so the method is able to detect more distant sequence-structure relationships than purely sequence-based methods. The authors describe this method as a 1-D–3-D profile method, in that a 3-D structure is translated into a 1-D string, which can then be aligned using traditional dynamic programming algorithms. Bowie *et al.* have applied the 1-D–3-D profile method to the inverse folding problem and have shown that the method can indeed detect remote matches, but in the cases shown the hits still retained some weak sequence similarity with the search protein. Environment-based methods appear to be incapable of detecting structural similarities between extremely divergent proteins, and between proteins sharing a common fold through convergent evolution—environment only appears to be conserved up to a point. Consider a buried polar residue in one structure that is found to be located in a polar environment. Buried polar residues tend to be functionally important residues, and so it is not surprising then that a protein with a similar structure but with an entirely different function would choose to place a hydrophobic residue at this position in an apolar environment. A further problem with environment-based methods is that they are sensitive to the multimeric state of a protein. Residues buried in a subunit interface of a multimeric protein will not be buried at an equivalent position in a monomeric protein of similar fold.

2.2 Threading: Jones *et al.* (1992)

The method which introduced the term 'threading' (1) went further than the method of Bowie, Lüthy, and Eisenberg in that instead of using averaged residue environments, a given protein fold was modelled in terms of a 'network' of pairwise interatomic energy terms, with the structural role of any given residue described in terms of its interactions. Classifying such a set of interactions into one environmental class such as 'buried alpha helical' will inevitably result in

the loss of useful information, reducing the *specificity* of sequence–structure matches evaluated in this way. Thus, in true threading methods, a sequence is matched to a structure by considering detailed pairwise interactions, rather than averaging them into a crude environmental class. However, incorporation of such non-local interactions means that simple dynamic programming string-matching methods cannot be used. There is therefore a trade-off to be made between the complexity of the sequence–structure scoring scheme and the algorithmic complexity of the problem.

Jones *et al.* (1) proposed a novel dynamic programming algorithm (now commonly known as 'double' dynamic programming) to the problem of aligning a given sequence with the backbone co-ordinates of a template protein structure, taking into account the detailed pairwise interactions. The problem of matching pairwise interactions is somewhat similar to the problem of structural comparison methods. The *potential environment* of a residue i can be defined as being the sum of all pairwise potential terms involving i and all other residues $j \neq i$. This is an analogous definition to that of a residue's *structural environment*, as described by Taylor and Orengo (11). In the simplest case, structural environment of a residue i may be defined as the set of all inter-Cα distances between residue i and all other residues $j \neq i$. Taylor and Orengo propose a novel dynamic programming algorithm for the comparison of such residue structural environments, and this method proved to be effective for the comparison of residue potential environments. A detailed description of the algorithm has recently been published (12).

For a sequence–structure compatibility function, Jones *et al.* chose to use a set of statistically derived pairwise potentials similar to those described by Sippl (13). Using the formulation of Sippl, short (sequence separation, $k \leq 10$), medium ($11 \leq k \leq 30$), and long ($k > 30$) range potentials were constructed between the following atom pairs: C$\beta \rightarrow$ Cβ, C$\beta \rightarrow$ N, C$\beta \rightarrow$ O, N $\rightarrow$ Cβ, N $\rightarrow$ O, O $\rightarrow$ Cβ, and O $\rightarrow$ N. For a given pair of atoms, a given residue sequence separation and a given interaction distance, these potentials provide a measure of energy, which relates to the probability of observing the proposed interaction in native protein structures. In addition to these pairwise terms, a 'solvation potential' was also incorporated. This potential simply measures the frequency with which each amino acid species is found with a certain degree of solvation, approximated by the residue solvent accessible surface area.

By dividing the empirical pair potentials into sequence separation ranges, specific structural significance may be tentatively conferred on each range. For instance, the short range terms predominate in the matching of secondary structural elements. By threading a sequence segment onto the template of an alpha helical conformation and evaluating the short range potential terms, the possibility of the sequence folding into an alpha helix may be evaluated. In a similar way, medium range terms mediate the matching of super-secondary structural motifs, and the long range terms, the tertiary packing.

