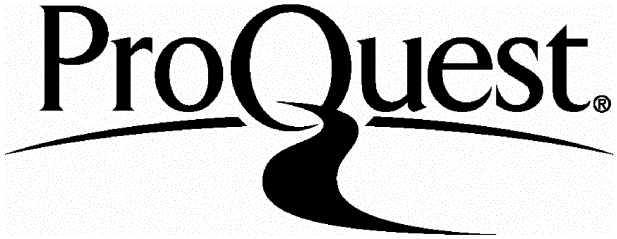
Structural approaches to protein sequence analysis

by David Tudor Jones

Department of Biochemistry and Molecular Biology
University College
Gower Street
London
WC1E 6BT

This thesis is submitted in fulfilment of the requirements for the degree of Doctor of Philosophy from the University of London.

April 17, 1993



Abstract

Various protein sequence analysis techniques are described, aimed at improving the prediction of protein structure by means of pattern matching.

To investigate the possibility that improvements in amino acid comparison matrices could result in improvements in the sensitivity and accuracy of protein sequence alignments, a method for rapidly calculating amino acid mutation data matrices from large sequence data sets is presented. The method is then applied to the membrane-spanning segments of integral membrane proteins in order to investigate the nature of amino acid mutability in a lipid environment.

Whilst purely sequence analytic techniques work well for cases where some residual sequence similarity remains between a newly characterized protein and a protein of known 3-D structure, in the harder cases, there is little or no sequence similarity with which to recognize proteins with similar folding patterns. In the light of these limitations, a new approach to protein fold recognition is described, which uses a statistically derived pairwise potential to evaluate the compatibility between a test sequence and a library of structural templates, derived from solved crystal structures. The method, which is called *optimal sequence threading*, proves to be highly successful, and is able to detect the common TIM barrel fold between a number of enzyme sequences, which has not been achieved by any previous sequence analysis technique.

Finally, a new method for the prediction of the secondary structure and topology of membrane proteins is described. The method employs a set of statistical tables compiled from well-characterized membrane protein data, and a novel dynamic programming algorithm to *recognize* membrane topology models by expectation maximization. The statistical tables show definite biases towards certain amino acid species on the inside, middle and outside of a cellular membrane.

Acknowledgements

In many ways this section was the hardest of all to write. It seems that I have received help and encouragement in one way or another from just about everyone I have met during my PhD course. Unfortunately, due to the limits of both space and memory, I can only hope to mention a fraction of those who have played an important part in the completion of this project. My sincerest apologies to anyone I have omitted.

It goes without saying that I wouldn't be writing this were it not for the help, encouragement, trust, friendship and of course education I have received from my two supervisors, Professor Janet Thornton and Dr Willie Taylor. In many senses, it is hard for a PhD student to actually choose a supervisor (let alone TWO supervisors) - they are pretty much chosen for you by the constraints of geography, project availability, timing, luck and of course whether or not the supervisor in question actually wants you to work on their project. All I can say is that if I could have chosen from every supervisor in the field, I could not have made a better choice. I greatly look forward to our continued friendship and collaboration in the future.

Thanks to the following:

Dr Christine Orengo for all the general help, encouragement, technical discussions on algorithms over the years. Above all, thanks for never physically ejecting me from your office when I've been looking for someone to talk to and you've had a desk-full of work to do!

Dr Tom Flores for all the help and advice, and the many discussions we've had on topics too numerous to mention - most of which left me out of my depth after about 2 milliseconds.

Dr Mark Swindells for the coffee, ice-cream, The Simpsons™, extremely stimulating technical arguments, and humour beyond the call of duty. Don't stay in Japan for ever!

Dr Simon Hubbard for even more coffee, ice-cream, discussions, the brilliant wit, and keeping a grip on reality even when the rest of the world appeared to be auditioning for a part in a sequel to *One Flew Over the Cuckoo's Nest.*

Dr Nigel Brown for teaching me about Un*x, and the world of *real* computers.

Dr Richard Mott for much helpful advice on statistics and many useful suggestions.

Dr Tom Kirkwood for making me welcome in the Laboratory of Mathematical Biology at the National Institute for Medical Research (NIMR), where some of this work was carried out.

And the following people for help, discussion, computer code, jokes but above all just for making my workplace fun to be in: Dr Andras Azodi, Steve Gardner, Dr Michael Green, Dr Gail Hutchinson, Dr Axel Kowald, Dr Roman Laskowski, Malcolm MacArthur, Dr Laurence Pearl, Frances Richardson, Dr Jus Singh, Helen Stirk, Dr Dek Woolfson, and Dr Marketa Zvelebil.

This work was supported by a SERC CASE studentship with the MRC National Institute for Medical Research, Mill Hill, London.

To my parents - Rose and Tudor Jones.

Contents

Though it may be hard to believe, the unusual random pattern on the final page is in fact a stereogram, which can be viewed by defocussing the gaze in a way similar that used for normal stereograms. Stereo glasses will not work in this case, however.

Materials in cover pocket

1. Taylor, W.R. & Jones, D.T. Templates, consensus patterns and motifs. (1991) *Curr. Opin. Struct. Biol.* **1**, 327-333.

2. Jones, D.T., Taylor, W.R. & Thornton, J.M. (1992) The rapid generation of mutation data matrices from protein sequences. *Comput. Applic. Biosci.* **8**, 275-282.

3. Jones, D.T., Taylor, W.R. & Thornton, J.M. (1992) A new approach to protein fold recognition. *Nature* **358**, 86-89.

List of Figures

List of Tables

Commonly used abbreviations

Å	Ångstrom
ATP	Adenosine triphosphate
C-terminal	Carboxy terminal
DNA	DeoxyriboNucleic Acid
MDM	Mutation Data Matrix
N-terminal	Amino terminal
NW	Needleman & Wunsch
RMSD	Root Mean Square Deviation
UPM	Unitary Protein Matrix (identity matrix)

Standard amino acid codes

Ala (A)	Alanine	Leu (L)	Leucine
Arg (R)	Arginine	Lys (K)	Lysine
Asn (N)	Asparagine	Met (M)	Methionine
Asp (D)	Aspartic acid	Phe (F)	Phenylalanine
Cys (C)	Cysteine	Pro (P)	Proline
Gln (Q)	Glutamine	Ser (S)	Serine
Glu (E)	Glutamic acid	Thr (T)	Threonine
Gly (G)	Glycine	Trp (W)	Tryptophan
His (H)	Histidine	Tyr (Y)	Tyrosine
Ile (I)	Isoleucine	Val (V)	Valine

Chapter 1

Introduction

Pattern Matching Methods in Protein Structure Prediction

Deep in the sea
all molecules repeat
the patterns of one another
till complex new ones are formed.
They make others like themselves
and a new dance starts.

Growing in size and complexity
living things
masses of atoms
DNA, protein
dancing a pattern ever more intricate.

- Richard P. Feynman

1.1 Protein structure prediction from first principles

Evidence for a direct relationship between the primary structure of a protein and higher levels was first presented by Anfinsen *et al.* (1961). It is generally thought that the amino acid sequence contains all the information required to permit a polypeptide to 'self-assemble' itself into a biologically active three-dimensional structure. Though this folding of the protein is context sensitive, in that its correct execution is highly dependent on the ambient conditions and the possible presence of cofactors or even 'molecular chaperones', the mechanisms that decode the structural information are in principle very simple. The structure of a folded protein is maintained through a complex interplay between a handful of physico-chemical forces: electrostatic effects, hydrogen bonding, hydrophobic effects, solvent entropic effects, salt bridging and so on. Most of these forces are well defined and have a solid backbone of theory behind them. In theory, the effects of hydrogen bonding can be modelled by simply solving the relevant quantum mechanical equations, solvent effects modelled by molecular dynamics, the list could continue. Predicting the structure and function of a protein sequence is therefore conceptually simple. A set of differential equations could be constructed where each physical force is simulated and the folding process itself simulated as a result. Given a sufficiently powerful computer an acceptable solution could be found. However, the number of these simultaneous equations that must be solved to predict the final protein conformation is astronomical. Though the number of atoms in the polypeptide chain itself is small, it is also necessary to consider the multiple interactions between the surrounding solvent molecules and the chain. By using suitable simplifications some progress has been made along these lines, however even where theoretical solutions have been within reach, it has proven difficult to find the correct energy functions and even more difficult to find functions that allow a *convergent* solution to be found (Levitt, 1976; Robson & Osguthorpe, 1979). Suffice it to say that though a direct theoretical approach to solving the folding problem could be envisaged, the problem may well turn out to be intractable via this route alone.

Given that we cannot now, and may never be able to predict protein structure from first principles, other 'heuristic' methods must be found. An obvious approach to solving vastly complex systems of equations is to merely observe the macroscopic properties exhibited by a wide range of different final solutions. In this case, a sensible approach is to analyze the final folded states of different proteins statistically. Amongst the attempts at analyzing the relationship between protein sequence and structure statistically were those of Chou & Fasman (1974) and Garnier *et al.* (1978). These attempts were strictly aimed at predicting the secondary structure of proteins. The basic idea behind these techniques is to assign a structural 'propensity' to either individual residues (e.g. Chou & Fasman) or short sequence segments (Garnier *et al.*). Though in a sense a degree of pattern matching is being carried out in these methods, they are not strictly pattern matching approaches. For a review of these essentially statistical approaches see Taylor (1987).

1.2 Pattern matching methods in the prediction of protein structure and function

Taylor (1988b) explains that the applicability of different pattern matching methods to given situations depends on the degree of homology found between the sequence under scrutiny and the database of known structures. For cases where the percentage identity (percentage of conserved residues between sequences) is 50% or above, the case is clear-cut. The function of a protein will in all probability have been identified by the location of such a close homologue. The structure will also follow closely, using the available techniques for modelling homologous proteins (e.g. Blundell *et al.*, 1988). Highly conserved residues can be spotted easily by the construction of suitable conservation plots, and using this information to 'pin' the unknown structure to the known structural 'template', a reasonable model can be produced. This approach is currently the most successful structure prediction method known.

Where a significant degree of homology (> 50% identity) exists between two sequences, pattern matching merely requires their optimal alignment. Two classes of global sequence

comparison techniques exist: one based on the concept of *dynamic programming*, the other based on the identification of common subsequences. Dynamic programming provides the most general and rigorous solution to this particular problem. The application of dynamic programming methods to the alignment of biological sequences was first described by Needleman and Wunsch (1970), and improved upon by other groups (Sankoff, 1972; Sellers, 1974; Smith & Waterman, 1981; Gotoh, 1982). Alignment begins by the construction of a similarity matrix thus:

	A	**C**	**D**	**E**	**F**	**G**	**H**	**L**
A	1	0	0	0	0	0	0	0
D	0	0	1	0	0	0	0	0
D	0	0	1	0	0	0	0	0
E	0	0	0	1	0	0	0	0
F	0	0	0	0	1	0	0	0
G	0	0	0	0	0	1	0	0
P	0	0	0	0	0	0	0	0

In this simple case exact residue matches are given a score of 1, and any other match a score of zero. More 'lenient' scoring schemes are commonly used, particularly that of Dayhoff (1968, 1978).

The Needleman & Wunsch method continues by the dynamic calculation of all possible paths through the matrix starting at the bottom right hand corner, finishing at top left. Values in the similarity matrix are replaced by the maximum score obtainable *from that point on.* At the end of the summing procedure a maximum value should be found along the top row or the leftmost column[1] which indicates the starting point for the optimal alignment path (the optimal alignment path for the above matrix is shown by the underlined scores). A further refinement to the method is the addition of a gap penalty

[1] This is true only if the initial similarity matrix is biased towards positive scores, in which case a global alignment results. If the matrix is biased towards negative scores, then a local alignment will result, and the highest scoring cell can be found anywhere in the matrix.

which prevents the insertion of a ridiculous number of gaps in order to maximize the overall score.

Dynamic programming, though providing a rigorous, optimal alignment, suffers from a severe lack of speed, even when implemented on a fast computer architecture. Though several attempts have been made to improve the performance of the standard Needleman & Wunsch methods (see Taylor, 1988a), faster approximate alignment techniques are more commonly used for 'front-line' pattern matching with large databases. These approximate techniques are all based on the identification of common subsequences or 'tuples'. The original technique based on tuple comparison was the crude but effective 'dot-matrix' technique (Staden, 1982 - review). In this technique, the two sequences are written along the x and y axes respectively of a matrix, and a dot placed at each location where the corresponding residue pairs match. The resulting matrix is inspected visually in order to detect diagonal runs of dots which are indicative of homologous stretches between the two sequences. Though the output from the dot-matrix technique is nicely visual, it is obviously not an automatic technique.

Automatic tuple comparison programs have been produced, notably by Wilbur and Lipman (1983) and Lipman and Pearson (1985). The FASTx programs of Lipman and Pearson (Pearson, 1990 - review) have become a *de facto* standard in the front-line database searching field. In all these methods the principle is simple. Each sequence is split into its respective tuples (subsequences of lengths 2..n) and these subsequences stored in a single hashed lookup table. The structure of this table is such that tuples found to be in both sequences are stored in a linked list. An alignment is produced by locating *significant diagonals*, which are simply the diagonals of the dot-matrix that have an above-average number of common tuple matches. As already stated, the main advantage of these methods is their speed: performance being around 40-50 times greater than that of dynamic programming approaches. The major disadvantage with tuple methods is that they rely on finding a fairly high number of common tuples, which in the case of fairly dissimilar sequences (< 50% ID) will not often be the case. Related to this problem, is the

difficulty in providing a decent 'similarity' scoring scheme where amino-acid similarity is scored, rather than amino-acid identity.

More recently a very rapid, and relatively sensitive, databank searching program (BLAST - Basic Local Alignment Search Tool) has been described (Altschul *et al.*, 1990). Though the speed of BLAST is impressive (searching around 500,000 amino-acid residues per second on a SUN 4/280), and its statistical properties are well characterized (Karlin & Altschul, 1990), it still falls far short of dynamic programming with respect to sensitivity. An interesting question that often arises when considering future developments in databank searching is whether or not priority should be given to the sensitivity of the method over the speed at which it executes. Heuristic methods are only worthwhile if a rigorous search is impractical using available computing resources. Extremely powerful computers are now readily available, and it is presently the case that a modern workstation can perform a rigorous dynamic programming search of the protein sequence databank in a few hours of CPU time. Whether this will remain the case is a matter of debate, though a glance at Figure 1.1 suggests that at the time of writing information technology is keeping ahead of biotechnology.

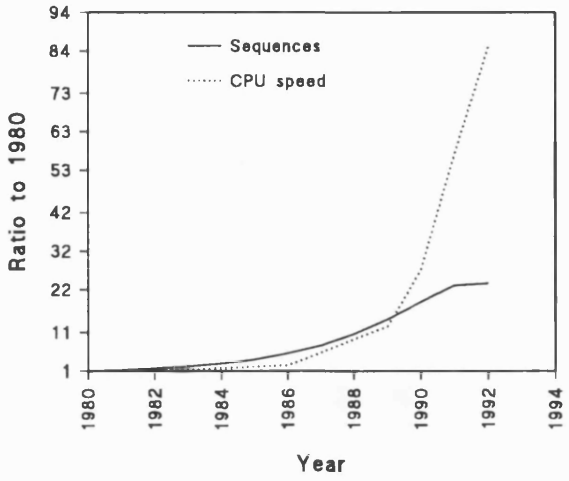


Figure 1.1

The size of the PIR databank (George *et al.*, 1992) is plotted against the speed of commonly available computers i.e. computers which would be readily available to the average biochemist. Values are shown relative to the state in 1980.

Simple pairwise alignment methods as described above are good at comparing sequences where the homology exceeds 50% or so. In some cases two sequences may in fact have an *overall* homology much less than 50% and yet have a common segment with extremely high homology. Terminal and loop regions may well be highly variable in sequence, or may even be deleted or extended across a single protein family. In these cases a global alignment across the full length of both sequences may fail to provide a significant result. Both the dynamic programming and tuple alignment techniques have been modified to allow the identification of regions of maximum *local* homology. In the case of dynamic programming, the single-stretch best local alignment method of Smith and Waterman (1981) has proven popular, however a more complicated method by Sellers (1974, 1984)

allows multiple subsequences to be detected in one operation. For an example of tuple-based local alignment see Waterman (1986).

1.3 Consensus methods

As the degree of homology between two sequences drops below 50% it becomes difficult to locate the *biologically* optimum pairwise alignment between them. Usually, however, more than two sequences are available. Given several examples from a single family of proteins it is possible to construct a *consensus* alignment between them. The principle of consensus methods is seen mirrored in many fields of science and mathematics. In an abstract sense, consensus alignment is simply an example of *statistical sampling*. The basic idea is that instead of a single amino-acid code at each alignment position, a histogram is constructed where the numbers of each of the 20 amino acids occurring at that position are tallied. This allows each sequence to 'vote' for the appropriate residue at any given alignment position. Usually the total alignment is iterated, using the consensus patterns built in the previous pass to direct the alignments of the following pass. The concept of consensus pattern matching will be discussed in more detail later.

Multiple consensus alignment techniques are useful for average homologies above around 30% at which point all alignments become statistically insignificant. The 30% identity cut-off very roughly marks the outermost boundary of the so-called 'Twilight Zone' (Doolittle *et al.*, 1986). The Twilight Zone is simply defined as the region of homology where an optimal alignment between random sequences is found to be no worse or better than the alignment between the trial protein sequences. It is unfortunate that in many documented cases useful alignments lie well inside the Twilight Zone. The problem is simply down to the extreme variability of protein sequence even where the higher levels of structure are seen to be highly conservative. Two very different sequences can quite easily have extremely similar folded conformations. Of course we might reasonably expect certain key residues to be conserved across the evolutionary tree, but how can these key residues be

located amid the extreme evolutionary 'noise'. The number of key residues may well be very small in comparison to the lengths of the sequences, but methods have been developed to enable such residues to be rapidly located. See Taylor (1988a, 1990) for a complete description of consensus alignment techniques.

The methods employed to detect remote homology between sequences are the truest forms of pattern matching. The earliest example of a strictly pattern matching approach to sequence comparison was that of the helical wheels of Schiffer and Edmundson (1967, 1968). Schiffer and Edmundson studied the distribution of hydrophobic residues along the lengths of α-helical sequence segments by plotting each residue in a circular fashion corresponding to the pitch of the helix. α-helical regions generally exhibit clustering of hydrophobic residues along a single sector of the wheel. Though of some vintage, this technique remains a powerful means for identifying and comparing α-helical regions. The method of helical wheels was refined by Palau and Puigdomenech (1974) who analyzed the zonal distribution of hydrophobic residues in helical regions. They found that in helical regions hydrophobic residues at sequence offsets n, n+1 and n+4 or n, n+3, n+4 acted to stabilize the helices.

Shortly after Palau and Puigdomenech's article two articles by Lim (1974a,b) were published. The first of Lim's article presented a general study of the stereochemistry of globular proteins, showing which residue types could be associated with regular (α or ß) or irregular protein conformations. In the second article, observations made on proteins from the contemporary structural database (25 structures in total), as outlined in the first

article, were distilled into 22 rules. These rules tended to be verbose and highly general. For example, Rule 3 for α-helix formation is as follows:

> *"Let the hydrophobic pair (1-2) or the hydrophobic triplet (1-2-5) formed from positions (i, i+1) and (i, i+1, i+4) respectively, and position i+1 contain phenylalanine. Such a hydrophobic pair (1-2) or a hydrophobic triplet (1-2-5) ... will be α-helical if Phe situated in position i+1 forms within the limits of the obtained α-helix a hydrophobic-hydrophilic pair (1-5) with G_L which is situated in position i-3."*

Lim claimed a prediction accuracy of 80-85% when these rules were applied to the same structures on which the rules were based. Though the accuracy of Lim's method is certainly not as high as 80% when applied to proteins absent from the original data set, it does in fact score higher, on average, than many of the more popular statistical techniques. Not surprisingly rules of this type were not easy to convert into computer code, which contributed to their lack of popularity. Taylor (1988b) points out that modern logic based languages seem to be suitable vehicles for encoding Lim's rules, and that work along these lines has been started by C. Rawlings and P. Stockwell (personal communication).

All the above pattern matching methods are concerned with detecting simple patterns primarily designed to predict secondary structure. The trend in protein sequence pattern matching has moved towards the construction of more complex templates, capable of matching higher levels of structure than basic α-helices or ß-sheet regions. Given that long-range interactions play an important part in the direction of protein folding, even at the purely secondary structural level, the necessity for these complex patterns is reasonably obvious. Nagano (1977) first described the use of a super-secondary structural motif, where the ßαß unit was analyzed. Following on from this work Nagano (1980) extended the algorithm to a generalized structure prediction system. This approach splits the sequence into pentapeptides to reduce the number of degrees of freedom in the folding simulation. The folding is further constrained by considering the packing in a 2D matrix

(3 x 11 boxes) rather than the 3D atomic coordinate space. The nub of the method is simple in that each pentapeptide is labelled as being α or ß depending on the total of a statistically determined secondary structure propensity for each. Likely ßαß units are then located by considering pentapeptide patterns that neighbour strongly predicted α or ß segments, with an appropriate distance filter that excludes ß-α/α-ß pairs that are too far apart on the grid. An important part of this method was a combinatorial analysis of all the possible permutations of predicted structural segments, this is analogous to scanning through all the possible three-dimensional packings given a mixture of well-defined and ill-defined structural units, except that in this case packing is performed on a 2D grid.

Richmond and Richards (1979) described a fairly involved technique for tertiary structure prediction. Though the method used a wide range of techniques outside the realm of pattern matching, a pattern matching approach was used at the core of the method to identify hydrophobic residues important for the packing of secondary structural elements. The number of packing permutations matching the given patterns tended to be enormous, though this number was quickly reduced by means of simple distance filters applied in particular to the ends of helices. Richmond and Richards only considered proteins in the αα folding class (e.g. myoglobin) where the hydrophobic patterns were well defined, and the distance filters easy to construct and apply. Cohen *et al.* (1980, 1982, 1986) extended this combined pattern-matching and packing technique to ßß and ßα folding types. This extension demanded the construction of hydrophobicity patterns for ß structure similar to those already well-recognized for α-helices. Success in this case depended on the provision of some external knowledge to the prediction problem. The fundamental limitation here is the lack of accuracy in secondary structure prediction. Of course, given knowledge of the protein's folding type, secondary structure prediction can be weighted towards reasonable accuracy, and indeed where this information is available, the method of Cohen *et al.* provides at least a fair guess at the overall protein topology.

The methods of Richmond and Richards and Cohen *et al.* essentially fail due to a lack of accuracy in secondary structure prediction. The above methods are *combinatorial* and as

such any uncertainties in the elements playing a part in the prediction process are grossly magnified by the time a complete tertiary structure is produced. Reducing the vast number of possible packing permutations relies on the provision of suitable filters applied to the component secondary structural elements. If the secondary structural elements are ill-defined (or rather mis-predicted) then we should not be surprised if the final predicted state is far from reality. These techniques are guilty of building skyscrapers on top of wooden foundations.

Taylor and Thornton (1983) attempted to improve the accuracy of secondary structure prediction by constructing templates capable of detecting super-secondary structural elements, or more specifically (but not exclusively) the ßαß unit. Using 62 examples of the ßαß unit from the Brookhaven database (Bernstein *et al.*, 1977) an ideal secondary structure sequence template was constructed. This ideal ßαß pattern was matched at each residue position of the test sequence matching the template profiles to the Garnier secondary structure prediction probability profiles (Garnier *et al.*, 1978; Gibrat *et al.*, 1987), and a score calculated. Different length variants of the ideal template were created by scaling the master template so as to accommodate the length variations observed in the available examples. Apart from the statistical sequence template, templates were also constructed for matching patterns of hydrophobicity (as in Lim's method) - one template scored highly for buried ß regions, the other for α-helical regions. The strongest fitting template was selected, and other matching templates selected according to various rules (for example forbidding overlapping α and ß regions). On a test set of 16 ß/α proteins, a prediction score of 70% was achieved, bettering the raw GOR secondary structure prediction technique by some 7.5%.

As an extension to the work on ßαß templates, Taylor (1986a) went on to produce a generalized consensus template method. The first major improvement to the original template method of Taylor and Thornton was to move over to 2D templates that could match more than one physico-chemical criterion at each alignment position. The second major improvement was to contrive a means for generating the templates automatically,

given a suitably well-defined sequence alignment. The method starts with a seed alignment, generally based on available structural information. A consensus pattern is created from this initial alignment such that each alignment position contains a count of each of the 20 amino acids. This ties back to the earlier discussion of multiple sequence alignments using consensus methods. One important advantage of consensus patterns is that they are insensitive to the odd misalignment. Consider the case where, for example, in 10 alignments a glycine residue is found at a particular alignment position and in the eleventh, due to a misalignment, a proline is aligned with the glycine consensus. Evidently, the proline match will be recorded, but in future alignments, the proline will be scored 10 times lower than the glycines - the 'glycine-like' tendency of that particular template position will be more-or-less conserved.

The initial consensus template, preferably solidly based on structural knowledge, is then used to collect further matching sequences from the sequence database. The new set of sequences is then aligned to the consensus, to produce a new expanded consensus template, ready for a further cycle of sequence collection, alignment, and template generation. When no new sequences are collected, the alignment cycle is exited and the final minimal set template constructed. The nature of this final template is totally different from the consensus templates used to direct the collection/alignment cycle. Rather than recording the number of each of the 20 amino acids observed at each position, the observed amino acid identities are used to pick a minimal covering class of amino acid from a Venn diagram. This Venn diagram comprises three major sets: Hydrophobic, Polar and Small (with subsets Aromatic, Aliphatic, Tiny, Charged and Positive). The smallest subset that contains all the residues observed at a particular alignment position is selected, and is recorded in the final template at the corresponding position. There are two main advantages to using minimal covering classes over consensus patterns. Firstly, minimal covering class patterns are *predictive* in that they are able to predict possible amino acid substitutions that may not have been observed in the limited data set that was used in the template's construction. The second advantage is that bias in the data-set is eliminated. Consider the previous case of 11 alignments where one proline is matched against 10

glycines. If this alignment proves to be valid and indeed a proline *is* a valid residue at that position then in the final template we *do not* wish to score it lower than the more common glycine residues. Indeed, it may turn out that the sequences that contributed the 10 glycines at that position were in fact highly homologous and that, were the sample of sequences less biased we might observe just as many cases where proline is present as glycine. Taylor's method concludes by aligning each of the contributing sequences against the final template using a set-based scoring scheme rather than a consensus-based scheme. If the whole process has been successful then we expect the set-based alignment to concur with the consensus alignment.

Pearl and Taylor (1987) used the above consensus template program suite to identify common features in the retroviral protease and aspartyl protease families. The method was able to detect the few conserved residues that formed the proposed active site even though the sequences were extremely non-homologous with respect to normal alignment methods.

A method ('profile analysis') similar in many respects to the template method of Taylor has been devised independently by Gribskov *et al.* (1987; 1990 - review). An important aspect of the profile analysis method, which has led to its popularity, is the reliability of the underlying statistics. In particular, the use of Z-scores (see page 134) calculated for different profile length ranges, not only provides a useful measure of statistical significance, but also highlights matches that might otherwise go unnoticed were the raw match scores to be used.

To enhance their previous work on secondary structural packing, a pattern matching approach to predicting structure in ß/α proteins was formulated by Cohen *et al.* (1983). Their algorithm concentrated on turn prediction, achieving a prediction score of 98%, though by using the reliable turn prediction a complete structure prediction was produced as a final result. The method is essentially based on Lim-like patterns, though by considering segments separated by predicted turns, further constraints could be applied to cope with ambiguous predictions. The first stage of the method (TURNGEN module)

involves locating turns by their characteristically high polar residue content. These predicted turn regions delimit independent sequence segments that are to be assigned as α, ß or 'null' (null structures can be irregular, or just isolated regular structures that do not interact with the central ß-sheet). Using a large array of small patterns, each delimited region is analyzed, and labelled as possibly α or ß depending on the overall pattern matching score (ABGEN module). The next stage attempts to label some segments as being *definitely* α, ß or null (ADEF, BDEF and NULLDEF modules) by using pattern combinations (for example if a segment has been previously labelled as a potential α-helical segment and the hydrophobic diamond pattern 'S' matches, then the segment is assigned as definitely α). Remaining uncertainty in the prediction is resolved by the application of high-level 'expert system' rules containing well-founded knowledge on the structure of ß/α proteins (ADJUST module). Processing continues (DELIMIT module) by using specific N and C terminal patterns to try to accurately determine the boundaries of the secondary structural elements now known to be contained in the 'definite' segments (for example a 'stop signal' for α-helices might well be one or more prolines or three hydrophobic residues). The final prediction is achieved by means of a scoring and ranking procedure (SCORE module) and by another expert system analysis of all the remaining combinatorial possibilities of definite segments and possibles (COMBINATORICS module). This combinatorial step is somewhat similar to the method of Nagano (1980) as described previously. The final icing is provided by filtering out the remaining statistical anomalies (OUTLIER module).

The method of Cohen *et al.* described above is complicated, but is interesting in that it uses earlier pattern matching approaches, using a rigorous artificial intelligence approach to handle ambiguous predictions. Another very interesting aspect is the use of expert rules to direct the global prediction process. Such a combination of methods will no doubt play a vital part in successful future prediction schemes, drawing predictive power from both blind pattern matching methods, and the hard-earned knowledge of experts in the rules of protein structure. From current evidence it would appear that prediction techniques drawing from one approach or the other (but not both) are only achieving prediction

scores of 60-70% or so. See also King & Sternberg (1990) for a more recent example of an artificial intelligence system applied to the problem of structure prediction - in this case using a machine learning algorithm.

1.4 Regular expressions

Another type of protein sequence pattern that has recently become popular is that based on *regular expressions*. A regular expression is simply a linear sequence pattern that permits the use of wildcards (matching any residues), set closures (matching residues in a particular set), gaps (matching a number of residues or none at all). To fit in with the complete definition of a regular expression, it must also be possible to define nested sub-patterns. For example in the pattern XYZ(P(QR))XYZ, the possible matches are XYZPQRXYZ or the sub-pattern PQR, or the sub-sub-pattern QR.

Probably the earliest example of true regular expressions being used for matching complex biological sequence patterns was the QUEST system designed by Abarbanel *et al.* (1984). QUEST is a rapid search tool that implements a concise pattern language, very closely modelled on the syntax of the Unix EGREP (Extended Global Regular Expression search and Print) program. QUEST was designed to be able to handle the kinds of patterns thought useful in sequence analysis and structure prediction, and allowed patterns to be defined in terms of named or explicitly defined sub-patterns. For example in turn-prediction a sub-pattern would simply be the set of single residues with a predilection for turn-formation ([PGQNSTEDRKH]) called for example 'tphilic'. A further sub-pattern would be [YPGQNSTEDRKH] (tphilic plus tyrosine) called 'yturnphilic'. A QUEST meta-pattern for a turn could then be defined as "tphilic{3} yturnphilic", or in English : three turn-formers, plus one turn-former or tyrosine.

A derivative of QUEST (which was coded in the commercial IBM pattern language MAINSAIL) was used by Cohen *et al.* for further work on turn prediction. The derivative

pattern system is known as PLANS (Pattern Language for Amino and Nucleic acid Sequences), and was coded in Lisp. With the convenience of a well-tailored pattern matching language, Cohen *et al.* analyzed turn patterns in all structural classes of proteins : αα, ßß, ß/α and ß+α. By virtue of the QUEST/PLANS meta-pattern capabilities, the pattern library was greatly extended and a high prediction accuracy was achieved (95% over all examples, 90% on homologous proteins extracted from the sequence database). On the debit side, this accuracy was achieved by the inclusion of many parameters based on global knowledge extracted from all available protein structures, so until a sufficiency of new structures becomes available on which to test the algorithm in an unbiased fashion, the true accuracy remains uncertain. Nevertheless, the current indications are that this method may very well provide an impressively powerful means for turn prediction (especially if the folding type of the protein can be independently determined).

1.5 Specific structural patterns

All the above methods are essentially generally applicable automatic methods for protein pattern analysis. Although some of the methods are more generally applicable than others, none of them attempt to apply rigid pattern matching rules based on comprehensive structural knowledge. Of course, pattern matching of this sort requires one of two things: either the structural unit is extremely simple, or the structural unit is fairly high-level (but then limited to a specific family of proteins).

The earliest analysis of a specific protein fold was that of the dehydrogenase nucleotide binding fold by Wootton (1974). The subject of nucleotide binding motifs has been repeatedly studied time after time. Well known work in this area includes the ATP binding patterns of Walker *et al.* (1982), and the ßαß dinucleotide binding motif of Wierenga and Hol (1983). Other 'macro-patterns' include calcium-binding patterns (for example the EF-hand helix-loop-helix pattern analyzed by Kretsinger, 1980), DNA binding patterns, and cyclic nucleotide binding patterns.

A library of such patterns is now being compiled in the form of the PROSITE database (Bairoch, 1991). To provide a useful tool for sequence and structure analysis, pattern matching *methods* must be linked to reliable and comprehensive libraries of pattern *data*, and PROSITE represents a first step in providing such a data resource. PROSITE is a well-documented library of semi-automatically generated protein sequence patterns (of the regular expression class), now integrated into the SWISS-PROT protein sequence database, and distributed by EMBL. Pattern libraries such as PROSITE will almost certainly play an important part in the identification and analysis of the vast number of sequences that will be produced as the various large-scale sequencing efforts begin to bear fruit.

Another approach to pattern analysis is to closely examine the sequence preferences for sub-structures at an extreme level of detail. This is the approach that has been followed by Thornton *et al.* (1988 - review) where the turns and loops between ß-sheets and α-helices have been rigorously classified into different structural groups, and then analyzed for sequence preferences at each position in the resulting patterns.

At the other end of the spectrum, Bashford, Chothia and Lesk (1987) have studied the pattern of sequence preference for a very large-scale sub-structure: the globin fold. Having comprehensively analyzed the structural features of the globin family, a pattern matching approach was use to correlate conserved sequence patterns with the known structural roles of each position in the fold. Conservation of amino-acid properties formed the core of the pattern method (comparable to the set method of Taylor, 1986a, 1986b). The resulting patterns derived from the globin analysis were as might be expected, highly specific to globins managing to accurately discriminate between globins and non-globins when applied to the available NBRF-PIR sequences. This method of laboriously constructing patterns manually from structural studies of protein families contrasts sharply with the automated approaches of for example Taylor (1986a). Both methods have both plus and minus points. Automatic procedures are obviously quicker, easier to use, and can work on families for which no structural information is available. Manual procedures allow the

problems of multiple alignment that can cause problems for the automatic methods to be circumvented, and produce patterns with a true structural significance.

1.6 Pattern assessment

The 'quality' of patterns is currently an ill-defined parameter. Certainly to date every pattern matching study has considered to a degree the sensitivity and specificity of the pattern under discussion, but few attempts have been made to quantify this information in a rational form. By sensitivity we mean the ability of a pattern to extract all the sequences used in its construction, and by specificity we mean the ability of a pattern to match the constituent sequences *and no others*. A straightforward approach is to consider a pattern as a collection of statistically independent strings of symbols, and to evaluate the expected frequency of occurrence of the total pattern. Hodgman (1989) shows how such probabilities can be calculated for patterns based on amino-acid property sets (e.g. Taylor, 1986a, 1986b).

An attempt at accurately quantifying the statistical significance of automatically generated protein sequence patterns forms an important part of the work by Smith and Smith (1990) on the PLSEARCH package. The automatic pattern method of Smith and Smith is a somewhat simpler version of the pattern generation scheme of Taylor (1986a). In place of a Venn diagram, in this method a strictly hierarchical grouping of amino-acids is used, subdividing the set of amino-acids into small mutually exclusive property sets. Whereas Taylor uses a seed alignment to initiate a cyclic alignment/clustering process, the method of Smith and Smith attempts to boot-strap the pattern creation by aligning each sequence in a clustered family of sequences to a rigid non-consensus pattern string, according to a binary tree ordered by the pairwise similarity score between each sequence and all others. The examples of the PLSEARCH pattern generation algorithm quoted in the paper are all relatively trivial, in that the alignments are fairly clear-cut. How such a rigid pattern generation scheme fairs when presented with highly divergent families of proteins

is an interesting question. The danger of misalignment is present when the alignment is performed without the use of consensus patterns. In a sense the method pre-empts the problem by using an insensitive clustering scheme to collect sequences together into families.

Rooman and Wodak (1988,1990) analyzed the predictive power of sequence motifs in an attempt to gain some insight into the current limitations in structure prediction accuracy. Their work involved the generation of short motifs that regularly occur in identical conformations. By extracting such motifs from different sized data-sets they surmise that the current lack of success for pattern based structural prediction is simply due to the limited size of the current structural database. Though within the scope of the study the results are valid enough, it may be unduly pessimistic in that it does not consider the possibilities of combinations of different prediction techniques. For example there is the possibility of using logic based expert knowledge to fill in gaps in the current database (e.g. Cohen *et al.*, 1983).

Following on from this, Rooman *et al.* (1989) analyzed the predictive power of recurrent turn motifs (Thornton *et al.*, 1988) in a similar way to that of Rooman and Wodak's study of automatically generated motifs. The general conclusion of the study was yet again disappointingly pessimistic showing their standalone predictive power to be poor. On a positive note the study demonstrates a sound statistically based method for validating consensus patterns, which is certainly an area of pattern generation and matching that must be further explored.

1.7 Prediction of progress at last[2]

Recently, some significant developments have been made with respect to pattern matching and protein structure. Two important new research avenues have been identified by a number of groups: the idea of *fold recognition*, and the extraction of secondary structural information from *aligned families of protein sequences.*

It is now well-known that proteins with no obvious sequence similarity can show remarkable similarities in their native folds. Examples of such proteins include the various TIM barrel enzymes, interleukin 1β/soybean trypsin inhibitor, and actin/hexokinase. The rate at which newly solved protein structures are perceived to have previously observed folds suggests that the number of protein topologies may be limited. Indeed, some estimates put the number of observed topologies at 50% of the total number of naturally occurring topologies. Given the significant possibility that a newly sequenced protein will have a previously observed fold, it is clearly useful to be able to recognize protein folds in sequences.

The methods described earlier for recognizing protein folds have been based purely on sequence information. These sequence analytic techniques work well for cases where some residual sequence similarity remains between the newly characterized protein and a protein of known 3-D structure. In the harder cases, such as those mentioned earlier, there is little or no sequence similarity with which to recognize proteins with similar folding patterns.

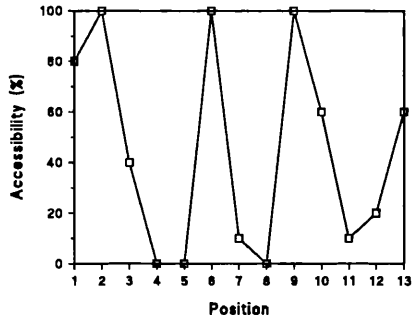
In the light of these limitations, a number of groups have developed new approaches to protein fold recognition which work by evaluating the compatibility between a test sequence and a library of structural patterns derived from either known crystal structures (Bowie *et al.*, 1991; Jones *et al.*, 1992b), or hypothetical model folds (Finkelstein and

[2] The title of a Nature "News and Views" by Thornton *et al.* (1991)

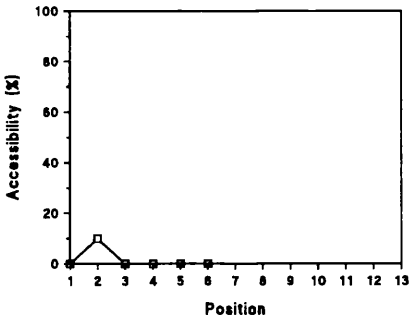
Reva, 1991). As much of the work described herein is concerned with this topic, further discussion will be left until later.

Often, rather than a single sequence a whole family of related sequences is available for analysis. By multiply aligning the sequence family, additional information may be obtained from the observed sequence conservation patterns, and the location of insertions and deletions. A prime example of the power of multiple sequence data was the successful secondary structure prediction of the cAMP-dependent kinases by Benner and Gerloff (1991). At the most basic level, the likely location of loop regions in the sequence data can be derived by observing where insertions and deletions occur. This process, which Benner and Gerloff call *parsing*, neatly circumvents the problems of traditional secondary structure prediction algorithms which are notably inaccurate in predicting the exact end-points of secondary structures. Further information can be obtained by observing that the most conserved regions of a protein sequence are those regions which are either functionally important, and/or buried in the protein core. Conversely, the more variable regions can be fairly confidently assumed to be on the surface of the protein, where few constraints are imposed on the nature of the amino acids in question, save a bias towards hydrophilic species. By clustering the sequences in an aligned family, and assessing the degree of sequence variability observed between very similar pairs, Benner and Gerloff demonstrate that the solvent accessibility (Lee and Richards, 1971; Chothia, 1976) of an amino acid residue can be predicted. Using these predicted degrees of accessibility, secondary structure can be predicted by comparing the accessibility patterns generally associated with specific secondary structures when packed against a hydrophobic protein core (Figure 1.2).

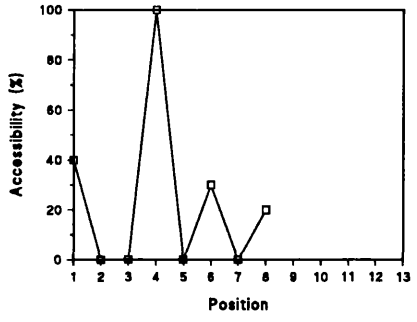
Limited topological information for proteins with a definite biological activity can be obtained from the observation that conserved polar residues tend to be functionally important, and thus close together forming part of the protein's active site.



Surface helix



Buried strand



Surface strand

Figure 1.2

Typical accessibility patterns for three common classes of secondary structural element. Buried helices infrequently occur, but are difficult to distinguish from buried strands.

Whilst it is clear that additional information is available from analyzing multiple sequence alignments, the quality and quantity of this information remain ill-determined. The fact that many very similar protein sequence pairs are required to apply the method reliably is born out by the lack of success in more recent predictions (Benner *et al.* 1992; Robson & Garnier, 1993), which have been performed on families with far fewer close pairs than that of the cAMP dependent kinases.

With the work now being done to sequence entire genomes, the imbalance between the number of known sequences and the number of known structures will become greater over the next few years. Methods such as those described in this introduction and perhaps those

developed as part of the project described in the next few chapters will play a vital role in the utilization of the available data. One day it is hoped that we will have a full understanding of just how proteins fold into a uniquely determined 3-D structure, but even when we have this knowledge, pattern matching methods will still be important tools.

Chapter 2

The Calculation of New Amino Acid Comparison Matrices

It is comparison that makes men miserable.

- Proverb

2.1 Introduction

Despite the great diversity of methods devised for the alignment and comparison of protein sequences, all of these depend at some point on the simple comparison of two amino acid residues. The most popular method for measuring the similarity between amino acids is to use a scoring matrix of some form. At its simplest, a typical scoring matrix comprises 20 x 20 elements, each element representing some metric that relates two residues. In view of this central dependence on a scoring matrix, a reasonable route towards possibly increasing the sensitivity of standard sequence analysis methods is to investigate improvements to the matrix itself, rather than the algorithms. It was this assumption that instigated the work presented in this chapter.