Recent features added to the method allow sequence information and predicted secondary structure information to be considered in the fold-recognition

process. Sequence information is weighted into the fold recognition potentials using a transformation of a mutation data matrix (12). By carefully selecting the weighting of the sequence components in the scoring function it is possible to balance the influence of sequence matching with the influence of the pairwise and solvation energy terms. In contrast to this, secondary structure information is not incorporated into the sequence–structure scoring function. In this case, secondary structure information is used to mask regions of the alignment path matrix so that the threading alignments do not align (for example) predicted β strands with observed α helices. A confidence threshold is applied to the secondary structure prediction data so that only the most confidently predicted regions of the prediction are used to mask the alignment matrix.

2.3 Protein fold recognition using secondary structure predictions: Rost (1997)

Although most fold recognition methods employ potentials of one kind or another, it is quite easy to design a useful fold recognition approach that at first sight does not employ potentials of any kind. Although not the first example of this approach, as a good recent example the PHD secondary structure prediction service (14) has recently been extended to offer a fold recognition option. In this case the system predicts the secondary structure and accessibility of each residue in the protein of interest, encodes this information in the form of a string (similar to the scheme employed by Bowie *et al.*) (10) and then matches this string against a library of strings computed from known structures. A number of other similar methods are also in development in other labs, though all based on the initial prediction of secondary structure by PHD. Clearly no explicit potentials are being employed in these methods, but potentials are implicitly coded into the neural network weights used to predict secondary structure in the first place.

2.4 Combining sequence similarity and threading: Jones (1999)

Jones (15) has recently proposed a hybrid fold recognition method which is designed to be both fast and reliable, and is particularly aimed at automated genome annotation. The method uses a sequence profile-based alignment algorithm to generate alignments which are then evaluated by threading techniques. As a last step, each threaded model is evaluated by a neural network in order to produce a single measure of confidence in the proposed prediction. The speed of the method, along with its sensitivity and very low false-positive rate makes it ideal for automatically predicting the structure of all the proteins in a translated bacterial genome. The method has been applied to the genome of *Mycoplasma genitalium*, and analysis of the results shows that as many as 46% (now 51%) of the proteins derived from the predicted protein coding regions have a significant relationship to a protein of known structure. The fact that alignments are generated by a sequence alignment step means that the method

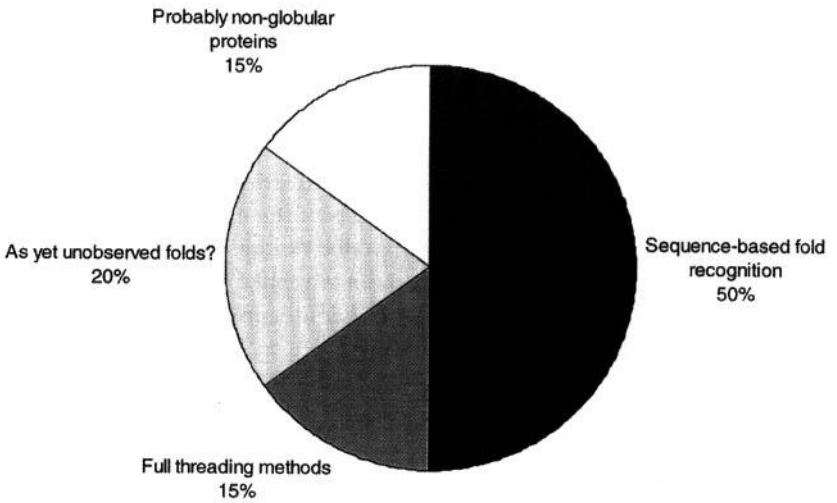

Figure 5 Hypothetical applicability of different categories of fold-recognition methods to the Open Reading Frames of small bacterial genomes. At present sequence-based fold recognition (e.g. GenTHREADER) is successful for around 50% of the ORFs. Structures for a further 15% of ORFs can probably be assigned by full threading methods such as THREADER, and the remaining 35% cannot currently be recognized either because the fold has not yet been observed, or because the ORF encodes a non-globular protein (e.g. a transmembrane protein).

is only expected to work for family or superfamily level similarities between the target and template proteins. This is both a positive and negative feature of the method. The negative aspect is that, of course, many purely structural similarities will not be detected by the method. The positive aspect is that superfamily relationships produce the most reliable results, and also allow some aspects of the function of the target protein to be inferred from the matched template structure. This latter point is particularly useful when annotating unknown genome sequences. *Figure 5* shows the current applicability of different types of fold recognition method to a genome such as that of *M. genitalium*.