2.2 Types of scoring matrix

The least sophisticated matrix is the 'Unitary Protein Matrix' (UPM), also known as the 'identity matrix'. The UPM scores a 1 for exactly matching residues and a 0 for every other combination. Obviously this matrix lacks sensitivity, being unable to detect the possibility of phenotypically "quiet" mutational events between two sequences. One advantage of the UPM is that it is wholly unbiased, providing a very easily understood alignment metric. The 'percentage identity' between two sequences is often offered as a universal means of describing the mutual degree of 'homology' between them. Though a low identity score can in no way prove or disprove the existence of homology, it has proved easier to provide rules of thumb for identity scoring than for any other scheme. In general for two sequences of reasonable length (say 50 residues or more), a percentage identity of greater than 20% points to a significant structural homology between them. Doolittle *et al.* (1986) have described a fuzzy region around 20% identity which they call the 'Twilight Zone'. Within this zone and below, it is not possible to tell the difference between real sequence similarity implying a common structural framework, and accidental similarity providing no useful structural information.

Probably the next simplest amino acid scoring matrix is the 'Genetic Code Matrix' (GCM). This matrix scores amino acid similarity by the maximum number of common nucleotide bases between their closest matching representative codons. Identical residues of course share a maximum of 3 bases, whereas non-identical residues may have only 0, 1 or 2 bases in common. This matrix has a pleasantly 'genetic flavour' to it, but it must be realized that the bulk of the selection pressure is on the protein sequence and not on the underlying DNA sequence. Although there does seem to be a reasonable correlation between the nucleotide codons associated with amino acids and the degree of chemical similarity between them (Wolfenden *et al.*, 1979, for example), the rather limited range of match-scores puts the GCM somewhat in the shade. To detect weak homologies between sequences a more accurate amino acid comparison table is required.

McLachlan (1972) published a scoring matrix that attempted to explicitly quantify the degree of chemical similarity between amino acids. This matrix, known as the Structure-Genetic Matrix (SGM), incorporated two sources of information in evaluating the similarities of amino acids. The first source was a statistical analysis of observed amino acid exchanges in available families of proteins, the second was from the assignment of transition values for each pair of amino acids depending on the number of overlapping physico-chemical properties between them. These data were used to 'bias' the UPM in such a way that only 20 of the 190 possible interchanges were significantly preferred (Feng, Johnson and Doolittle, 1985). The problem with the SGM and other matrices that attempt to incorporate 'real' amino acid similarities is that the groupings used are artificial, there is no guarantee that an arbitrary common amino acid property is at all important for structural and functional conservation between proteins. A better approach is to concentrate on the observed exchanges between amino acids in very similar aligned sequences. Evidently amino acids that share the *appropriate* properties will exchange more frequently than ones that do not. McLachlan's earlier attempt to compare amino acids (McLachlan, 1971) was based entirely on such a statistical approach.

Recently, matrices based on the principles of structural comparison have been described (Risler *et al.*, 1988; Overington *et al.*, 1990, 1992). These matrices essentially contain statistics on all N^2 pairwise substitutions observed at equivalent positions in structurally aligned families of proteins. In the case of Overington *et al.*, a range of matrices is calculated, one from each class of structural environment, an example of one such class being 'buried coil' for example. These matrices show great promise in increasing the accuracy of sequence-to-sequence, and sequence-to-structure alignments, though the sparsity of structural data presently available is a significant disadvantage of this approach.

The most widely used comparison matrix today is the Log-Odds Matrix and the very closely related Mutation Data Matrix (MDM) published by Dayhoff *et al.* (1968, 1972, 1978). The MDM was calculated from a study of the exchange probabilities (or odds) derived from an analysis of the evolutionary changes seen in groups of very similar proteins. A strictly Markovian model (i.e. the current probabilities are independent of previous events) of amino acid exchange is assumed in the Dayhoff model. This model has been criticized (see George *et al.*, 1990, for a review), but comparisons of different scoring schemes have tended to hesitantly recommend the MDM over other matrices (Feng *et al.*, 1985).

In this chapter a straightforward and automatic procedure for generating mutation data matrices is presented, in order that very large sets of sequences may be processed without using inordinate amounts of computing resources. In particular using this method it is possible to improve the generality of the MDM, in that there is now access to a much greater variety of protein sequences than that available to Dayhoff and her co-workers in 1978, and it is hoped that the matrices presented here will more clearly express the general nature of the underlying amino acid similarities.

The original mutation data matrix (MDM68) was presented in the original *Atlas of Protein Sequence and Structure* (1968), and the method (outlined below) remained virtually

unaltered through each of the subsequent updates. There are 5 main steps required for the creation of a mutation data matrix, which will be briefly described.

2.3 Construction of the raw PAM matrix

The basic unit of molecular evolution expressed in a MDM is the *Accepted Point Mutation* or with a little licence to ease pronunciation: PAM. One PAM is simply the mutation of a single amino acid in a sequence such that the new amino acid may be accommodated in the structure and function of the protein. In general therefore amino acid residues that are frequently seen to exchange in a PAM matrix typically have similar physico-chemical properties.

The raw PAM substitution matrix is created by considering the possible mutational events that could have occurred between two closely related sequences. Ideally we would like compare every present day sequence with its own immediate predecessor and thus accurately map the evolutionary history of each sequence position. Of course this is impossible, and so two main courses of action may be taken to approximate this information. The method used by Dayhoff was the *common ancestor* method. Here closely homologous pairs of present day sequences are taken and a common ancestral sequence inferred. Given only a pair of present day sequences, an unambiguous inferred common ancestor cannot be generated. A complete phylogenetic tree is required in this case to allow the *most probable* common ancestors to be inferred for each tree node. The important thing to realize is that the inference of common ancestors must consider the overall topology of the tree. Every suggested common ancestor must be traced-back to higher level nodes and evaluated in order to determine whether or not that ancestral sequence is the most probable for the tree as a whole.

An alternative to the common ancestor method is to relate present day sequences by their pairwise alignment distances, estimating a possible phylogenetic tree from this distance

matrix. This method was first described by Fitch and Margoliash (1967). Though construction of the distance matrix is a trivial exercise, the generation of an *optimal* phylogenetic tree from this data again requires an exhaustive iterative analysis such that the total number of mutations required to produce the present day set of sequences is minimized. Though both of the above methods have advantages and disadvantages, matrix methods are now most widely used.

No matter which method is finally used to infer the phylogenetic tree, construction of the PAM matrix is the same. The raw matrix is generated by taking pairs of sequences, either a present day sequence and its inferred ancestor, or two present day sequences, and tallying the amino acid exchanges that have apparently occurred. Given the following alignment:

ACDEFL
AGDEAL

we count four PAMs (C→G, G→C, F→A and A→F). The raw PAM matrix is obviously symmetric given the fact that we cannot know whether for example C mutated to G or G mutated to C, there is no harm in this as we are interested in discerning the extent of similarity between amino acids here, 'similarity' is generally thought of as being symmetric. Treatment of gaps/insertions in an alignment is arbitrary, one possibility is to count gap characters as another type of amino acid, another possibility that is probably the safer of the two is to simply ignore gaps. We are after all only interested in the *exchange* of amino acids, the deletion of a particular amino acid tells us nothing of its relative similarity to other amino acids, though it does provide information as to the amino acid's characteristic 'mutability'.

2.4 Calculation of relative mutabilities

Evidently if we are to estimate the probability of a given mutational event, we must consider two pieces of information. Firstly how likely is it that a given amino acid A

changes at all, secondly how likely is it that the given amino acid changes to amino acid B given that A *does* change. We are therefore interested in the conditional probability that amino acid A changes to amino acid B given that A is seen to change. The probability of amino acid A changing at all in a given unit of time is usually expressed as the *relative mutability* of A. Relative mutability is simply calculated as the number of observed changes of an amino acid divided by its frequency of occurrence in the aligned sequences. From the alignment show earlier, A is seen to change once, but occurs 3 times in the alignment. The relative mutability of A from this alignment alone is therefore calculated as $^1/_3$. An overall measure of relative mutability must allow for the different evolutionary distances and different sequence lengths found in a non-specific collection of sequences. Mutability is normalized by defining the basic unit of evolutionary distance as being a single accepted point mutation in a sequence of length 100. The average relative mutability of an amino acid given this definition is therefore the total number of changes observed for this amino acid in *all* the families of proteins considered, divided by the total sum of all local frequencies of occurrence of the amino acid multiplied by the numbers of mutations per 100 residues in each of the branches of all the family trees.

2.5 Calculation of the mutation probability matrix

The basic matrix in the generation of MDM type matrices is the *mutation probability matrix.* Elements of this matrix give the probability that a residue in column *j* will mutate to the residue in row *i* in a specified unit of evolutionary time. Evidently a diagonal

element of this matrix represents the probability of residue $i=j$ remaining unchanged, and hence being easily calculated according to the following formula:

$$M_{ji} = 1 - \lambda m_j$$

where

m_j is the average relative mutability of residue j, and

λ is a proportionality constant.

Nondiagonal elements are given by:

$$M_{ij} = \frac{\lambda m_j A_{ij}}{\sum_i A_{ij}}$$

where

A_{ij} is a (nondiagonal) element of the raw PAM matrix

The value of λ relates to the evolutionary distance represented by the probability matrix, accordingly:

$$\sum_i f_i M_{ii} = 1 - \frac{P}{100}$$

where

f_i is the normalized frequency of occurrence of residue i, and

P approximates the evolutionary distance (in PAMs) represented by the matrix. This relationship breaks down for $P \gg 5$.

P is usually given the value 1 so that the basic mutation probability matrix represents a distance of 1 PAM. Matrices representing larger evolutionary distances may be derived from the 1 PAM matrix by matrix multiplication. Squaring the 1 PAM matrix gives a 2 PAM matrix, cubing it a 3 PAM matrix and so forth.

2.6 Calculating the log-odds matrix

Of more use than the mutation probability matrix in the alignment of protein sequences is the *relatedness odds matrix.* This symmetric matrix represents the probability of residue j being replaced by residue i per occurrence of i, and is derived from the mutation probability matrix simply by dividing each element M_{ij} by the normalized frequency of occurrence of i, f_i. For the purposes of sequence comparison the relatedness odds for each alignment position should be multiplied together in order to arrive at a total 'alignment odds' value. To avoid slow floating point multiplications, the relatedness odds matrix is usually converted to the log odds-matrix (also known as the Mutation Data Matrix) thus:

$$MDM_{ij} = 10 \log_{10} R_{ij}$$

where

R_{ij} are elements of the Relatedness Odds Matrix

(MDM_{ij} values are rounded to the nearest integer)

2.7 Automating the procedure

Although computational tools were used in constructing the original MDMs, in particular for the inference of common ancestral sequences and the generation of phylogenetic trees, the whole process was only partially automated. This was hardly of consequence considering the small number of available sequences in the 1970s, but as at the time of writing some 30000 protein sequences are available for analysis it is evident that a more streamlined approach is now required.

The method described here for generating MDMs is in fact very similar in essence to that described by Dayhoff *et al.* (1978). The method involves three steps: a) clustering the sequences into homologous families, b) tallying the observed mutations between highly similar sequences, and c) relating the observed mutation frequencies to those expected by

pure chance. The main difference here is in the use of an approximate method (a pairwise present-day ancestor scheme) for inferring the phylogenetic relationships amongst the sequences in the data set. A program was written to compute all the relevant data automatically from a file of protein sequences.

In view of the relative inefficiency of standard methods for inferring maximum parsimony phylogenetic trees it was found to be necessary to implement an approximate method to find the reasonable family trees by means of cluster analysis of the sequence data. Though the limitations of using such simple means alone for the inference of phylogenetic trees are well known (Czelusniak *et al.,* 1990), and the large-scale structure of such crude phylogenetic trees tends to be somewhat incorrect, the relationships between closely related sequences are inferred correctly. To verify the methodology, an attempt was made to recreate the set of sequences used to construct MDM78. Using these sequences it was found that this mutation data closely approximated those in the original work with 164 of the 400 mutation frequencies (number of mutations occurring per 10000 observations) being identical, and 350 differing by 5 or less. It should be pointed out that though these results very closely match those of Dayhoff *et al.*, the matrices here are not derived from the same explicit evolutionary model outlined in the original work. The practical significance of this fact depends on the intended application of the matrices. In terms of sequence analysis applications, a derivation independent of the choice of evolutionary model might well be preferred due to the reduced possibility of bias (in particular, maximum parsimony nucleotide substitution methods will tend to produce results biased towards the exchanges expected from the genetic code rather than generally observed amino acid similarities). A further justification for determining relationships via a pairwise scheme is that of the 2621 families of proteins in the current release of SWISS-PROT, 79% contain fewer than 5 sequences. With such small families the results of simple clustering and those of rigorous maximum parsimony analysis are indistinguishable with respect to the present application.

In generating the initial distance matrix, no assumption is made that the input sequences are in any way pre-clustered into family groups, and are therefore forced to calculate the entire distance matrix to sort the sequences into families, and thereafter produce trees for each family. Evidently the vast majority of pairwise comparisons are unnecessary, so some simple (and quick) means is needed to filter out sequence pairs that have no chance of producing alignment identity scores > 85%.

Dynamic programming methods, though inherently rigorous, are certainly not efficient means for comparing biological sequences. Less rigorous, yet much faster, methods based on the detection of common subsequences between two strings have been developed notably by Wilbur and Lipman (1983) and Lipman and Pearson (1985). These methods have their roots in the earlier dot-plot methods for visually determining whether or not sequence similarity exists, for example see Staden (1982) for a review. These common subsequence, or 'tuple' methods as they are more commonly known, are very efficient being linearly dependent on the sum of the lengths of the two sequences under comparison. This efficiency is mostly provided by the use of hashing techniques where the two tuple lists are indexed by a unique integer derived from the tuples themselves. This means that the presence of a particular tuple (the 3-tuple DSG for example) in one of the sequences may be determined by a single index calculation and subsequent array access.

Proposed here is a simple approximate algorithm for 'estimating' the percentage identity between two protein sequences without prior alignment. The method described here conceptually involves the comparison of two histograms, where these histograms depict the distributions of 3-tuples in the two sequences being considered. Obviously two identical sequences will have identical 3-tuple distributions, and so the histograms will exactly match. As the number of differences between the two sequences increases the degree of overlap between the two 3-tuple histograms will decrease. For two random sequences there will of course always be some overlap between the two corresponding

histograms, but provided that there are sufficient identical triplets between both sequences it may be assumed that the sequences show a *potential* homology.

This approximation algorithm can be coded in just 20 lines of standard C, and runs very quickly. To begin, the longest sequence is taken and a hash table constructed containing the frequencies of occurrence of the constituent triplets. The triplet frequencies of the shorter sequence are then compared with those of the longer. A comparison score is calculated thus:

$$S = \frac{\sum_{pqr=AAA}^{VVV} MIN\left(f_a^{pqr} , f_b^{pqr} \right)}{MIN\left(n_a , n_b \right) - 2}$$

where f_a^{pqr} and f_b^{pqr} are the frequencies of occurrence of triplet pqr in sequence a and b, and n_a and n_b are the respective sequence lengths.

This normalized score (S) is effectively the fractional area of overlap between the two triplet histograms. Scatter plots based on all possible pairwise alignment scores in a set of 200 protein sequences (containing a mixture of related and unrelated sequences) plotted against the proposed scoring metric were produced (a subset of this data is shown plotted in Figure 2.1). The raw triplet scores were thus compared with Needleman-Wunsch scores (above 40% ID), and the following relationship (correlation coefficient 0.986) was observed:

$$I \approx 100 S^{0.3912}$$

where

S is the normalized triplet frequency score, and

the result I is in units of percentage identity.

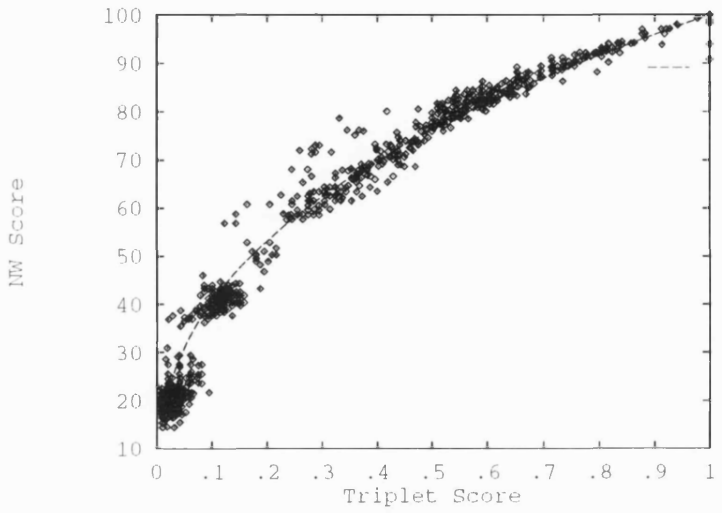


Figure 2.1

A plot showing the relationship between the heuristic triplet similarity score, and the scores obtained from a number of rigorous alignments.

By aligning only those sequence pairs with corrected triplet scores indicating sequence identity ≥ 45% and subsequently excluding sequence pairs with alignment scores of ≤ 85% identity it was possible to rapidly generate a sparse distance matrix complete enough for these purposes. By combining this very rapid heuristic measure of identity with an efficiently coded dynamic programming algorithm as a 'second level filter' construction of the distance matrix proceeded at an average rate of over 1000 similarity score calculations per second on a Sun 4/280 (standard Sun C compiler). Out of the 130 million pairwise alignments that would normally be required, only 559692 passed the initial similarity filter, speeding up the process nearly 200 fold. The overall procedure is illustrated in Figure 2.2.

Using this matrix of identity scores, the sequences were subjected to an efficient single-linkage clustering algorithm, with mutation statistics being generated for each sequence by aligning it with the sequence that offers the highest pairwise alignment score. For each sequence pair, amino acid substitutions are tallied, with alignment positions containing at least one non-standard residue code (B,Z,X or 'Gap') being ignored.

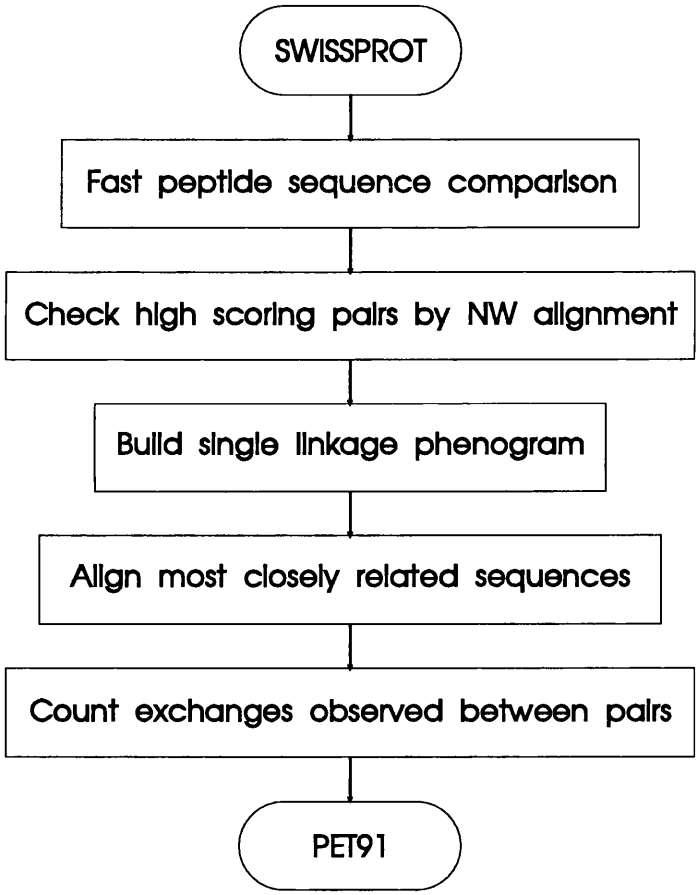


Figure 2.2

An outline of the described method for generating mutation data matrices.

Dynamic programming based sequence alignment algorithms have been improved by many workers since its initial conception. Gotoh (1982) recoded the algorithm to trade space efficiency for time efficiency, resulting in a pairwise alignment algorithm dependent on the product of the two sequence lengths. This method is now considered to be the standard algorithm. Further efficiency may be gained by ignoring improbable alignment paths that would require very large gaps to be inserted. This is achieved by windowing the initial score matrix, thus restricting the alignment path search to biologically plausible regions (Kruskal & Sankoff, 1983). Taylor (1988a,1990) has coded an efficient global alignment routine as part of the multiple alignment program, MULTAL, and a modified form of this code is used to perform the final pairwise alignments prior to counting the observed exchanges.

2.8 Program implementation

The matrix generation program MAKEPET is coded in standard Sun C, and should be portable to most platforms supporting a C compiler. The required matrix PAM distance and other control parameters are specified as command line arguments. MAKEPET takes as input a single file of sequences in 'compact PIR' format, where each sequence is preceded by two description lines and terminated by a '*' character. A simple keyword searching program SEQGREP allows specific sets of sequences to be compiled from the complete sequence databank, permitting the easy generation of matrices biased towards particular structural or functional classes (membrane-bound proteins for example).

2.9 Results

The upper half of Table 2.1 shows how many of each of the possible 190 exchanges were observed, with the lower half of Table 2.1 showing the equivalent of the widely used MDM78 matrix ($\log_{10}$ relatedness-odds matrix for 250 PAMs), which is called PET91

(Pairwise Exchange Table 1991). The 1 PAM mutation probability matrix required to generate mutation data matrices for evolutionary distances other than 250 PAMs is shown in Table 2.2. PET91 was generated from Release 15.0 of the SWISS-PROT protein sequence database (Bairoch & Boeckmann, 1991), containing 16941 sequences, though sequences shorter than 20 residues were excluded to avoid insignificant alignments. It should be noted that the 250 PAM matrix is shown here for reasons of comparison with the most common variant of the original matrix, and that matrices calculated for evolutionary distances other than 250 PAMs are often found to perform better for some sequence comparisons. The sequence databank search program, BLAST (Altschul *et al.*, 1990), for example, uses a 120 PAM Dayhoff matrix by default.

Of particular interest here are the differences between these results and those of the original work, a rough impression of which may be gained from a comparison of the relative mutabilities shown in Table 2.3 with those observed by Dayhoff (1978). A value of 0.76 is obtained for the Spearman rank correlation coefficient between the old and new relative mutabilities, indicating little overall change. Ser (serine) and Thr (threonine) are found to be the most mutable residues in this work, as opposed to asparagine and serine in the 1978 table. Trp (tryptophan) and Cys (cysteine) are found to be least mutable here, which agrees with the earlier findings, though the mutability of Cys found here is double the original value. The frequencies of occurrence of the amino acid residues (Table 2.3) show no significant differences from the earlier values.

Table 2.1

The 250 PAM PET91 Matrix ($\log_{10}$ relatedness odds), based on 59190 accepted point mutations found in 16130 protein sequences. Values have been multiplied by 10 and rounded to the nearest integer. The upper half of the matrix shows the actual numbers of exchanges observed.

Table 2.2

PET91 Mutation Probability Matrix for an evolutionary distance of 1 PAM. Values are scaled by a factor of 10^5. Elements of this matrix give the probability that a residue in column j will mutate to the residue in row i in an evolutionary distance of 1 PAM. A diagonal element of this matrix represents the probability of residue i=j remaining unchanged.

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	2	247	216	386	106	208	600	1183	46	173	257	200	100	51	901	2413	2440	11	41	1766
R	-1	5	116	48	125	750	119	614	446	76	205	2348	61	16	217	413	230	109	46	69
N	0	0	3	1433	32	159	180	291	466	130	63	758	39	15	31	1738	693	2	114	55
D	0	-1	2	5	13	130	2914	577	144	37	34	102	27	8	39	244	151	5	89	127
C	-1	-1	-1	-3	11	9	8	98	40	19	36	7	23	66	15	353	66	38	164	99
Q	-1	2	0	1	-3	5	1027	84	635	20	314	858	52	9	395	182	149	12	40	58
E	-1	0	1	4	-4	2	5	610	41	43	65	754	30	13	71	156	142	12	15	226
G	1	0	0	1	-1	-1	0	5	41	25	56	142	27	18	93	1131	164	69	15	276
H	-2	2	1	0	0	2	0	-2	6	26	134	85	21	50	157	138	76	5	514	22
I	0	-3	-2	-3	-2	-3	-3	-3	-3	4	1324	75	704	196	31	172	930	12	61	3938
L	-1	-3	-3	-4	-3	-2	-4	-4	-2	2	5	94	974	1093	578	436	172	82	84	1261
K	-1	4	1	0	-3	2	1	-1	1	-3	-3	5	103	7	77	228	398	9	20	58
M	-1	-2	-2	-3	-2	-2	-3	-3	-2	3	3	-2	6	49	23	54	343	8	17	559
F	-3	-4	-3	-5	0	-4	-5	-5	0	0	2	-5	0	8	36	309	39	37	850	189
P	1	-1	-1	-2	-2	0	-2	-1	0	-2	0	-2	-2	-3	6	1138	412	6	22	84
S	1	-1	1	0	1	-1	-1	1	-1	-1	-2	-1	-1	-2	1	2	2258	36	164	219
T	2	-1	1	-1	-1	-1	-1	-1	-1	1	-1	-1	0	-2	1	1	2	8	45	526
W	-4	0	-5	-5	1	-3	-5	-2	-3	-4	-2	-3	-3	-1	-4	-3	-4	15	41	27
Y	-3	-2	-1	-2	2	-2	-4	-4	4	-2	-1	-3	-2	5	-3	-1	-3	0	9	42
V	1	-3	-2	-2	-2	-3	-2	-2	-3	4	2	-3	2	0	-1	-1	0	-3	-3	4

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	98759	27	24	42	12	23	66	129	5	19	28	22	11	6	99	264	267	1	4	193
R	41	98962	19	8	21	125	20	102	74	13	34	390	10	3	36	69	38	18	8	11
N	43	23	98707	284	6	31	36	58	92	26	12	150	8	3	6	344	137	0	23	11
D	63	8	235	98932	2	21	478	95	24	6	6	17	4	1	6	40	25	1	15	21
C	44	52	13	5	99450	4	3	41	17	8	15	3	10	28	6	147	28	16	68	41
Q	43	154	33	27	2	98955	211	17	130	4	64	176	11	2	81	37	31	2	8	12
E	82	16	25	398	1	140	99042	83	6	6	9	103	4	2	10	21	19	2	2	31
G	135	70	33	66	11	10	70	99369	5	3	6	16	3	2	11	129	19	8	2	32
H	17	164	171	53	15	233	15	15	98867	10	49	31	8	18	58	51	28	2	189	8
I	28	12	21	6	3	3	7	4	4	98722	212	12	113	31	5	28	149	2	10	630
L	24	19	6	3	3	29	6	5	12	122	99328	9	90	101	53	40	16	8	8	117
K	28	334	108	14	1	122	107	20	12	11	13	99101	15	1	11	32	57	1	3	8
M	36	22	14	10	8	19	11	10	8	253	350	37	98845	18	8	19	123	3	6	201
F	11	3	3	2	14	2	3	4	11	41	230	1	10	99357	8	65	8	8	179	40
P	150	36	5	7	3	66	12	16	26	5	97	13	4	6	99278	190	69	1	4	14
S	297	51	214	30	44	22	19	139	17	21	54	28	7	38	140	98548	278	4	20	27
T	351	33	100	22	9	21	20	24	11	134	25	57	49	6	59	325	98670	1	6	76
W	7	65	1	3	23	7	7	41	3	7	49	5	5	22	4	21	5	99684	24	16
Y	11	12	30	23	43	10	4	4	134	16	22	5	4	222	6	43	12	11	99377	11
V	226	9	7	16	13	7	29	35	3	504	161	7	71	24	11	28	67	3	5	98772

	Relative Mutability* (1991)	Relative Mutability* (1978)	Relative Frequency of Occurrence (1991)	Relative Frequency of Occurrence (1978)
Ala (A)	100	100	0.077	0.087
Arg (R)	83	65	0.051	0.041
Asn (N)	104	134	0.043	0.040
Asp (D)	86	106	0.052	0.047
Cys (C)	44	20	0.020	0.033
Gln (Q)	84	93	0.041	0.038
Glu (E)	77	102	0.062	0.050
Gly (G)	50	49	0.074	0.089
His (H)	91	66	0.023	0.034
Ile (I)	103	96	0.053	0.037
Leu (L)	54	40	0.091	0.085
Lys (K)	72	56	0.059	0.081
Met (M)	93	94	0.024	0.015
Phe (F)	51	41	0.040	0.040
Pro (P)	58	56	0.051	0.051
Ser (S)	117	120	0.069	0.070
Thr (T)	107	97	0.059	0.058
Trp (W)	25	18	0.014	0.010
Tyr (Y)	50	41	0.032	0.030
Val (V)	98	74	0.066	0.065

* Relative to Ala which is arbitrarily assigned a mutability of 100.

Table 2.3

Relative mutabilities and normalized frequencies of occurrence for the 20 amino acid residues, calculated from the PET91 data set, compared with the values from Dayhoff *et al.* (1978).

Table 2.4 shows the pattern of changes between the MDM78 and the PET91 matrices. Both Cys and Trp show very different patterns of mutability, both now showing a much greater tendency to exchange with other amino acid residues than in the previous study. This can be attributed mainly to the paucity of mutational events involving Cys and Trp in the original data set. Overall, in Dayhoff's data 35 amino acid exchanges were never observed at all (for example Cys and Trp), here however all possible exchanges have been observed (Cys and Trp exchanging 38 times in the current data set). PET91 incorporates 442 Trp exchanges and 1292 Cys exchanges, whereas only 7 Trp exchanges and 28 Cys exchanges were recorded for the MDM78 matrix. Interestingly, however, the average absolute change of the Cys matrix elements is higher than that of Trp, even though the Cys sample was larger than that of Trp in the 1978 data set. This anomaly is attributable to the fact that Cys residues occur in three very different chemical roles in proteins: as free sulphydryl groups (-S-H), in disulphide bridges (-S-S-), and as ligands for metals (-S..X). The number of observed Cys exchanges in the original work would have been insufficient to effectively sample these three situations. In addition, the Cys residue exchanges observed in the original work were mostly from the metallothionein sequences included in the data set.

It is also interesting to note that even with the very large amount of data collected here, some amino acid exchanges are still very seldom observed: Trp and Asn (asparagine) for example were only seen to exchange twice. Indeed it is hard to be certain whether these highly infrequent exchanges are real observations or artefacts caused by errors in the sequence database.

A common method for interpreting the complex trends in a similarity matrix is to project the 20x20=400-dimensional pattern onto a plane via multidimensional scaling (French & Robson, 1983; Taylor, 1986b; Taylor & Jones, 1993). The plot in Figure 2.3 shows such a projection, which clearly delineates the relationships between the 20 amino acids found in PET91. The general trends shown in the PET matrix are essentially those found in the original Dayhoff matrix: hydrophobicity and size being the most significant factors.

2.10 Summary

In general, the most significant differences (PET91 matrix elements differing from MDM78 elements by +/- 2 or more) correspond almost exactly to exchanges that were observed no more than once in Dayhoff's sequence alignments. Despite these few anomalous differences, however, it is interesting (if somewhat disappointing) to see how little the bulk of PET91 differs from MDM78. The fundamental amino acid similarities remain unchanged, and given that enough data has now been collected to iron out the residual sampling errors in the mutation data matrix, it is possible to feel confident that PET91 represents a relatively unbiased measure of amino acid similarity in sequence data and should be used in preference to the MDM78 in sequence analysis applications.

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	0	+1	0	0	+1	-1	-1	0	-1	+1	+1	0	0	+1	0	0	+1	+2	0	+1
R	+1	-1	0	0	+3	+1	+1	+3	0	-1	0	+1	-2	0	-1	-1	0	-2	+2	-1
N	0	0	+1	0	+3	-1	0	0	-1	0	0	0	0	+1	0	0	+1	-1	+1	0
D	0	0	0	+1	+2	-1	+1	0	-1	-1	0	0	0	+1	-1	0	-1	+2	+2	0
C	+1	+3	+3	+2	-1	+2	+1	+2	+3	0	+3	+2	+3	+4	+1	+1	+1	+9	+2	0
Q	-1	+1	-1	-1	+2	+1	0	0	-1	-1	0	+1	-1	+1	0	0	0	+2	+2	-1
E	-1	+1	0	+1	+1	0	+1	0	-1	-1	-1	+1	-1	0	-1	-1	-1	+2	0	0
G	0	+3	0	0	+2	0	0	0	0	0	0	+1	0	0	0	0	-1	+5	+1	-1
H	-1	0	-1	-1	+3	-1	-1	0	0	-1	0	+1	0	+2	0	0	0	0	+4	-1
I	+1	-1	0	-1	0	-1	-1	0	-1	-1	0	-1	+1	-1	0	0	+1	+1	-1	0
L	+1	0	0	0	+3	0	-1	0	0	0	-1	0	-1	0	+3	+1	+1	0	0	0
K	0	+1	0	0	+2	+1	+1	+1	+1	-1	0	0	-2	0	-1	-1	-1	0	+1	-1
M	0	-2	0	0	+3	-1	-1	0	0	+1	-1	-2	0	0	0	+1	+1	+1	0	0
F	+1	0	+1	+1	+4	+1	0	0	+2	-1	0	0	0	-1	+2	+1	+1	-1	-2	+1
P	0	-1	0	-1	+1	0	-1	0	0	0	+3	-1	0	+2	0	0	+1	+2	+2	0
S	0	-1	0	0	+1	0	-1	0	0	0	+1	-1	+1	+1	0	0	0	-1	+2	0
T	+1	0	+1	-1	+1	0	-1	-1	0	+1	+1	-1	+1	+1	+1	0	-1	+1	0	0
W	+2	-2	-1	+2	+9	+2	+2	+5	0	+1	0	0	+1	-1	+2	-1	+1	-2	0	+3
Y	0	+2	+1	+2	+2	+2	0	+1	+4	-1	0	+1	0	-2	+2	+2	0	0	-1	-1
V	+1	-1	0	0	0	-1	0	-1	-1	0	0	-1	0	+1	0	0	0	+3	-1	0

Table 2.4

The difference matrix ($PET91_{ij}$ - $MDM78_{ij}$) between the 250 PAM PET91 matrix and the MDM78 matrix. A positive matrix element indicates that the PET91 value is higher than the related value in MDM78. Absolute differences greater than 2 are shown shaded.

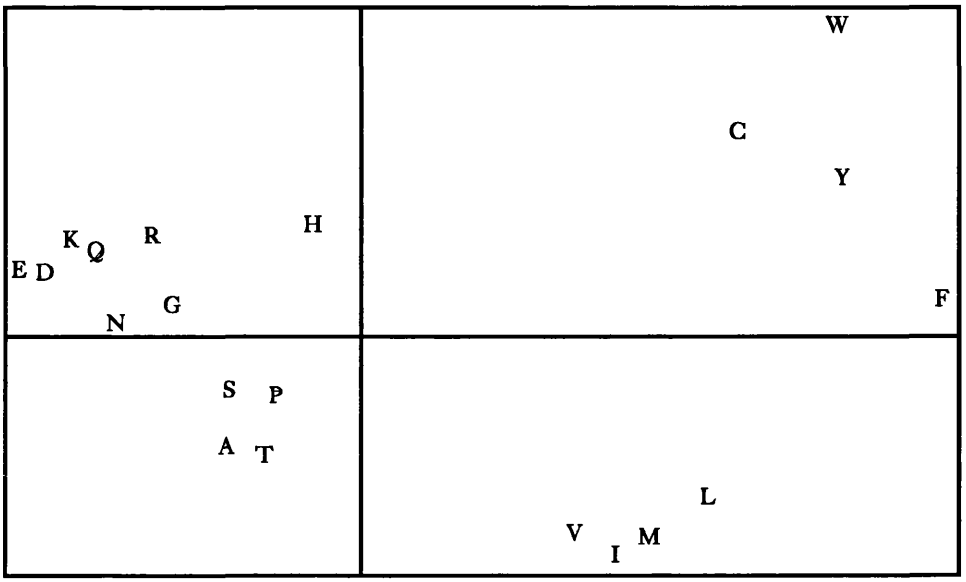


Figure 2.3

The general trends in amino acid residue similarity shown in the PET91 relatedness odds matrix, visualized by means of multidimensional scaling.

2.11 Alignment parameter optimization

To successfully use any given amino acid scoring matrix in a typical sequence alignment application, the control parameters for the algorithm in question must be tailored to the matrix. Generally, the most important control parameters in alignment algorithms pertain to the penalty imposed on the insertion of gaps in the alignment. The most common form of the gap penalty function giving the total gap penalty p is a simple linear expression:

$$p = mk + c$$

where k is the length of the proposed gap, m is the gap extension penalty (length dependent term), and c is the gap initiation penalty (length independent term). The choice of a gap penalty function is fairly arbitrary, and a linear gap penalty function is mostly selected for convenience, though it is important to note that for the common Gotoh (1982) implementation of the NW algorithm the gap penalty must either be constant or at least monotonically increase with the length of the gap.

A standard way to determine the best values of m and c, is to take a number of sequence pairs for which the optimum alignments are known, align them using a range of gap penalty terms and then compare the resulting alignments with the correct ones. Determination of the optimum alignment can be a rather subjective process, unless the structures of the proteins in question are known, in which case the alignments obtained through optimal structure alignment can be used as references. This strategy was used here. The pairs of protein chains used are listed in Table 2.5. The pairs selected all have significant sequence similarity, though not exceeding 35%, which ensures that the alignment problems are non-trivial.

Reference alignments in each case were derived by structural alignment using the program SSAP (Taylor and Orengo, 1989; Orengo and Taylor, 1990).

Table 2.6 and Figure 2.4 show the results of searching for optimal values for the gap-extension parameter *m* and the gap-initiation parameter *c*. For each alignment the number of residues equivalenced in both the structurally derived alignment and the sequence alignment was tallied. These individual values were summed, and the two totals ($N_{seq.}$ and $N_{struc.}$) divided and scaled by 100 to give a percentage:

$$100\frac{N_{seq.}}{N_{struc.}}\%$$

A score of 100% would indicate that every sequence alignment was in perfect accord with the appropriate structural alignment.

The highest accuracy (83.63%) was obtained with the values *m=1* and *c=12* i.e. a penalty of 12 for initiating a gap and 1 for each gap position. An equivalent analysis using the MDM78 matrix produced a maximum accuracy of 81.36% using the values *m=1* and *c=14*.

PDB Code 1	Description	PDB Code 2	Description
2FB4 (L)	Immunoglobulin FAB (human)	2FB4 (H)	Immunoglobulin FAB (Human)
1MBA	Myoglobin (sea hare)	2LHB	Hemoglobin V (sea lamprey)
1PAZ	Pseudoazurin (Alcaligenes Faecalis)	7PCY	Plastocyanin (green algae)
1YCC	Cytochrome c (yeast)	3C2C	Cytochrome c2 (Rhodospirillum Rubrum)
3DFR	Dihydrofolate reductase (Lactobacillus Casei)	5DFR	Dihydrofolate reductase (E. Coli)
4FXN	Flavodoxin (Clostridium MP)	1FX1	Flavodoxin (Desulfovibrio Vulgaris)
4PTP	β-Trypsin (bovine)	1SGT	Trypsin (Streptomyces Griseus)
3APP	Acid protease (Penicillium Janthin-ellum)	4CMS	Chymosin B (bovine)

Table 2.5

The protein chain pairs used to optimize the alignment parameters for the PET91 matrix. Where appropriate, chain identifiers are shown proceeding the PDB code in brackets.

The PET91 matrix does appear to offer a marginal improvement over the original MDM78 matrix in terms of alignment accuracy, but unfortunately not on the scale that might have been hoped. This is not an unexpected result considering the fact that the new matrix is relatively unchanged from the original, apart from the values relating to the most infrequently occurring residue exchanges. Whilst it is important to determine improved values for the probabilities of these infrequent exchanges, such improvements will not improve the bulk of sequence alignments as the infrequent exchanges by definition will

not often arise. The most the can be expected is minor improvements in alignments, particularly those which contain a high proportion of the residues for which exchange propensities were ill-defined in the original work by Dayhoff (1978).

Table 2.6

The gap-penalty optimization matrix for the PET91 matrix. The values shown denote the percentage of all residue pairs which are equivalenced both by the sequence and structural alignment.

	0	1	2	3	4	5	6	7	8	9	10	Ext
0	51.33	63.09	70.70	70.42	73.29	77.04	78.24	73.26	74.08	72.32	69.49	
1	56.25	69.02	71.86	74.17	75.80	77.26	78.61	74.13	75.17	71.51	68.72	
2	62.31	71.48	73.92	76.83	75.32	78.32	79.75	77.38	73.70	68.72	68.60	
3	66.67	73.09	76.71	76.42	77.64	79.46	81.02	77.56	74.42	68.60	68.07	
4	71.42	75.34	75.56	77.23	78.20	79.68	79.69	76.31	70.37	67.92	68.07	
5	73.01	76.09	76.34	78.06	79.68	79.32	80.00	74.74	70.52	68.14	68.17	
6	76.70	77.14	78.00	81.68	79.18	79.18	79.79	74.82	70.52	68.17	68.17	
7	78.34	77.86	78.76	80.38	79.18	79.72	78.97	74.82	68.77	68.17	68.17	
8	78.58	80.10	80.12	80.19	79.12	79.66	78.97	74.82	68.77	68.17	67.59	
9	78.92	81.31	81.86	79.41	79.66	79.96	78.71	73.50	69.06	67.59	67.59	
10	80.19	81.66	81.72	79.35	79.96	79.19	78.71	73.50	68.19	67.59	66.59	
11	80.25	81.97	81.96	79.96	79.96	79.19	78.67	70.85	68.19	66.59	66.59	
12	80.42	83.63	81.91	79.96	79.96	79.19	77.79	68.19	67.20	66.59	66.84	
13	80.40	82.03	80.98	79.96	79.19	77.72	76.22	67.20	67.20	66.84	66.43	
14	77.97	82.03	79.96	79.96	78.32	77.72	75.22	67.20	67.44	66.84	66.43	
15	78.67	81.32	79.96	79.07	78.27	76.72	75.73	67.44	67.44	66.43	66.43	
16	79.54	81.94	79.88	78.32	78.27	75.22	74.18	67.44	67.44	66.43	66.43	
17	78.47	80.98	79.88	78.27	76.72	75.46	70.09	67.44	67.03	66.43	66.43	
18	77.81	80.98	79.88	78.27	76.95	75.46	70.09	67.44	67.03	66.43	61.46	
19	77.23	81.13	79.84	76.95	75.46	75.04	70.09	67.03	67.03	66.43	60.81	
20	77.03	79.88	80.07	76.95	75.46	75.04	67.44	67.03	67.03	66.43	59.57	
21	77.62	80.07	78.51	76.95	75.04	74.18	67.03	67.03	67.03	60.04	58.30	
22	77.96	80.07	76.95	76.95	75.04	74.18	67.03	67.03	64.86	58.30	58.30	
23	77.92	78.51	76.95	75.04	75.04	73.76	67.03	66.23	63.73	58.30	58.30	
24	76.71	78.51	76.95	75.04	75.04	73.76	66.23	64.96	58.30	58.30	58.30	
25	75.09	78.51	76.53	75.04	74.62	68.87	64.96	64.96	58.30	58.30	54.94	
26	75.09	77.63	75.66	75.04	73.82	67.59	64.96	64.62	58.30	58.30	51.73	
27	75.09	77.21	75.66	73.82	72.53	67.59	64.96	63.92	58.30	54.94	48.63	
28	75.09	75.66	74.85	72.53	71.17	64.96	63.78	62.55	54.94	51.73	48.63	
29	74.66	72.89	71.18	70.57	69.21	62.31	61.83	60.46	52.85	51.73	48.63	
Init												

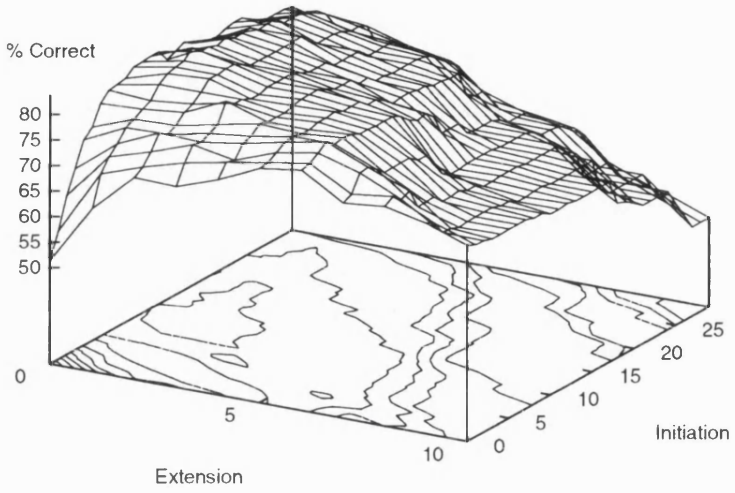


Figure 2.4

The gap-penalty optimization data plotted in the form of a 3-D surface, projected to form a set of contours. A high scoring ridge is apparent between 10 and 20 (initiation penalty) and 0 and 2 (extension).

2.12 A mutation data matrix for transmembrane proteins

The widely used Mutation Data Matrix (MDM), is an amino acid comparison matrix calculated from a study of the exchange probabilities (or odds) derived from an analysis of the evolutionary changes seen in groups of very similar proteins. The previous section described the construction of such a matrix using all the available sequences in a sequence databank. However, specific sets of sequences can be used to create matrices biased towards a particular protein family or structural class. Furthermore, matrices can be constructed for particular elements of secondary structure or residue environments (Overington *et al.*, 1990, 1992; Lüthy *et al.*, 1991). In this section, a mutation data matrix is calculated for membrane spanning segments. This new mutation data matrix is found to be very different from matrices calculated from general sequence sets which are biased towards water-soluble globular proteins, and the differences are discussed in the context of specific structural requirements of membrane spanning segments. This new matrix will help improve the accuracy of integral membrane protein sequence alignments, and could also be of use in the rational design of site directed mutagenesis experiments for this class of proteins.

Given the extreme difference between the typical environment of integral membrane associated proteins and that of globular proteins, it is not surprising that the relationship between protein sequence and structure is different for these two important classes of proteins. A very obvious example of this is the difference in the structural roles played by the 20 standard amino acids in transmembrane segments and in globular domains. A simple way to analyze amino acid properties is to observe the frequencies of amino acid exchanges in closely related sequences, a technique typified by the ubiquitous Dayhoff matrix calculated by Dayhoff *et al.* (1978), described earlier. The previous section described a highly efficient method for generating mutation data matrices from very large sequence sets (Jones *et al.*, 1992a), and this method was applied to the generation of a matrix based on mutations occurring in transmembrane segments of integral membrane proteins.

Three possible methods were considered for selecting a suitable transmembrane data set. Ideally, the data set would be constructed from all the segments *experimentally determined* to be transmembranal (either where the 3-D structure is known or where the membrane topology has been studied by chemical or immunological means). Failing that ideal, putative transmembrane segments could be included i.e. segments which have been identified as probably transmembranal by the sequence depositors, either through a knowledge of the relevant biochemistry, by homology or analogy with a related protein, or through standard prediction techniques (e.g., Rao & Argos, 1986; von Heijne, 1992). The third option would be to extend the data set further by applying a standard prediction algorithm to undocumented sequences expected to include transmembrane segments. The first approach is at present not feasible due to the very limited experimental data on integral membrane proteins. Despite the success of current prediction techniques (von Heijne, 1992; Jahnig, 1990) they are still not reliable enough to apply blindly, and therefore we rejected this option in favour of using documented transmembrane segments, including those which are experimentally determined and those which have been essentially predicted, but which have at least been vetted by the sequence depositors.

The source data for this work was a set of documented transmembrane segments extracted from Release 23.0 of SWISS-PROT (Bairoch & Boeckmann, 1991). This derived databank comprised 1765 sequences, containing 5662 transmembrane segments. This data set was extended by searching for sequences closely related ($\geq$ 85% sequence identity) to this initial set in a minimally redundant sequence databank (D.T. Jones, unpublished results). This databank comprises all the non-identical protein sequences extracted from SWISS-PROT Release 23, PIR Release 33 (Barker *et al.*, 1992) and an automatic translation of GenBank Release 73 (Bilofsky & Burks, 1988), totalling 72000 sequences. Using the MAKEPET program (Jones *et al.*, 1992a), a mutation data analysis was performed on this final data set, and a set of mutation data matrices calculated. The final matrix was generated from 3155 pairwise alignments (in total, 1.27×10^8 sequence comparisons were performed), providing 4845 accepted point mutations (PAMs). Separate analyses were performed for both single-spanning (1765 alignments, 1765 PAMs) and multiple-spanning

transmembrane segments (1405 alignments, 3612 PAMs). The combined transmembranal matrix is based on 3 times as many PAMs as the Dayhoff matrix, but in view of the fact that some amino acids occur very infrequently in transmembrane segments, such a large data set is essential to provide sufficient samples across the entire matrix.

The previously observed amino acid biases in transmembrane segments (von Heijne, 1981) are evident in Table 2.7. The most commonly occurring residue in transmembrane helices is leucine both for single and multi-spanning segments. Valine is the next most common residue in single-spanning segments, and isoleucine the next most common residue in multi-spanning segments. As expected, the polar residues are not frequent in transmembrane segments, with the negatively charged amino acids being the most clearly disfavoured residues. Single-spanning segments are significantly more hydrophobic in nature than multi-spanning segments with a total frequency of occurrence of hydrophobic amino acids (alanine, isoleucine, leucine, methionine, phenylalanine, tryptophan, valine) of 68% compared with the multi-spanning frequency of 55%.

Table 2.7

Relative mutabilities and normalized frequencies of occurrence for the 20 amino acid residues, calculated from transmembrane protein segments compared with the PET91 values.

	*Relative Mutability (General)	Frequency of Occurrence (General)	*Relative Mutability (Transmem)	Frequency of Occurrence (Transmem)	*Relative Mutability (Single)	Frequency of Occurrence (Single)	*Relative Mutability (Multi)	Frequency of Occurrence (Multi)
Ala (A)	100.0	0.0767	100.0	0.1051	100.0	0.1137	100.0	0.1026
Arg (R)	82.7	0.0515	134.3	0.0157	182.1	0.0217	106.8	0.0135
Asn (N)	103.0	0.0427	60.2	0.0185	115.8	0.0109	50.4	0.0211
Asp (D)	84.1	0.0518	76.3	0.0089	56.9	0.0057	79.9	0.0100
Cys (C)	46.2	0.0196	98.7	0.0219	71.8	0.0193	106.5	0.0229
Gln (Q)	84.5	0.0405	80.2	0.0141	121.7	0.0066	74.6	0.0168
Glu (E)	76.3	0.0617	72.3	0.0097	163.9	0.0047	57.0	0.0113
Gly (G)	52.0	0.0733	50.7	0.0758	49.6	0.0888	51.3	0.0712
His (H)	91.9	0.0228	63.9	0.0168	79.3	0.0113	61.2	0.0188
Ile (I)	102.5	0.0539	135.4	0.1188	127.0	0.1326	138.2	0.1137
Leu (L)	54.0	0.0919	69.2	0.1635	58.3	0.1769	72.7	0.1583
Lys (K)	72.5	0.0588	79.7	0.0112	129.2	0.0120	62.6	0.0111
Met (M)	95.6	0.0239	146.3	0.0333	193.7	0.0284	132.6	0.0351
Phe (F)	51.1	0.0402	65.7	0.0777	89.6	0.0554	59.9	0.0856
Pro (P)	58.4	0.0508	42.4	0.0260	84.9	0.0173	35.0	0.0291
Ser (S)	116.4	0.0685	110.2	0.0568	99.0	0.0478	113.8	0.0597
Thr (T)	107.1	0.0586	127.9	0.0523	161.1	0.0499	119.2	0.0531
Trp (W)	25.1	0.0143	38.8	0.0223	80.0	0.0168	28.7	0.0242
Tyr (Y)	48.8	0.0322	48.3	0.0324	79.7	0.0235	40.9	0.0353
Val (V)	100.1	0.0661	144.4	0.1195	121.6	0.1565	155.1	0.1065

* Relative to Ala which is arbitrarily assigned a mutability of 100.

The upper half of Table 2.8 shows how many of each of the possible 190 exchanges were observed in all the transmembrane segments, with the lower half of Table 2.8 showing the transmembranal counterpart of the PET91/MDM78 matrix (250 PAM $\log_{10}$ relatedness-odds matrix). The 250 PAM matrix is again shown here for comparison with the most common variant of the original matrix. The 1 PAM mutation probability matrix is shown in Table 2.9.

Table 2.8

The 250 PAM transmembrane protein exchange matrix ($\log_{10}$ relatedness odds), based on 4845 accepted point mutations found in 5662 transmembrane segments. Values have been multiplied by 10 and rounded to the nearest integer. The upper half of the matrix shows the actual numbers of exchanges observed.

Table 2.9

Mutation Probability Matrix for an evolutionary distance of 1 PAM. Values are scaled by a factor of 10^5. Elements of this matrix give the probability that a residue in column j will mutate to the residue in row i in an evolutionary distance of 1 PAM. A diagonal element of this matrix represents the probability of residue i=j remaining unchanged.

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	2	21	2	7	13	4	6	160	6	44	43	5	10	21	34	198	202	0	1	292
R	-1	7	0	1	2	21	3	22	21	4	8	53	19	0	1	5	5	28	0	0
N	-1	2	11	14	1	7	0	0	8	4	5	11	3	1	2	32	19	1	1	2
D	0	1	6	12	0	0	12	15	4	1	0	2	1	0	1	0	6	0	1	4
C	0	-1	-1	-3	6	0	0	13	2	4	11	0	1	34	0	48	13	8	23	47
Q	-2	6	3	2	-3	11	16	1	26	1	16	6	3	0	5	7	2	0	0	0
E	0	2	1	8	-3	7	13	21	0	0	0	0	0	0	0	4	2	0	0	7
G	1	0	-2	3	-1	-1	3	6	1	10	0	0	3	4	7	64	12	5	0	53
H	-3	5	3	3	-1	7	2	-3	11	3	2	0	1	0	0	0	4	0	29	2
I	0	-3	-3	-3	-1	-4	-4	-2	-4	2	273	0	161	66	4	22	150	1	4	883
L	-2	-3	-4	-5	-1	-2	-5	-4	-4	1	3	1	153	251	37	43	26	20	6	255
K	-2	9	5	3	-3	6	1	-1	4	-4	-4	12	4	0	0	1	2	0	5	1
M	-1	0	-2	-3	-1	-2	-3	-3	-3	1	1	-1	3	8	0	1	32	1	5	89
F	-2	-4	-4	-6	1	-4	-6	-4	-3	-1	1	-5	0	5	0	32	9	2	54	37
P	0	-3	-2	-2	-4	0	-3	-2	-4	-3	-1	-4	-3	-4	11	9	10	0	1	1
S	2	-1	2	0	1	-1	0	1	-2	-1	-2	-1	-2	-1	-1	3	134	1	22	13
T	1	-1	1	0	0	-2	-1	0	-2	0	-1	-2	0	-2	-1	2	3	1	3	48
W	-4	5	-3	-4	1	0	-3	-2	-1	-3	-2	3	-2	-3	-6	-3	-4	12	2	18
Y	-3	-1	-1	-2	3	0	-5	-5	6	-4	-3	1	-3	2	-5	0	-3	-2	10	2
V	0	-2	-3	-3	0	-4	-2	-1	-4	2	0	-4	1	-1	-3	-1	0	-2	-4	2

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	98950	138	11	81	61	29	64	218	37	38	27	46	31	28	135	360	399	0	3	252
R	21	98590	0	12	9	154	32	30	129	3	5	487	59	0	4	9	10	129	0	0
N	2	0	99368	162	5	51	0	0	49	3	3	101	9	1	8	58	38	5	3	2
D	7	7	78	99200	0	0	128	20	25	1	0	18	3	0	4	0	12	0	3	3
C	13	13	6	0	98964	0	0	18	12	3	7	0	3	45	0	87	26	37	73	41
Q	4	138	39	0	0	99158	171	1	160	1	10	55	9	0	20	13	4	0	0	0
E	6	20	0	139	0	117	99241	29	0	0	0	0	0	0	0	7	4	0	0	6
G	157	145	0	174	61	7	225	99468	6	9	0	0	9	5	28	116	24	23	0	46
H	6	138	45	46	9	190	0	1	99329	3	1	0	3	0	0	0	8	0	92	2
I	43	26	22	12	19	7	0	14	18	98579	172	0	499	88	16	40	296	5	13	763
L	42	53	28	0	52	117	0	0	12	237	99274	9	475	333	147	78	51	92	19	220
K	5	349	62	23	0	44	0	0	0	0	1	99164	12	0	0	2	4	0	16	1
M	10	125	17	12	5	22	0	4	6	140	97	37	98465	11	0	2	63	5	16	77
F	21	0	6	0	160	0	0	5	0	57	158	0	25	99311	0	58	18	9	172	32
P	33	7	11	12	0	37	0	10	0	3	23	0	0	0	99555	16	20	0	3	1
S	194	33	179	0	226	51	43	87	0	19	27	9	3	42	36	98844	265	5	70	11
T	198	33	106	70	61	15	21	16	25	130	16	18	99	12	40	244	98657	5	10	41
W	0	184	6	0	38	0	0	7	0	1	13	0	3	3	0	2	2	99593	6	16
Y	1	0	6	12	108	0	0	0	178	3	4	46	16	72	4	40	6	9	99493	2
V	287	0	11	46	221	0	75	72	12	767	161	9	276	49	4	24	95	83	6	98485

2.13 Discussion

As might be expected, the transmembrane protein mutation data matrix is quite different from the matrix calculated from a general sequence set. The most obvious feature of the matrix is the high relative mutability of the hydrophobic residues: isoleucine, methionine, and valine. Interestingly, leucine (the most commonly occurring residue in transmembrane segments) is roughly half as mutable as the other hydrophobic residues, possibly as a result of its high propensity for helix formation (in globular proteins). It is possible that the presence of leucine (and alanine) helps stabilize the helical conformation both prior to, and after, membrane insertion and is thus more highly conserved than the other hydrophobic residues which are found to disfavour helix formation in solution. An alternative explanation for the relative immutability of leucine could be that it is particularly compatible with the aligned helix packing generally observed in transmembrane proteins, a situation perhaps somewhat akin to that of the leucine-zipper motif (O'shea *et al.*, 1991). However, the fact that the mutability of leucine is just as low for single-spanning segments would seem to point away from this explanation.

The high propensity for tryptophan to exchange with arginine (28 observed exchanges) is rather surprising. However, these exchanges occur in very few protein families (primarily cytochrome c oxidase polypeptide II, and the ATP synthase A chain) and could therefore be rather atypical of transmembrane segments as a whole. On the other hand, it is possible that tryptophan and arginine could participate in similar interactions with the apolar lipid and the polar head groups. In this situation, the polar epsilon nitrogens in both amino acids could interact favourably with the head groups whilst the preceding apolar sections of the respective side chains could interact favourably with the lipid. As a result of the high probability of subsequent arginine→lysine exchanges, tryptophan also scores highly with lysine in the 250 PAM log odds matrix despite the fact that no direct tryptophan-lysine exchanges were observed in the current data set.

As expected, proline residues appear to be highly conserved in transmembrane segments, presumably due to the special role of proline residues in "kinking" transmembrane helices, as noted by several groups (von Heijne, 1991; Woolfson *et al.*, 1991). It should be noted that the frequency of occurrence of proline in transmembrane segments is not much different from its frequency of occurrence in the general sequence set. However, if it is presumed that most of the transmembrane segments are in fact transmembrane helices, and the frequency of occurrence of proline in these segments (2.6%) is compared to the equivalent frequency of 1.9% in globular protein helices, proline appears somewhat more prevalent in transmembrane helices than in globular protein helices. The difference is even more striking when the occurrence of proline-containing helices is considered: only 19% of helices in globular proteins contain one or more proline residues, whereas 50% of the annotated transmembrane segments were found to incorporate this amino acid. These occurrences become 3.5% and 37% respectively if the first turn of the helix is excluded from the calculation. Thus proline occurs in the middle of transmembrane helices 10 times as often as it does in the middle of helices in globular domains.

Apart from serine and threonine, the polar residues in general are less mutable in transmembrane protein segments than their counterparts in globular proteins. Serine and threonine are unusual in that they are capable of satisfying the hydrogen bonding capacity of their single hydroxyl groups by interacting with the main chain carbonyl group of residue *i-3* or *i-4* in the previous turn of the helix, and are thus compatible with the lipid environment. In terms of their exchanges with their apolar equivalents (leucine and isoleucine), serine prefers to exchange with leucine whereas threonine prefers isoleucine. This is in accordance with the fact that both threonine and isoleucine have centres of asymmetry, and a similar exchange pattern is observed in the general sequence set. It would appear that for multi-spanning transmembrane segments, polar residues are fairly highly conserved. Polar residues in these transmembrane segments are generally associated with specific functionality, either binding required prosthetic groups, forming ion-channels or perhaps stabilizing the helical bundles by forming ion-pairs. Polar residues, and in particular charged residues, are so infrequently found in single-spanning segments that

mutation data for these residues are not statistically significant. The fact that arginine and lysine appear to be fairly mutable might be surprising considering their important role as topogenic signals (von Heijne, 1992). However, on closer inspection it is seen that despite being fairly mutable, they tend to exchange between themselves. Presumably, arginine and lysine are equally satisfactory in directing membrane insertion.

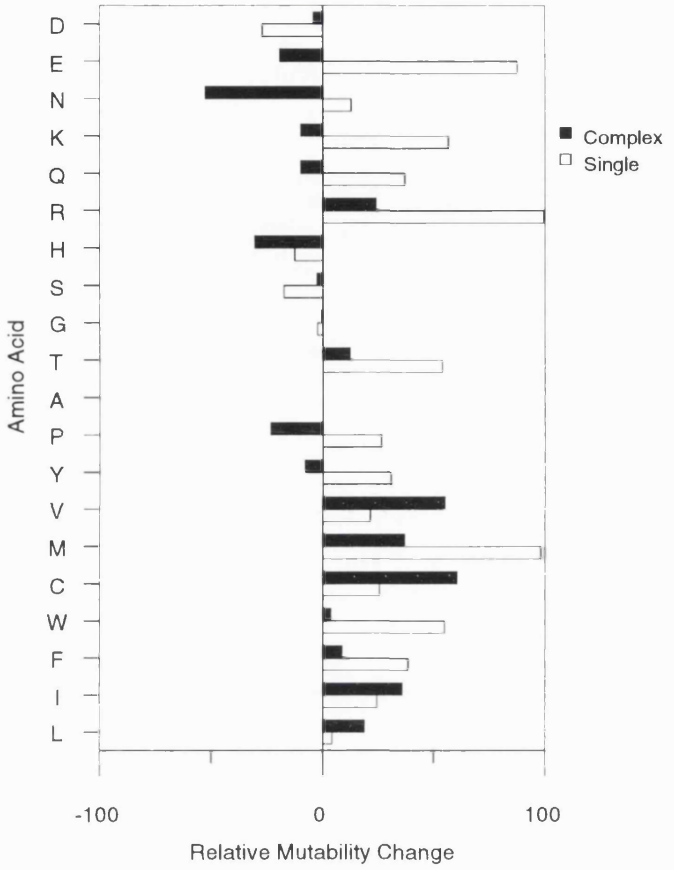


Figure 2.5

Changes in relative mutability between general proteins and integral membrane proteins. Data for both single-spanning and complex segments is shown. Positive values indicate that the mutability for transmembrane proteins is higher than that for general sequences. The amino acids are ordered along the y-axis by their polarity (Grantham, 1974), with the most polar amino acid at the top.

General trends in the mutability changes observed in transmembrane segments are clearly seen in Figure 2.5. For multi-spanning proteins, a clear distinction is seen between polar and apolar amino acids, where the apolar amino acids become highly variable and the

polar amino acids highly conserved. Perhaps the most notable example of this change is the change observed for asparagine, which changes from being one of the third most mutable residues in the general sequence set to being the fourth most highly conserved. Whether asparagine has a specific function in transmembrane segments or this is simply a spurious observation is not yet clear. In the case of single-spanning segments, there is a much higher background level of mutation than for multi-spanning segments. This is evidenced by the higher average mutation rate for these segments: 0.046 mutations per residue per unit time as opposed to 0.029 mutations per residue per unit time. Clearly there are far fewer sequence constraints on these segments, and it would appear that the only real requirements for these segments is that they be hydrophobic and contain strong helix-formers.

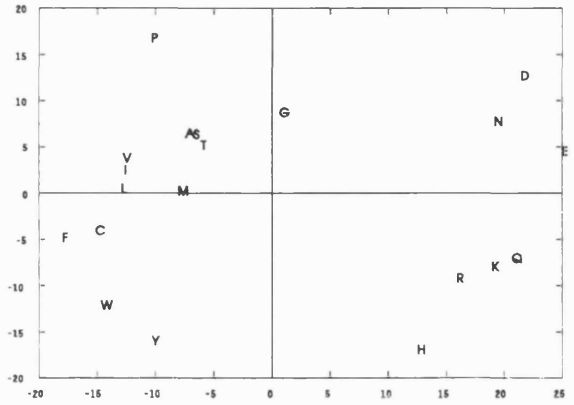
Despite the fact that the trends in the transmembrane mutation data are as expected from a knowledge of the lipid environment, one of the most important conclusions to be formed from this data is that comparison matrices calculated for general sequence sets do not adequately describe the conservation patterns observed in transmembrane segments. Of course the most important factor in amino acid similarity matrices is the groupings of the side chain chemical properties, which remains constant. However, the relative importance of these properties is seen to be very different for transmembrane segments. These similarities between amino acids are again visualized by a multi-dimensional projection of the mutation data matrix (French & Robson, 1983; Taylor, 1986b; Taylor & Jones, 1993). Figure 2.6 shows such a projection of the mutation data matrix in Table 2.8, and the equivalent matrix for the general sequence set. It is clear from the projections made, that the amino acid groupings are indeed well conserved, yet the separation between groups is somewhat different. In the general sequence set, hydrophobicity and size contribute equally to the conservation patterns observed, whereas size contributes very little to the transmembranal pattern. In the general set, alanine, serine, threonine and proline cluster with the polar residues, whilst in transmembrane segments they are seen to be more closely related to the hydrophobic group. Hydrophobicity is of course by far the most significant factor for transmembrane segments, but the next most important

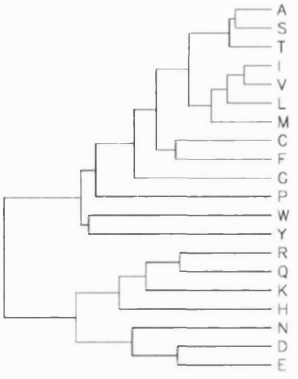
classification to make is whether the side chain is charged, and whether it is negatively charged or positively charged. In the general protein set the charged amino acids cluster together, with little distinction between oppositely charged groups (aspartic acid/lysine for example). In transmembrane segments, however, the sign of the charge is apparently more important, since charged amino acids in these segments are usually functionally-related, or involved in directing the orientation of the segments in the membrane. Clearly when trying to align distantly related transmembrane segments it is vital to bear these differences in mind. Alignment programs that use the transmembrane matrix for transmembrane regions (either experimentally determined or predicted) and a general mutation data matrix for the polar flanking regions are likely to perform much better than programs that use a single matrix. The benefits of such a bipartite scheme on a number of well-characterised membrane protein families are currently under investigation.

Figure 2.6

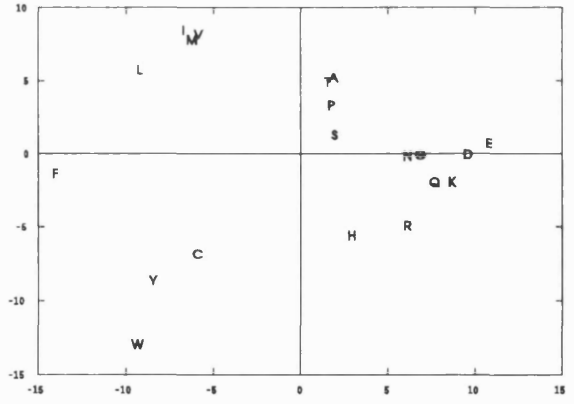
Multi-dimensional scaling projections of the 250 PAM log-odds matrices, and unweighted pair group mean analysis dendrograms for a) transmembrane sequences and b) a general set of sequences.

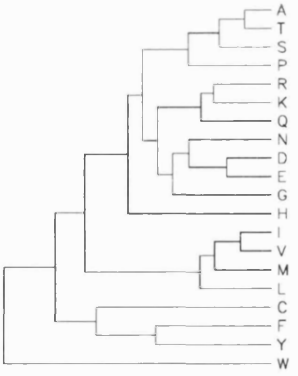
a)


P
G
D
N
E
A
S
T
V
I
L
M
F
C
W
Y
K
Q
R
H


A
S
T
I
V
L
M
C
F
G
P
W
Y
R
Q
K
H
N
D
E

b)


L
P
S
E
D
F
Q
K
H
R
C
Y
W


A
T
S
P
R
K
Q
N
D
E
G
H
I
V
M
L
C
F
Y
W

Chapter 3

Protein Tertiary Structure Prediction by Fold Recognition

In nature's infinite book of secrecy
A little can I read.

- *Antony and Cleopatra*, I. i. 15

3.1 Introduction

The prediction of protein tertiary structure from sequence may be expressed symbolically by expressing the folding process as a mathematical function:

$$\boldsymbol{C} = F(\boldsymbol{S})$$

where

$$\boldsymbol{S} = [s_1, s_2, \ldots, s_n]$$

$$\boldsymbol{C} = [\theta_1, \theta_2, \ldots, \theta_{3n-2}]$$

$$s \in \{Ala, Arg, \ldots, Val\}$$

In this case the main chain conformation of the protein chain $\boldsymbol{S}$ is represented as a vector of main chain torsion angles $\boldsymbol{C}$, with the chain itself being defined as a vector of elements corresponding to members of the set of 20 standard amino acids. The folding process is therefore defined as a function which takes an amino acid sequence and computes from it a sequence of main chain torsion angles. The choice of representation of the folded chain conformation in torsion space is arbitrary, and the problem can just as readily be expressed in terms of relative orthogonal 3-D coordinates, or with some indeterminacy in chirality, inter-atomic distances.

The protein folding problem can thus be considered a search for the folding function F. It is probable, however, that no simple representation of the folding function exists, and that even if the function exists in any form whatsoever, the only device capable of performing the required function evaluation is the protein chain itself. Conceptually, the simplest way to arrange for a protein sequence to code for its own native 3-D structure

is to arrange for the native structure to be the global minimum of the protein chain's free energy. The folding process is therefore transformed into an energy function minimization process, where the energy function could take as input the protein sequence vector $\boldsymbol{S}$, and the vector of torsion angles $\boldsymbol{C}$. Given a particular sequence $\boldsymbol{S}$, the folding process is therefore transformed into a *search* through the set of all corresponding vectors of torsion angles $\boldsymbol{C}$ for the minimum of an energy function E, where E is defined thus:

$$E(\boldsymbol{S}, \boldsymbol{C}_{native}) < E(\boldsymbol{S}, \boldsymbol{C}_{non-native})$$

The exact form of this energy function is as yet unknown, but it is reasonable to assume that it would incorporate terms pertaining to the types of interactions observed in protein structures, such as hydrogen bonding and van der Waals effects. The conceptual simplicity of this model for protein folding stimulated much research into *ab initio* tertiary structure prediction. A successful ab initio approach necessitates the solution of two problems. The first problem to solve is to find a potential function for which the above inequality at least generally holds. The second problem is to construct an algorithm capable of finding the global minimum of this function. To date, these problems remain essentially unsolved, though some progress has been made, particularly with the construction of efficient minimization algorithms (Kostrowicki & Scheraga, 1992; Bouzida *et al.*, 1992).

It is unlikely that proteins really locate the global minimum of a free energy function in order to fold into their native conformation. The case against proteins searching conformational space for the global minimum of free energy was argued by Levinthal (1968). The *Levinthal paradox*, as it is now known, can be demonstrated fairly easily. If we consider a protein chain of N residues, we can estimate the size of its conformational space as roughly 10^N states. This assumes that the main chain conformation of a protein may be adequately represented by a suitable choice from just 10 main chain torsion angle triplets for each residue. In fact, Rooman *et al.* (1991) have shown that just 7 states are sufficient. This of course neglects the additional conformational space provided by the

side chain torsion angles, but is a reasonable rough estimate, albeit an underestimate. The paradox comes from estimating the time required for a protein chain to search its conformational space for the global energy minimum. Taking a typical protein chain of length 100 residues, it is clear that no physically achievable search rate would enable this chain to complete its folding process. Even if the atoms in the chain were able to move at the speed of light, it would take the chain around 10^{82} seconds to search the entire conformational space, which compares rather unfavourably to the estimated age of the Universe (10^{17} seconds).

Clearly proteins do not fold by searching their entire conformational space. There are many ways of explaining away Levinthal's paradox. A highly plausible mechanism for protein folding is that of encoding a *folding pathway* in the protein sequence. Despite the fact that chains of significant length cannot find their global energy minimum, short chain segments (5-7 residues) could quite easily locate their global energy minimum within the average lifetime of a protein, and it is therefore plausible that the location of the native fold is driven by the folding of such short fragments (Moult & Unger, 1991). Levinthal's paradox is only a paradox if the free energy function forms a highly convoluted energy surface, with no obvious downhill paths leading to the global minimum. The folding of short fragment can be envisaged as the traversal of a small downhill segment of the free energy surface, and if these paths eventually converge on the global energy minimum, then the protein is provided with a simple means of rapidly locating it's native fold.

One subtle point to make about the relationship between the minimization of a protein's free energy and protein folding is that the native conformation need not correspond to the global minimum of free energy. One possibility is that the folding pathways initially locate a local minimum, but a local minimum which provides stability for the average lifetime of the protein. In this case, the protein in question would always be observed with a free energy slightly higher than the global minimum *in vivo*, but would eventually locate its global minimum if isolated and left long enough *in vitro* - though the location of the

global minimum could take many years. Thus, a biologically active protein could in fact be in a *metastable* state, rather than a stable one.

3.2 A limited number of folds

Many fragments of evidence point towards there being a limited number of *naturally occurring* protein folds. If we consider a chain of length 50 residues we might naively calculate the number of possible main chain conformations as 7^{50} ($\approx 10^{42}$). Clearly most of these conformations will not be stable folds, and many will not be even physically possible. In order to form a compact globular structure a protein chain necessarily has to form regular secondary structures (Chan & Dill, 1990; Dill & Chan, 1990; Gregoret & Cohen, 1991), and it is this constraint, along with the constraints imposed from a requirement to effectively pack the secondary structures formed that limit the number of stable conformational states for a protein chain. In addition to the constraints imposed from physical effects on protein stability, there are also evolutionary constraints on the number of occurring folds. Where do new proteins come from? The answer according to Doolittle (1992) is of course from other proteins. In other words the folding patterns we observe today are the result of the evolution of a set of ancestral protein folds.

If the number of possible folds is limited, then this fact should be apparent in the presently known protein structures. Do folds recur in apparently unrelated proteins? The answer appears to be a definite "yes": Table 3.1 lists most of the known examples where a definite and unexpected structural similarity has been observed between members of apparently unrelated protein families. Reports of these "fold analogies" are becoming more and more common in the literature, though whether this is due to a real saturation effect where the probability of the fold of a newly solved structure matching an existing one increases due to the increase in the number of known folds, or whether this is simply due to an increased awareness of the possibility (and the increased use of structural comparison programs) is a matter of debate.

Table 3.1

A summary of the currently known examples of proteins sharing little sequence similarity, but which have highly similar folds (Orengo *et al.*, 1993). Where no reference is given for the measured structural similarity, the similarity was calculated using the method of Orengo *et al.* (1992).

PDB Code	Title	Len	PDB Code	Title	Len	RMSD	No. Equivs.	Seq ID (%)	Reference
1hbb	Hemoglobin	140	1col	Colicin A	197	3.5	111	12	Orengo & Taylor (1992)
1thb	Hemoglobin	141	1cpc	C-phycocyanin	162	4.2	115	9	Pastore & Lesk (1990)
1cpc	C-phycocyanin	162	-	Diphtheria Toxin	171	4.6	100	1	Orengo & Taylor (1992)
256b	Cytochrome b562	106	1hmz	Hemerythrin	114	4.5	56	5	
1hmz	Hemerythrin	114	1le2	Apolipoprotein E2	144	3	91	7	
4ptp	Beta trypsin	223	1snv	Sindbis viral capsid protein	151	3.3	132	9	
1acx	Actinoxanthin	107	1hoe	Amylase Inhibitor	76	4.3	61	13	
2tim	Triosephosphate isomerase	249	1ald	Aldolase	363	3.2	109	10	
	" "		1gox	Glycolate Oxidase	350	3.8	186	8	
	" "		1fcb	Flavocytochrome B2	494	3.9	201	7	
	" "		1pii	Anthranilate Isomerase	452	3.4	176	7	
	" "		1wsy	Tryptophan Synthase (ch. A)	248	5.4	155	6	

PDB Code	Title	Len	PDB Code	Title	Len	RMSD	No. Equivs.	Seq ID (%)	Reference
	Triosephosphate isomerase		6xia	Xylose Isomerase	387	6	176	4	
	" "		5rub	Rubisco	436	4.5	108	5	
	" "		2taa	Taka-Amylase A	478	6.4	74	7	
	" "		4enl	Enolase	436	-	-	4	
4fxn	Flavodoxin	138	3chy	CheY Protein	128	3	98	8	
5p21	Ras p21 protein	166	5cpa	Carboxypeptidase A	307	6.2	105	3	
1etu	Elongation factor tu	177	1cse	Subtilisin Carlsberg	274	5.9	132	7	
1ubq	Ubiquitin	76	1fxi	Ferredoxin I	96	3.5	65	5	
1ubq	Ubiquitin	76	2gb1	Protein G	56	3	51	14	
1snc	Staphylococcal Nuclease	135	1bov	Verotoxin 1B	69	3.5	63	6	Murzin & Chothia (1992)
				Aspartyl tRNA synthetase N-terminal domain					Murzin & Chothia (1992)
-	H-Ferritin		-	Ribonucleotide Reductase (R2)					Murzin & Chothia (1992)

PDB Code	Title	Len	PDB Code	Title	Len	RMSD	No. Equivs.	Seq ID (%)	Reference
-	H-Ferritin		-	T-Catalase (TC)					Murzin & Chothia (1992)
-	" "		-	Methane Mono-oxygenase					Murzin & Chothia (1992)
-									
-	SH3 domain		-	R67 dihydrofolate reductase					Murzin (1992)
-	Hematopoietic cytokines		-	Interleukin-4					Murzin & Chothia (1992)
			-	Human Growth Hormone					Murzin & Chothia (1992)
			-	Interleukin-2					Murzin & Chothia (1992)
-	Nerve Growth Factor		-	Transforming growth factor β2					Swindells (1992)
-	Cystatin		-	Monellin					Murzin & Chothia (1992)
-	" "		-	Stefin B					Murzin & Chothia (1992)
-	P1 Nuclease		-	Phospholipase C					Murzin & Chothia (1992)
-	Catalytic domain of dihydrolipoyl transacetylase		-	chloramph. acetyltranferase (catalytic domain)					Murzin & Chothia (1992)

PDB Code	Title	Len	PDB Code	Title	Len	RMSD	No. Equivs.	Seq ID (%)	Reference
2cd4	CD4	176	2rhe	Immunoglobulin	114	1.9	92	17	
3hla	Class I Histocompatibility Antigen	99	2rhe	Immunoglobulin	114	4.1	85	6	
1cd8	CD8	114	2rhe	Immunoglobulin	114	2.1	105	20	
	papD-chaperonin		2rhe	Immunoglobulin	114				Holmgren & Branden (1989)
8ilb	Interleukin 1-beta	146	1tie	ETI I STI	165	2.7	119	9	Murzin *et al.* (1992)
	(beta trefoil fold)		3fgf	Fibroblast growth factor	124	2.5	119	14	Murzin & Chothia (1992)
			1aai	Ricin (ch. B)	262	3.6	83	5	Murzin *et al.* (1992)
				Hisactophilin					Murzin *et al.* (1992)
2stv	Jelly Roll virus coat proteins	184	1tnf	Tumour necrosis factor	152	3.9	105	4	
1atn	Actin		1hsc	70 kD Heat Shock Cognate Protein					Flaherty *et al.* (1991)
		372	2yhx	Hexokinase	457	5.9	256	5	Bork *et al.* (1992)

PDB Code	Title	Len	PDB Code	Title	Len	RMSD	No. Equivs.	Seq ID (%)	Reference
	α/β hydrolase family			Acetylcholinesterase					Ollis *et al.* (1992)
				Cutinase					
				Dienelactone hydrolase					
				Haloalkane dehalogenase					
				Serine carboxypeptidase					
				Human pancreatic lipase					
1lap	Leucine aminopeptidase	480	5cpa	Carboxypeptidase	307	4.4	220	4	

A limited number of folds and the recurrence of folds in protein which share no significant sequence similarity offer a "short-cut" to protein tertiary structure prediction. As already described, it is impractical to attempt tertiary structure prediction by searching a protein's entire conformational space for the minimum energy structure, but if we know that there could be as few as 1000 possible protein folds (Chothia, 1992), then the intelligent way to search a protein's conformational space would be to simply consider only those regions which correspond to this predefined set. This is analogous to the difference between an exam requiring the writing of an essay and an exam requiring multiple-choice questions to be answered. Clearly a person with no knowledge of the subject at hand has a much greater chance of achieving success with the multiple-choice paper than with the essay paper.

Suppose we had derived a practical potential function for which the native conformational energy was lower than that of any other conformation, and that we had identified M possible chain folds, then we would have the basis of a useful tertiary structure prediction scheme. In order to predict the conformation of a given protein chain S, the chain would be folded into each of the M known chain conformations ($C_1...C_M$), and the energy of each conformation calculated. The predicted chain conformation would be the conformation with the lowest value of the potential function. The term generally applied to schemes of this type is *fold recognition*, where instead of trying to predict the fold of a protein chain *ab initio*, we attempt to recognize the correct chain fold from a list of alternatives.

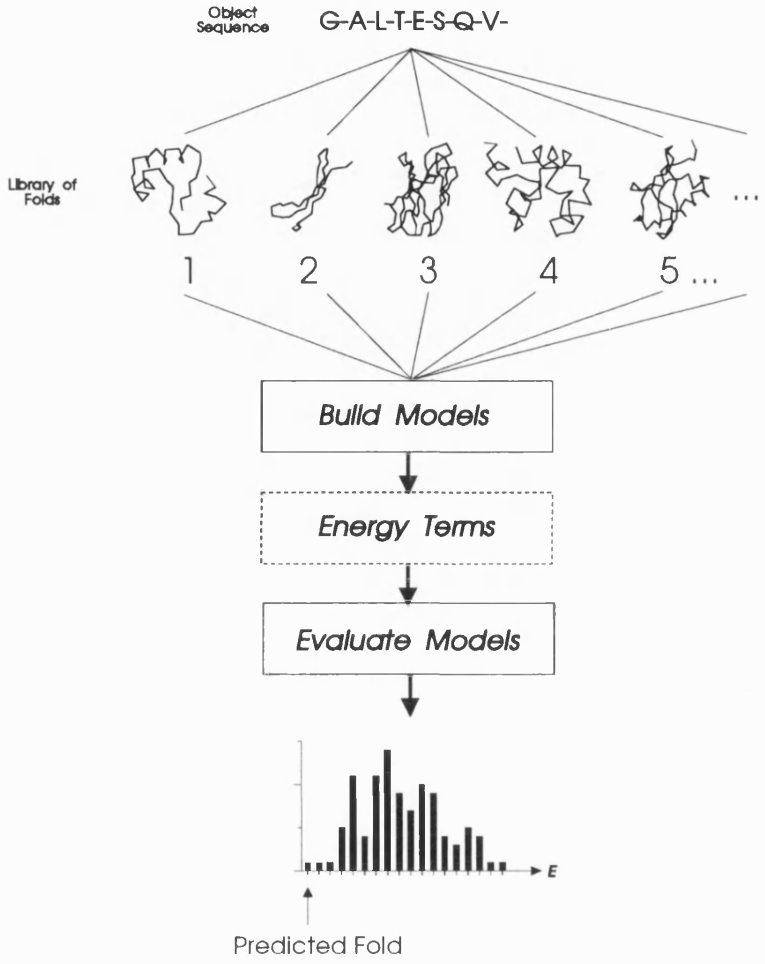


Figure 3.1

Diagrammatic representation of a possible approach to protein structure prediction by fold recognition.

Figure 3.1 shows 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.

3.3 The fold library

A suitable representative library of folds needs to be found. These folds can be observed crystal structures, NMR structures, or even theoretical model structures. If the library is limited to observed structures then the method will evidently be capable of recognizing only previously observed folds. As has been already discussed, the frequency of occurrence of similar folds between proteins sharing no significant sequence similarity would seem to indicate that creating a library entirely out of known folds is perfectly reasonable. This is the approach that has been used in this work. A future development of the method will entail the generation of putative model folds, based on folding rules derived from the known structures. Several groups have already attempted the generation of putative folds with some success (Cohen *et al.*, 1980, 1982; Taylor, 1991), however, the structures created were only very approximate models of real proteins. As an illustration of the limited accuracy to which folds can be synthesized, Figure 3.2 shows a model of sperm whale myoglobin superposed onto the crystallographically determined coordinates (Phillips, 1980). The model was initially based on the edges of a regular icosahedron (Murzin & Finkelstein, 1983; Taylor, 1991) and subjected to a number of refinement steps in order to generate a more realistic structure (Taylor, 1991). The Cα RMSD for the entire chain is 3.6 Å, which indicates a definite similarity between the two structures, though detailed aspects of the globin fold are not precisely reproduced.

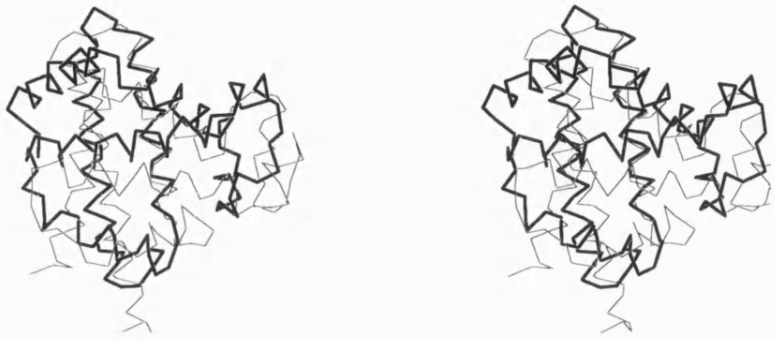


Figure 3.2

Model of myoglobin constructed on a polyhedral framework. The structure drawn with a thick line is the crystallographically determined structure for sperm whale myoglobin, onto this is superposed the model for myoglobin based on a regular polyhedral framework (thin line). The RMSD in this case is 3.6 Å over all 153 equivalent Cα atoms.