Unlike full threading methods, which require a great deal of computer power to run, this type of method can be made readily available to the public via a simple Web server. The GenTHREADER method is available from the following URL:

http://globin.bio.warwick.ac.uk/psipred

3 Assessing the reliability of threading methods

Although the published results for the fold recognition methods can look impressive, showing that threading is indeed capable of recognizing folds in the absence of significant sequence similarity, it can be argued that in all cases the correct answers were already known and so it is not clear how well they would

perform in real situations where the answers are not known at the time the predictions are made. It was not until these methods were tested in a set of blind trials—the Critical Assessment in Structure Prediction experiments (CASP)—that it became clear how powerful these methods could be when used without prior knowledge of the correct answer. The CASP experiment has now been run three times (CASP1 in 1994, CASP2 in 1996, CASP3 in 1998): and in the last meeting results from over 30 methods were evaluated by the independent assessors. Up to date information on all of the CASP experiments can be obtained from the following Web address:

http://predictioncenter.llnl.gov

3.1 Alignment accuracy

Most published methods are evaluated solely on the basis of fold assignment, i.e. the method is evaluated on its ability to correctly pick the correct fold. However, in practice, fold assignment is not sufficient in its own right. Given a correct fold assignment the next step is of course to generate an accurate sequence structure alignment and to use this alignment to generate an accurate 3-D model for the target protein. In cases where a fold has been assigned, the alignment can be passed to an automatic comparative modelling program (e.g. MODELLER3) so that loops and side chains can be built in.

The accuracy of alignments that can be produced by fold recognition methods can be measured in terms of the Root Mean Square Deviation (RMSD) between the implied prediction model and the observed experimental structure. Analysis of the results of the CASP experiments has shown that alignment accuracy correlates strongly with the degree of evolutionary and structural divergence between the available template structures and the target protein. The degree of model accuracy that can be expected can be broken down into three categories of structure relationship:

(a) **Family** (e.g. *Figure 1*). Evident sequence similarity. Threading models will be almost entirely accurate, with an RMSD of between 1.0 and 3.0 Angstroms, depending on the degree of sequence similarity.

(b) **Superfamily** (e.g. *Figure 2*). No significant sequence similarity, but evident common ancestry between the template and target structure. Models for this class of similarity will be partially correct (mostly in active site regions) and will have an RMSD of between 3.0 and 6.0 Angstroms typically (though sometimes more depending on the accuracy of the alignment produced).

(c) **Analogy** (e.g. *Figure 3*). No apparent common ancestry between the template and target structure. Low quality models are expected for this category of similarity. RMSD is not a good way to evaluate models of this quality as very large shifts in the alignment produce virtually random RMSD values. At best, alignments in this class are 'topologically correct', in that the correct elements of secondary structure are equivalenced, but frequently shifts in the alignment are so large as to render the models entirely incorrect.

3.2 Post-processing threading results

Perhaps one of the most significant observations that came from the CASP2 prediction experiment was that a great deal of success in fold recognition can be achieved purely from a deep background knowledge of protein structure and function relationships. Alexey Murzin (one of the authors of the SCOP protein structure classification scheme) identified a number of key evolutionary clues which led him to correctly assign membership of some of the target proteins to known superfamilies (16). Also in two cases he was able to confidently assign a 'null prediction' to targets with unique folds purely by considering their predicted secondary structure. These feats are quite remarkable, but not easily reproduced by non-experts in protein structure and function. Despite this, it is very clear that an important future development of practical fold-recognition is to take both structure and function into account when ranking the sequence–structure matches.

Even without new developments in fold recognition algorithms, information on function and other sources of information can be applied to the results of a threading method as a 'post-processing' step. Rather than simply taking the top scoring fold to be the assumed correct answer, a fold from, say, the top 10 matches can be selected by human intervention. Such intervention might involve visual inspection of the proposed alignment, inspection of the proposed 3-D structure on a graphics workstation, comparison of proposed secondary structure with that obtained from secondary structure prediction or even consideration of common function between the target and template proteins.