The fold library used in the work described in this and the following chapter was extracted from the July 1991 release of the Brookhaven protein databank (PDB) (Bernstein *et al.*, 1977). Selecting a representative set of chain folds from the PDB not only reduces the bias in any statistical calculation based on the data set, but also helps reduce the amount of computation required. The steps used to generate the representative set of chains were as follows:

i. Model structures, NMR entries and Cα-only chains were excluded from further consideration, along with chains shorter than 30 residues.

ii. The remaining chains were sorted in order of resolution (highest resolution, first), and stored in list A. There are of course many other factors that can be taken into account when evaluating the relative quality of protein structures determined by X-ray crystallography. In particular the crystallographic R-factor provides a measure of the goodness of fit between the observed reflections and those predicted from the refined structure. Generally a reasonable R-factor (as a percentage) should be less than 10 times the resolution (Å) of the collected data. A further criterion that should be considered is the ideality of the structure's stereochemistry (see Morris *et al.*, 1992 for example), and in particular the distribution of observed ϕ/ψ angles with respect to the core regions of the Ramachandran plot (Ramachandran, 1968). For the purposes of this work, however, selection by resolution alone was deemed adequate.

iii. Each chain was extracted from list A and compared with each entry in list B (initially an empty list). The candidate chain was aligned with each entry in list B using a standard alignment algorithm (Gotoh, 1982) with the unitary protein matrix and a constant gap penalty of 4. If any of the resultant sequence identity (normalized by the shorter of the two sequence lengths) exceeded 30%, then the candidate chain was rejected and the next entry in list A was processed. To ensure that the alignment was statistically significant, the sequences under consideration were randomly shuffled and re-aligned 100 times. If any of the randomized sequence alignment scores exceeded the score obtained with the initial sequences, then the match was rejected.

The resulting list of chains (list B) was taken as a representative set of chain folds. The above method ensures that the best resolved structures are chosen first, and that no pair in the remaining list of chains has a statistically significant sequence identity score of more than 30%. The choice of a 30% cut-off was made based on the analysis by Sander and Schneider (1991) where, by means of a rigorous comparison of structural fragments by rigid body superposition, a relationship between sequence similarity and structural

similarity was determined. For chains over 70 residues in length, the minimum score implying structural similarity was found to be around 25%. To allow for the scatter observed, this minimum cut-off was increased by 5%.

From the automatically generated list, chains incorporated a high proportion of unknown amino acids and frequent or large chain breaks were excluded. Also entries 1NXB and 2GN5 were excluded as these structures are probably incorrect (Morris *et al.*, 1992). Finally the light chain of 2FB4 (human immunoglobulin Fab) was added to the list, which would otherwise be rejected on the grounds of its similarity to 2RHE (Bence-Jones protein, lambda variable domain). This rather arbitrary decision was made to ensure that a complete light chain was present in the data set in addition to a single variable domain. This was the only effort made to take domain structure into account, but a move towards a domain library as opposed to a chain library is an important future development. The resulting list of 102 chains is shown in Table 3.2.

Table 3.2

The 102 chain folds used in this work. All structures have a resolution of at least 2.8 Å, and no pair of chains has >30% sequence identity (with the exception of the L chain of 2FB4 and 2RHE - see text).

ID	Protein	Chain	Res	Authors
1ABP	L-arabinose-binding protein - E. Coli		2.40	F.A.Quiocho,G.L.Gilliland
1BP2	Phospholipase a2 - bovine pancreas		1.70	B.W.Dijkstra et al.
1CC5	Cytochrome c5 - Azotobacter vinelandii		2.50	C.D.Stout & D.C.Carter
1CCR	Cytochrome c - rice embryos		1.50	H.Ochi et al.
1CD4	CD4 (1 - 183 plus asp - thr) - human		2.30	S.-E.Ryu et al.
1CRN	Crambin - Abyssinian cabbage		1.50	W.A.Hendrickson & M.M.Teeter
1CSE	Subtilisin carlsberg - Bacillus Subtilis	E,I	1.20	W.Bode
1CTF	L7/L12 50s ribosomal protein - E. Coli		1.70	M.Leijonmarck & A.Liljas
1CY3	Cytochrome c3 - Desulfovibrio Desulfuricans		2.50	R.Haser et al.
1DHF	Dihydrofolate reductase (dhfr) - human	A	2.30	J.F.Davies & J.Kraut
1ECA	Erythrocruorin - Chironomous Thummi Thummi		1.40	W.Steigemann & E.Weber
1FD2	Ferredoxin - Azotobacter Vinelandii		1.90	C.D.Stout
1FX1	Flavodoxin - Desulfovibrio Vulgaris		2.00	K.D.Watenpaugh et al.
1GCR	Gamma-ii crystallin - bovine		1.60	C.Slingsby et al.
1GD1	Glyceraldehyde-3-phosphate dehyd. - Bac. Stearo.	O	1.80	T.Skarzynski et al.
1HIP	Oxidized high potential iron protein - Chromatium Vinosum		2.00	C.W.Carter et al.
1HOE	Alpha-amylase inhibitor hoe - Streptomyces Tendae		2.00	J.W.Pflugrath et al.
1I1B	Interleukin-1 beta - human		2.00	B.C.Finzel et al.
1L01	Lysozyme (E.C.3.2.1.17) - bacteriophage T4		1.70	S.Dao-pin et al.
1LH1	Leghemoglobin (acetate,met) - yellow lupin		2.00	B.K.Vainshtein et al.
1LRD	Lambda repressor-operator complex - bacteriophage l.	3	2.50	S.Jordan & C.Pabo
1LZ1	Lysozyme (E.C.3.2.1.17) - human		1.50	P.J.Artymiuk & C.C.F.Blake
1MBA	Myoglobin - sea hare		1.60	M.Bolognesi et al.
1MBD	Myoglobin - sperm whale		1.40	S.E.V.Phillips
1PAZ	Pseudoazurin - Alcaligenes Faecalis		1.55	K.Petratos et al.
1PCY	Plastocyanin - poplar leaves		1.60	J.M.Guss & H.C.Freeman
1PFK	Phosphofructokinase (R-state) - E. Coli	A	2.40	Y.Shirakihara & P.R.Evans
1PHH	p-hydroxybenzoate hydroxylase - Pseudomonas Fluor.		2.30	H.A.Schreuder & J.Drenth
1RHD	Rhodanese - bovine		2.50	W.G.J.Hol et al.
1SN3	Scorpion neurotoxin - scorpion		1.80	R.J.Almassy et al.
1TGS	Trypsinogen/inhibitor complex - bovine and porcine	I	1.80	M.Bolognesi et al.
1TNF	Tumor necrosis factor-alpha (cachectin) - human	A	2.60	M.J.Eck & S.R.Sprang
1UBQ	Ubiquitin - human		1.80	S.Vijay-kumar et al.

ID	Protein	Chains	Res	Authors
1UTG	Uteroglobin (oxidized) - rabbit		1.30	I.Morize et al.
1WSY	Tryptophan synthase - Salmonella Typhimurium	A,B	2.50	C.Hyde et al.
1YPI	Triose phosphate isomerase (/tim$) - yeast	A	1.90	T.Alber et al.
256B	Cytochrome B-562 (oxidized) - E. Coli	A	1.40	K.Hamada et al.
2AAT	Aspartate aminotransferase - E. Coli		2.80	D.Smith et al.
2AZA	Azurin (oxidized) - Alcaligenes Denitrificans	A	1.80	E.N.Baker & G.E.Norris
2CA2	Carbonic anhydrase II - human		1.90	A.E.Eriksson et al.
2CCY	Cytochrome c' - Rhodospirillum Molischianum	A	1.67	B.C.Finzel et al.
2CDV	Cytochrome c3 - desulfovibrio $vulgaris		1.80	Y.Higuchi et al.
2CNA	Concanavalin A - jack bean		2.00	G.N.Reeke et al.
2CPP	Cytochrome P450cam - Pseudomonas Putida		1.63	T.L.Poulos
2CRO	434 cro protein - phage 434		2.40	A.Mondragon et al.
2CYP	Cytochrome c peroxidase - baker's yeast		1.70	B.C.Finzel et al.
2ER7	Endothiapepsin - chestnut blight fungus	E	1.60	B.Veerapandian et al.
2FB4	Immunoglobulin Fab - human	H,L	1.90	M.Marquart & R.Huber
2GBP	Galactose binding protein - E. Coli		1.90	N.K.Vyas et al.
2HLA	Class I histocompatibility antigen aw 68.1 - Human	A,B	2.60	T.P.J.Garrett et al.
2LBP	Leucine-binding protein - E. Coli		2.40	J.S.Sack et al.
2LTN	Lectin - garden pea	A	1.70	F.L.Suddath et al.
2MHR	Myohemerythrin - sipunculan worm		1.70	S.Sheriff & W.A.Hendrickson
2OVO	Ovomucoid third domain - silver pheasant		1.50	W.Bode & O.Epp
2PAB	Prealbumin (human plasma) - human	A	1.80	S.J.Oatley & C.C.F.Blake
2RHE	Bence-jones protein (variable domain) - human		1.60	W.Furey jnr. et al.
2RNT	Ribonuclease T1 - Aspergillus Oryzae		1.80	W.Saenger et al.
2SGA	Proteinase A - Streptomyces Griseus		1.50	M.N.G.James & A.R.Sielecki
2SNS	Staphylococcal nuclease - Staphylococcus Aureus		1.50	M.J.Legg et al.
2SOD	Cu,Zn superoxide dismutase - bovine	O	2.00	J.A.Tainer et al.
2SSI	Subtilisin inhibitor - Streptomyces Albogriseolus		2.60	Y.Mitsui et al.
2STV	Coat protein - satellite tobacco necrosis virus		2.50	T.A.Jones & L.Liljas
2TMN	Thermolysin - Bacillus Thermoproteolyticus	E	1.60	D.E.Tronrud et al.
2WRP	Trp repressor - E. Coli	R	1.65	C.L.Lawson & P.B.Sigler
351C	Cytochrome c551 - Pseudomonas Aeruginosa		1.60	Y.Matsuura et al.
3ADK	Adenylate kinase - porcine		2.10	G.E.Schulz

ID	Protein	Chain	Res	Authors
3BLM	Beta-lactamase - Staphylococcus Aureus		2.00	O.Herzberg & J.Moult
3CLA	Chloramphenicol acetyltransferase - E. Coli		1.75	A.G.W.Leslie
3DFR	Dihydrofolate reductase - Lactobacillus Casei		1.70	D.J.Filman et al.
3FXC	Ferredoxin - Spirulina Platensis		2.50	M.Kakudo et al.
3GAP	Catabolite gene activator protein - E. Coli	A	2.50	I.T.Weber & T.A.Steitz
3GRS	Glutathione reductase - human		1.54	G.E.Schulz & P.A.Karplus
3HHB	Hemoglobin (deoxy) - human	A	1.74	G.Fermi & M.F.Perutz
3ICB	Intestinal calcium-binding protein - bovine		2.30	D.M.E.Szebenyi & K.Moffat
3ICD	Isocitrate dehydrogenase - E. Coli		2.50	J.H.Hurley et al.
3PGK	Phosphoglycerate kinase - baker's yeast		2.50	P.J.Shaw et al.
3PGM	Phosphoglycerate mutase - baker's yeast		2.80	J.W.Campbell et al.
4CPA	Carboxypeptidase A-alpha - bovine	I	2.50	W.N.Lipscomb & D.C.Rees
4CPV	Calcium-binding parvalbumin - carp		1.50	V.D.Kumar et al.
4DFR	Dihydrofolate reductase - E. Coli	A	1.70	D.J.Filman et al.
4FXN	Flavodoxin - Clostridium MP		1.80	M.L.Ludwig
4MDH	Cytoplasmic malate dehydrogenase - porcine	A	2.50	J.J.Birktoft & I.J.Banaszak
4PTP	Beta trypsin - bovine		1.34	J.L.Chambers et al.
4RXN	Rubredoxin - Clostridium Pasteurianum		1.20	K.D.Watenpaugh et al.
4SGB	Serine proteinase B inhibitor - potato tuber	I	2.10	M.James & H.Greenblatt
4TNC	Troponin c - chicken		2.00	M.Sundaralingam
4XIA	D-xylose isomerase - arthrobacter	A	2.30	K.Henrick et al.
5ACN	Aconitase - pig	A	2.10	A.H.Robbins & C.D.Stout
5CPA	Carboxypeptidase A-alpha - bovine		1.54	W.N.Lipscomb
5PTI	Trypsin inhibitor - bovine		1.00	A.Wlodawer & R.Huber
6LDH	Lactate dehydrogenase - dogfish		2.00	C.Abad-zapatero & M.G.Rossmann
7CAT	Catalase - beef liver	A	2.50	M.R.N.Murthy et al.
7RSA	Ribonuclease A - bovine		1.26	A.Wlodawer & G.L.Gilliland
8ADH	Apo-liver alcohol dehydrogenase - horse		2.40	T.A.Jones & H.Eklund
8ATC	Aspartate carbamoyltransferase - E. Coli	A,B	2.50	H.Ke et al.
9PAP	Papain - papaya		1.65	I.G.Kamphuis & J.Drenth
9WGA	Wheat germ agglutinin - wheat	A	1.80	C.S.Wright

Of the 102 chains, it is useful to note which would be excluded were a more stringent set of quality selection criteria to be applied:

1ABP - Unrefined structure, poor stereochemical quality
1CY3 - R-factor 34%, poor stereochemical quality
1FX1 - Unrefined structure, though high stereochemical quality
3FXC - R-factor 31%, poor stereochemical quality
3PGK - Unrefined structure, moderately poor stereochemical quality
3PGM - R-factor 29%, poor stereochemical quality

3.4 The modelling process

The central process in structure-based fold recognition is the fitting of a given sequence onto a structural template. One way of visualizing this process is to imagine the side chains of the object protein being fitted onto the backbone structure of the template protein. This process is of course almost identical to the process of homology modelling (Blundell *et al.*, 1988). The standard modelling process consists of three basic steps. Firstly at least one suitable homologous template structure needs to be found. Secondly, an optimal alignment needs to be generated between the sequence of the template structure (the source sequence) and the sequence of unknown structure (the object sequence). Thirdly, the framework structure of the template is 'mapped' onto the object sequence. After several stages of energy minimization, the model is ready for critical evaluation.

Each step in the modelling process has its associated problems, though the first two steps are the most critical overall. Evidently, if no homologous structure can be found, the process cannot even be started, and even when a homologous structure is available (perhaps selected on the basis of functional similarity), the degree of homology may be so low as to render the alignment of the sequences impossible by normal means ("by eye"

or by automatic alignment). There exists a significant link between the detection of homology and the subsequent alignment of sequences in that both steps employ variants of the same algorithm: typically the algorithm of Needleman and Wunsch (1970) or Wilbur and Lipman (1983). More recently, pattern matching methods have been developed which offer far greater sensitivity than that offered by simple pairwise sequence alignment (Taylor, 1986a; Gribskov, 1987; Bashford *et al.*, 1987; Barton, 1990). These methods, recently reviewed (Taylor & Jones, 1991), in one way or another generate a consensus pattern based on the multiple alignment of several homologous sequences. For example, a globin template (Bashford *et al.*, 1987) may be constructed by aligning the many available globin sequences against a known globin structure, identifying the conserved amino acid properties at each position in the template. Though these methods are capable of inferring reasonably distance homologies, allowing, for example, the modelling of HIV protease based on the aspartyl proteinases (Pearl and Taylor, 1987), they are limited by their dependence on the availability of several homologous sequences, and on the ability of multiple alignment algorithms to successfully align them. In the July 1992 release of the Brookhaven database (Bernstein *et al.*, 1977) there are 142 unique protein chains (chains that show no significant homology with any other). Of these 142 chains, only 51 have suitable sets of homologous sequences in the sequence database to enable consensus templates to be constructed. In general, therefore, only two sequences are available: the sequence of the template structure and the sequence being modelled.

The previously described methods for detecting homology work by increasing the sensitivity of standard sequence comparison algorithms. The general assumption is that some residual sequence similarity exists between the template sequence and the sequence under investigation, which is often not the case. Clearly, therefore, the ideal modelling method would not make this assumption, and work with cases where there is no detectable sequence similarity between the object sequence and the source protein.

The development of a method capable of aligning a sequence with a structural template without reference to the sequence of the template protein formed a major part of this

project. For reasons that will be discussed in the next chapter, this is a computationally hard problem.

3.5 Evaluating the models

The inability of standard atomic force-fields to detect misfolded proteins was first demonstrated by Novotny and Karplus (1984). Their test problem was very simple, and yet is a good illustration. In this study, the sequences of myohemerythrin and an immunoglobulin domain of identical length were exchanged. Both the two native structures, and the two "misfolded" proteins were then subjected to energy minimization using the CHARMm (Brooks *et al.*, 1983) force-field. The results were somewhat surprising in that it was impossible to distinguish between the native and misfolded structures on the basis of the calculated energy sums. Novotny and Karplus correctly surmised that the reason for this failure was the neglect of solvation effects in the force-field. In a later study (Novotny *et al.* 1988), the force-field was modified to approximate the effects of solvent and in this case the misfolded structures could be identified. The work of Novotny and Karplus encouraged several studies into effective methods for evaluating the correctness of protein models, which will now be briefly reviewed.

Eisenberg and McLachlan (1986) were able to distinguish correct models from misfolded models by using an elegantly simple solvation energy model alone. By calculating a solvation free energy for each amino acid type and calculating the degree of solvent accessibility for each residue in a given model structure, the correctly folded models were clearly distinguished from the misfolded.

Baumann *et al.* (1989) also used a solvation term to recognize misfolded protein chains, along with a large number of other general statistical properties of sequences forming stable protein folds. Holm and Sander (1992) have recently proposed another solvation model, which appears to be very able at detecting misfolded proteins, even those proteins

which have shifts of their sequence on their correct native structure. Interestingly enough a sequence-structure mismatch can quite easily occur not just in theoretically derived models, but even in crystallographically derived models. For example one of the xylose-isomerase structures in the current Brookhaven database has in part a clearly mistraced chain. Such errors can be detected by use of a suitable solvation based model evaluation procedure.

Perhaps the most widely known method for testing the overall quality of a protein model is that proposed by Lüthy *et al.* (1992), who used a rather more precise definition of residue environment to assess models. This method will be discussed more fully later.

3.6 Statistically derived pairwise potentials

Several groups have used statistically derived pairwise potentials to identify incorrectly folded proteins. Using a simplified side chain definition, Gregoret and Cohen (1990) derived a contact preference matrix and attempted to identify correct myoglobin models from a set of automatically generated models with incorrect topology, yet quite reasonable core packing.

Hendlich *et al.* (1990) used a set of potentials of mean force, first described by Sippl (1990), not only to correctly reject the misfolded protein models of Novotny and Karplus, but also to identify the native fold of a protein amongst a large number of decoy conformations generated from a database of structures. In this latter case, the protein sequence of interest was blindly fitted to all contiguous structural fragments taken from a library of highly resolved structures, and the interatomic pairwise energy terms summed in each case. For example, consider a protein sequence of 50 residues being fitted to a structure of length 100 residues. The structure would offer 51 possible conformations for this sequence, starting with the sequence being fitted to the first 50 residues of the structure, and finishing with the sequence being fitted to the last 50. Taking care to

eliminate the test protein from the calculation of potentials, Hendlich *et al.* (1990) correctly identified 41 out of 65 chain folds. Using factor analysis, Casari and Sippl (1992) have found that the principal component of their potentials of mean force is a hydrophobic potential of simple form. This principal component potential alone is found to be almost as successful as the full set of potentials in identifying correct folds.

In a very similar study to that performed by Hendlich *et al.* (1990), Crippen (1991) used a simple discrete contact potential to identify a protein's native fold from all contiguous structural fragments of equal length extracted from a library of structures. The success rate (45 out of 56) in this case was marginally higher than that of Hendlich *et al.* (1990) due to the fact that the contact parameters in this case were optimized against a 'training set' of correct and incorrect model structures. Maiorov and Crippen (1992) improved upon these results using a continuous contact potential, with the new contact function correctly identifying virtually all chain folds defined as being 'compact'.

Both the work of Hendlich *et al.* and Crippen demonstrates a very restricted example of fold recognition, whereby sequences are matched against suitable sized contiguous fragments in a template structure. A much harder recognition problem arises when more complex ways of fitting a sequence to a structure are considered i.e. by allowing for relative insertions and deletions between the object sequence and the template structure. Suitable treatment of insertions and deletions is essential to a generalized method for protein fold recognition.

Ponder and Richards (1987)

The first true example of a fold recognition attempt was the template approach of Ponder and Richards (1987). Ponder and Richards concerned themselves with the inverse folding problem in that they tried to enumerate sequences that could be compatible with a given backbone structure. The evaluation potential in this case was a simple van der Waals

potential, and so models were effectively scored on the degree of overlap between side chain atoms. A further requirement was for the core to be well-packed, which was achieved by considering the conservation of side chain volume. In order to fit the side chains of a given sequence onto the backbone an exhaustive search was made through a "rotamer library" of side chain conformations. If after searching rotamer space the side chains could not be fitted successfully into the protein core, then the sequence was deemed incompatible with the given fold. As a sensitive fold recognition method, however, this method was not successful. Without allowing for backbone shifts, the packing requirement of a given protein backbone was found to be far too specific. Only sequences very similar to the native sequence could be fitted successfully to the fixed backbone.

Bowie *et al.* (1990)

A rather more successful attempt at fold recognition was made by Bowie *et al.* (1990). The first stage of this method involves the prediction of residue accessibility from multiple sequence alignments, which is itself another interesting recent development (discussed earlier in Chapter 1). In essence, alignment positions with high average hydrophobicity and high conservation are predicted to be buried and relatively polar variable positions predicted to be exposed to solvent. The degree of predicted exposure at each position of the aligned sequence family is then encoded as a string. This string is then matched against a library of similarly encoded strings, based, however, not on predicted accessibilities but on *real* accessibilities calculated from structural data. Several successful recognition examples were demonstrated using this method. Of particular note was the matching of an aligned set of Ef Tu sequences with the structure of flavodoxin. The similarity between Ef Tu and flavodoxin is not readily apparent even from structure (Orengo *et al.*, 1992) and so this result is quite impressive.

Bowie *et al.* (1991)

Bowie, Lüthy and Eisenberg (1991) have attempted to match sequences to folds by describing the fold not just in terms of solvent accessibility, but in terms of the *environment* of each residue location in the structure. In this case, the environment is 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 environment of a particular residue thus defined tends to be more highly conserved than the identity of the 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 1D-3D profile method, in that a 3D structure is encoded as a 1D string, which can then be aligned using traditional dynamic programming algorithms (e.g., Gotoh, 1982). Bowie *et al.* have applied the 1D-3D 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 have still retained some sequence similarity with the search protein, even though in the case of actin and the 70 kD heat-shock protein the sequence similarity is very weak (Bork *et al.* 1992). 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. In a rather roundabout way, the same authors went on to use this method to successfully evaluate protein models (Lüthy *et al.* 1992), and with a commendable degree of frankness

demonstrated that the method was capable of detecting a previously identified chain tracing error in a structure solved in their own laboratory.

Finkelstein and Reva (1991)

Finkelstein and Reva (1991) have used a simplified lattice representation of protein structure for their work on fold recognition. The problem they consider is that of matching a sequence to one of the 60 possible 8-stranded β-sandwich topologies. Each strand has 3 associated variables: length, position in the sequence and spatial position in the lattice Z direction. The force-field used by Finkelstein and Reva includes both short range and long range components. The short range component is simply based on the beta-coil transition constants for single amino acids, similar in many respects to the standard Chou-Fasman (1974) propensities. The long range interaction component has a very simple functional form. For a pair of interacting (contacting) residues, it is defined simply as the sum of their solvent transfer energies as calculated by Fauchere and Pliska (1983).

The configurational energy of the 8 strands in this simple force field is minimized by a simple iterative method. At the heart of the method is a probability matrix (a 3-dimensional matrix in this case) for each of the strands, where each matrix cell represents one triplet of the strand variables i.e. length, sequence position and spatial position. The values in each cell represent the probability of observing the strand with the values associated with the cell. The novel aspect of this optimization strategy is that the strands themselves do not physically move in the force field, only the probabilities change. At the start of the first iteration the strand coordinate probabilities are assigned some arbitrary value, either all equal, or set close to their expected values (the first strand is unlikely to be positioned near the end of the sequence for example). A new set of probabilities is then calculated using the current mean field and the inverse Boltzmann equation. As more iterations are executed it is to be hoped that most of the probabilities will collapse to zero, and that eventually a stable "self-consistent" state will be reached. Finkelstein and Reva

found that the most probable configurations corresponded to the correct alignment of the 8 stranded model with the given sequence, and that when the process was repeated for each of the 60 topologies, in some cases the most probable configuration of the native topology had the highest probability of all.

The simplicity of the lattice representation and the uncomplicated force field are critical to the success of this method. A more detailed inter-residue potential would prevent the system from reaching a self-consistent state, and would be left either in a single local minimum or more likely oscillating between a number of local minima. In addition, whilst it is quite practical to represent β-sheets on a lattice, it is not clear how α-helices could be reasonably represented. It will be interesting to see whether this method can be extended to classes of protein structure other than the all-β class.

3.7 Optimal sequence threading

The method described in the rest of this chapter and that following, has something in common both with the method of Bowie, Lüthy and Eisenberg, and that of Finkelstein and Reva. Despite the obvious computational advantages of using residue environments, it is clear that the fold of a protein chain is governed by fairly specific protein-protein and protein-solvent atomic interactions. A given protein fold is therefore better 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. The main difficulty in the use of environments alone for recognizing protein folds is that helices look like other helices, and strands like other strands. A sequence that folds into one helix of particular structure, will probably easily fold into any other helix of similar length. A very good example of two topologies which cannot be distinguished after encoding into environmental classes is an $(\alpha\beta)_8$ barrel (a "TIM barrel") and a parallel

αβ sandwich (a Rossmann fold). In this case both topologies comprise alternating α and β structure, where the strands are mostly inaccessible to solvent. Providing that the αβ sandwich is of sufficient size, or if flanking domain regions provide additional secondary structural elements (the Rossmann domain itself typically has only 6 strands), then the 1-D descriptors of the two structures are almost identical. This is illustrated in Figure 3.3, where the secondary structure and accessibility of TIM (triose phosphate isomerase) has been manually aligned with those of lactate dehydrogenase.

The factor that limits the scope of the search for a stable threading is packing. Whilst the sequence of any isolated helix could substitute for any other, the sequences for a packed pair of helices are much more highly constrained. For a complete protein structure, solvation effects also come into play. In general, then, for a globular protein, the threading of its sequence onto its structure is constrained by local interactions (in the example given, the required formation of a helix), long-range pairwise interactions (helix-helix packing for example) and solvation effects, which are primarily governed by the periodic accessibilities of exposed helices and strands.

In view of this, we should like to match a sequence to a structure by considering the plethora of detailed pairwise interactions, rather than averaging them into a crude environmental class. However, incorporation of such non-local interactions into standard alignment methods such as the algorithm of Needleman and Wunsch (1970), has hitherto proved computationally impractical. The next chapter will be concerned with possible solutions to this computational problem. The remainder of this chapter will describe the formulation of a potential function capable of distinguishing correctly folded from misfolded proteins.

Figure 3.3
Manually derived alignment of triose phosphate isomerase (TIM) with lactate dehydrogenase based on residue environments. Line 1 (TIM) / 3 (LDH) : amino acid sequence, Line 2 : residue accessibility (0 = 0-9%, 9 = 90-99%, * > 99%), Line 3 (TIM) / 1 (LDH) : secondary structure (H = α-helix, A = antiparallel strand, P = parallel strand, G = 3/10 helix, otherwise coil).

3.8 Formulating a model evaluation function

The general approach described here employs a set of information theoretic potentials similar to the recently described potentials of mean force (Sippl, 1990; Hendlich et al., 1990). These potentials associate event probabilities with statistical free energy. If a certain event is observed with probability p (say the occurrence of a leucine residue alpha-carbon and an alanine alpha-carbon at a separation of 5 Å) we can associate an 'energy' with this event by the application of the inverse Boltzmann formula:

$$E = -kT \ln[p] \ .$$

The constant $-kT$ may be ignored, in which case the units are no longer those of free energy but of *information* (in units of nats). For simplicity, we have also ignored the additional term Z, known as the Boltzmann sum. A clear explanation of why this is acceptable is given by Sippl (1990). The important point about both free energy and information entropy formulations of probability is that the resulting values are additive. Consider two independent events with probabilities p and q respectively. The probability of both events occurring together is simply pq, but multiplication is difficult to implement in pattern matching algorithms. Transforming the combined probability pq by taking logs provides the following useful result:

$$\ln[pq] = \ln[p] + \ln[q] \ .$$

Therefore the important part of the calculation of potentials of mean force, and the related techniques of information theory is simply converting probabilities to log-likelihoods.

Typically we are interested in relative rather than absolute probabilities. Taking the above example, it is of little interest to know how probable it is that a leucine alpha-carbon and an alanine alpha-carbon are found to be separated by 5 Å. Of much greater interest is the

question of how probable this leucine-alanine separation is in comparison with other residue pairs. If the probability of *any* residue pair having an alpha-carbon separation of s is $f(s)$ and the frequency of occurrence for residue pair ab is $f_{ab}(s)$ then we can write down the potential of mean force as follows:

$$\Delta E_{ab}(s) = -kT \ln[\frac{f_{ab}(s)}{f(s)}] .$$

Sippl divides this potential into a set of potentials relating to different topological levels $1..k$, which is simply the residue pair sequence separation. For the tripeptide sequence MFP, $k=1$ for residue pairs MF and FP, with $k=2$ for residue pair MP. In reality, probability density functions $f_k(s)$ and $f_k^{ab}(s)$ are unknown and must be replaced by the relative frequencies observed in the available structural database denoted $g_k(s)$ and $g_k^{ab}(s)$ respectively, where s is typically divided into 20 intervals for sampling. As there are 400 residue pairs (sequence asymmetry is assumed) and only some 15000-20000 residues in the set of non-homologous protein structures, the observed frequency distributions $g_k^{ab}(s)$ are only weak approximations of the true probability densities and must therefore be corrected to allow for the very small sample size. By considering the observation process as the collection of information quanta, Sippl suggests the following transformation:

$$f_k^{ab}(s) \approx \frac{1}{1+m\sigma} g_k(s) + \frac{m\sigma}{1+m\sigma} g_k^{ab}(s)$$

where m is the number of pairs ab observed at topological level k and σ is the weight given to each observation. As $m \to \infty$ this transformation has the required property that the right and left-hand sides of the equation become equal as $g_k^{ab}(s) \to f_k^{ab}(s)$. Given the number of residues in the database and the small number of histogram sampling intervals

it is assumed that $f_k(s) \approx g_k(s)$. From the previous two equations the following formula may be derived:

$$\Delta E_k^{ab} = kT \ln[1+m\sigma] - kT \ln[1+m_{ab}\sigma\frac{g_k^{ab}(s)}{g_k(s)}] .$$

The potentials used in this work are calculated exactly as described by Hendlich *et al.* (1990) where pairwise interatomic potentials are derived from a set of non-homologous proteins. The following interatomic potentials are calculated between the main chain N, O, and side chain Cβ: Cβ → Cβ, Cβ → N, Cβ → O, N → Cβ, N → O, O → Cβ, and O → N. In all, 7 pairwise interactions are considered between each pair of residues *i,j*. By excluding interactions between atoms beyond the Cβ atom in each residue, the potentials are rendered independent of specific side chain conformation. The proteins used for the generation of the potentials are listed in Table 3.2. Dummy Cβ atoms were constructed for glycine residues and other residues with missing Cβ atoms using a tetrahedral bond angle of 52.3° and Cα-Cβ distance of 1.538 Å (Hazes & Dijkstra, 1988).

A possible criticism of the mean force potentials proposed by Sippl is that there exists in the force field a dependence on protein size. The problem lies in the fact that interactions even as distant as 80 Å are taken into account in the calculation of the potentials, and so consequently, the bulk of data for these large distances is derived from large proteins. This was recognized by Hendlich *et al.* (1990), where it was suggested that the ideal case would be for the potentials to be calculated from proteins of roughly equal size to the protein of interest. Unfortunately, this simple solution is generally impractical. The data set used to generate the mean force potentials is already sparse, even before subdivision into size ranges.

In order to render the mean force potentials less dependent on protein chain length, these long distance interactions must be replaced by a size independent parameter. The first requirement in replacing these interaction parameters is to determine a suitable dividing line which separates short distance from long distance interactions. The next step is then

to determine the nature of the information encoded by these interactions. Finally, a suitable size independent parameter can be sought to replace this information.

Consider two protein atoms separated by a distance d. Clearly if d is large there will be no significant physical interaction between these atoms. Conversely, if d is small then we might expect there to be some influence, whether it be a hydrophobic effect, an electrostatic effect or even a covalent interaction. If such an influence exists, then we might also expect there to be some residue preferences for the residues containing these atoms, and consequently we would expect some kind of correlation between the two residue identities. This provides a possible way to determine a cut-off distance for meaningful residue-residue interactions. If the identities of two residues can be considered to be independent variables, then these residues (or more correctly, the residue side chains) will probably not be involved in a significant physical interaction.

To determine the degree of dependency between residues separated by a particular distance, some measure of statistical association is required. The method selected here is based on *statistical entropy*, a common concept in statistical physics and information theory. The entropy of a system with I states, where each state occurs with probability p_i is defined as:

$$H(x) = -\sum_{i=1}^{I} p_i \ln p_i$$

Consider two experiments x and y with I and J possible outcomes respectively, each of which occurs with a probability $p_{i.}$ ($i = 1 \ldots I$), and $p_{.j}$ ($j = 1 \ldots J$). The entropy H of these systems is defined as:

$$H(x) = -\sum_{i=1}^{I} p_{i.} \ln p_{i.}$$

and

$$H(y) = -\sum_{j=1}^{J} p_{.j} \ln p_{.j}$$

Entropy in this case is essentially defined as the degree of freedom of choice, or more strictly in this case, the degree of equiprobability. If the outcome probabilities in each experiment are equal then the statistical entropy is maximized, as this represents the maximum freedom of choice. If the probability of one outcome is unity (the others of course being zero) then zero entropy is achieved, corresponding to a lack of choice whatsoever.

If we link both experiments, then we can represent the overall outcomes in the form of a *contingency table*. An example of such a linked pair of experiments is the throwing of a pair of dice, in which case the contingency table would have 6 rows and 6 columns, representing the 6 possible outcomes for each die.

The entropy of the combined experiment is:

$$H(x,y) = -\sum_{i,j} p_{ij} \ln p_{ij}$$

To determine the statistical association between experiments x and y, the entropy of y *given* x and x *given* y may be derived. If a knowledge of the outcome of experiment x

allows a wholly accurate prediction of the outcome of experiment y, then the entropy of y *given* x must be zero. Conversely, if a knowledge of experiment x is found to be of no benefit whatsoever in the prediction of y, then the conditional entropy in this case is maximized.

The entropy of y *given* x is as follows:

$$H(y|x) = \sum_i p_i \sum \frac{p_{ij}}{p_{i\cdot}} \ln \frac{p_{ij}}{p_{i\cdot}} = \sum_{i,j} p_{ij} \ln \frac{p_{ij}}{p_{i\cdot}}$$

and the entropy of x *given* y:

$$H(x|y) = \sum_i p_i \sum \frac{p_{ij}}{p_{\cdot j}} \ln \frac{p_{ij}}{p_{\cdot j}} = \sum_{i,j} p_{ij} \ln \frac{p_{ij}}{p_{\cdot j}} \quad .$$

Finally a suitable symmetric measure of interdependence (known as the *uncertainty*) between x and y is defined thus:

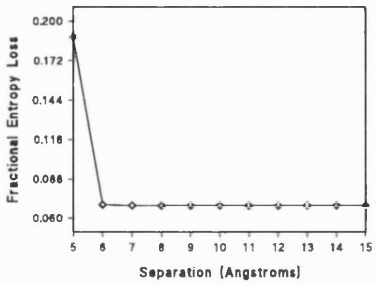
$$U(x,y) \equiv 2\left[\frac{H(y)+H(x)-H(x,y)}{H(x)+H(y)}\right]$$

An uncertainty between x and y of zero indicates that the two experimental variables are totally independent ($H(x,y) = H(x) + H(y)$), whereas an uncertainty of one ($H(x) = H(y) = H(x,y)$) indicates that the two variables are totally dependent. One would hope that in the case of the two dice experiment previously described, that $U(x,y)$ would be found to be close to zero for a large number of trials, though gluing the dice together would be a sure way of forcing $U(x,y)$ to unity.

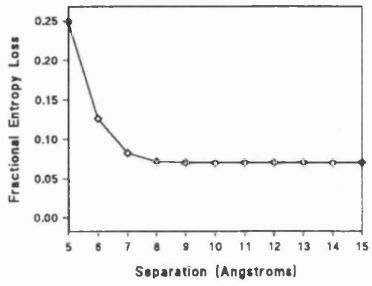
Using the uncertainty measure, it is now possible evaluate residue correlations in protein structures. In this case, the two experimental variables are the identities of two residues separated by a given distance in a particular structure. Using the 102 chains listed in Table 3.2, six 20 x 20 contingency tables were set up for each distance range. The first

5 tables were constructed by counting residue pairs with sequence separations of 1 to 5 (short *range* interactions), the other being constructed by counting all pairs with sequence separations > 10 (long range). Values in each table were converted to relative frequencies by normalization such that:

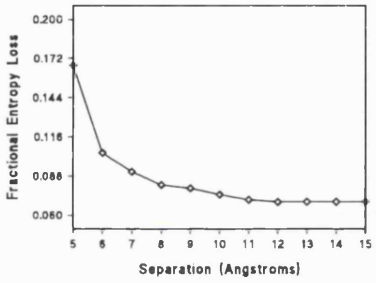
$$\sum_{i,j} p_{ij} = 1$$



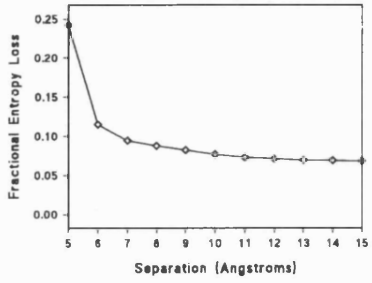
a)



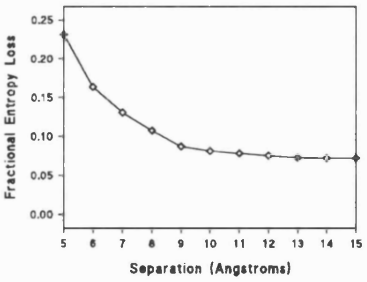
b)



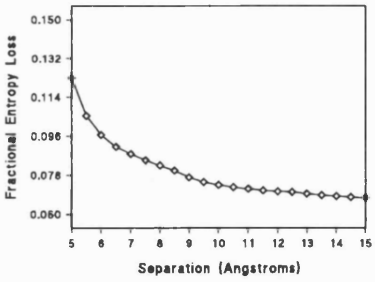
c)



d)



e)



f)

Figure 3.4

Uncertainty coefficient (fractional loss of statistical entropy) for residue identities over sequence separations: a) 1, b) 2, c) 3, d) 4, e) 5, f) >10. The maximum observed distances for each sequence separation are as follows: a) 6.36 Å, b) 9.72 Å, c) 13.08 Å, d) 16.45 Å, e) 19.43 Å, f) > 32.97 Å. Points beyond these distances have no meaning.

The plots in Figure 3.4 clearly show the ranges over which short range and long range effects can be detected statistically. As might be expected, the strongest sequence specific effects are observed across short sequential separations, where steric and covalent effects predominate. Most importantly, both the short and long range interactions become undetectable when averaged over distances greater than around 12 Å (though it must be realized that for the very short separations, 1 and 2, it is impossible for the separation to exceed 10 Å). It must be stressed that this doesn't necessarily imply that the physical forces themselves do not act beyond this distance, only that the effects do not manifest themselves in the selection of amino acid residues.

Bearing in mind the observable extent of detectable sequence specific effects, the calculation of mean force potentials in this work was modified from the method described by Sippl (1990). Rather than taking into account all interactions up to around 80 Å, only atom pairs separated by 10 Å or less were used. However, much useful information remains in the long distance distributions. Considering a protein molecule as a globule comprising an inner hydrophobic core it is readily apparent that the bulk of the longer pairwise distances will originate from residue pairs distributed on the surface of the globule, which is illustrated in Figure 3.5.

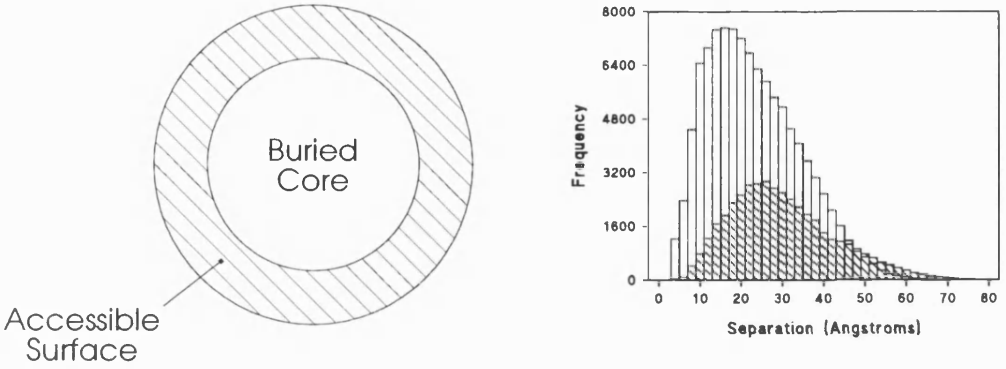


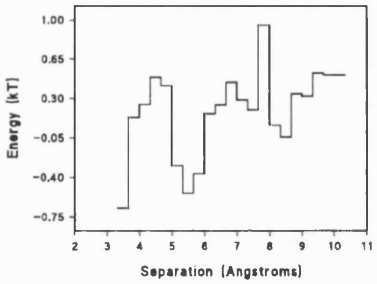
Figure 3.5

Distance distributions for accessible (shaded) and inaccessible residue pairs in monomeric protein structures. Buried residues are taken to be those with relative accessibilities < 5%; accessible residues with relative accessibilities > 50%.

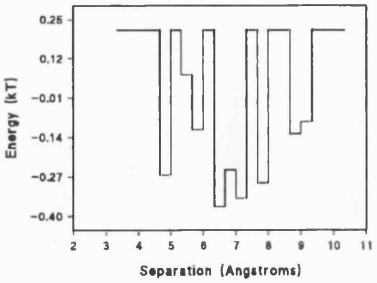
As the excluded long distance potentials clearly only encode information pertaining to the hydrophobic effect, the most logical replacement for these interactions must be a potential based on the solvent accessibility of the amino acid residues in a structure.

3.9 Calculation of potentials

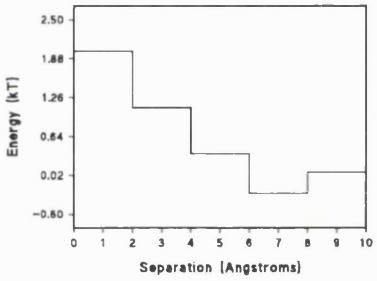
For the short range potentials, minimum and maximum pairwise distances were determined for each type of atomic pairing at each topological level from a small set of very highly resolved crystal structures. These distance ranges were subdivided into 20 intervals. For the medium and long range potentials, interactions were sampled over the range 0-10 Å with a fixed sampling interval of 2 Å. A small selection of the pairwise interaction potentials is show in Figure 3.6.



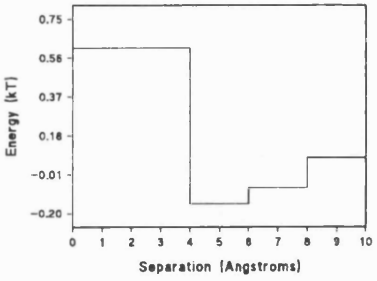
a)



b)



c)



d)

Figure 3.6

Sample pairwise potentials: a) Short range (k=3) Ala-Ala Cβ-Cβ, b) Short range (k=3) Phe-Tyr Cβ-Cβ, c) Long range (30 < k < 50) Ala-Ala Cβ-Cβ, d) Long range (30 < k < 50) Arg-Glu Cβ-Cβ.

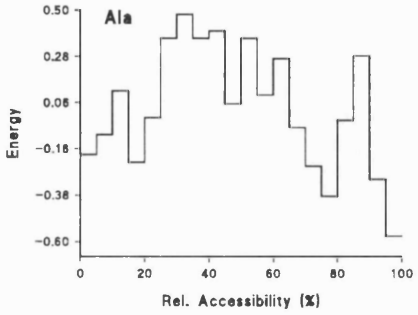
As discussed in the previous section, in addition to the pairwise potentials (and in place of the long range, long distance interactions), a solvation potential was also incorporated. This potential simply measures the frequency with which each amino acid species is found

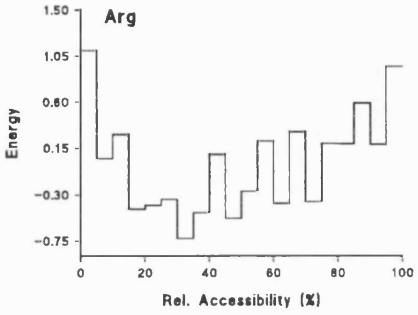
with a certain degree of solvation, approximated by the residue solvent accessible surface area. The solvation potential for amino acid residue a is defined as follows:

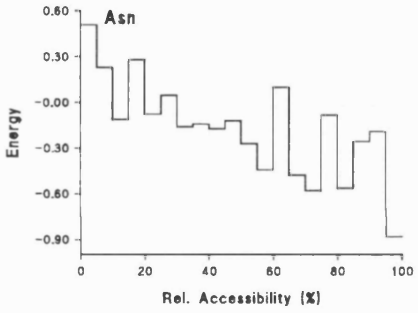
$$\Delta E^{a}_{solv.}(r) = -kT \ln[\frac{f^{a}(r)}{f(r)}]$$

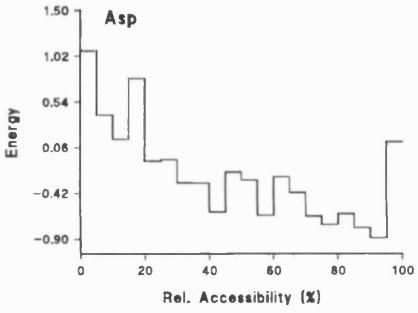
where r is the % residue accessibility (relative to residue accessibility in GGXGG extended pentapeptide). Residue accessibilities were calculated using the DSSP program of Kabsch and Sander, 1983, which uses an algorithm similar to that described by Shrake and Rupley (1973) for the calculation of surface area. The solvation potentials were generated with a histogram sampling interval of 5 %. To ensure that subunit or domain interactions did not affect the results, only monomeric proteins were used in the calculation. As can be seen in Figure 3.7, the solvation potentials clearly show the hydropathic nature of the amino acids and prove to be a more sensitive measure of the likelihood of finding a particular amino acid with a given relative solvent accessibility than the long distance interaction potentials they are designed to replace.

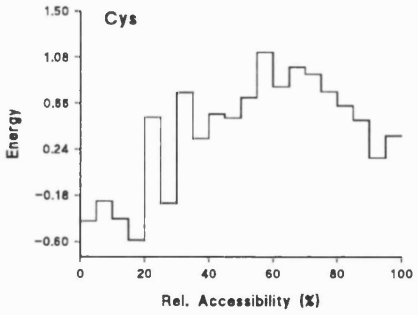
Physical descriptors other than residue accessibility may be taken into account by means of similar potentials. In particular, 8 Å and 14 Å Ooi numbers (Nishikawa & Ooi, 1986) have been examined, which correlate with the radial location of amino acids in globular protein structures. Being defined purely in terms of Cα distances, Ooi numbers have the advantage of being totally independent of the sequence of amino acids, which is not the case for relative accessibilities. Unfortunately, in most cases Ooi numbers were found to be less useful than accessibilities for improving sequence-to-structure alignments. This can be attributed to the fact that though Ooi numbers are more highly conserved than relative accessibilities across equivalent sites in a related set of protein structures, their ability to sharply define surface residues is poor. This lack of definition is due to the fact that protein surfaces are highly convoluted, whereas the underlying assumption in the calculation of Ooi numbers is that proteins are smooth globules.

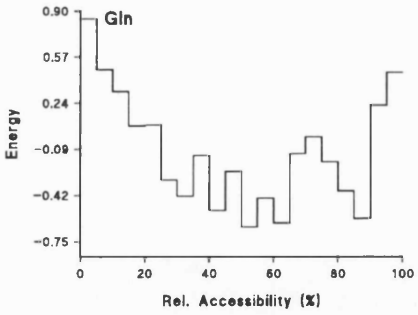

Ala
Energy
0.50
0.28
0.06
-0.16
-0.38
-0.60
0
20
40
60
80
100
Rel. Accessibility (%)

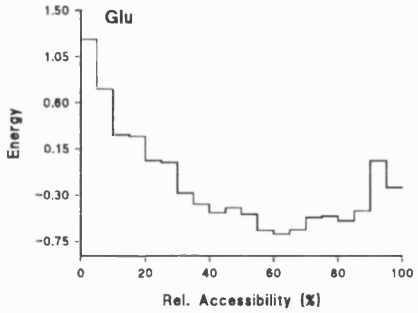

Arg
Energy
1.50
1.05
0.60
0.15
-0.30
-0.75
0
20
40
60
80
100
Rel. Accessibility (%)

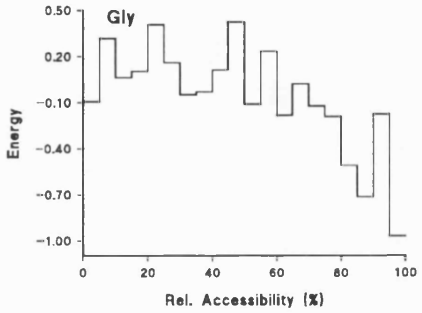

Asn
Energy
0.60
0.30
-0.00
-0.30
-0.60
-0.90
0
20
40
60
80
100
Rel. Accessibility (%)

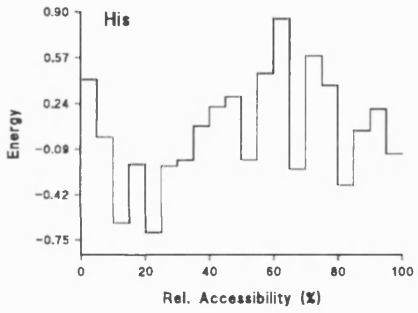

Asp
Energy
1.50
1.02
0.54
0.06
-0.42
-0.90
0
20
40
60
80
100
Rel. Accessibility (%)

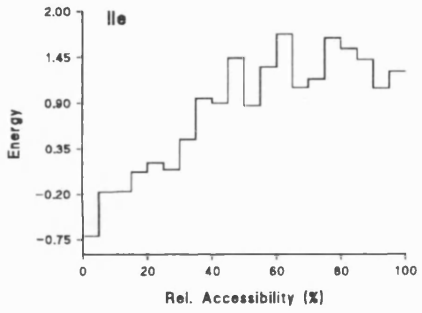

Cys
Energy
1.50
1.08
0.65
0.24
-0.18
-0.60
0
20
40
60
80
100
Rel. Accessibility (%)

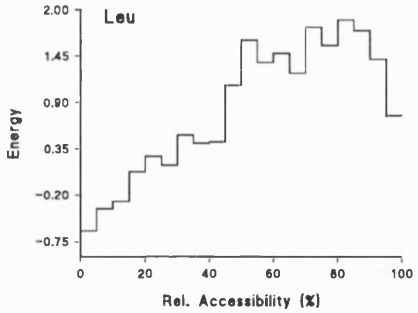

Gln
Energy
0.90
0.57
0.24
-0.09
-0.42
-0.75
0
20
40
60
80
100
Rel. Accessibility (%)

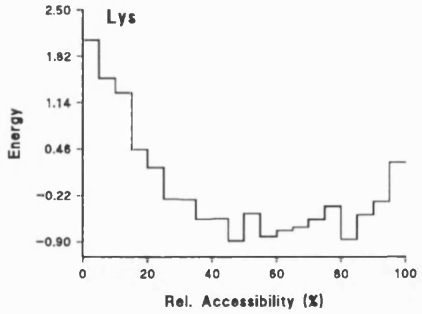

Glu
Energy
1.50
1.05
0.60
0.15
−0.30
−0.75
0
20
40
60
80
100
Rel. Accessibility (%)

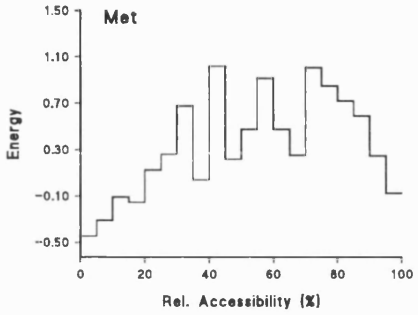

Gly
Energy
0.50
0.20
−0.10
−0.40
−0.70
−1.00
0
20
40
60
80
100
Rel. Accessibility (%)

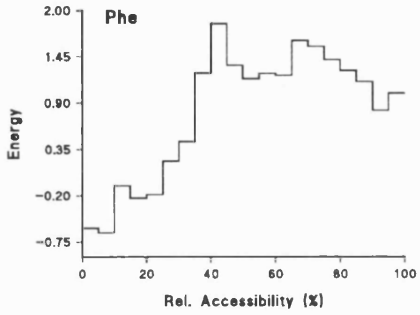

His
Energy
0.90
0.57
0.24
−0.09
−0.42
−0.75
0
20
40
60
80
100
Rel. Accessibility (%)

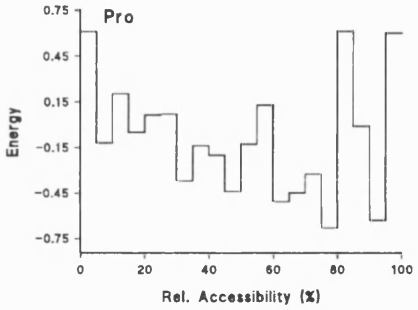

Ile
Energy
2.00
1.45
0.90
0.35
−0.20
−0.75
0
20
40
60
80
100
Rel. Accessibility (%)

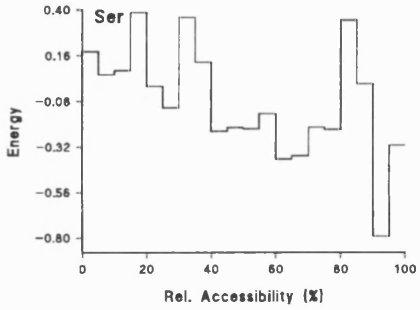

Leu
Energy
2.00
1.45
0.90
0.35
−0.20
−0.75
0
20
40
60
80
100
Rel. Accessibility (%)

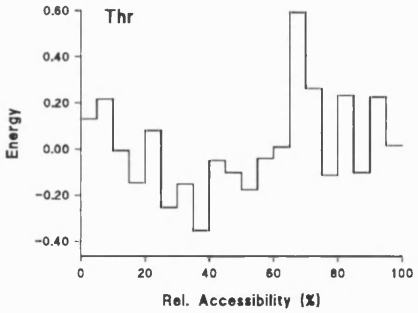

Lys
Energy
2.50
1.82
1.14
0.46
−0.22
−0.90
0
20
40
60
80
100
Rel. Accessibility (%)

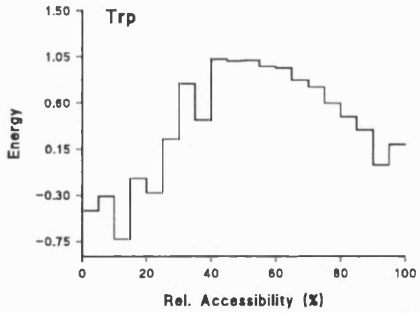

Met
Energy
1.50
1.10
0.70
0.30
−0.10
−0.50
0
20
40
60
80
100
Rel. Accessibility (%)


Phe
Energy
2.00
1.45
0.90
0.35
−0.20
−0.75
0
20
40
60
80
100
Rel. Accessibility (%)


Pro
Energy
0.75
0.45
0.15
−0.15
−0.45
−0.75
0
20
40
60
80
100
Rel. Accessibility (%)


Ser
Energy
0.40
0.16
−0.08
−0.32
−0.56
−0.80
0
20
40
60
80
100
Rel. Accessibility (%)


Thr
Energy
0.60
0.40
0.20
0.00
−0.20
−0.40
0
20
40
60
80
100
Rel. Accessibility (%)


Trp
Energy
1.50
1.05
0.60
0.15
−0.30
−0.75
0
20
40
60
80
100
Rel. Accessibility (%)

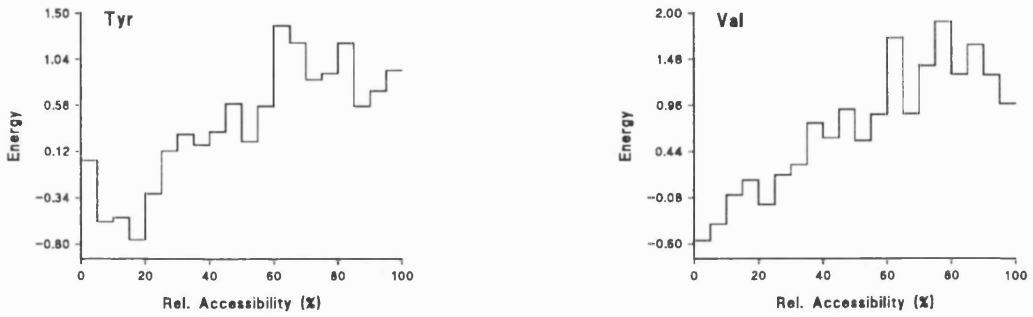


Figure 3.7

Solvation potentials for the 20 standard amino acids based on monomeric proteins. The potentials (in units of kT) are shown plotted against residue relative accessibility, as described in the text.

3.10 A sequence similarity potential

Whilst in the general case it is not possible to assume any sequence similarity between the template and the sequence being modelled, in the cases where there in fact is a degree of sequence similarity, it is advantageous to weight this factor into the evaluation function. A third set of potentials may therefore be defined, which allow sequence similarity (residue similarity to be more precise), the mainstay of traditional alignment methods, to be considered in the overall methodology. Many residue similarity measures are available based on physico-chemical characteristics (Nakai et al., 1988), sequence alignments (Dayhoff, 1978) or structural superposition (Risler, 1988; Overington et al., 1990, 1992). Of these methods, mutational statistics based on sequence alignments are the easiest to

express in potential energy terms. A suitable statistical analysis of amino acid mutation was carried out by Dayhoff, 1978, who expressed the mutations observed across a family of proteins in terms of a simple Markovian model. By considering alignments between very similar sequences, the probabilities of exchange for all 400 pairs of amino acid residues could be estimated. Dividing these raw exchange probabilities $p(A \rightarrow B)$ by the relative frequencies of occurrence of the resultant amino acids $f(B)$, produces a matrix of *relatedness odds*. These odds represent the ratio between the observed frequencies of exchange and the frequencies expected by chance. A relatedness odds value significantly less than 1.0 represents an unfavourable amino acid exchange, and a value significantly greater than 1.0, a favourable exchange. There is a clear analogy between this calculation and that used to calculate the relative frequencies prior to calculation of the potentials of mean force using the inverse Boltzmann equation, and indeed Dayhoff suggests transforming the relatedness odds matrix into a *log odds* matrix to circumvent the requirement of multiplying the relatedness odd values during sequence alignment. Dayhoff simply used base-10 logarithms which resulted in the values contained in the log odds matrix having units of decimal digits (again information entropy). By transforming the relatedness odds matrix R_{ij} using the following formula, we can express the exchange preferences for the corresponding 400 amino acid pairs *ab* in terms of free energy changes:

$$\Delta E_{mut.}^{a \rightarrow b} = -kT \ln[R_{ij}] \quad .$$

A positive mutational free energy thus represents an unfavourable apparent amino acid exchange (Trp and Asn for example) and negative free energy, a favourable one (Ser and Thr for example).

As the Dayhoff matrix had not been updated since 1978, a more up to date matrix was computed from the current sequence databank (see Chapter 2 and Jones *et al.*, 1992a), and use was made of a 250 PAM relatedness odds matrix based on 71000 accepted point mutations found in Release 17 of the SWISS-PROT sequence databank (Bairoch & Boeckmann, 1990). A further possibility would be to use environment-specific mutation

data matrices, which have been found to outperform generic matrices in some cases (Chapter 2; Overington *et al.*, 1990, 1992; Lüthy *et al.*, 1991). It must be stressed that in the examples presented in this and the following chapter, free energy terms relating to sequence similarity were totally excluded from the threading evaluation function, and thus the structural template encoded no sequence information from the template protein.

3.11 The final evaluation function

The final evaluation function is expressed as follows:

$$\Delta E_{(TOTAL)} = W_l \Delta E_{(PAIRWISE\ INTERACTION)} + W_s \Delta E_{(SOLVATION)} + W_m \Delta E_{(MUTATION)}$$

where W_l, W_s, and W_m are weighting factors for the three free energy components. The first term may be thought of as relating to the likelihood of the two peptides folding into a similar conformation, the second term to the likelihood of finding hydrophilic residues in solvent accessible sites (and hydrophobics in buried sites), with the third term expressing the degree of similarity between the two peptide sequences. For all the results shown here, the following weights were used: W_l = 1.0, W_s = 15.0, and W_m = 0.0. In practice the mutability term would be given a positive weight, but in order to evaluate the threading algorithm proposed here it must be excluded so as to avoid biasing the threading towards simple sequence alignment rather than structural matching. The choice of 15.0 for W_s was made for the reason that this has been found to be the average ratio between the sum of pairwise potentials terms and the sum of residue solvation potentials for a number of test proteins. This factor succeeds in balancing the overall contributions of the pairwise and solvation potentials.

3.12 Accommodating structural variation

The discrete nature of the potentials used in the evaluation function, and the small size of the knowledge base sample used to construct them results in the evaluation function being rather too sensitive to structural variation. This is of little consequence when attempting to thread a sequence onto its native structural template as the interatomic distances will be exactly correct, within the bounds of standard crystallographic errors. When attempting the threading of a sequence onto a non-native template, the set of pairwise distances will not be correct for the given sequence. This is of course the fundamental limitation of homology modelling. A protein model can only ever be an approximation of the correct structure. In this work it is naively assumed that we can simply thread the sequence of one protein onto the framework of another even though we know that the structure cannot be an accurate model. Clearly if we are going to identify the correct threading of a sequence onto a non-native structure we must accommodate the structural variation that would have occurred during the evolution of the object sequence and the native template sequence. To do this we can either 'relax' the structural template or 'soften' the evaluation function. Relaxing the structure would mean varying the coordinates of every atom under consideration so as to minimize the overall threading energy. Clearly this is not feasible as for every trial threading a complete energy minimization procedure would need to be performed. A more tractable approach is to locally minimize each pairwise interaction over a small range. An additional distance variation parameter, δ, is defined for the evaluation function. This parameter effectively softens the mean potential field by making allowances for subtle variations between the distance matrix of the template structure and the unknown distance matrix of the model structure. The larger δ becomes, the greater the permitted variation. In practice for any given pairwise interaction potential $\Delta E_k^{ab}(d)$, we locally optimize this potential over the range $d\text{-}\delta \leq d \leq d\text{+}\delta$. Of course, physically, each element in a distance matrix is not independent of the others. We ideally should like to minimize the evaluation function in a true 3 dimensional space, but the computational cost of this is too high. We end up resorting to minimizing the evaluation function for N atoms along a single direction vector

in an *(N-1)-dimensional* space. Despite these shortcomings, such a simplistic means for softening the mean potential field is surprisingly effective for small values of δ. For the results shown in this chapter, δ was set to zero, though for the examples of non-native fold recognition presented in the next chapter, where a need for a greater degree of tolerance in structural variation is recognized, δ was set to 0.20 Å unless otherwise noted.

Table 3.3 shows the results of evaluating the potentials described here using the method previously described by Hendlich *et al.* (1990). In this test the protein sequence of interest is blindly fitted to all contiguous structural fragments taken from the entire library of chain folds, and the energy terms summed in each case. As 5ACN is the largest chain in the library, no alternative contiguous chain conformations are available for it. The two aspects of the potentials described here (the truncated pairwise terms and the solvation terms) are tested independently, with the total energy values being calculated as E_{pair} + 15 E_{solv}. Z-scores were calculated using the formula:

$$Z = \frac{E_{native} - \bar{E}}{\sigma}$$

where E_{native} is the energy of the native threading, $\bar{E}$ the mean energy, and σ the unbiased standard deviation of the distribution of threading energies.

Table 3.3

An evaluation of the potentials by fitting the sequences of each member of the fold library onto every contiguous section of the same set of structures. See text for details.

PDB Code	Epair (kT)	Pair Z-score	Pair rank	Esolv (kT)	Solv Z-score	Solv rank	Etotal	Tot. Z-score	Tot. rank	Structures
1ABP	11.9	-3.9	1	-28.03	-9.2	1	-408.59	-9.5	1	1786
1BP2	-14.94	-6.5	1	-3.76	-4	2	-71.3	-4.4	2	8901
1CC5	-10.03	-5.3	1	-2.81	-3.3	3	-52.22	-3.8	1	12051
1CCR	-8.91	-5.4	1	-11.37	-5.9	1	-179.49	-6.3	1	9752
1CD4	2.03	-3.8	1	-17.89	-7.6	1	-266.26	-7.9	1	6081
1CRN	0.45	-1.9	511	1.81	-0.4	5322	27.61	-0.6	4458	15554
1CTF	-10.16	-4.2	1	-5.73	-4.3	1	-96.11	-4.7	1	13399
1CY3	14.67	-0.3	3679	-5.13	-4.1	1	-62.34	-4.1	1	9246
1ECA	-12.06	-6	1	-16.85	-7.1	1	-264.8	-7.6	1	8059
1FD2	-3.94	-3.9	1	-8.48	-5.2	1	-131.17	-5.5	1	10131
1FX1	-6.64	-4.9	1	-16.13	-7.1	1	-248.58	-7.5	1	7394
1GCR	-7.26	-5.4	1	-19.42	-8.6	1	-298.63	-9	1	6035
1HIP	-3.21	-3.3	10	-4.25	-3.8	2	-66.89	-4.1	1	11878
1HOE	-5.04	-4	1	2.34	-1	2087	30.05	-1.3	1178	12849
1I1B	0.63	-3.6	2	-12.51	-6.2	1	-186.95	-6.4	1	7169
1L01	-4.64	-4.4	1	-18.24	-7.5	1	-278.24	-7.8	1	6504
1LH1	-3.27	-4	1	-17.7	-7.2	1	-268.84	-7.5	1	7059
1LZ1	-14.37	-6.3	1	-12.27	-6.3	1	-198.43	-6.8	1	8438
1MBA	-15.21	-6.1	1	-15.62	-6.8	1	-249.46	-7.3	1	7453
1MBD	-19.87	-6.6	1	-19.78	-7.7	1	-316.64	-8.2	1	7059
1PAZ	-10.84	-5.3	1	-10.64	-5.7	1	-170.47	-6.1	1	9106

PDB Code	Epair (kT)	Pair Z-score	Pair rank	Esolv (kT)	Solv Z-score	Solv rank	Etotal	Tot. Z-score	Tot. rank	Structures
1PCY	-7.45	-5	1	-10.74	-6.3	1	-168.54	-6.7	1	10698
1PHH	1.61	-6	1	-30.54	-10.2	1	-456.44	-10.8	1	593
1RHD	-17.48	-7.3	1	-24.69	-8.6	1	-387.78	-9.2	1	2087
1SN3	-6.06	-4.4	1	-2.01	-2.9	23	-36.18	-3.2	8	13680
1UBQ	-5.98	-4.2	2	-9.17	-5.4	1	-143.58	-5.7	1	12668
1UTG	-0.7	-2.6	54	1.94	-1.7	607	28.46	-1.9	412	13214
2AAT	23.77	-3.6	1	-30.23	-9.7	1	-429.72	-10	1	578
2CA2	-12.9	-6.9	1	-23.19	-8.7	1	-360.71	-9.2	1	3059
2CDV	-4.71	-4.1	1	-9.46	-4.9	1	-146.54	-5.2	1	10052
2CNA	8.66	-3.6	1	-15.95	-6.9	1	-230.65	-7.2	1	3641
2CPP	-23.11	-8.9	1	-36.4	-11.2	1	-569.15	-11.9	1	523
2CRO	-11.53	-5.1	1	-1.92	-2.9	18	-40.26	-3.3	6	13680
2CYP	-30.55	-9.3	1	-20.17	-7.6	1	-333.07	-8.2	1	2087
2GBP	-33.57	-9.1	1	-26.37	-9.2	1	-429.07	-10	1	1720
2LBP	-14.75	-7.7	1	-37.79	-10.8	1	-581.66	-11.4	1	1122
2MHR	-2.71	-3.7	1	-10.16	-5.6	1	-155.19	-5.9	1	9246
2OVO	0.73	-2	352	-2.87	-3.1	12	-42.33	-3.2	8	14548
2RHE	-6.23	-4.8	1	-6.98	-4.7	2	-110.96	-5.1	2	9532
2RNT	-10.33	-5.4	1	-8.14	-5.4	1	-132.38	-5.9	1	10291
2SGA	-16.95	-6.9	1	-8.91	-4.8	1	-150.54	-5.6	1	5726
2SNS	-2.91	-3.8	1	-11.86	-5.7	1	-180.79	-5.9	1	7750

PDB Code	Epair (kT)	Pair Z-score	Pair rank	Esolv (kT)	Solv Z-score	Solv rank	Etotal	Tot. Z-score	Tot. rank	Structures
2SSI	3.31	-2.6	37	-5.59	-4.1	1	-80.54	-4.3	1	10052
2STV	3.51	-3.6	2	4.27	-2.4	46	67.55	-2.7	21	5410
31LRD	-11.15	-4.7	2	-6.41	-4.4	1	-107.32	-4.8	1	11421
351C	-11.25	-4.9	1	-7.28	-5	1	-120.52	-5.4	1	11848
3ADK	-13.1	-6.4	1	-21.98	-8.1	1	-342.82	-8.5	1	5019
3BLM	-5.8	-5.7	1	-26.95	-8.3	1	-409.97	-8.6	1	2913
3CLA	-14.94	-6.4	1	-17.12	-7.1	1	-271.81	-7.6	1	4300
3DFR	-11.28	-5.9	1	-13.29	-6.3	1	-210.6	-6.8	1	6390
3FXC	3.51	-2.2	144	-4.85	-4.2	1	-69.17	-4.4	1	10507
3GRS	-19.29	-8.6	1	-26.98	-9.8	1	-424.06	-10.4	1	333
3ICB	-13.92	-5.8	1	-10.88	-5.6	1	-177.13	-6	1	12460
3ICD	-23.81	-8.4	1	-24.91	-9.8	1	-397.42	-10.5	1	477
3PGK	41.28	-2.5	6	-28.01	-10.2	1	-378.84	-10.3	1	472
3PGM	13.32	-2.6	18	-16.33	-7.1	1	-231.67	-7.3	1	3723
4CPV	-13.08	-5.3	1	-12.69	-6.5	1	-203.49	-6.9	1	9710
4FXN	-2	-3.9	1	-17.34	-7.3	1	-262.17	-7.6	1	7699
4PTP	-15.64	-7.2	1	-13.2	-6.5	1	-213.7	-7.2	1	3955
4RXN	-1.41	-2.7	61	-2.5	-2.8	22	-38.95	-3	11	14428
4TNC	-24.39	-7.3	1	-14.92	-6.5	1	-248.12	-7.1	1	6487
5ACN	-60.29	-	1	-41.44	-	1	-681.86	-	1	1
5CPA	-7.57	-6.3	1	-16.24	-7.3	1	-251.1	-7.8	1	1697

PDB Code	Epair (kT)	Pair Z-score	Pair rank	Esolv (kT)	Solv Z-score	Solv rank	Etotal	Tot. Z-score	Tot. rank	Structures
5PTI	-11.11	-6.1	1	-2.44	-2.9	15	-47.78	-3.4	4	14037
6LDH	-2.05	-5.7	1	-17.52	-7.5	1	-264.84	-8	1	1313
7RSA	-5.02	-4.2	1	-6.76	-4.7	1	-106.35	-5	1	8584
8ADH	-11.23	-7.1	1	-27.35	-9.3	1	-421.41	-9.9	1	786
9PAP	-14.52	-6.5	1	-6.35	-5	1	-109.81	-5.6	1	4337
A1DHF	-12.3	-6.1	1	-15.42	-7	1	-243.63	-7.4	1	5491
A1PFK	-28.61	-8.5	1	-27.56	-9.1	1	-442.04	-9.8	1	1458
A1TNF	7.95	-2.3	70	-14.42	-6.5	1	-208.32	-6.6	1	6892
A1WSY	-38.56	-9.1	1	-19.42	-7.4	1	-329.93	-8.2	1	2696
A1YPI	-20.96	-7.6	1	-18.97	-7.6	1	-305.45	-8.2	1	3204
A256B	-24.04	-6.8	1	-13.1	-6	1	-220.48	-6.6	1	9864
A2AZA	-12.11	-6	1	-13.01	-6.7	1	-207.28	-7.1	1	8259
A2CCY	-14.19	-5.1	1	-9.16	-5.4	1	-151.59	-5.9	1	8388
A2HLA	-0.53	-5.1	1	-22.29	-8.1	1	-334.92	-8.4	1	2565
A2LTN	-4.82	-4.9	1	-11.26	-5.7	1	-173.74	-6.1	1	5534
A2PAB	-8.89	-5.1	1	-4.5	-3.8	1	-76.35	-4.2	1	9273
A3GAP	-1.62	-4.4	1	-21.05	-8.1	1	-317.42	-8.4	1	4486
A3HHB	-16.36	-6.3	1	-9.21	-5.1	1	-154.48	-5.7	1	7518
A4DFR	-10.12	-5.4	1	-12.35	-6.2	1	-195.4	-6.6	1	6537
A4MDH	-25	-8	1	-28.68	-9.5	1	-455.26	-10.1	1	1254
A4XIA	-35.48	-9.6	1	-15.7	-7.5	1	-270.97	-8.3	1	602

PDB Code	Epair (kT)	Pair Z-score	Pair rank	Esolv (kT)	Solv Z-score	Solv rank	Etotal	Tot. Z-score	Tot. rank	Structures
A7CAT	-13.52	-8.2	1	-1.34	-7.4	1	-33.6	-8.1	1	258
A8ATC	-19.69	-7.3	1	-21.98	-8.3	1	-349.42	-8.9	1	1636
A9WGA	-22.57	-8.4	1	-6.25	-4.7	1	-116.25	-5.5	1	5972
B1WSY	-10.49	-6.9	1	-14.53	-7.7	1	-228.42	-8.3	1	675
B2HLA	-8.92	-5.2	1	-12.6	-6.2	1	-197.9	-6.5	1	10424
B8ATC	-2.9	-4.1	2	-12.24	-6.6	1	-186.5	-6.9	1	7226
E1CSE	-25.97	-8.6	1	-21.71	-8	1	-351.57	-8.9	1	2464
E2ER7	-3.81	-6.4	1	-18.85	-7.4	1	-286.55	-8.1	1	1297
E2TMN	-3.69	-6.2	1	-10.71	-6.3	1	-164.27	-6.8	1	1527
H2FB4	-10.52	-6.5	1	-17	-7	1	-265.59	-7.5	1	3756
I1CSE	-6.7	-4.3	1	-3.19	-3.2	10	-54.57	-3.5	3	13561
I1TGS	-1.98	-2.9	40	-2.59	-2.9	17	-40.81	-3.1	8	14231
I4CPA	-0.95	-2.2	276	-3.55	-3.1	7	-54.18	-3.2	4	16137
I4SGB	-4.09	-3.8	2	-0.68	-2	279	-14.34	-2.2	130	14726
L2FB4	-12.11	-6.5	1	-18.59	-7.8	1	-290.96	-8.4	1	4194
O1GD1	-36.98	-9.2	1	-19.94	-8.3	1	-336.1	-9.2	1	1240
O2SOD	-9.11	-5.7	1	-10.81	-5.6	1	-171.27	-6.1	1	6947
R2WRP	-10.9	-4.9	1	-2.21	-3.4	4	-44.06	-3.9	1	10022

Excluding 5ACN, of the remaining 101 chains, 86 of the native conformations are recognized from the set of decoys. This success rate (85%) is much higher than that achieved by Hendlich *et al.* (1990), where only 63% of the native chain conformations could be distinguished. The success is also higher than that achieved by Crippen (1991) where the potentials were actually fitted to a training set of folds such that the potentials were forced to recognize the native conformations in the training set. In fact, both contributions to the total energy are very able at distinguishing the native conformations. The pairwise potential alone recognizes 82 out of 101 (81%) chain conformations, and the solvation potential alone recognizes 84 out of 101 (83%). It must be stressed that in all the trials described, the native chain *and any detectable sequence homologues* (sequences > 25% identical to the native) were excluded from the calculation of the mean force potentials.

In order to test the validity of the proposed method for aligning a sequence with a structure it is vital to determine whether or not the structural parameters alone are capable of determining the correct threading when a sequence is being threaded on its correct native structural template. Ideally, in this case, we should be able to identify the correct threading amongst all possible decoys. This is a harder test than the contiguous conformation test. It is not possible to explore the space of all possible threadings for any protein sequence-structure relationship. Even for a small protein such as PTI (pancreatic trypsin inhibitor), there are an enormous number of possible threadings, far too many to search through by an exhaustive search (this point is discussed more fully in the following chapter). For small structures, we can of course limit indels to the loop regions and simply slide the secondary structures along the sequence being threaded, as shown in Figure 3.8. As a first step, therefore, all possible threading scores were computed for a number of sufficiently small proteins, with indels limited to the loop regions.

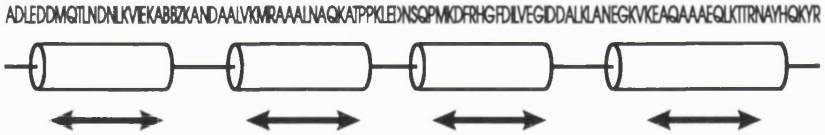
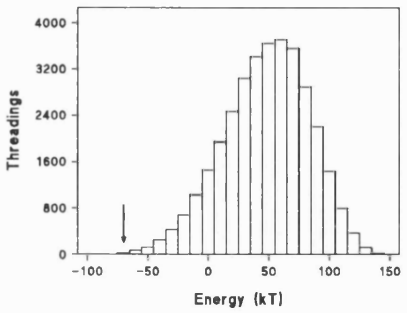
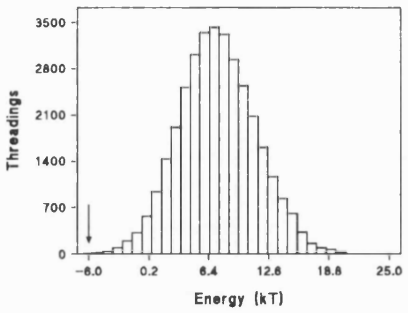


Figure 3.8

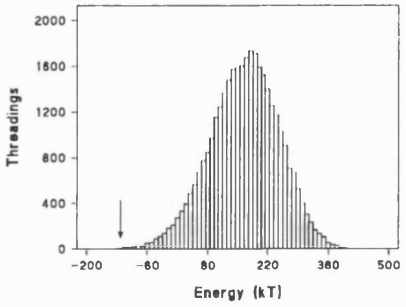
Diagram illustrating the exhaustive threading procedure. Secondary structures are effectively moved along the sequence like abacus beads.



1R69 - pairwise



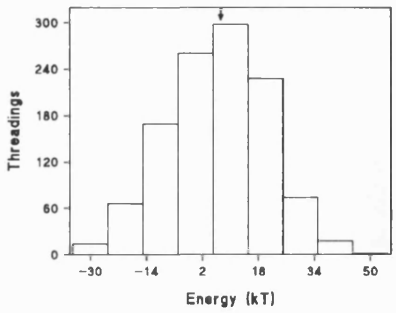
1R69 - solvation



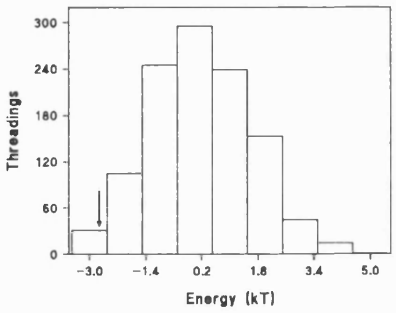
1R69 - pairwise + solvation

Figure 3.9

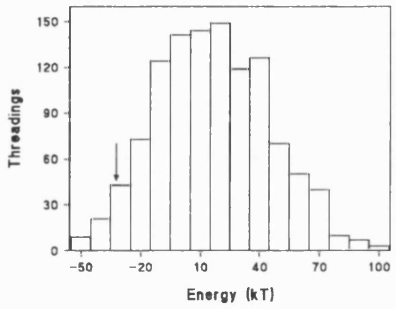
Exhaustive threading energy histograms for repressor 1R69. The position of the native threading is indicated by an arrow.



4RXN - pairwise



4RXN - solvation



4RXN - pairwise + solvation

Figure 3.10

Exhaustive threading energy histograms for rubredoxin 4RXN. The position of native threading is indicated with an arrow.

Table 3.4

The exhaustive native threading results for a number of small proteins using the following potentials: a) Cβ → Cβ, b) Cβ → O, c) Cβ → N, d) O → Cβ, e) O → N, f) N → Cβ, g) N → O, h) all pairs, i) solvation and j) all pairs + solvation (x 15). Where a letter precedes the PDB code, only the specified chain or subunit was threaded, though calculated accessibilities in these cases were calculated on the entire multimer or complex.