3.3 Why does threading work?

Although many different formulations of energy function have been used for fold recognition, it has been shown that the principal factor in the most successful of these empirical potentials essentially encodes the general 'hydrophobic effect', rather than specific interactions between specific side chains. (e.g. the interaction potential between like charges is the same as that derived for unlike charges, reflecting not the specific interaction between side chains, but their overall preference to lie on the surface). Despite this observation that specific pair interactions are not vital to successful fold recognition, threading methods based on pairwise interactions do seem to work better than profile methods (as evidenced for example in the predictions made during the CASP experiments). This might at first sight seem contradictory, particularly as it is apparent that specific pairwise interactions are not conserved between analogous fold families (17). Nevertheless, threading methods do seem to be picking up signals which are not detected by simple 1-D profile methods. Why might this be the case?

One reasonable explanation may be that profile-based fold-recognition methods make the assumption that the pattern of accessibility between two divergent protein structures is perfectly conserved, and it is this assumption that results in their relatively poor performance. Threading methods, on the other hand, are able to model the environment of a residue by summing the hydrophobic pair

interactions surrounding a particular residue. These pair interaction environments of course change as the threading alignment changes, and it is this sensitivity of residue environments to changes in the sequence–structure alignment that results in the increased predictive power of threading methods. Although this explains why threading works even when specific contacts are not being conserved, it also explains why sequence–structure alignments are generally of poor quality when compared with known structure–structure alignments.

4 Limitations: strong and weak fold-recognition

What are the limitations of current fold-recognition methods? Let's consider two forms of the fold-recognition problem. In the first form of the problem we seek a set of potentials (and a method for performing the sequence–structure alignment) which will reliably recognize the closest matching fold for a given sequence from the thousands of alternatives—as many as 7000 naturally occurring folds have been estimated (18). This form of the problem is referred to as the 'strong' fold-recognition problem. It is possible that the strong fold-recognition problem is actually insoluble because, quite simply, the real protein free energy function is itself almost certainly incapable of satisfying this requirement. In other words, given the physical 'unreality' of threaded models, there may exist no energy function which is capable of uniquely recognizing the correct fold in all cases. One possible avenue for moving towards this goal may be to consider simulated folding pathways for each fold in the fold library, but for the time being, perfect fold recognition is but a distant dream.

The 'weak' fold-recognition problem is a far more practical formulation of the problem. Here the goal is to recognize and exclude folds which are not compatible with the given sequence with the eventual aim of arriving at a shortlist of possible conformations for the protein being modelled. At first sight this may not seem different from the goal of strong fold recognition, but the distinction is quite important. Even without a sophisticated fold-recognition method, weak recognition can be achieved by the application of simple common-sense rules. For example, if it is known that a protein is comprised entirely of alpha-helices (which might be known from circular dichroism spectroscopy, for example) then a large number of possible folds can be eliminated immediately (the correct fold could not be the all-beta immunoglobulin fold, for example). By applying a set of such rules, the 7000 or so possible folds could quickly be whittled down to a shortlist of say 10.

In reality, most, if not all of the published fold recognition methods really implement weak fold recognition. In the hands of an experienced user, however, who can make use of functional or structural clues in the prediction experiment, even weak fold recognition can be very powerful.

4.1 The domain problem in threading

Perhaps the main practical limitation of most 'weak' threading methods is that they are aimed at recognizing single globular protein domains, and perform very

poorly when tried on proteins which comprise multiple domains. Unfortunately, threading cannot be used for identifying domain boundaries with any degree of confidence, and indeed the general problem of detecting domain boundaries from amino acid sequence remains an unsolved problem in structural biology. If the domain boundaries in the target sequence are already known, then of course the target sequence can be divided into domains before threading it, with each domain being threaded separately. Predictions can be attempted on very long multi-domain sequences, but in these cases the results will not be reliable unless it is clear that the matched protein has an identical domain structure to the target. For example, the periplasmic small-molecule binding proteins (e.g. leucine–isoleucine–valine binding protein) are two domain structures (two doubly wound parallel alpha/beta domains), but they all have identical domain organization. Proteins within this superfamily can thus be recognized by threading methods despite their multi-domain structure.