PDB Code	Energy (kT)	Z-score	Rank	Total Threadings
1CSE	-11.86	-2.9	4	9140
1CTF	-33.22	-3.5	1	925
1R69	-13.75	-4	1	33650
1SN3	-7.13	-2.6	90	16216
1UTG	-2.27	-1.4	341	4846
2CRO	-8.42	-1.9	1466	53131
2OVO	-2.45	-2	244	9140
3ICB	-14.98	-4.2	2	35961
4RXN	1.15	-0.2	505	1129
5PTI	-26.41	-4.9	1	35961
A256B	-45.65	-5.2	1	10627
I1TGS	-1	-1.7	513	9881
I4SGB	-6.1	-2.4	87	10661

a)

PDB Code	Energy (kT)	Z-score	Rank	Total Threadings
1CSE	-12.28	-2.9	17	9140
1CTF	-17.86	-2.6	7	925
1R69	-7.3	-3.3	2	33650
1SN3	-5.31	-2.1	346	16216
1UTG	-13.47	-2.3	17	4846
2CRO	-17.59	-3.7	1	53131
2OVO	3.01	-0.3	3487	9140
3ICB	-11.94	-3.2	8	35961
4RXN	2.04	0	595	1129
5PTI	-12.49	-3	13	35961
A256B	-37.43	-4.4	1	10627
I1TGS	0.18	-1.5	718	9881
I4SGB	-10.24	-3.3	5	10661

b)

PDB Code	Energy (kT)	Z-score	Rank	Total Threadings
1CSE	-13.42	-3.3	2	9140
1CTF	-28.56	-2.4	4	925
1R69	-13.91	-3.5	16	33650
1SN3	0.26	-1.2	2133	16216
1UTG	-13.33	-2.2	62	4846
2CRO	-21.76	-3.4	14	53131
2OVO	-2.67	-2	192	9140
3ICB	-19.27	-3.9	4	35961
4RXN	0.12	-0.6	299	1129
5PTI	-16.1	-3.6	5	35961
A256B	-39.24	-4.1	2	10627
I1TGS	-3.47	-2.3	138	9881
I4SGB	-9.08	-2.4	47	10661

c)

PDB Code	Energy (kT)	Z-score	Rank	Total Threadings
1CSE	-12.52	-2.5	31	9140
1CTF	-24.94	-2.4	11	925
1R69	-17.8	-3.5	5	33650
1SN3	-5.86	-2.5	68	16216
1UTG	-14.34	-2.2	30	4846
2CRO	-21.46	-3.3	53	53131
2OVO	-3.1	-2.6	19	9140
3ICB	-20.03	-3.9	4	35961
4RXN	2.52	0.4	744	1129
5PTI	-16.18	-3.2	35	35961
A256B	-42.82	-4.4	1	10627
I1TGS	0.99	-1.5	824	9881
I4SGB	-6.49	-2.6	64	10661

d)

PDB Code	Energy (kT)	Z-score	Rank	Total Threadings
1CSE	-11.13	-2.6	14	9140
1CTF	-12.31	-1.2	102	925
1R69	-10.43	-3.2	56	33650
1SN3	-3.73	-1.7	529	16216
1UTG	-5.14	-1.3	481	4846
2CRO	-11.64	-2.1	608	53131
2OVO	-1.21	-2.1	137	9140
3ICB	-7.18	-1.9	879	35961
4RXN	1.07	0	557	1129
5PTI	-9.78	-2.1	497	35961
A256B	-33.91	-3.2	23	10627
I1TGS	-2.85	-2.1	180	9881
I4SGB	-6.59	-2.1	147	10661

e)

PDB Code	Energy (kT)	Z-score	Rank	Total Threadings
1CSE	-15.35	-3.1	2	9140
1CTF	-41.03	-3.3	1	925
1R69	-14.2	-3.1	39	33650
1SN3	-11.09	-3.1	8	16216
1UTG	-4.44	-1.7	168	4846
2CRO	-19.13	-2.4	529	53131
2OVO	-0.79	-2	269	9140
3ICB	-17.59	-3.3	42	35961
4RXN	0.02	-0.7	284	1129
5PTI	-17.82	-3.3	14	35961
A256B	-43.32	-3.6	4	10627
I1TGS	-0.74	-1.6	534	9881
I4SGB	-5.9	-1.9	340	10661

f)

PDB Code	Energy (kT)	Z-score	Rank	Total Threadings
1CSE	-13.87	-3.2	2	9140
1CTF	-20.59	-2.6	1	925
1R69	-11.69	-2.9	46	33650
1SN3	-3.68	-1.9	374	16216
1UTG	-9.06	-1.9	84	4846
2CRO	-7.15	-1	9125	53131
2OVO	-0.86	-1.6	557	9140
3ICB	-4.01	-1.5	1816	35961
4RXN	1.12	-0.3	453	1129
5PTI	-16.63	-3.3	3	35961
A256B	-29	-3.6	1	10627
I1TGS	-0.93	-1.8	472	9881
I4SGB	-10.78	-3.2	3	10661

g)

PDB Code	Energy (kT)	Z-score	Rank	Total Threadings
1CSE	-90.43	-3.5	1	9140
1CTF	-178.52	-3.2	1	925
1R69	-89.08	-4	1	33650
1SN3	-36.54	-2.5	29	16216
1UTG	-62.06	-2	53	4846
2CRO	-107.14	-3.2	25	53131
2OVO	-8.07	-2.3	132	9140
3ICB	-94.99	-3.8	1	35961
4RXN	8.05	-0.2	440	1129
5PTI	-115.42	-4	1	35961
A256B	-271.38	-4.7	1	10627
I1TGS	-7.82	-2.1	245	9881
I4SGB	-55.18	-3.2	5	10661

h)

PDB Code	Energy (kT)	Z-score	Rank	Total Threadings
1CSE	-3.66	-2	41	9140
1CTF	-9.2	-3	1	925
1R69	-5.92	-3.5	1	33650
1SN3	-2.45	-2.1	149	16216
1UTG	-0.1	-1.3	520	4846
2CRO	-2.46	-2.7	177	53131
2OVO	-2.18	-2.3	107	9140
3ICB	-13.2	-3.3	1	35961
4RXN	-2.52	-2.2	4	1129
5PTI	-2.59	-1.7	1639	35961
A256B	-17.43	-3.7	1	10627
I1TGS	-4.62	-2.8	6	9881
I4SGB	0.88	0.1	6056	10661

i)

PDB Code	Energy (kT)	Z-score	Rank	Total Threadings
1CSE	-145.33	-3.1	1	9140
1CTF	-316.52	-3.2	1	925
1R69	-177.88	-4.4	1	33650
1SN3	-73.29	-3	1	16216
1UTG	-63.56	-1.8	176	4846
2CRO	-144.04	-3.5	20	53131
2OVO	-40.77	-2.8	27	9140
3ICB	-292.99	-4.3	1	35961
4RXN	-29.75	-1.8	31	1129
5PTI	-154.27	-3.3	14	35961
A256B	-532.83	-4.7	1	10627
I1TGS	-77.12	-3.2	3	9881
I4SGB	-41.98	-1.9	100	10661

j)

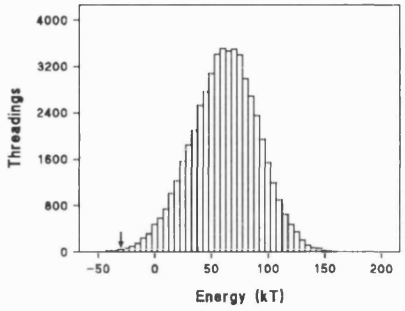
Of the atom pairs tested it is evident from these results on small proteins that the O → N and N → O potentials are of little use in locating the native threading. The implication here is that main chain hydrogen bonding is not greatly dependent on the amino acid species involved, which is in agreement with previous work (Lifson & Sander, 1979). Best results were obtained with the CB → CB potential (4 out of 14 native threadings in top place), a sum of all potentials (6 out of 14), the solvation potential (4 out of 14) and a combination of the summed pairs and the solvation potential (6 out of 14). None of the potentials were able to identify the native threadings of uteroglobin (1UTG), ovomucoid 3rd domain (2OVO), rubredoxin (4RXN), eglin-C (1TGS, chain I) or potato proteinase inhibitor (4SGB, chain I). Two factors contribute to the lack of success in threading these proteins. Firstly, in two cases, the folds are stabilized by atypical forces. For example, the structure of rubredoxin is constrained by the 'cage' formed by the interactions between the iron centres and the cysteine residues. In the case of uteroglobin, the structure is stabilized by the formation of a tight homodimer, with disulphide bridges between the two

subunits. Perhaps more important than the problem of unusual stabilization, is the fact that in all these cases, the protein chains are not folded into a compact globular structure. As will be born out in the next chapter, when threadings of larger proteins will be considered, a key factor in the recognition of folds is detecting the hydrophobic interactions in the core of a protein fold. In the case of all these small proteins, little hydrophobic core is formed, and so consequently the solvation effects are not clearly defined.

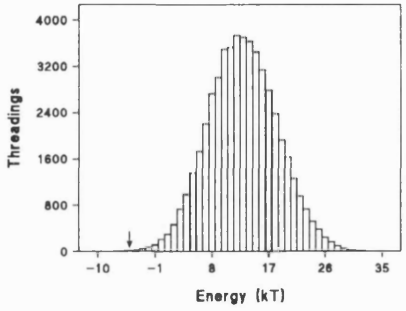
Clearly, for small proteins, in some cases at least, the evaluation function is capable of discriminating between many decoy threadings and the native. Given that the smallest difference between the native and non-native threaded models is the shift of a single strand or helix by as little as a single residue position, these results are quite remarkable. It has been found that both the pairwise potential terms *and* the solvation potential terms are needed to effectively separate the native from the non-native threadings; one or other is generally insufficient.

Though interesting at a theoretical level, the threading of a sequence onto its native structural template is of little practical use. Can the evaluation function successfully direct the threading of a sequence onto a non-native structural template? This is the basic requirement for this methodology to be useful as a modelling and structure prediction tool. As a first step to answering this question, the exhaustive threading of hemerythrin (2HMQ) on the structural template of the homologous protein, myohemerythrin (2MHR) was investigated. The sequences of these proteins are 45% identical, and the structures are found to have a Cα root mean square deviation (RMSD) of 2.3 Å.

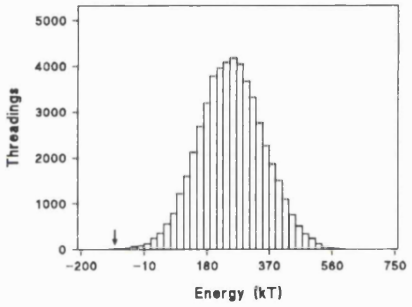
The threading histogram for the sequence of 2HMQ on the structure of 2MHR is shown in Figure 3.11. The structures of 2HMQ and 2MHR were structurally aligned using SSAP (Taylor and Orengo, 1989; Orengo and Taylor, 1990, 1991), and using this alignment as a reference, the expected optimal threading of 2HMQ onto 2MHR was deduced. This structurally optimal threading of 2HMQ onto 2MHR was found to be the lowest energy threading of all 51799 possible threadings.



MHR/HMQ - pairwise



MHR/HMQ - solvation



MHR/HMQ - pairwise + solvation

Figure 3.11

Exhaustive threading energy histograms for hemerythrin sequence (HMQ) on the structure of myohemerythrin (MHR).

Given the success of previous attempts at identifying misfolded proteins (Gregoret and Cohen, 1989; Heindlich *et al.,* 1990; Crippen, 1991) it is easy to be mislead into believing that the identification of mis*threaded* proteins should be as easy. This is not the case, however. Heindlich *et al.* and Crippen have shown that for most protein sequences, the best matching fold out of all suitably sized fragments from the crystallographic database is the protein's native fold. In these earlier experiments, no attempt was made to modify

each fragment to accommodate the sequence. Naturally, most of the fragments would have conformations hopelessly incompatible with the given sequence, the majority not even resembling globular domains. Even when matched against a fold similar to the native fold of the sequence (myoglobin sequence/hemoglobin structure for example), the loop lengths will of course not be correct, and consequently every interatomic interaction will be wrong. If the effects of insertions and deletions are not taken into account, therefore, the only likely match for a sequence will naturally be its native fold, and this fold will be easily distinguished amongst the decoy fragment folds.

The evidence so far suggests that against the odds, the native threading of a protein may be identified from a simple set of pairwise interaction parameters, and a crude solvation potential. Furthermore, even for the case of a non-native threading of a sequence onto a related structure, where the interactions will only approximate those that would occur in the native structure, the optimal threading may again be identified by the same evaluation function.

Even with the limited number of structures small enough to be exhaustively threaded and evaluated it is clear that there is some considerable promise here. By means of the described evaluation function, and using the simplest of search algorithms, a plausible automatically generated model for hemerythrin has been generated, based on the structure of myohemerythrin, and this model is found to have a lower threading potential than all the other decoys in the search space. Given this base, it is now possible to consider the more complex cases that would need to be tackled should this methodology be applied to real-life problems.

Chapter 4

Protein Tertiary Structure Prediction by Fold Recognition - Algorithms and Results

Attempt the end, and never stand to doubt;
Nothing's so hard but search will find it out.

- Robert Herrick,
"Seek and find"

4.1 Searching for the optimal threading

Given an efficient means for the evaluation of a hypothetical sequence threading relationship, the problem of finding the optimal threading must be considered. For a protein sequence of length L and a template structure of which M residues are in regular secondary structures, the total number of possible threadings is given by

$$\binom{L}{M} \equiv \frac{L!}{(L-M)!\ M!} \ .$$

The scale of the search problem for locating the optimal threading of a sequence on a structure amongst all possible threadings may be appreciated by considering bovine pancreatic trypsin inhibitor (Brookhaven code 5PTI) as an example. Of the 58 residues of 5PTI, 30 are found to be in regular secondary structure. Using the above expression the total number of threadings for 5PTI is calculated as 2.9×10^{16}. Given that 5PTI is a very small protein, it is clear that the search for optimal threading is a non-trivial problem.

One way to reduce the scale of the problem is to restrict insertions and deletions (indels) to the loop regions. By excluding indels from secondary structural elements, the problem reduces to a search for the optimal set of loop lengths for a protein. Under these conditions threading a sequence on a structure may be visualized as the sliding of beads on an abacus wire, where the beads are the secondary structures, and the exposed wire the remaining loop regions. The restricted threading of a small four helix bundle protein (myohemerythrin - Hendrickson *et al.*, 1975) is depicted in Figure 4.1. Restricting indels to loop regions in this way reduces the search space in this case from 1.3×10^{33} to just 44100.

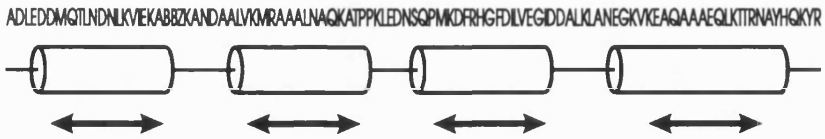


Figure 4.1

Diagram illustrating the exhaustive threading procedure. Secondary structures are effectively moved along the sequence like abacus beads. On average, loop lengths vary between 2 and 12.

Unfortunately, even with this extreme restriction on the threading patterns, the search space again becomes unreasonably large for proteins of even average size. The exact number of threadings with secondary structure indels disallowed is a complex function of both sequence length and the number and sizes of constituent secondary structural elements. As a rule of thumb, however, it is found that for N secondary structural elements, with loops of average length (say between 2 and 12 residues), the number of threadings is $O(10^N)$. For typical proteins comprising 10-20 secondary structures, it is clearly not possible to locate the optimal threading by an exhaustive search in a reasonable period of time.

4.2 Methods for combinatorial optimization

Various means have been investigated for locating the optimal threading of a sequence on a structure. The methods are briefly detailed below.

1. Exhaustive Search

As demonstrated in the previous chapter, for small proteins of < 6 secondary structural elements, and disallowing secondary structure indels, it is practical to simply search

through all possible threadings in order to locate the threading of lowest energy. Unfortunately, the evaluation function is tailored towards average sized globular proteins with hydrophobic cores, whereas the small proteins tend to be less globular and typically lack a hydrophobic core.

2. Monte Carlo Methods

Monte Carlo methods have been often exploited for conformation calculations on proteins. Two *directed* search procedures have been used in this project: *simulated annealing*, and *genetic algorithms*. Simulated annealing has been recently exploited in the alignment of protein structures (Šali & Blundell, 1990), and in the optimization of side chain packing in protein structures (Lee & Subbiah, 1991). Simulated annealing is a simple random search process. In this instance, random threadings are generated and evaluated using the evaluation function described earlier. Where a proposed threading has a lower energy than the current threading, the proposed threading is accepted. In the case where a proposed threading has a higher energy than the current, it is accepted with probability p where

$$p = e^{-\Delta E/kT}$$

and where ΔE is the difference between the current and the proposed threading energy and T is the current annealing 'temperature'. After a predefined number of accepted changes, the temperature is slightly reduced. This whole procedure is repeated until no further reduction in threading energy is achieved, at which point the system is said to be frozen. The schedule of cooling is critical to the success of simulated annealing.

Genetic algorithms (Goldberg, 1989) are similar in concept to simulated annealing, though their model of operation is different. Whereas simulated annealing is loosely based on the principles of statistical mechanics, genetic algorithms are based on the principles of natural selection. The variables to be optimized are encoded as a string of binary digits, and a *population* of random strings is created. This population is then subjected to the genetic operators of selection, mutation and crossover. The probability of a string surviving from one generation to the next relates to its fitness. In this case, low energy

threadings are deemed to be fitter than those with higher energies. Each string may be randomly changed in two ways. The mutation operator simply selects and changes a random bit in the string. An alternative means for generating new strings is the crossover operator. Here a randomly selected portion of one string is exchanged with a similar portion from another member of the string population. The crossover operator gives genetic search the ability to combine moderately good solutions so that 'super individuals' may be created.

In use, these methods prove to be capable of locating the optimal threading, but with no guarantee that they will do so in any given run of the threading program. Ideally the results from many runs should be pooled and the best result extracted, which is of course time consuming. A further problem is that the control parameters (the cooling schedule in the case of simulated annealing and the selection, mutation and crossover probabilities in the case of genetic search) need adjustment to match each threading problem individually. Parameters found suitable for threading a protein with 10 secondary structures will generally not be suitable for threading a protein with 20 secondary structures for example. The methods are typically plagued by 'unreliability', yet are found to be highly robust. Given a sufficiently slow cooling rate in the case of simulated annealing, or a sufficiently large population of strings in the case of genetic algorithms, and in both cases a sufficient number of runs, very low energy threadings will be found providing they exist at all in the given search space.

3. Dynamic Programming

It should be apparent that there exists a clear similarity between optimizing the threading of a sequence on a structural template and finding the optimal alignment of two sequences. In such terms, threading is simply the alignment of an amino acid sequence against a sequence of positions in space. At first sight it might well appear that the same dynamic programming methods used in sequence alignment (Needleman & Wunsch, 1970; Gotoh, 1982) could easily be applied to the threading problem. Unfortunately, this is not

the case. In a typical alignment algorithm a score matrix is constructed according to the following recurrence formula:

$$S_{ij} = D_{ij} + \max \begin{cases} S_{i+1,j+1} ; \\ \max\limits_{k=i+2 \rightarrow N_A} S_{k,j+1} - g ; \\ \max\limits_{l=j+2 \rightarrow N_B} S_{i+1,l} - g ; \end{cases}$$

where S_{ij} is an element of the score matrix, D_{ij} is a measure of similarity between residues i and j in sequences of length N_A and N_B respectively and g is a gap penalty which may be either a constant or a function of, for example, gap length. By tracing the highest scoring path through the finished matrix the mathematically optimum alignment between the two sequences may be found for the given scoring scheme. In the special case where D_{ij} is a function only of i and j dynamic programming alignment algorithms have execution times proportional to the product of the sequence lengths. However, if D_{ij} is defined in terms of non-local sequence positions in addition to i and j, dynamic programming no longer offers any advantage; the alignment effectively requires a full combinatorial search of all possible pairings. In the case of the evaluation function defined here, in order to determine the energy for a particular residue, all pairwise interactions between the residue in question and every other residue in the protein need to be calculated. In other words, in order to evaluate the threading potentials in order to fix the location of a single residue, the location of every other residue needs to have been fixed beforehand.

By excluding the medium and long range ($k > 10$) pairwise terms from the evaluation function and by considering only interactions between residues in the same secondary structural element, dynamic programming can be applied to the problem. For example, consider a case where a template comprising a single 10 residue helical segment is being

matched against a 100 residue sequence. Discounting the possibility of indels in the helix itself, there are 91 possible alignments between the helical template and the sequence. As indels may not occur in the helix, for any given position in the sequence (i = 1...91), all possible inter-helical pairwise interactions are defined. However, this simplification allows only local conformational effects to be considered. Packing between secondary structures may be evaluated only by means of the solvation potentials and not by any pairwise terms. Clearly it would be ideal to devise an efficient dynamic programming method capable of taking non-local pairwise terms into account.

4. Double Dynamic Programming

The requirement here to match pairwise interactions relates to the requirement of structural comparison methods. The *potential environment* of a residue i is defined here as being the sum of all pairwise potential terms involving i and all other residues $j \neq i$. This is a similar definition to that of a residue's *structural environment*, as described by Taylor & Orengo, 1989. In the simplest case, a residue's structural environment is defined as being the set of all inter-Cα distances between residue i and all other residues $j \neq i$. Taylor & Orengo, 1989, propose a novel dynamic programming algorithm (known as double dynamic programming[3]) for the comparison of residue structural environments, and it is a derivative of this method that is proposed here for the effective comparison of residue potential environments.

Let T_m (m = 1...M) be the elements of a structural template, and S_n (n = 1...N) be the residues in the sequence to be optimally fitted to the template. We wish to determine a score $Q(T_m,S_n)$ for the location of residue n at template position m. In order to achieve this, the optimal interaction between residue n and all residues $q \neq n$ conditional on the matching of T_m *and* S_n is calculated by the application of the standard dynamic programming algorithm. We define two matrices: a low-level matrix $\boldsymbol{L}$ (more precisely

[3] In fact the original algorithm was unnamed, and the rather descriptive term "double dynamic programming" was coined by Dr Chris Sander at EMBL.

a *set* of low-level matrices), and a high-level matrix $\boldsymbol{H}$ into which the best paths through each of the low-level matrices are accumulated.

For each m,n, the total potential of mean force, $Z(m,n,p,q)$, may be calculated for each p,q where p is again an element in the structural template, and q a residue in the object sequence:

$$Z(m,n,p,q) = \Delta E_{solv}^{S_q}(A_p) + \begin{cases} \Delta E_{(q-n)}^{S_n S_q}(d_{mp}) & ; q>n, p>m \\ \Delta E_{(n-q)}^{S_q S_n}(d_{pm}) & ; q<n, p<m \\ 0 & ; q=n, p=m \\ U & ; \begin{cases} q=n, p\neq m \\ q<n, p>m \\ q>n, p<m \end{cases} \end{cases}$$

where U is a large positive constant penalty which forces the final path to incorporate pair m,n, A_p is the accessibility of template position p, d_{mp} and d_{pm} are elements of the template interatomic distance matrix and the pairwise, $\Delta E_k^{ab}(r)$, and solvation, $\Delta E_{solv}^{a}(s)$, terms are as defined in the previous chapter. Pairwise terms are summed over all required atom pairs (Cβ → Cβ, Cβ → N for example), using appropriate values from the distance matrix, though typically, for computational efficiency, the low-level matrices are calculated using the Cβ → Cβ potential alone.

The low-level matrix $\boldsymbol{L}$ is then calculated using the standard NW algorithm:

$$\boldsymbol{L}_{pq} = Z(m,n,p,q) + \min\begin{cases} \boldsymbol{L}_{p+1,q+1} \ ; \\ \min\limits_{r=p+2\rightarrow N_A} \boldsymbol{L}_{r,q+1} + g(S_p) \ ; \\ \min\limits_{s=q+2\rightarrow N_B} \boldsymbol{L}_{p+1,s} + g(S_p) \ ; \end{cases}$$

where S_p is the secondary structural class (helix, strand, coil) of template position p. $g(S_p)$ is a simple secondary structure dependent gap penalty function:

$$g(S_p) = \begin{cases} G_s & ; \ S_p = helix,strand \\ G_c & ; \ S_p = coil \end{cases}$$

where G_s and G_s are both positive constants, and $G_s >> G_c$. The low-level matching procedure is illustrated in Figure 4.2. As with the application of the algorithm to structure comparison, it is found to be advantageous to accumulate the low-level matrix scores along each suggested low-level path, and so the paths from each low-level matching, conditional on each proposed match between T and S, for which the path scores exceed a preset cut-off, are accumulated into $\boldsymbol{H}$ thus:

$$\acute{\boldsymbol{H}}_{pq} = \boldsymbol{H}_{pq} + \min\begin{cases} \boldsymbol{L}_{p+1,q+1} \ ; \\ \min\limits_{r=p+2\rightarrow N_A} \boldsymbol{L}_{r,q+1} \ ; \\ \min\limits_{s=q+2\rightarrow N_B} \boldsymbol{L}_{p+1,s} \ ; \end{cases}$$

for all p,q along the optimum traceback path in $\boldsymbol{L}$.

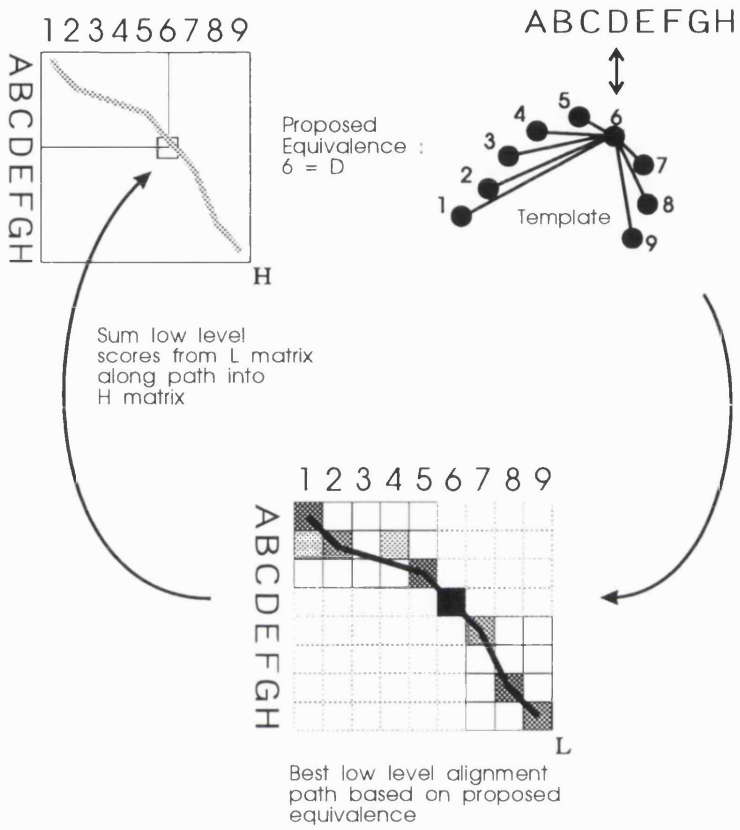


Figure 4.2

An outline of the double dynamic programming algorithm. For each proposed sequence-structure equivalence an optimal path is calculated based on interactions with the equivalenced residue. Scores are summed along each path into the high level matrix.

The overall operation may be thought of as the matching of a distance matrix calculated from the template coordinates with a *probability* matrix (in practice, an information matrix) calculated from the object sequence.

The final alignment (matrix $\boldsymbol{F}$) is generated by finding the best path through the final high-level matrix thus:

$$\boldsymbol{F}_{pq} = \boldsymbol{H}_{pq} + \min\begin{cases} \boldsymbol{F}_{p+1,q+1} \ ; \\ \min\limits_{r=p+2\rightarrow N_A} \boldsymbol{F}_{r,q+1} + g(S_p) \ ; \\ \min\limits_{s=q+2\rightarrow N_B} \boldsymbol{F}_{p+1,s} + g(S_p) \ ; \end{cases}$$

In the above expressions, each instance of the NW algorithm has been formulated in terms of *minimizing* a cost function, as the scores in this application are energies. In practice, however, these expressions could be converted trivially to a form where a score is maximized, simply by changing the sign of the calculated energy values, or by just leaving the interaction propensities in units of information. Where mention is made of a *high-scoring* path (which is rather more familiar terminology in sequence comparison) in the following pages, then this should be taken as referring to a path with *low* energy.

As described, the double dynamic programming algorithm is too slow to be useful for fold recognition. The efficient algorithm by Gotoh (1982) for calculating the NW score matrix is $O(MN)$ where MN is the product of the two sequence lengths. Double dynamic programming involves the use of this algorithm for all MN possible equivalent pairs of residues, giving an overall algorithmic complexity of $O(M^2N^2)$. On a typical present day workstation, a single instance of the Gotoh algorithm for two sequences of length 100 can be performed in around 0.25 CPU seconds. Multiplying this time by 100^2 provides an

estimate of 2500 CPU seconds to complete a single double dynamic programming comparison, which is clearly too slow to be applied to the fold recognition problem, where at least 100 instances of the double dynamic programming algorithm would be required (totalling 3 CPU days). Furthermore, many comparisons would involve sequences or structures longer than 100 residues in length. The absurdity of the problem becomes apparent when it is realized that to compare a single sequence 500 residues in length with a structure of similar size, roughly 17 CPU hours would be required.

Clearly, if the double dynamic programming algorithm is to be of use, short-cuts must be taken. The most straightforward short-cut to take is to apply a window to both levels of the algorithm, as described in Chapter 2. This is very helpful, but still insufficient to make the algorithm convenient for general use. Orengo and Taylor (1990) proposed the use of a pre-filtering (*residue selection*) pass to exclude unlikely equivalences before going on to calculate a low-level path. In the case of structural comparison, Orengo and Taylor initially select pairs of residues from the two structures primarily on the basis of similarities in their relative solvent accessibility and main chain torsion angles. Residue pairs found to be in different local conformations, and with differing degrees of burial are clearly unlikely to be equivalenced in the final structural alignment, and consequently should be excluded as early as possible in the double dynamic programming process. Unfortunately for sequence-structure alignment, it is not possible to select residue pairs on real measured quantities such as accessibility or torsion angles. However, these quantities could in principle be *predicted* for the sequence under consideration, and these predicted values then compared with the real values observed in the template structure. In practice, these values are not actually predicted directly, but the method proposed here certainly makes use of the same principles that might be employed in their prediction.

4.3 Residue selection for sequence-structure alignments

The residue selection stage of optimal sequence threading proposed here involves the summation of local interaction potential terms ΔE_k^{ab} over all residue pairs in overlapping windows of length L, where L is a small constant odd-number over the range, say, 5-31. Similarly the solvation terms are summed for each of the L residues. The window is clipped appropriately if it spans either the N or C-terminus of either the sequence or the structure, for example the window length for the first and last residue in either cases would be $\frac{1}{2}(L+1)$. To equalize the contribution of the pairwise terms and the solvation terms, the average energy is calculated and summed for both, giving a total energy for the sequence-structure fragment of:

$$S_{mn} = E(fragment) = \frac{2\sum_{i=1}^{L-1}\sum_{j=i+1}^{L} E_{pair}^{ij}}{L(L-1)} + \frac{\sum_{k=1}^{L} E_{solv}^{k}}{L}$$

Energies are calculated for every sequence fragment threaded onto every structure fragment, and the results stored in a selection matrix S. The process is illustrated in Figure 4.3. Using the residue selection step, the number of initial pairs (m,n) for which low-level paths need be calculated is reduced up to 100 fold.

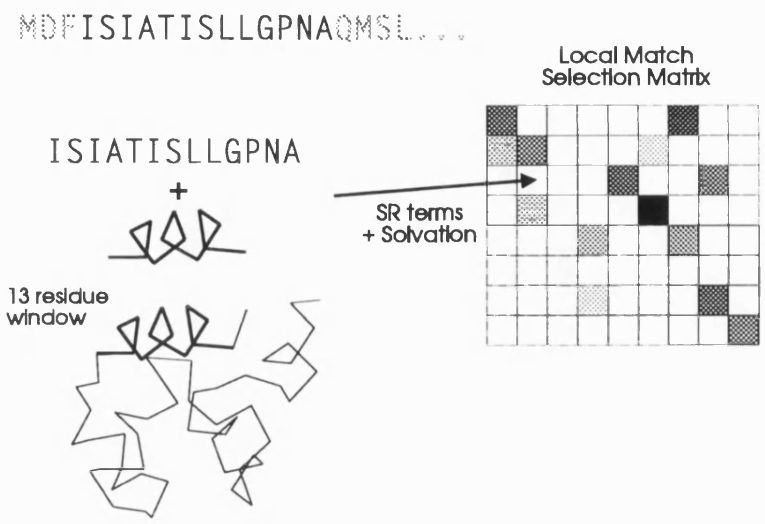


Figure 4.3

The initial residue selection process, illustrated here for a fragment length of 13.

At present, both Monte Carlo methods and double dynamic programming are used to solve large threading problems. The advantage of dynamic programming is that it takes a fixed amount of time to find the same threading each time it is used on a particular problem, and is generally found to be faster than simulated annealing for example. Unfortunately, in some cases, consistent multiple paths cannot be found, and the alignment corresponding to the final consensus path is not much better than random. These cases are easily identified by observing when the energy of the final consensus threading is unusually high ($E \gg 0$ as a rule). In the case of Monte Carlo methods, good solutions ($E \ll 0$) are always found, although as discussed above, the final solutions are not always the best possible. An ongoing avenue of research is the development of a method that combines the best aspects of both classes of algorithm, where good non-optimal solutions found by random search are refined by the application of dynamic programming.

4.4 Double dynamic programming summary

To show how the double dynamic programming algorithm is applied to the sequence-to-structure alignment problem, a simple example will be presented. The example presented here is the simple case of threading the sequence of sipunculid worm hemerythrin (Holmes & Stenkamp, 1991) onto the structure of the homologous myohemerythrin (Hendrickson *et al.*, 1975) from the same organism: four-helix bundles which carry oxygen.

Figure 4.4 shows the resulting residue selection matrix for Cβ → Cβ interactions, calculated with a window length of 13, and a softening parameter of 0.2 Å. From this matrix, pairs scoring above 1.5 standard deviations above the mean score are selected (Figure 4.5).

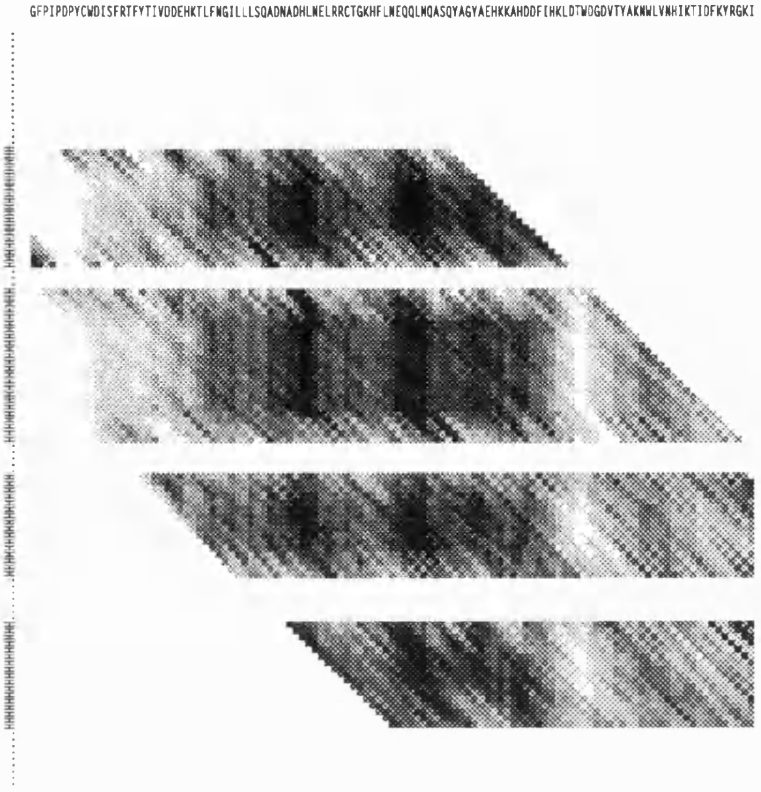


Figure 4.4

The calculated residue selection matrix for the threading of hemerythrin onto a structural template derived from myohemerythrin. Dark squares indicate high scoring (low energy) pairs.

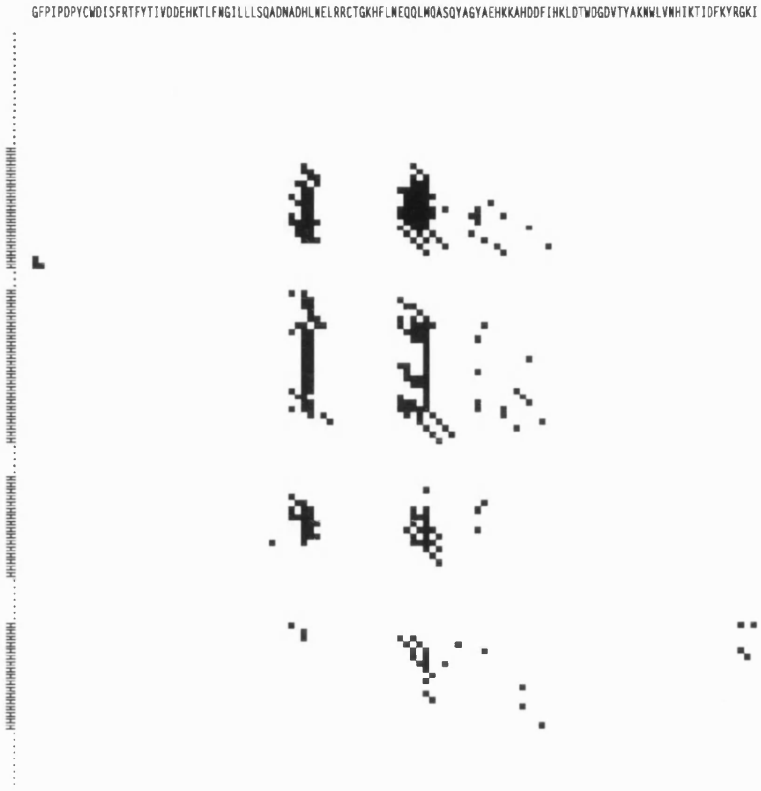


Figure 4.5

Locally equivalent pairs scoring over 1.5 standard deviations above the mean are shown.

Having identified a reasonable number of residue-site pairs for which the local structure (over a 13 residue window) and solvent accessibility appears to be compatible, the next step is to take each pair in turn, and calculate the best alignment path conditional on the proposed equivalence.

It might be helpful to step back and consider the reason for the calculation of these paths. The initial selection phase has identified a possible match between a residue in the sequence and a site in the structural template. If we make this equivalence, and fit the

proposed residue into the proposed site, then we may now calculate pairwise interactions between this site (now filled) and all other sites in the structure. The next problem that must be solved is to fit the remaining (M-1) residues in the sequence into the remaining (N-1) sites, which is achieved using a standard dynamic programming alignment algorithm.

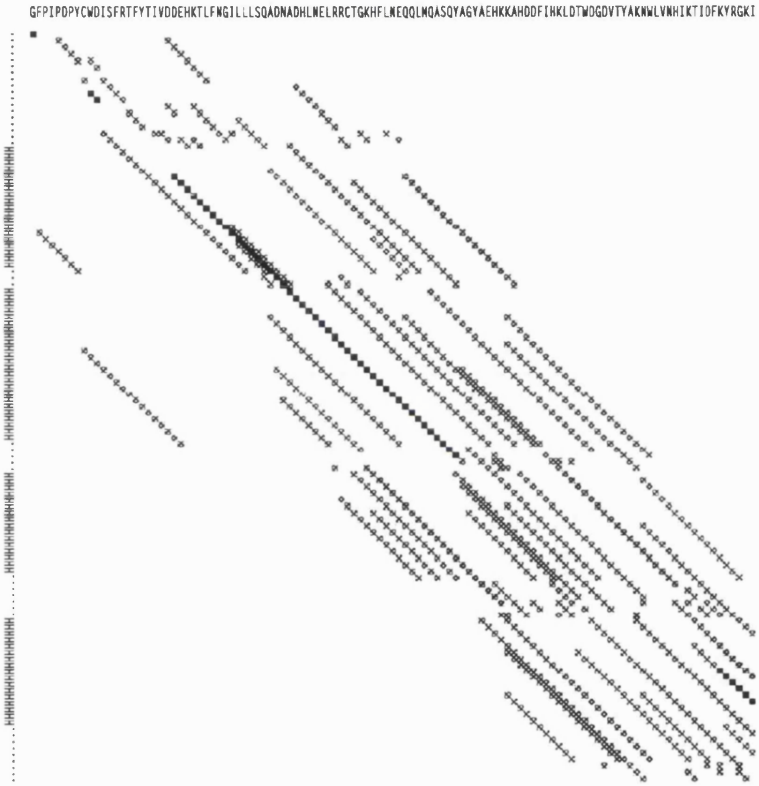


Figure 4.6

The final state of the high level matrix, showing the accumulated low-level paths.

Along each calculated low level path for which the path score exceeds the prescribed cut-off (30.0), scores from the low level matrix (***L***) are accumulated in the high level matrix (***H***). Figure 4.6 shows the final state of the high level matrix (***H***) after each of the 56 high

scoring paths have been accumulated, note that some regions of the matrix have been emphasized by the reinforcement of several paths: the assumption being that these regions are likely to be the most reliable. Where more than 100 paths exceed the cut off, only the top 100 are taken into account. One final application of the NW algorithm on the H matrix provides the best consensus path, which is taken as the optimum threading.

4.5 Threading large structures

It has been shown that in most cases, the native relationship between the secondary structure coordinates and the sequence of a small protein is easily identified by means of the proposed evaluation function. The next obvious question to ask is whether this holds true for larger structures. Though the larger structures will relate better to the knowledge-base, clearly a more complex structure has many more degrees of freedom in which it can satisfy the evaluation criteria, and thus the likelihood of identifying the *correct* threading should be reduced.

To test the ability of the algorithms to locate the minimum energy threading, and to test whether this minimum corresponds to the correct native threading, attempts were made to thread a number of protein sequences onto their native structures. The pairwise and solvation terms were investigated separately.

The most sophisticated of the previously described algorithms, double dynamic programming, is unfortunately inapplicable to the problem of testing the potentials on native threading problems. The algorithm can certainly be used to thread a protein sequence onto it's native structure, but the native threading is unfairly selected above all others not only by means of the threading energy, but also from the nature of the algorithm itself (particularly from the intricate arrangement of secondary structure biased gap penalties). In order to properly evaluate the potentials on native threading problems, therefore, it is necessary to select *intrinsically unbiased* algorithms for the task.

In order to optimize the threading in the field consisting of the pairwise energy terms[4], a simulated annealing procedure was used, as described earlier. A slow cooling rate (2% per cycle) was used, along with a large number of reconfigurations per cycle (200 successful reconfigurations, with a maximum of 2000 reconfigurations) to ensure the greatest chance of locating the optimum threading. Of course, there is no guarantee that the values found correspond to the true global minimum threading, but the values may at least be considered close to this minimum.

As the solvation terms are 1-D parameters, optimization of the threading in the solvation field may be carried out using a straightforward dynamic programming algorithm, in which case the threading obtained is guaranteed to be the global minimum. One point of caution here is in the use of gap penalties. If in the alignment of the sequence with the vector of accessibilities gaps are penalized, then the native threading, which naturally incorporates no gaps, will be automatically favoured. To circumvent this problem, the gap penalty is set to zero in between secondary structures, and set to a very high value (1000.0) otherwise. This constrains the threadings searched to be similar to those explored by the simulated annealing algorithm, though in this case, an even wider search space is covered. With this gap penalty scheme, the native threading proffers no immediate advantage over the alternatives.

A summary of the results is shown in Table 4.1. The difference in performance between the pairwise and the solvation potentials in this case could hardly be more marked. In almost every case, the native threading is found to be the global minimum of the solvation potential, no matter whether loop residues are included or not. In contrast to this, in only one case (1MBD - sperm whale myoglobin) does the native threading correspond to the global minimum of the short distance pairwise potential. Looking at the optimum threadings obtained in the pairwise field, it is clear that the threadings are fairly minor

[4] Interactions were summed over all topological ranges, using the atom pairs: Cβ → Cβ, Cβ → N, Cβ → O, N → Cβ, and O → Cβ. Interactions with residues not in regular secondary structure were ignored.

variants of the native. Most of the shifts are less than 4 residues, and correspond generally to a single turn of a helix or a strand shift of one or two residues.

The difficulty in locating the native threading using the pairwise term alone may be attributed to two things. Firstly the loop residues are not taken into account. Secondly, and more importantly, is the fact that the pairwise terms are self-referential. In other words, favourable contacts for a particular secondary structure cannot be determined without reference to another secondary structure contact surface. In other words shifts in one secondary structure threading can be easily compensated by shifts in the secondary structures with which it interacts. Given the large number of secondary structures in larger protein chains, providing many degrees of freedom, the fact that most secondary structural elements interact with only one or two other elements, and the inherent softness of the potentials[5], it is really not surprising that slightly different threadings can be found with lower energy. The results of Ponder and Richards (1987) show that by using a hard pairwise potential (in the form of a simple Lennard-Jones interaction), the native threading can be easily recognized, but in this case of course, the template is unable to match sequences other than those closely related to the native.

The solvation potential appears to easily recognize the native threading over all others. Why this should be so is illustrated in Figure 4.7. As stated above, the pairwise terms can only be defined with reference to elements other than that under consideration. However, in the case of the solvation terms, the relative accessibilities are predefined over the entire protein chain. A shift in one secondary structural element cannot be compensated by shifts in other elements, and so in this case too it should not be surprising that the solvation terms *can* recognize the native threading.

[5] Remember that the histogram sampling interval for the medium and long-range potentials is 2 Å, which means that, at worst, distance variations of up to 2 Å will not change the interaction between two atoms.

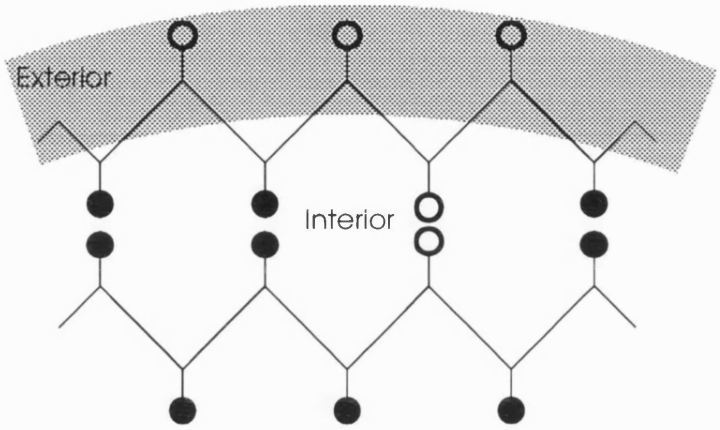


Figure 4.7

A depiction of how solvation terms and contact terms constrain the matching of a sequence to a structure in different ways. Taking 'H' to represent a hydrophobic residue (shaded side chains), and 'P' to represent a polar residue (unshaded), the top strand is constrained by solvation to have the sequence (HPHPHPH). Without reference to solvation, many sequences can form stable contacts between the two strands, as long as the sequence of the other strand changes to compensate.

		Pairwise				Full Solv			SS Solv		
PDB Code	No. SS	Threadings	E_{native}	$E_{opt.}$	Correct SS	E_{native}	E_{opt}	Correct SS	E_{native}	E_{opt}	Correct SS
1ABP	13	0	-1.83	-51.25	2	-48.16	-48.16	13	-22.36	-22.36	13
1CCR	4	20491	-27.99	-38.35	2	-19.53	-19.53	4	-9.95	-9.95	4
1CD4	15	34305	-11.65	-37.62	3	-30.73	-30.73	15	-14.71	-14.71	15
1FX1	9	22319	-15.40	-47.92	1	-27.71	-27.71	9	-19.90	-19.90	9
1GCR	16	33954	1.44	-38.55	4	-33.37	-33.37	16	-22.31	-22.31	16
1HIP	4	23685	-11.65	-27.75	0	-7.29	-7.29	4	1.69	-4.11	0
1I1B	13	48252	50.79	-24.83	0	-21.48	-21.48	13	-5.36	-5.36	13
1MBD	5	14656	-148.21	-148.21	5	-33.99	-33.99	5	-26.08	-26.08	5
1PAZ	9	18315	-41.11	-56.46	4	-18.28	-18.28	9	-11.38	-11.38	9
1RHD	16	36089	-22.42	-76.15	0	-42.41	-42.41	16	-24.21	-24.21	16
2AAT	20	37360	32.19	-49.22	1	-51.94	-51.94	20	-34.63	-34.63	20
2CNA	14	41613	33.82	-29.69	0	-27.41	-27.41	14	-15.21	-15.21	14
2CPP	21	47976	-170.26	-170.65	6	-62.54	-62.54	21	-43.87	-43.87	21
2CYP	15	40224	-71.41	-102.83	5	-34.65	-34.65	15	-19.61	-19.61	15
2GBP	21	41227	-155.80	-169.56	6	-45.30	-45.30	21	-34.09	-34.09	21
2RHE	9	18000	16.23	-28.11	2	-11.99	-11.99	9	-8.60	-8.60	9
2RNT	7	21172	-5.55	-40.44	2	-13.98	-13.98	7	-6.15	-6.15	4
2SGA	17	38371	8.96	-34.47	3	-15.30	-15.30	17	-12.27	-12.27	17
3ADK	14	45288	-32.89	-65.70	5	-37.76	-37.76	14	-35.04	-35.04	14
3BLM	17	35647	-59.03	-88.83	4	-46.29	-46.29	17	-30.99	-30.99	17
3DFR	13	43961	-40.75	-61.06	3	-22.83	-22.83	13	-14.34	-14.34	13
3GRS	33	26238	-112.27	-105.02	7	-46.36	-46.36	33	-35.86	-35.86	33
3PGM	10	22877	1.42	-57.52	1	-28.06	-28.06	10	-18.01	-18.01	10

		Pairwise				Full Solv			SS Solv		
PDB Code	No. SS	Threadings	E_{native}	$E_{opt.}$	Correct SS	E_{native}	E_{opt}	Correct SS	E_{native}	E_{opt}	Correct SS
4CPV	6	23284	-77.46	-95.44	1	-21.81	-21.81	6	-13.46	-13.46	6
4FXN	9	23545	-53.54	-69.43	4	-29.80	-29.80	9	-25.74	-25.74	9
A1TNF	7	21946	5.04	-21.31	1	-24.77	-24.77	7	-12.40	-12.40	7
A1YPI	19	32418	-76.58	-111.16	5	-32.58	-32.58	19	-30.48	-30.48	19
A7CAT	21	31885	-50.00	-80.81	3	-2.30	-17.23	2	-12.20	-12.20	21
H2FB4	17	23776	-21.24	-45.06	2	-29.21	-29.21	17	-13.95	-13.95	17
O2SOD	8	18713	-20.07	-44.50	3	-18.57	-18.57	8	-8.55	-8.55	8

Table 4.1

Results of native sequence threading for large structures. KEY: **No. SS** - the number of secondary structural elements in the chain, $\mathbf{E_{native}}$ - energy for native threading, $\mathbf{E_{opt.}}$ - lowest energy threading found, **Correct SS** - number of secondary structural elements correctly located for the lowest energy threading found. **Full Solv** results include solvation terms for loop residues, whereas **SS Solv** only include terms for residues in secondary structures.

From these results it is clear that in nearly all cases the minimum of the solvation threading energy evaluation function does in fact correspond to the correct native threading. Whether the proposed double dynamic programming algorithm can locate similarly optimal threadings will be investigated in the following section.

4.6 Non-native threading of large structures

Having shown the ability of the potentials (in particular the solvation potentials) to recognize to find the optimal native threadings for large structures, we consider now the final test: the ability of the double dynamic programming algorithm to find the optimal *non-native* threadings for large structures. The algorithm employed in the following sections is as described earlier, though with a few minor additional details.

a) Pairwise terms involving loop residues are ignored, with the low-level matrix elements for these pairs being defined solely on the basis of the solvation terms.

b) The gap penalties for the low-level matrix are $G_s = 3.0$, $G_c = 0.05$. For the high-level matrix, $G_s = 5M$, $G_c = 5/M$ where M is the highest element in the final matrix $\boldsymbol{H}$. This provides a degree of autoscaling for the high-level matrix gap penalties, to take into account the fact that the range of the value in $\boldsymbol{H}$ depends on the number and quality of accumulated paths.

c) The cut-off used to prevent low scoring paths from being accumulated into the high-level matrix was set at 28.0.

To demonstrate the ability of the optimal threading method for aligning sequences with non-native structures, two alignments, commonly used as alignment 'benchmarks', are shown in Figures 4.8 and 4.9. Figure 4.8 shows the alignment of the sequence of lupin leghemoglobin (Arutiunian *et al.*, 1980) and the structure of sperm whale myoglobin

(sequence identity 17% - Phillips, 1980). Similarly, Figure 4.9 shows the alignment of the second domain sequence of rhodanese (Ploegman *et al.*, 1978) with the structure of the first domain (sequence identity 9%). From inspection on a graphics workstation, these two structural pairs show significant structural similarity, yet differ in some minor details such as a small additional helix in MBD at residue 52. Standard pairwise sequence alignments are ineffective in either case, and so this is an excellent chance to test the sensitivity of sequence threading. Each alignment is compared to the structurally determined alignment as determined by the method of Orengo and Taylor (1990). Lines are drawn between structurally equivalent residues. It should be noted that no sequence information was incorporated in the alignment process. The structure is considered simply as a chain of anonymous placeholders, through which the given sequence is threaded. As before, the calculation of potentials was 'jack-knifed', in that sequences related to the one being threaded (> 25% identity) were excluded from the data set. It is clear that the proposed method is capable of aligning sequences and structures with great accuracy. The alignment of leghemoglobin and myoglobin is 100% accurate in the core regions, differing from the structural alignment only in a few arbitrarily equivalenced loop positions. Whilst the harder rhodanese alignment is not as accurate as the globin alignment, apart from one misaligned helical region, all secondary structural elements are correctly equivalenced. After alignment, it is possible to evaluate the contributions from each residue to the total threading energy, and thus identify regions of the alignment that are 'unstable'; i.e. regions in which the sequence does not appear to fit the equivalenced structure. By far the most unstable region of the rhodanese alignment is found to coincide with the misaligned helix at residue 43 of domain 1. Ways to refine given alignments, so as to correct for locally misaligned regions are being investigated. In general, it is found that α proteins are aligned with higher accuracy than β proteins, which probably relates to the greater influence of local interactions in α-helical structures.

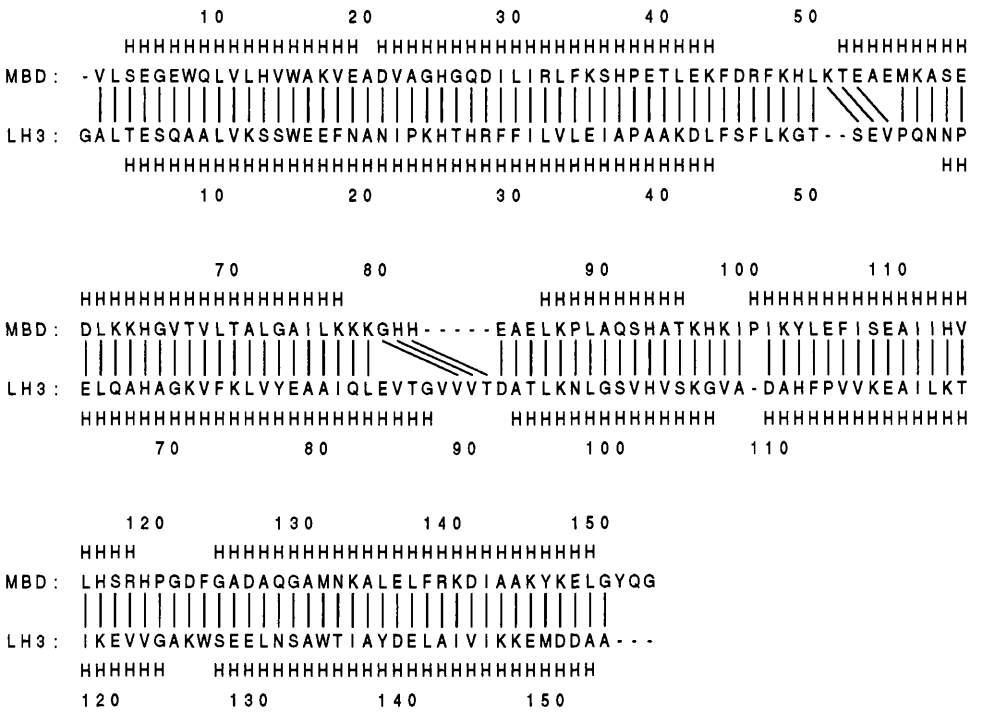


Figure 4.8

Optimal threading of yellow lupin leghemoglobin on the structure of sperm whale myoglobin (Brookhaven code 1MBD). Lines are drawn between structurally equivalent residue pairs as determined by alignment of the protein structures.

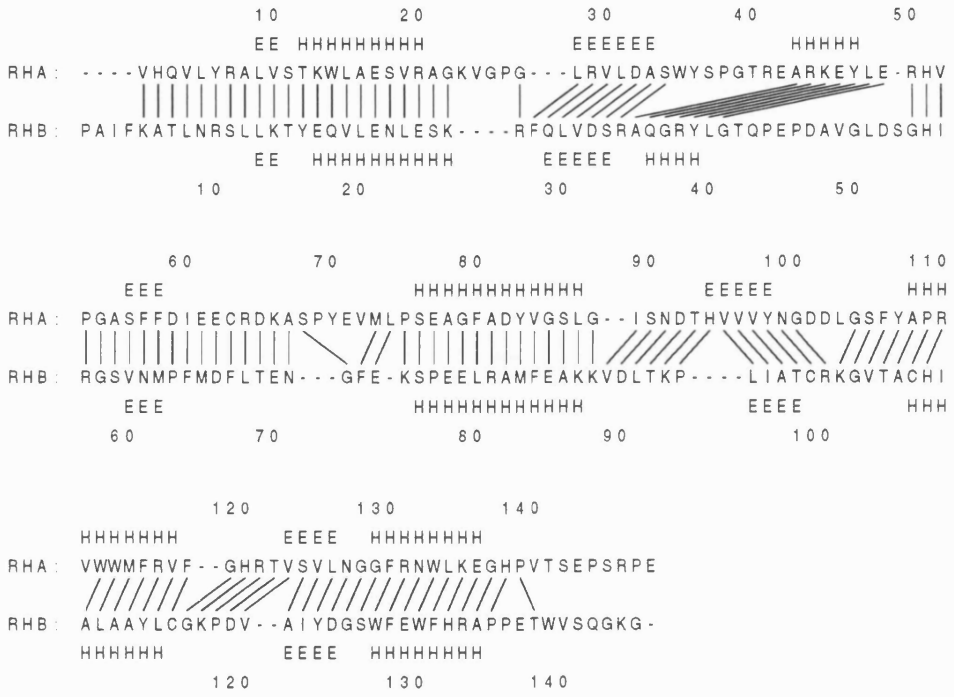


Figure 4.9

Optimal threading of rhodanese domain 2 sequence on the structure of rhodanese domain 1.

4.7 Identifying chain folds

Having verified the ability of double dynamic programming for sequence-structure alignment, we now consider its ability to *recognize* protein folds. Given a sequence, can we identify its native fold from a choice of those stored in the protein structure database?

The results of trial searches performed with the method of optimal sequence threading are shown in Figure 4.10 and Table 4.2. A constant set of alignment parameters (gap penalty for example) was used for all databank searches shown. It is probable that a better set of parameters will be found as the method is developed further. Typical execution times for a single search of 102 chains are around 100 CPU minutes on a Unix workstation (Solbourne 5/602). The method is clearly better at identifying large structures (> 100 residues) than small ones, and performs optimally for structures with some α-helical content. This clearly points to a statistical bias in the potentials towards such structures, which is in accordance with the bias observed in traditional secondary structure prediction.

Templates for each chain were constructed as described earlier, with residues not in helices or strands (as calculated by DSSP - Kabsch & Sander, 1983) assigned as loop residues. For the 70 kD heat shock cognate protein and hexokinase searches, the coordinates for actin were also included in the fold library (coordinates deposited under the code 1ATN). Proteins >25% sequence identical to the test protein were again excluded from the calculation of potentials.

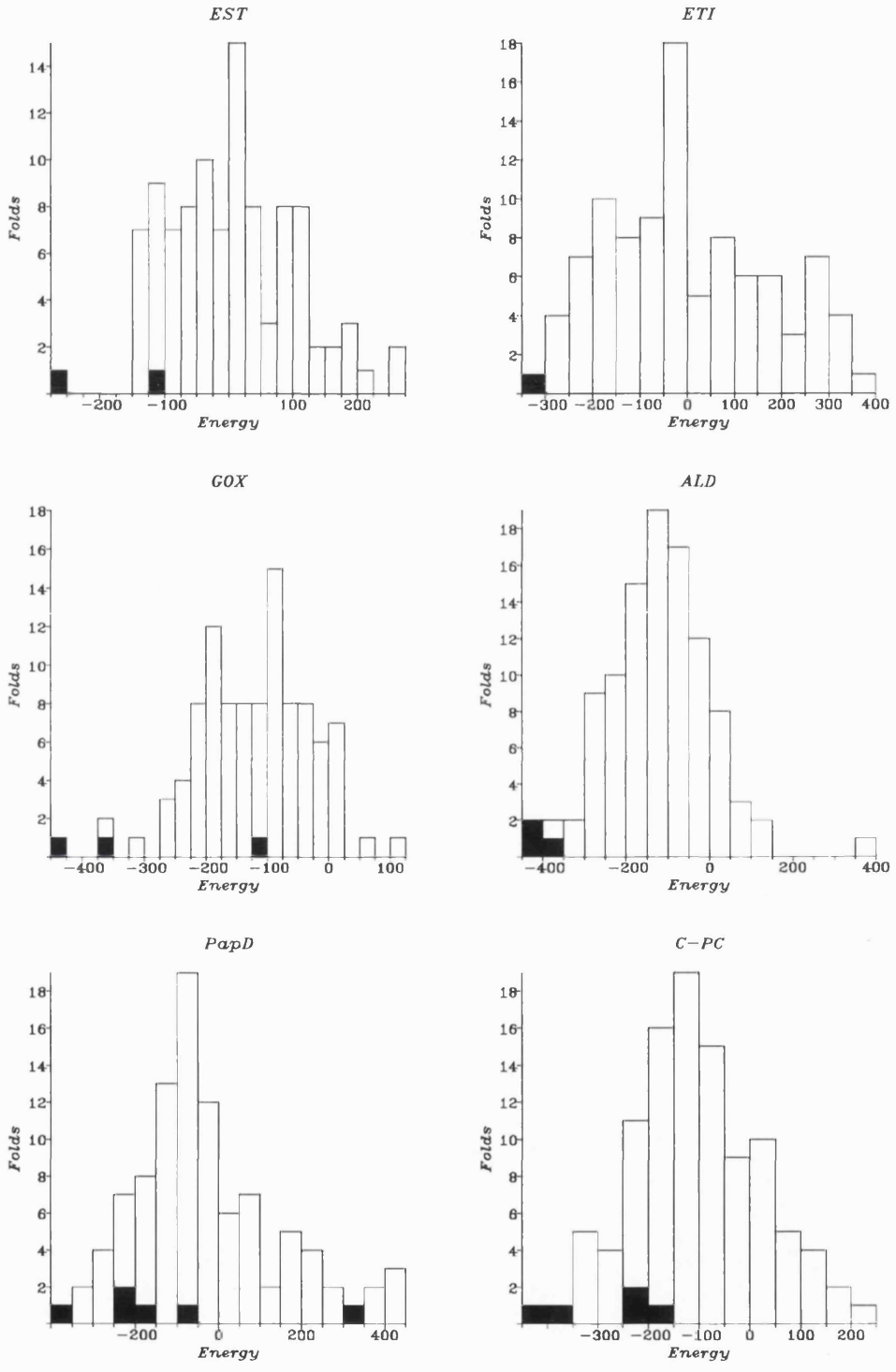
After determining the optimal threading, the energy of the final threaded model was calculated using a *zero* distance variation parameter δ, and over the following atom pairs: $C\beta \rightarrow C\beta$, $C\beta \rightarrow N$, $N \rightarrow C\beta$, $C\beta \rightarrow O$, $O \rightarrow C\beta$ rather than just the $C\beta \rightarrow C\beta$ potential used to calculate the threading. Hardening the potentials, and adding more atom pair terms for the final energy calculation improved the selectivity of the searches, by deepening and sharpening the threading energy minimum. When this was done at the start of the

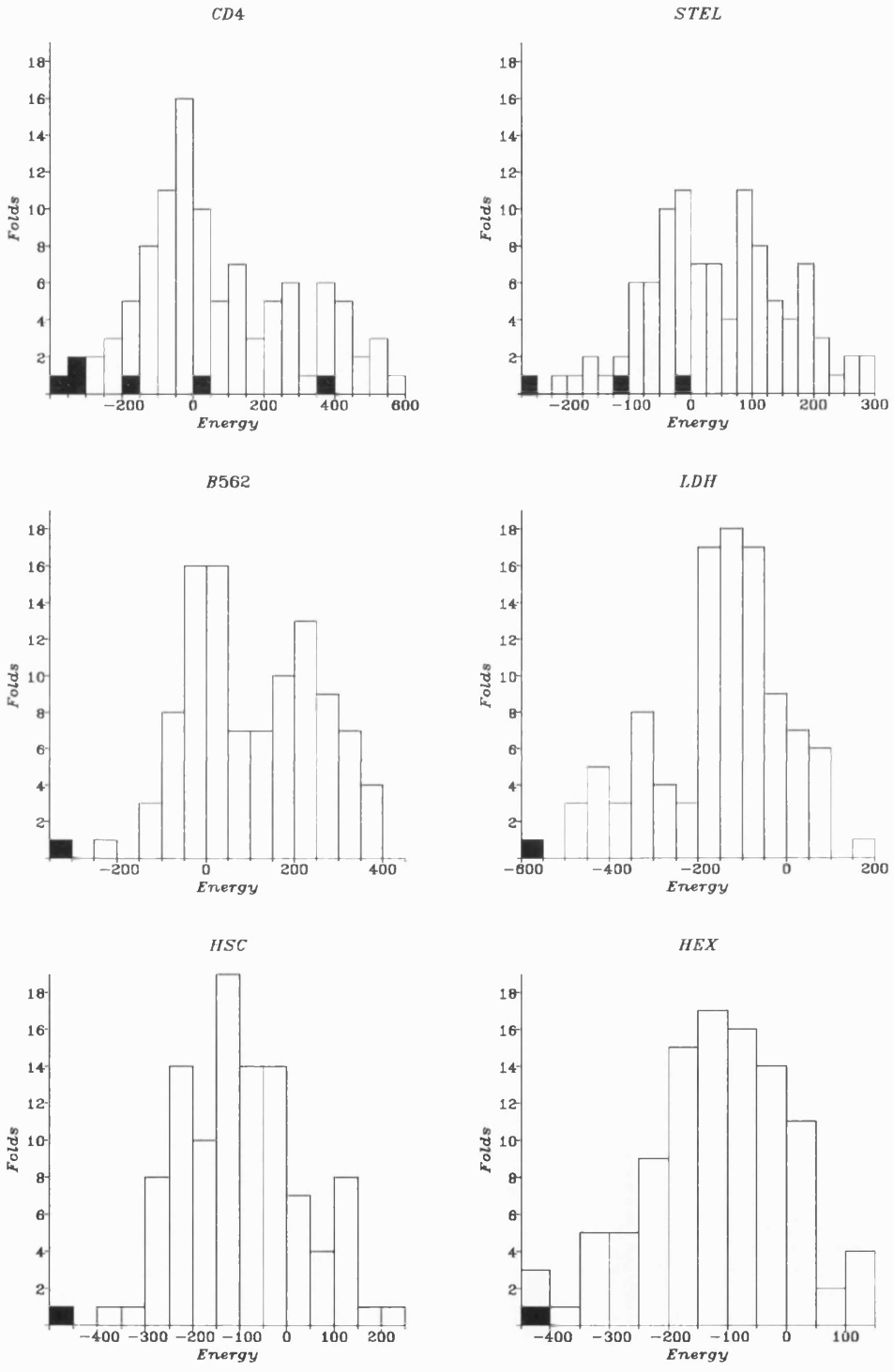
threading procedure, the double dynamic programming algorithm proved to be unable to locate a reasonable alignment.

In order to equalize the contribution of the solvation terms with the pairwise terms, the pairwise and solvation terms were summed and stored separately, and standard deviations (SD_{pair} and SD_{solv}) for the two contributing factors calculated over the set of 102 folds. The final energy was taken as: $E = E_{pair} + W E_{solv}$, where $W = (SD_{pair} / SD_{solv})$. In Table 4.2, the 'confidence' of the match (ΔE) is given in terms of the absolute energy difference between the top scoring fold and the next highest scoring, different, fold.

Figure 4.10

Threading histograms for the trial fold recognition searches. In each histogram, the positions of folds expected to match the given sequence (i.e. those folds similar to the known fold of the test sequence) are shown as shaded bars. For example, in the case of LDH (lactate dehydrogenase), the expected match in the database of folds is MDH (malate dehydrogenase). This match is shown as a single shaded bar representing an energy of -577 (kcal/mol), an energy which is lower than that achieved by any other fold.


EST
Folds
Energy
ETI
Folds
Energy
GOX
Folds
Energy
ALD
Folds
Energy
PapD
Folds
Energy
C−PC
Folds
Energy


CD4
Folds
Energy
STEL
Folds
Energy
B562
Folds
Energy
LDH
Folds
Energy
HSC
Folds
Energy
HEX
Folds
Energy

Test Protein	Source	Fold	Best Match	ΔE	%Seq. ID	Matches
C-phycocyanin β (C-PC)	Red Algae	Globin	1MBA	101	7	1,2,9,18,25
Glycolate Oxidase (GOX)	Spinach	TIM Barrel	1WSY(A)	52	10	1,3,49
Muscle Aldolase (ALD)	Human	TIM Barrel	4XIA(A)	80	6	1,2,3
Lactate Dehydrogenase (LDH)	Dogfish	Rossmann	4MDH(A)	87	15	1
Elastase (EST)	Pig	Trypsin	4PTP	110	35	1,14
CD4	Human	Ig	2FB4(H)	87	10	1,2,10,31,98
Stellacyanin (STEL)	Varnish Tree	Cu Binding	2AZA(A)	18	14	1,6,20
Cytochrome B562 (B562)	E. Coli	4-helix bundle	2MHR	78	6	1
Trypsin Inhibitor DE-3 (ETI)	Kaffir Tree	Interleukin 1β	1I1B	14	5	1
papD - chaperonin	E. Coli	Ig	2FB4(H)	64	15	1,5,9,16,35,93
70 kD H.S. Cognate (HSC)	Cow	Actin	1ATN(A)	94	9	1
Hexokinase B (HEX)	Yeast	Actin	1ATN(A)	0	12	1

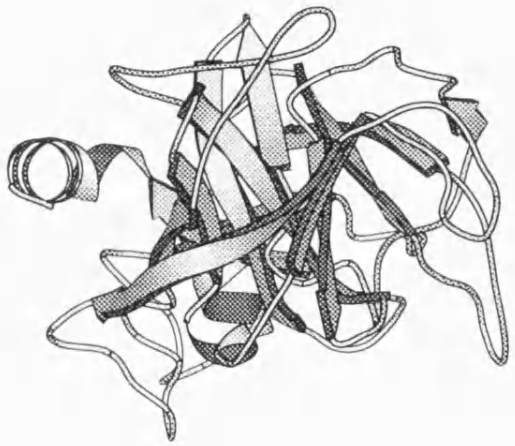
Table 4.2

Summary of results of a set of trial fold recognition searches. In each case the database included 102 protein chains as described in the text, except where the test protein was itself in the database, in which case it was excluded. The 'Best Match' column gives the Brookhaven ID of the best matching chain fold (including chain ID where appropriate), along with the sequence identity between the best matching chain and the test protein. Positions in the sorted list of threading energies of similar folds are also shown.

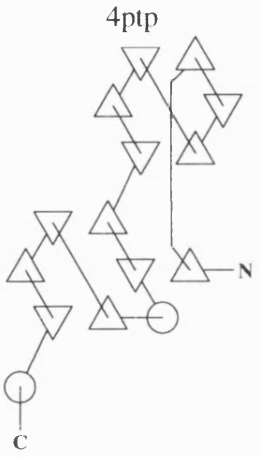
4.8 Elastase

The first example is a relatively easy problem in fold recognition. Elastase is a member of the trypsin-like serine protease family, sharing about 35% sequence identity with trypsin itself. The trypsin fold comprises two antiparallel β-barrel domains, with loops from both domains forming the enzyme's active site, comprising a catalytic triad of His and Asp from domain 1 and the key Ser from domain 2. The fold and topology of bovine trypsin (Bode & Schwager, 1975) is shown in Figure 4.11.

Despite the fact that the similarity between elastase and trypsin is sufficiently high that it can be detected by a standard sequence search algorithm, this example is a useful initial test, which allows the score of a highly significant hit to be determined. However, apart from bovine β-trypsin which is clearly shown to be the most compatible fold for elastase, one additional relative of trypsin is in the library of 102 folds, proteinase A from *Streptomyces Griseus* (Sielecki *et al.*, 1979), but which is positioned some way down the list (position 14). This failure is probably attributable to the significant difference in the lengths of the loops between elastase and proteinase A, and also the fact that the most highly conserved portions of the structures are in coil regions, which are generally ignored by the alignment algorithm. Clearly future developments of the method will have to tackle the problem of such misalignments.



a)



b)

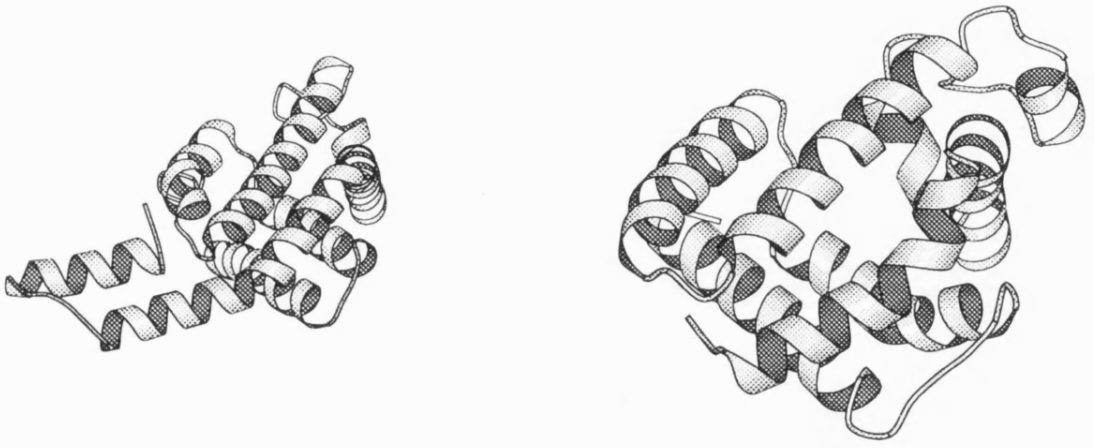
Figure 4.11

a) A Molscript (Kraulis, 1991) ribbon drawing of the structure of bovine trypsin. b) TOPS (Flores *et al.*, 1993) topology schematic.

4.9 C-Phycocyanin

Perhaps the most exciting demonstration of the method, is the example of C-phycocyanin (C-PC), a phycobiliprotein forming part of the phycobilisomes of blue-green and red algae. The protein occurs as a $(\alpha\beta)_3$ trimer, each monomeric unit comprising one α chain and one β chain. The striking feature of the C-PC chain fold is the fact that the globular portion (helices A-H) closely resembles the globin fold (Schirmer *et al.* 1985; Figure

4.12). Despite the similarity in fold, the sequence homology between the globins and C-PC is very low, with only 14 of the 174 β-chain residues of C-PC having identical counterparts in myoglobin; after careful structural comparison, however, Pastore and Lesk (1990) have proposed a very distant evolutionary link between the phycocyanins and the globins. To date, sequence analysis methods have proven unable to reliably detect the globin fold in C-PC. For example, despite great success in constructing templates to select almost every available globin sequence, Bashford *et al.* (1987) were not successful in matching these templates against the phycocyanins. Using the optimal threading algorithm, the 102 protein chains were searched for folds compatible with a single C-PC sequence. The top two folds were found to be sea hare myoglobin (-451 kcal/mol - Bolognesi *et al.*, 1989) and midge erythrocruorin (-356 kcal/mol - Weber *et al.*, 1978) followed by a number of other α-rich protein folds. Figure 4.13 shows the alignment corresponding to the optimal threading of the C-PC β-chain sequence onto the best matching fold (sea hare myoglobin). The fact that the optimal threading algorithm finds sea hare myoglobin to be the best model for C-PC is in accordance with the findings of Pastore and Lesk (1990), who find the helix geometry of this globin to be closest to that of C-PC. Not only has the method correctly identified its globin fold, but has accurately located it in the C-PC sequence, and has generated an alignment close to that obtained by careful structural alignment. It is clear that the method has identified the related folds in the database.



C-phycocyanin Myoglobin

Figure 4.12

Ribbon diagrams of C-phycocyanin and myoglobin.

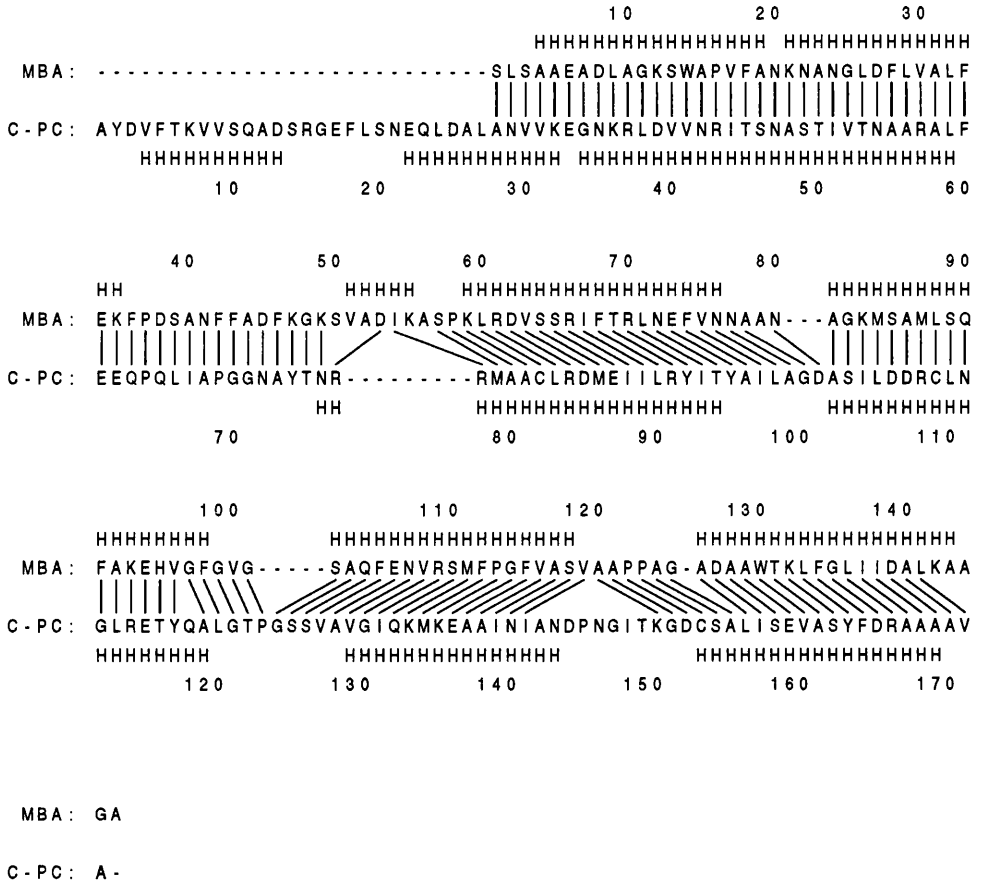


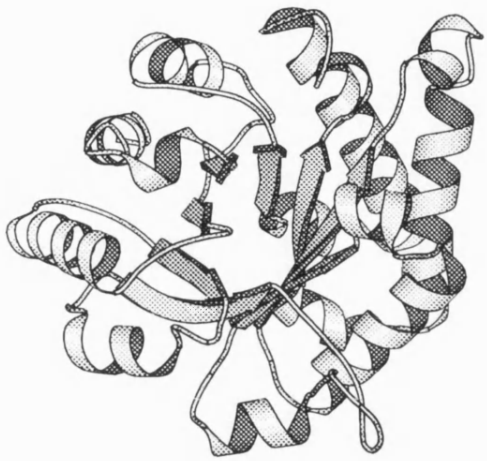
Figure 4.13

The alignment of sea hare myoglobin (1MBA) with C-phycocyanin β chain from *Mastigocladus laminosus* (SWISSPROT code PHCB$MASLA), found by optimal threading. Author assigned secondary structure codes are shown. The alignment is compared to the structurally determined alignment by Pastore and Lesk (1990).

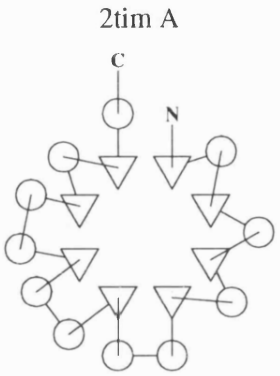
4.10 TIM barrels

There are two main classes of parallel α/β folding pattern: the αβα sandwich (described in the next section), and the α/β barrel, generally known as the "TIM barrel" after the first enzyme found to have this fold, Triosephosphate IsoMerase (Banner *et al.*, 1975), which is depicted in Figure 4.14. Both folding patterns are built up out of right-handed β-α-β units, though with different strand connectivities. Barrel structures comprise sequential β-α-β motifs, with strictly sequential strand ordering (12345678), leaving the helices on the outside of the closed sheet, whereas the units alternate in direction for the sandwich structures, producing a strand ordering such as 654123 in the case of the mononucleotide-binding domain of lactate dehydrogenase, and resulting in the helices packing against both faces of the open sheet.

The α/β barrel folding pattern has been observed in around 20 different enzymes to date, and in most cases the barrel comprises 8-strands. The active site in these α/β barrel enzymes is generally formed out of the loops connecting the α-helices to the carboxy-terminal ends of the strands, and it is this remarkable consistency coupled with the high structural similarity that has led to the proposed notion of a distant common ancestral α/β barrel. It now seems more likely, however, that the α/β barrel has been "re-invented" many times over the course of evolution, and it is the simplicity of the fold coupled with it's very high stability that has led to its common occurrence in present day protein domains.



a)



b)

Figure 4.14

a) A Molscript ribbon drawing of the structure of triosephosphate isomerase. b) Topology schematic for the same structure.

Given the ubiquity of the α/β barrel fold, it is a very tempting target for the application of fold recognition. Two α/β barrel searches were performed: one with the sequence of spinach glycolate oxidase (Lindqvist, 1989), and one with the sequence of human muscle aldolase (Gamblin *et al.*, 1991). The folds of both enzymes contain classic 8-stranded barrels. In both cases, the fold is recognized, and in the case of human aldolase, the three α/β barrels present in the library, triosephosphate isomerase, xylose isomerase (Henrick *et al.*, 1989, and the A chain of tryptophan synthase (Hyde *et al.*, 1988) are found to have lower threading energies than the other 99 chain folds. The lowest energy match for glycolate oxidase was found to be tryptophan synthase, with triosephosphate isomerase

at position 3 and xylose isomerase dropping down to position 49. For human muscle aldolase, the best matching structure was found to be xylose isomerase, followed immediately by triosephosphate isomerase and the A chain of tryptophan synthase.

4.11 Lactate dehydrogenase

Lactate dehydrogenase was the first NAD-dependent dehydrogenase protein to have its structure determined (Adams *et al.*, 1970). Two years later, the structure of malate dehydrogenase was solved (Hill *et al.*, 1972), which was found to have an almost identical structure. Rossmann (1974) identified the α/β mononucleotide-binding domain as a frequently occurring motif, and it is now most commonly known as the "Rossmann Fold". The general arrangement of this motif is to have a central open twisted parallel β-sheet, with helices packed on both sides of the sheet. The mononucleotide-binding motif has been found in a large number of different proteins, often as a separate domain from the functional domain in larger proteins. Lactate dehydrogenase and malate dehydrogenase share more structural similarity than a common mononucleotide-binding motif. Not only is the mononucleotide-binding domain identical in both cases, but their *functional* domains are very similar as well, and they are clearly evolutionarily related, albeit distantly.

Not surprisingly, a search of the 102 chains with malate dehydrogenase identifies lactate dehydrogenase as by far the most favourable fold. It is interesting to note where the other chain folds incorporating mononucleotide-binding domains are located in the list of threading energies. These topologically similar parallel $\alpha\beta$ domains were positioned at 3,7,11,12,13,17,19,31,34 and 82 in the list. Given the gross differences in structure observed in these domains, such a distribution is reasonable. However, this does raise an interesting question that will be addressed again in the concluding sections later, and that is the question of how to handle protein domains. Obviously in the case of malate and lactate dehydrogenase the similarity is quite evident over the entire chain, but quite clearly the possibility of matches between constituent domains will often arise. As implemented

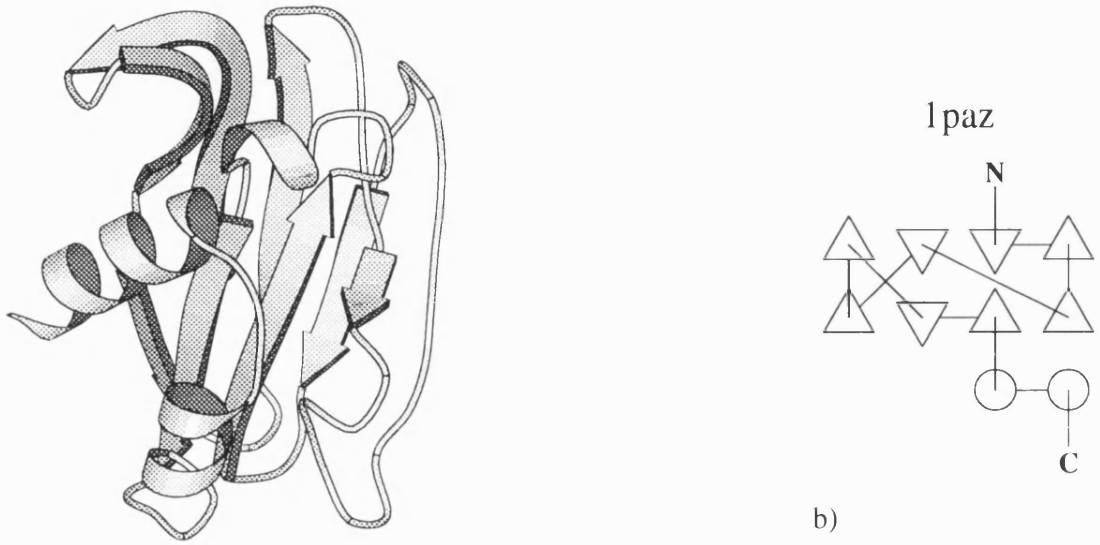
at present, the threading algorithm is based on a *global alignment* algorithm, which strives to match an entire sequence with an entire structure. Clearly if domains are to be recognized reliably, then a *local alignment* algorithm will be required.

4.12 Stellacyanin

Stellacyanin is a remote member of the family of copper-II binding electron transport proteins, a family which includes azurin (Norris *et al.*, 1983), pseudoazurin (shown as the representative example in Figure 4.15 - Petratos *et al.*, 1987) and plastocyanin (Garrett *et al.*, 1984). Despite the fact that the structure of stellacyanin has yet to be solved[6], the structure of it's closest relative in the family (cucumber basic protein) has been determined, and there has been keen interest in the construction of models for stellacyanin (Fields *et al.*, 1991).

A search of the fold library using the stellacyanin sequence revealed azurin to be the most compatible fold, with pseudoazurin at position 6 and plastocyanin at position 20.

[6] In fact, the structure of stellacyanin has just been solved - Guss, J.M., *personal communication.*



a)

b)

Figure 4.15

a) A Molscript ribbon drawing of the structure of pseudoazurin. b) Topology schematic for the same structure.

4.13 Cytochrome B562

Cytochrome B562 (Mathews *et al.*, 1979) is one of the many examples of perhaps the simplest of all the recurrent folds: the four-helix bundle. Despite the apparent simplicity of the fold, there are 48 distinct topologies possible for four-helix bundles (Presnell & Cohen, 1989), of which only around 6 have been observed in structures solved so far. The motif is found in a wide range of proteins of different function, including myohemerythrin

(non-haem oxygen carrier - shown in Figure 4.16), ferritin (iron storage), Interleukin 4 and other similar cytokines, and even the tobacco mosaic virus coat protein.

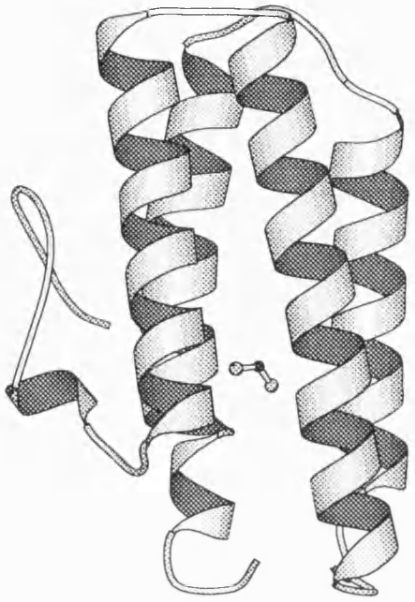


Figure 4.16

Ribbon diagram of myohemerythrin.

A search of the 102 folds with the sequence of cytochrome B562 produces the expected result that myohemerythrin (Hendrickson *et al.*, 1975), the only other four-helix bundle in the library, is the most compatible fold. It is interesting to note that though both proteins share a common folding pattern, their cores are different. Two iron atoms are bound in the core of myohemerythrin to enable oxygen binding, whereas in the case of the cytochrome, a single oxygen carrying iron is present as part of a heme group pincered by two helices. Clearly the detailed internal packing must be very different between these proteins, but nonetheless, the common topology is clearly recognized.