5 The future

One major difference between the academic challenge of protein structure prediction and the practical applications of such methods is that in the latter case there is an eventual end in sight. As more structures are solved, more target sequences will find matches in the available fold libraries—matched either by sequence comparison or threading methods. In terms of practical application, the protein-folding problem will thus begin to vanish. There will of course still be a need to better understand protein-folding for applications such as *de novo* protein design, and the problem of modelling membrane protein structure will probably remain unsolved for some time to come, but nonetheless, from a practical viewpoint, the problem will be effectively solved. How long until this point is reached? Given the variety of estimates for the number of naturally occurring protein folds, it is difficult to come to a definite conclusion, but taking an average of the published estimates for the number of naturally occurring protein folds and applying some intelligent guesswork, it seems likely that when threading fold libraries contain around 1500 different domain folds it will be possible to build useful models for almost every globular protein sequence in a given proteome. At the present rate at which protein structures are being solved, this point is possibly 15–20 years away. However, pilot projects are now underway to explore the possibility of crystallizing every globular protein in a typical bacterial proteome. If such projects get fully under way, which seems likely, then a complete domain fold library may be only five years away.

References

1. Jones, D. T., Taylor, W. R., and Thornton, J. M. (1992). A new approach to protein fold recognition. *Nature*, **358**, 86.
2. Godzik, A. and Skolnick, J. (1992). Sequence-structure matching in globular proteins: Application to supersecondary and tertiary structure determination. *Proc. Natl. Acad. Sci. USA*, **89**,12098.

3. Ouzounis, C., Sander, C., Scharf, M., and Schneider, R. (1993). Prediction of protein structure by evaluation of sequence-structure fitness. Aligning sequences to contact profiles derived from three-dimensional structures. *J. Mol. Biol.*, **232**, 805.
4. Abagyan, R., Frishman, D., and Argos, P. (1994). Recognition of distantly related proteins through energy calculations. *Proteins: Struct. Funct. Genet.*, **19**, 132.
5. Overington, J., Donnelly, D., Johnson, M. S., Sali, A., and Blundell, T. L. (1992). Environment-specific amino-acid substitution tables—tertiary templates and prediction of protein folds. *Protein Sci.*, **1**, 216.
6. Matuso, Y., Nakamura, H., and Nishikawa, K. (1995). Detection of protein 3-D-1-D compatability characterised by the evaluation of side-chain packing and electrostatic interactions. *J. Biochem. (Japan)*, **118**, 137.
7. Madej, T., Gilbrat, J.-F., and Bryant, S. H. (1995). Threading a database of protein cores. *Proteins*, **23**, 356.
8. Lathrop, R. H. and Smith, T. F. (1996). Global optimum protein threading with gapped alignment and empirical pair score functions. *J. Mol. Biol.*, **255**, 641.
9. Taylor, W. R. (1997). Multiple sequence threading: An analysis of alignment quality and stability. *J. Mol. Biol.*, **269**, 902.
10. Bowie, J. U., Lüthy, R., and Eisenberg, D. (1991). A method to identify protein sequences that fold into a known three-dimensional structure. *Science*, **253**,164.
11. Taylor, W. R. and Orengo, C. A. (1989). Protein structure alignment. *J. Mol. Biol.*, **208**, 1.
12. Jones, D. T. (1998). THREADER : Protein Sequence Threading by Double Dynamic Programming. In *Computational methods in molecular biology* (ed. S. Salzberg, D. Searls, and S. Kasif). Elsevier, Amsterdam.
13. Sippl, M. J. (1990). Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins *J. Mol. Biol.*, **213**, 859.
14. Rost, B. (1997). Protein fold recognition by prediction-based threading. *J. Mol. Biol.*, **270**, 1.
15. Jones, D. T. (1999). GenTHREADER: An efficient and reliable protein fold recognition method for genomic sequences. *J. Mol. Biol.*, **287**, 797.
16. Murzin, A. G. and Bateman, A. (1997). Distant homology recognition using structural classification of proteins. *Proteins Suppl.*, **1**, 105.
17. Russell, R. B. and Barton, G. J. (1994). Structural features can be unconserved in proteins with similar folds—an analysis of side-chain to side-chain contacts, secondary structure and accessibility. *J. Mol. Biol.*, **244**, 332.
18. Orengo, C. A., Jones, D. T., and Thornton, J. M. (1994). Protein superfamilies and domain superfolds. *Nature*, **372**, 631.
19. Kraulis, P. J. (1991). Molscript—a program to produce both detailed and schematic plots of protein structures. *J. Appl. Crystallogr.*, **24**, 946.
20 Jones, D. T., and Thornton, J. M. (1996). Potential energy functions for threading. *Curr. Opin. Struct. Biol.*, **6**, 210.