4.14 Trypsin inhibitor DE-3

Trypsin inhibitor DE-3 (Onesti *et al.*, 1991) from *Erythrina caffra* (ETI) is a Kunitz-type inhibitor, similar to soybean trypsin inhibitor (STI). The structure of ETI consists of 12 antiparallel β-strands joined by long loops, which can be thought of either as a tetrahedral arrangement of 4 sheets, or a 6-stranded antiparallel β-barrel, closed at one end by the other six strands coupled in pairs. The structure of ETI is remarkably similar to that of interleukin-1β (Finzel *et al.*, 1989), shown in Figure 4.17, although the two protein families show no detectable sequence similarity.

Despite the fact that the total threading energy for interleukin-1β is the lowest out of the 102 folds, it is interesting to note that the fold is essentially recognized by its solvation pattern alone. In fact the contribution of the pairwise terms was only barely negative (-1.2 kcal/mol), with several other all-β folds providing somewhat lower energies, although none of energies were particularly low. Conversely, the solvation energy for interleukin-1β clearly separated it from the other folds. A reasonable explanation for this is the fact that these trefoil structures incorporate very large exposed loops, which on the one hand contribute nothing to the pairwise energy terms, and on the other hand provide a very clear solvation signal which is easily detected.

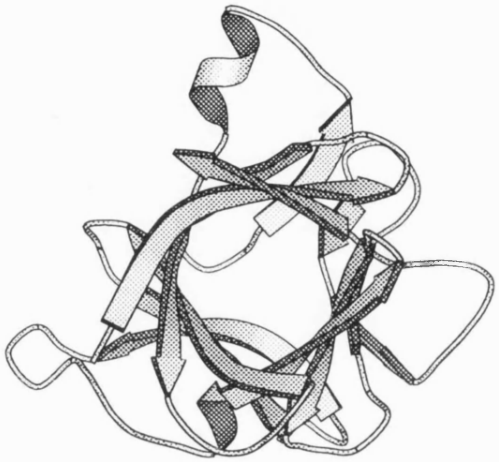
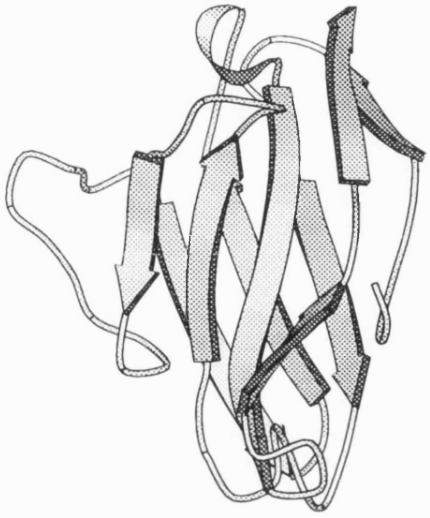


Figure 4.17
Ribbon diagram of interleukin 1β.

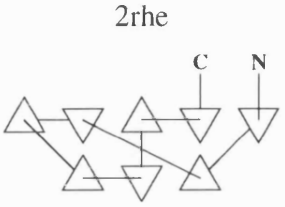
4.15 CD4 and chaperonin papD

The immunoglobulin fold recurs with astonishing regularity in many aspects of the immune system. The best known immunoglobulin molecule (IgG) comprises 12 immunoglobulin domains of similar overall structure, with 2 heavy-chains each containing 3 constant domains, and 1 variable domain (the variable domains contain the hyper-variable binding loops), and two light chains each with 1 constant and 1 variable domain. Other Ig molecules have different constant/variable domain arrangements. The constant domains comprise 7 strands arranged as two antiparallel β-sheets of 3 and 4 strands. The variable domains include an extra β-hairpin providing the second complementarity

determining region (CDR2), forming a 9-stranded (4+5) β-sandwich. Apart from the Ig molecule itself, similar domains are found in the HLA molecule, and in T-cell surface glycoprotein molecules.



a)



b)

Figure 4.18

a) A Molscript ribbon drawing of the structure of Bence-Jones protein variable-lambda domain. b) Topology schematic for the same structure. As can be seen in the ribbon diagram, the precise topology in this case is rather ambiguous. The TOPS program has (fairly reasonably) interpreted the structure as a 8-stranded sandwich (3+5).

The first Ig search involves a search using the sequence of the N-terminal fragment of the T-cell surface glycoprotein CD4 (Wang *et al.*, 1990; Ryu *et al.*, 1990), which is itself a member of the 102 fold library. This fragment comprises two Ig-like domains, though the second domain is somewhat distorted. Of the other 101 folds, the lowest energy threading

is obtained for the IgG Fab heavy chain (2FB4, chain H - Marquart *et al.*, 1980), with the next best match being the Bence-Jones protein variable-lambda domain (shown in Figure 4.18 - Furey *et al.*, 1983). The Fab light chain does not appear to match particularly well, appearing at position 31 in the ranked list of energies. The algorithm is clearly not capable of detecting the α3 Ig domain found in the heavy (A) chain of HLA (Bjorkman *et al.*, 1987), as it is found at position 98 (the B chain is at position 10). As will be discussed later, due to the global nature of the method, it is currently not capable of identifying domains that form small parts of much larger structures, as is the case for the α3 domain of HLA chain A. In particular, in this case, the solvent accessibility pattern for this HLA Ig-like domain is very specific to the HLA structure, as it is mostly buried. As a final comment, it is interesting to note that the topologically similar copper-binding proteins pseudoazurin, plastocyanin and azurin appear at positions 3, 6 and 21.

The uropathogenic Escherichia coli papD gene product, required for the biogenesis of digalactoside-binding P pili, is quite remarkable in that its fold comprises two Ig-like domains which are similar in sequence to the human lymphocyte differentiation antigen Leu-1/CD5A (Holmgren & Branden, 1989). Whilst this folding pattern is not unusual for a protein from a higher organism, the Ig fold had not been previously observed in prokaryotes before the structure of papD was solved.

Searching the fold library with the papD sequence again identified the H chain of 2FB4 to be the best matching fold, with CD4 being the next best matching Ig fold at position 5, followed by the Bence-Jones protein domain at position 9. Again matching with the two Ig domains in HLA was poor, with the B chain appearing at position 16, and the A chain at 93. Again this is attributable to the unusual accessibility pattern for this HLA domain.

4.16 70 kD heat shock cognate protein and hexokinase

The 70 kD heat shock cognate protein (Flaherty *et al.*, 1990), actin (Kabsch *et al.*, 1990), and hexokinase (Anderson *et al.*, 1978) are a functionally diverse family of proteins, which share a common ATPase domain. Despite there being little sequence similarity between them, the 44 kD N-terminal ATPase fragment of HSC70 has an almost identical structure to that of actin (illustrated in Figure 4.19). The structures of rabbit skeletal muscle actin and bovine HSC70 can be superposed with an RMSD of 2.3 Å over 241 equivalent C-α positions (Flaherty *et al.*, 1991). The similarity between hexokinase and actin is more topological than at the level of specific structural detail.

The two degrees of similarity are born out by the threading results for these proteins in that although actin is the lowest energy fold for hexokinase, the separation between the actin fold and the next best matching fold (aspartate transcarbamylase - ATC) is almost zero (0.1 kcal/mol), the rather weak structural similarity between hexokinase and actin would therefore appear to be just at the limits of the method. In contrast, the match between HSC70 and actin is clearly significant.

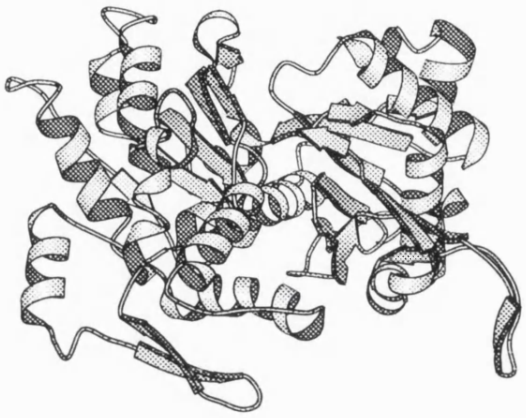


Figure 4.19
Ribbon diagram of actin.

4.17 Other examples

To conclude the description of optimal sequence threading two further examples will be presented, which arose after the initial development of the optimal threading method. Both problems were in fact posed by researchers interested in the protein family in question, and when the problems were posed, the structures of the proteins were unknown, and in fact the structure of one of these proteins is still unknown.

The first example was that of bovine aldose reductase, which catalyzes the NADPH-dependent reduction of D-glucose to D-sorbitol. In view of the fact that aldose reductase binds a nicotinamide adenine dinucleotide coenzyme, it was predicted that the enzyme would exhibit the classic parallel αβ sandwich structure described in section 4.11. A search through the fold library produced the results shown in Figure 4.20.

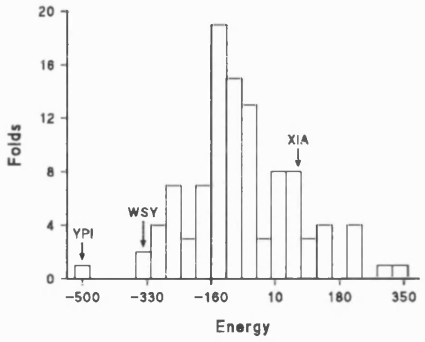


Figure 4.20

Optimal threading histogram for bovine aldose reductase. The three TIM barrel folds are indicated (YPI - triose phosphate isomerase, WSY - tryptophan synthase and XIA xylose isomerase).

Clearly, the fold of yeast triosephosphate isomerase is highly compatible with the sequence of bovine aldose reductase. Unfortunately, this result wasn't taken seriously in view of the fact that a TIM barrel fold was not at all likely for a NAD(P)H binding protein. When the 2.5 Å resolution structure of porcine aldose reductase was published however (Rondeau *et al.*, 1992), it was clear that the threading procedure had been successful. Aldose reductase, did indeed have a TIM barrel fold, the first NAD binding protein observed to have this folding pattern. This anecdotal example in fact raises an apparently trivial, yet important point about protein structure prediction: in that without

experimentally determining the structure, it is impossible to know for sure whether a prediction is right or wrong. Of course, the important test will be for a structure to be predicted and for useful experimental results to be obtained from the predicted structure ahead of its solution.

The final example is just such a case, where experimental work is being carried out on the predicted fold before its final solution. The protein encoded by the amiC gene of *Pseudomonas aeruginosa*, regulates the expression of inducible aliphatic amidase activity (Wilson & Drew, 1992). The problem of identifying the fold of this protein was posed by Dr L. Pearl, and application of the optimal threading algorithm (Figure 4.21) provided two strong candidates for matches: LBP (leucine binding protein - Sack *et al.*, 1989b), and GBP (galactose binding protein - Vyas *et al.*, 1988). An even better match than LBP was later found to be its close relative, leucine-isoleucine-valine binding protein (Sack *et al.*, 1989a), Brookhaven code 2LIV, which was not in the initial library of folds. Both these proteins are in the class of small-molecule binding proteins found in the periplasmic space between the inner and outer cell membranes of Gram negative bacteria, which comprise two parallel $\alpha\beta$ domains of similar structure (see Figure 4.22). Arabinose binding protein (ABP) also has a similar folding pattern to LIV, LBP and GBP, though the threading does not appear to match it to amiC.

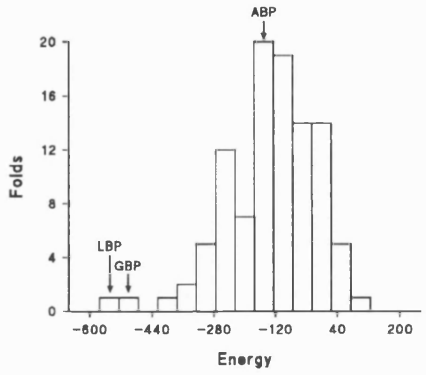


Figure 4.21

Optimal threading histogram for amiC. The three small-molecule binding folds are indicated.

Close scrutiny of multiple alignments between amiC and a set of LIV/LBP sequences showed that a distant evolutionary relationship between amiC and the small-molecule binding proteins was at least possible, particularly in view of the fact that the indels between amiC and the LIV/LBP families coincided with the indels observed in the LIV/LBP alignment alone (Wilson *et al.*, 1993). Furthermore, experimental evidence has now been obtained for the binding of acetamide with amiC (Wilson *et al.*, 1993), which not only supports its role as the regulator of the amidase system, but which makes its kinship to the small-molecule binding proteins all the more reasonable. From the combined experimental and theoretical work, it now seems highly probable that amiC shares a common fold with the periplasmic small-molecule binding proteins, which is remarkable considering the fact that amiC is located in the cytosol and has a unique regulatory role. If the predicted structure proves correct, then this will be the first example of the small-molecule binding fold outside the periplasm of gram-negative bacteria.

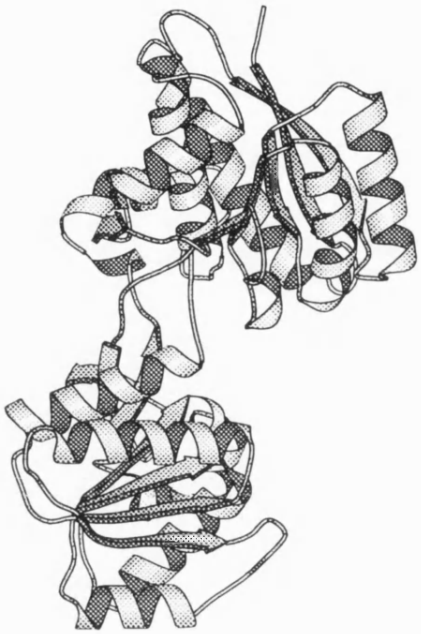


Figure 4.22

Ribbon diagram of leucine-isoleucine-valine (LIV) binding protein.

Subsequent to the successful crystallization of the amiC gene product (Wilson *et al.*, 1992), native crystal diffraction data has already been obtained to 3.5 Å resolution. Attempts at solving the structure by molecular replacement based on LIV/LBP models have unfortunately not been successful, but in view of the extremely distant relationship and the contrary requirement for very accurate models for successful molecular replacement this is not too surprising. Nevertheless, it is hoped that the structure will be solved in the near future by isomorphous replacement, and so the prediction will be either confirmed or refuted.

4.18 Future developments

These results are intended as preliminary verification of the potential of optimal sequence threading to fold identification and, possibly, prediction. Further investigation into this application of the methodology is in progress, though even at this stage, optimal sequence threading would appear capable of not only being able to assist in the alignment of a sequence to a structure, but also of suggesting *which* structure should provide the most likely matching template.

The methods described here are currently being integrated into a complete structure prediction package, which takes as input a sequence (or a family of sequences) and produces as output a set of model coordinates. Whether or not these coordinates relate to reality will depend on whether or not any similarities can be detected between the input sequence and any protein of known structure. From the results shown in this work, optimal sequence threading seems to provide a useful means of both detecting remote structural relationships, and utilizing these relationships to begin construction of a putative model.

For the cases considered here only a single sequence has been used in the threading procedure, however much additional information can be derived from consideration of multiple sequences. Instead of threading one sequence onto a structural template, it is possible to thread a *consensus sequence* onto the template. A consensus sequence is typically derived from a multiple sequence alignment and conveys information on the pattern of sequence conservation observed over a family of proteins. Using consensus sequences, secondary structure signals become more prominent due to the averaging out of 'evolutionary noise', and in addition, regions of the consensus sequence containing gaps may be matched preferentially to the template's loop regions.

One other problem with the optimal threading procedure as described, that might indirectly benefit from the use of multiple sequence data, is how to properly deal with

structural domains. It is widely recognized that polypeptide chains longer than about 150 residues are very often divided into more or less obvious structural domains (Janin & Chothia, 1985). In the simplest case of *continuous* domain structure, the chain will only make one crossing into a domain, resulting in a sequential ordering of domain assignments along the sequence. In the *discontinuous* case, however, the chain will make two or even more crossings into a particular domain. A good example of this more complicated case is the case of the periplasmic small molecule binding proteins, where the chain makes three crossings between the two domains (very clearly shown in Figure 4.22).

As mentioned earlier, the threading method as described works on the basis of global similarity between a sequence and a structure. For a multidomain structure, therefore, the method will only reliably detect a match between *all* domains and a given sequence, rather than any of the domains individually. In some fortunate cases, the match between a sequence and a component domain may be so strong that the domain structure will be correctly identified, but this clearly cannot be generally relied upon.

There are two routes towards the better handling of multidomain structures. Firstly the chain fold library could be converted into a *domain* fold library, whereby the chains are divided into their constituent domains prior to a fold recognition search. Ideally such a division would be performed automatically, using a suitable domain assignment method. Several methods have been proposed for the automatic assignment of continuous domain boundaries (for example see Rose, 1985), though the results frequently disagree with the "by-eye" assignments made by the crystallographers, and results for discontinuous domains are very poor indeed. An automatic method has been developed, based on the detection of discontinuities in a graph incorporating main chain and side chain hydrogen bonding and hydrophobic contacts, which is currently under evaluation. Preliminary results seem to indicate that such an approach works more reliably for discontinuous domains, and about as well as previous approaches for the continuous examples.

The second approach to the handling of domains, which is perhaps of greater theoretical interest, but which is a far less tractable problem, is the detection of domain boundaries *from sequence*. In some cases, domain boundaries in sequences can be assigned by means of sequence homology (Barker *et al.*, 1987, 1988), where local similarity is detected between part of a newly characterized protein sequence, and another sequence of known function and/or structure, but this is only readily applicable in a limited number of cases. Intron/exon boundaries can sometimes correlate with structural domain boundaries, though this is not reliable, and is of course strictly limited to "new" eukaryotic proteins. Other approaches to the detection of domain boundaries have not proven successful (Busetta & Barrans, 1984; Vonderviszt & Simon, 1986). The detection of domains from sequence remains, therefore, an important unsolved problem in molecular biology. It might well be expected that a useful solution to this problem will prove a major stepping-stone towards a solution to the ultimate problem of predicting protein structure from amino acid sequence.

4.19 Conclusions

It is now well-known that proteins such as the various TIM barrel enzymes, interleukin 1β/soybean trypsin inhibitor, and actin/hexokinase can show remarkable similarities in their native folds with no apparent sequence similarities. Furthermore, the rate at which newly solved protein structures are perceived to have previously observed folds suggests that the number of protein topologies may be limited. Indeed, some estimates put the number of observed topologies at 50% of the total number of naturally occurring topologies (though 10% is a more likely estimate). Given the significant possibility that a newly sequenced protein will have a previously observed fold, it is essential that methods for the recognition of protein folds in sequences be developed.

Despite the threading problem being computationally hard, a number of methods have been presented (including a novel derivation of a structural alignment algorithm) for effectively locating the optimal threading, and these methods appear to work very well. These methods can easily be extended to take account of more complex atomic interactions such as explicit hydrogen bonding, disulphide bridge formation, or indeed any physical effect that can be expressed as a function of interatomic distances.

By means of a crude set of potentials, encompassing local and long-range pairwise residue-residue terms, along with a simple solvation potential, it is clearly possible to determine the optimal threading of a sequence onto a given structural template. Thus, in order to align a sequence with a template structure, only a knowledge of the main chain atom positions is required; there is no need to consider the template's sequence, and no need to consider the detailed aspects of side chain interactions. By threading a sequence onto each of a library of structural templates and evaluating the total energy of each threading, the fold of the sequence may be identified. Sequence threading therefore provides a holistic means of both identifying structural relationships and extrapolating these relationships towards the automatic generation of a low resolution model structure.

Current success of the method described here and methods developed in other labs is extremely encouraging, and it is to be hoped that these individual approaches to the problem will cross-fertilize and lead to even more successful methods in the future. A good example of this cross-fertilization is the method recently described by Godzik *et al.* (1992) which incorporates aspects of both our optimal threading and the lattice-based approach of Finkelstein and Reva. Wilmanns and Eisenberg (1993) also report favourable results from the addition of a pairwise mean force potential to the original environment-based parameters of Bowie *et al.* (1991). With luck, by the time we have observed every possible fold, we will have the wherewithal to recognize these folds in our sequence data.

Chapter 5

A Model Recognition Approach to the Prediction of Membrane Protein Structure and Topology

He killed the noble Mudjokivis.
Of the skin he made him mittens,
Made them with the fur inside
Made them with the skin side outside.
He, to get the warm side inside,
Put the inside skin side outside.
He, to get the cold side outside,
Put the warm side fur side inside.
That's why he put the fur side inside,
Why he put the skin side outside,
Why he turned them inside outside.

- The Modern Hiawatha

5.1 Introduction

Integral membrane proteins represent an important, yet functionally diverse class of protein structure. From the observations made on the structure of bacteriorhodopsin (determined by electron diffraction; Henderson *et al.* 1990) and the photosynthetic reaction centre (determined by X-ray crystallography; Deisenhofer *et al.*, 1985), it has been concluded that transmembranal segments are typically apolar helices, 17-25 residues in length. Such an arrangement allows the apolar side chains to interact favourably with the lipid environment, whilst fully satisfying the hydrogen bonding potential of the peptide units in a regular secondary structure. An alternative arrangement has been observed in the crystallographically determined structure of a bacterial outer membrane porin (Weiss *et al.*, 1992), where the transmembranal segments are β-strands arranged in an 16-stranded barrel. The porin transmembranal segments are again apolar but due to the extended conformation can be as short as 6-7 residues in length. It is generally believed (or perhaps more correctly "hoped") that the structure of porin is the exception rather than the rule, and so typical methods for the prediction of membrane-spanning segments of integral membrane proteins (von Heijne, 1981, 1992; Argos *et al.*, 1982; Eisenberg, 1984; Engelman *et al.*, 1986) implicitly assume the predicted segments to be helices. How reasonable this assumption is cannot be determined until more integral membrane protein structures are solved, however circular dichroism studies at least support the notion, indicating that most transmembranal proteins have a very high helix content.

Methods for the prediction of transmembrane spanning segments are typically based on hydropathy analysis (Kyte & Doolittle, 1982). The simplest scheme is to generate a hydrophobicity plot for a given sequence, with transmembrane segments being centred at the peaks of the plot. For an initial hydrophobicity plot a window of 17-22 residues is taken, and using a suitable hydrophobicity scale (Kyte & Doolittle, 1982; Engelman *et al.*, 1986; Cornette *et al.*, 1987) the average residue hydrophobicity calculated. Plots based on smaller windows (5-11 residues) are helpful to delineate the end points of the segments. Despite being generally apolar, transmembranal segments often exhibit a degree of

amphipathicity and this can be used in addition to the simple hydrophobicity plot to improve predictive accuracy (Eisenberg 1984, Stirk *et al.* 1992).

Recent studies (von Heijne & Gavel, 1988; Nakashima & Nishikawa, 1992) have indicated the presence of topogenic signals in integral membrane proteins, i.e. sequence patterns which correlate with the topology of the membrane-spanning segments. The most evident of these signals is the prevalence of positively charged residues in the interior (cytoplasmic) loops which is now familiarly known as the 'positive inside rule' (von Heijne & Gavel, 1988). Such topogenic signals can be used to evaluate the plausibility of predicted integral membrane structures, a fact which has been very elegantly demonstrated by von Heijne (1992).

In this work a method is described that simultaneously takes into account the prediction of transmembrane secondary structure and the location of topogenic signals. For any given topology and scoring scheme, a mathematically optimal solution is found, which enables the likelihood of each suggested topology to be objectively assessed. The basic idea here is the idea of *expectation maximization*, a simple statistical method which is concerned with the generation and fitting of models to data. Traditional prediction schemes attempt to determine the most reasonable underlying model based on an analysis of one or more sequences. In contrast, expectation maximization attempts to search for the model which best explains the given data. Given a function which calculates the total probability for the match of a given model with a given sequence, the resulting model from expectation maximization should correspond to the maximum of this function.

The first requirement for expectation maximization is the definition of a model (used here in the statistical sense, rather than the biomolecular sense). In the case of transmembrane prediction such a model includes parameters for the number of membrane-spanning segments n, the topology t (N-terminus in or out), and the length l, and location i in the sequence of each segment.

In the case of the work shown here, residues are classified as being in 5 structural states, as shown in Figure 5.1. The 5 states are as follows: L_i (inside loop), L_o (outside loop), H_i (inside helix end), H_m (helix middle), and H_o (outside helix end). For a helix of length *l*, the number of residues taken to be in the end caps was arbitrarily taken as being *4*. Other cap definitions could be used, and it is possible that a more rigorous definition of the cap/middle boundaries might improve results.

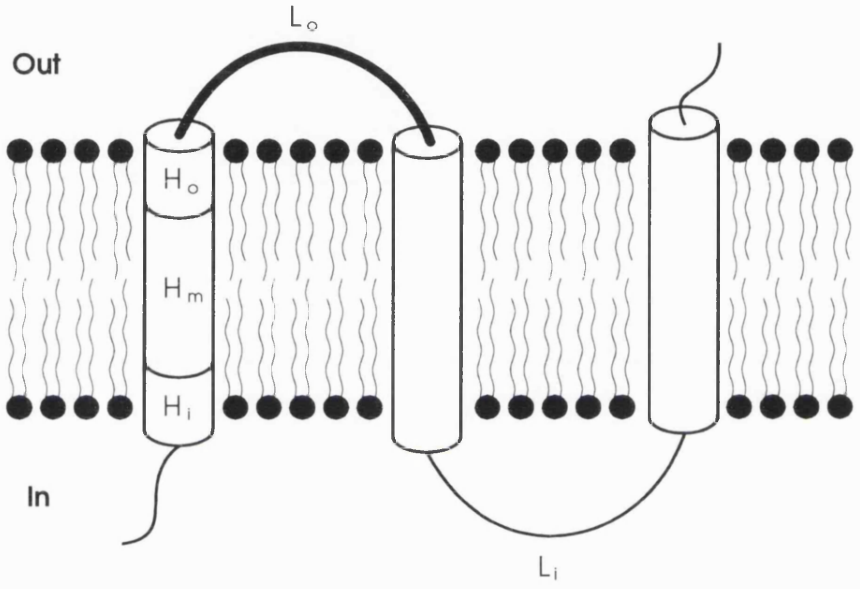


Figure 5.1

The 5 structural states defined for a typical helical transmembrane protein.

The source data for this work was a set of documented transmembrane proteins extracted from Release 23.0 of SWISS-PROT (Bairoch & Boeckmann, 1991). The initial set of 1765 membrane sequences was split into two subsets, one containing the proteins with a single membrane-spanning segment, and the other with multiple membrane-spanning segments. Both sets were further reduced to include just those sequences for which the membrane topology was given, in addition entries for which any transmembrane segment was listed as shorter than 17 or longer than 25 were eliminated. In view of the difficulty

in assigning membrane-spanning segments for multi-spanning proteins, only the multi-spanning segments for which at least some experimental data were available were included (listed in Table 5.1). Much of this structure and topology information must be taken as being hypothetical, but it is to be hoped that the majority of the data are correct. Unfortunately, until more experimental data become available on membrane protein structure, there is little option but to take the authors' descriptions at face value. The final single and multi-spanning data sets comprised 285 and 35 sequences respectively.

RCEL_CHLAU	REACTION CENTER PROTEIN L CHAIN. 5/92
RCEM_CHLAU	REACTION CENTER PROTEIN M CHAIN. 5/92
5HT2_CRIGR	5-HYDROXYTRYPTAMINE 2 RECEPTOR (5-HT-2). 5/92
5HT3_MOUSE	5-HYDROXYTRYPTAMINE 3 RECEPTOR PRECURSOR (5-HT-3). 3/92
5HTA_HUMAN	5-HYDROXYTRYPTAMINE 1A RECEPTOR (5-HT-1A). 5/92
A1AA_HUMAN	ALPHA-1A ADRENERGIC RECEPTOR. 5/92
A2AA_HUMAN	ALPHA-2A ADRENERGIC RECEPTOR (SUBTYPE C10). 5/92
EDG1_HUMAN	PROBABLE G PROTEIN-COUPLED RECEPTOR EDG-1. 8/92
MOTA_ECOLI	CHEMOTAXIS MOTA PROTEIN. 11/91
MALF_ECOLI	MALTOSE TRANSPORT INNER MEMBRANE MALF PROTEIN. 11/90
SECY_BACSU	SECY PROTEIN. 3/92
OPS1_CALVI	OPSIN RH1 (OUTER R1-R6 PHOTORECEPTOR CELLS OPSIN). 8/91
OPSB_HUMAN	BLUE-SENSITIVE OPSIN (BLUE CONE PHOTORECEPTOR PIGMENT). 8/91
OPSD_BOVIN	RHODOPSIN. 8/91
AA1R_CANFA	ADENOSINE A1 RECEPTOR. 8/92
AA2R_CANFA	ADENOSINE A2 RECEPTOR. 8/92
C561_BOVIN	CYTOCHROME B561. 7/89
ADT_RICPR	ADP,ATP CARRIER PROTEIN (ADP/ATP TRANSLOCASE). 8/91
CYOB_ECOLI	CYTOCHROME O UBIQUINOL OXIDASE SUBUNIT I (EC 1.10.3.-). 11/90
CYOC_ECOLI	CYTOCHROME O UBIQUINOL OXIDASE SUBUNIT III (EC 1.10.3.-). 11/90
CYOD_ECOLI	CYTOCHROME O UBIQUINOL OXIDASE OPERON PROTEIN CYOD. 11/90
SECE_ECOLI	INNER MEMBRANE PROTEIN SECE. 8/91
GAPB_HUMAN	GAP JUNCTION BETA-1 PROTEIN (CONNEXIN 32). 3/92
LEP_ECOLI	SIGNAL PEPTIDASE I (EC 3.4.-.-) (SPASE I). 8/92
PT2M_ECOLI	PHOSPHOTRANSFERASE ENZYME II, MANNITOL-SPECIFIC. 3/92
ATHP_NEUCR	PLASMA MEMBRANE ATPASE (EC 3.6.1.35). 12/92
LACY_ECOLI	LACTOSE PERMEASE (LACTOSE-PROTON SYMPORT). 12/92
OPPB_SALTY	OLIGOPEPTIDE PERMEASE PROTEIN OPPB. 12/92
TAPA_HUMAN	CELL SURFACE PROTEIN TAPA-1. 8/92
DHSC_BACSU	SUCCINATE DEHYDROGENASE CYTOCHROME B-558. 7/89
LSPA_ECOLI	LIPOPROTEIN SIGNAL PEPTIDASE. 5/91
IMM1_ECOLI	IMMUNITY PROTEIN FOR COLICIN E1. 11/90
IMMA_CITFR	IMMUNITY PROTEIN FOR COLICIN A. 8/91
TCR1_ECOLI	TETRACYCLINE RESISTANCE PROTEIN. 2/91
UHPT_ECOLI	HEXOSE PHOSPHATE TRANSPORT PROTEIN. 8/92

Table 5.1

Multi-spanning protein sequences used to calculate topogenic parameters. Codes are from SWISS-PROT Release 23. To reduce bias in the calculated parameters, the full list was reduced so that no remaining pair of sequences is significantly more than 60% sequence identical.

	Single	Multi
Sequences	285	35
Transmembrane segments	285	174
Inside loop residues	5699	2032
Outside loop residues	830	1498
Inside helix residues	1140	696
Outside helix residues	1140	696
Middle helix residues	4003	2037
Total residues	179037	8730

Table 5.2

Composition of data sets used to calculate topogenic parameters. Note that the total residue counts include residues not assigned to any structural state i.e. oversized loops.

For each of the 5 structural classes, log likelihoods for each of the 20 amino acids were calculated:

$$s_i = \ln(q_i/p_i)$$

where p_i is the relative frequency of occurrence (or fraction) of amino acid i in all the sequences in the data set, and q_i is the relative frequency of occurrence of amino acid i in a particular structural class. A positive score indicates a higher than expected frequency for a given amino acid to be found in a particular structural class, a negative score a lower than expected frequency, and a score close to zero indicates that the frequency of occurrence of the given amino acid in a particular class is no different from that expected from chance alone. The scores calculated for the previously described set of sequences are shown for both single-spanning (Table 5.3 and Figure 5.2) and multi-spanning segments (Table 5.4 and Figure 5.3).

The log likelihood values clearly encode a variety of topogenic signals. The preference for positively charged residues to be found in the inside loops is clearly seen. It is interesting to note that a similar effect is seen between the inside and outside helix caps, though this could be due to the indeterminate boundaries between the author-defined membrane-spanning segments and their flanking regions. Of more interest are the signals that cannot be attributed to the simple positive-inside rule. The most striking of these is the preference both tryptophan and tyrosine exhibit for outside positions. The unusual abundance of tryptophan residues in outside locations of the photosynthetic reaction centre (and tyrosine in bacteriorhodopsin) has been noted by Schiffer *et al.* (1992), but it would appear from the results presented here that this is a general feature of transmembrane proteins as a whole. Without further experimental evidence, the question of whether these residues help in the direction of membrane topology, or merely act to stabilize the final topology remains open.

	p	$q(L_i)$	$s(L_i)$	$q(L_o)$	$s(L_o)$	$q(H_i)$	$s(H_i)$	$q(H_m)$	$s(H_m)$	$q(H_o)$	$s(H_o)$
Ala	0.065	0.058	**-0.111**	0.104	**0.474**	0.081	**0.23**	0.12	**0.622**	0.1	**0.434**
Arg	0.047	0.088	**0.633**	0.049	**0.042**	0.003	**-2.803**	0.004	**-2.548**	0.004	**-2.58**
Asn	0.048	0.045	**-0.064**	0.051	**0.059**	0.006	**-2.137**	0.005	**-2.219**	0.010	**-1.577**
Asp	0.052	0.055	**0.046**	0.039	**-0.295**	0.002	**-3.204**	0.003	**-2.711**	0.004	**-2.511**
Cys	0.030	0.017	**-0.546**	0.016	**-0.600**	0.038	**0.256**	0.023	**-0.269**	0.008	**-1.335**
Gln	0.043	0.050	**0.164**	0.029	**-0.378**	0.010	**-1.459**	0.005	**-2.17**	0.010	**-1.459**
Glu	0.063	0.072	**0.135**	0.059	**-0.058**	0.002	**-3.392**	0.003	**-2.898**	0.006	**-2.411**
Gly	0.068	0.054	**-0.226**	0.054	**-0.228**	0.040	**-0.522**	0.093	**0.316**	0.066	**-0.021**
His	0.023	0.029	**0.261**	0.021	**-0.093**	0.010	**-0.823**	0.006	**-1.282**	0.004	**-1.853**
Ile	0.049	0.033	**-0.389**	0.031	**-0.438**	0.128	**0.968**	0.130	**0.982**	0.154	**1.154**
Leu	0.090	0.062	**-0.367**	0.078	**-0.147**	0.237	**0.968**	0.231	**0.944**	0.186	**0.727**
Lys	0.054	0.102	**0.645**	0.041	**-0.269**	0.007	**-2.029**	0.004	**-2.516**	0.001	**-3.638**
Met	0.019	0.026	**0.311**	0.028	**0.383**	0.030	**0.462**	0.024	**0.231**	0.028	**0.364**
Phe	0.038	0.033	**-0.124**	0.033	**-0.120**	0.102	**0.991**	0.061	**0.476**	0.061	**0.476**
Pro	0.057	0.058	**0.023**	0.080	**0.338**	0.008	**-1.906**	0.017	**-1.2**	0.048	**-0.172**
Ser	0.078	0.076	**-0.025**	0.081	**0.036**	0.029	**-0.993**	0.050	**-0.447**	0.045	**-0.548**
Thr	0.064	0.055	**-0.142**	0.090	**0.34**	0.025	**-0.948**	0.045	**-0.351**	0.044	**-0.361**
Trp	0.015	0.015	**-0.014**	0.030	**0.708**	0.052	**1.241**	0.007	**-0.789**	0.044	**1.078**
Tyr	0.034	0.029	**-0.136**	0.019	**-0.544**	0.083	**0.903**	0.013	**-0.919**	0.031	**-0.075**
Val	0.067	0.041	**-0.483**	0.067	**0.002**	0.107	**0.474**	0.155	**0.84**	0.147	**0.788**

Table 5.3

Topogenic parameters for single-spanning transmembrane segments. Relative frequencies of occurrence for all structural states (p), relative frequencies for specific structural states (L_i : loop inside, L_o : loop outside, H_i : helix inside, H_o : helix outside and H_m : helix middle), and log-likelihoods for specific structural states (in bold) are shown for the 20 standard amino acids.

	p	$q(L_i)$	$s(L_i)$	$q(L_o)$	$s(L_o)$	$q(H_i)$	$s(H_i)$	$q(H_m)$	$s(H_m)$	$q(H_o)$	$s(H_o)$
Ala	0.078	0.073	**-0.064**	0.061	**-0.245**	0.106	**0.316**	0.097	**0.222**	0.092	**0.169**
Arg	0.044	0.083	**0.631**	0.044	**-0.006**	0.01	**-1.437**	0.004	**-2.406**	0.008	**-1.758**
Asn	0.041	0.046	**0.1**	0.055	**0.287**	0.015	**-1.025**	0.025	**-0.492**	0.02	**-0.732**
Asp	0.037	0.038	**0.016**	0.048	**0.266**	0.003	**-2.559**	0.01	**-1.358**	0.008	**-1.563**
Cys	0.021	0.018	**-0.156**	0.026	**0.191**	0.025	**0.169**	0.031	**0.373**	0.016	**-0.302**
Gln	0.033	0.041	**0.235**	0.039	**0.192**	0.01	**-1.219**	0.009	**-1.289**	0.012	**-0.971**
Glu	0.045	0.058	**0.251**	0.054	**0.178**	0.006	**-2.058**	0.007	**-1.885**	0.009	**-1.614**
Gly	0.065	0.052	**-0.235**	0.072	**0.092**	0.056	**-0.161**	0.067	**0.017**	0.076	**0.152**
His	0.018	0.026	**0.339**	0.027	**0.396**	0.006	**-1.111**	0.007	**-1.009**	0.01	**-0.617**
Ile	0.07	0.047	**-0.399**	0.044	**-0.476**	0.131	**0.622**	0.115	**0.493**	0.098	**0.338**
Leu	0.11	0.075	**-0.392**	0.083	**-0.283**	0.158	**0.36**	0.164	**0.396**	0.163	**0.391**
Lys	0.042	0.086	**0.714**	0.044	**0.045**	0.008	**-1.674**	0.005	**-2.145**	0.006	**-1.997**
Met	0.029	0.028	**-0.053**	0.029	**-0.016**	0.044	**0.405**	0.036	**0.204**	0.044	**0.413**
Phe	0.054	0.039	**-0.337**	0.051	**-0.061**	0.077	**0.349**	0.093	**0.545**	0.104	**0.655**
Pro	0.046	0.045	**-0.029**	0.062	**0.291**	0.017	**-1.002**	0.036	**-0.262**	0.031	**-0.413**
Ser	0.075	0.083	**0.105**	0.078	**0.047**	0.057	**-0.271**	0.065	**-0.144**	0.055	**-0.304**
Thr	0.06	0.065	**0.074**	0.066	**0.088**	0.049	**-0.204**	0.056	**-0.064**	0.051	**-0.166**
Trp	0.017	0.015	**-0.179**	0.022	**0.235**	0.026	**0.393**	0.023	**0.264**	0.034	**0.656**
Tyr	0.036	0.029	**-0.227**	0.04	**0.106**	0.065	**0.592**	0.032	**-0.122**	0.062	**0.552**
Val	0.077	0.056	**-0.314**	0.055	**-0.326**	0.132	**0.547**	0.12	**0.453**	0.102	**0.284**

Table 5.4

Topogenic parameters for multi-spanning transmembrane segments.

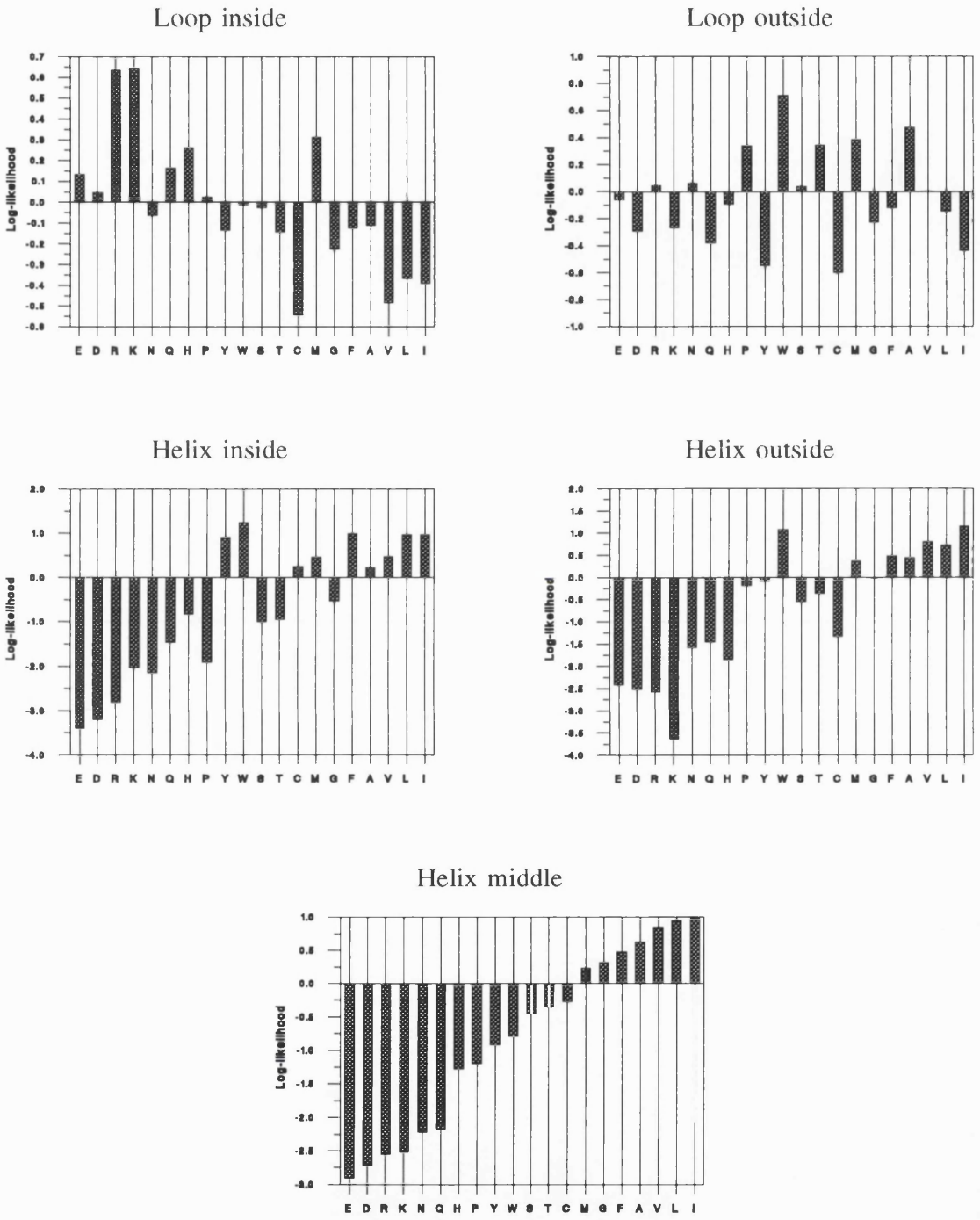


Figure 5.2

Plots of the topogenic parameters for single-spanning segments. Amino acids are ordered by the helix middle values.

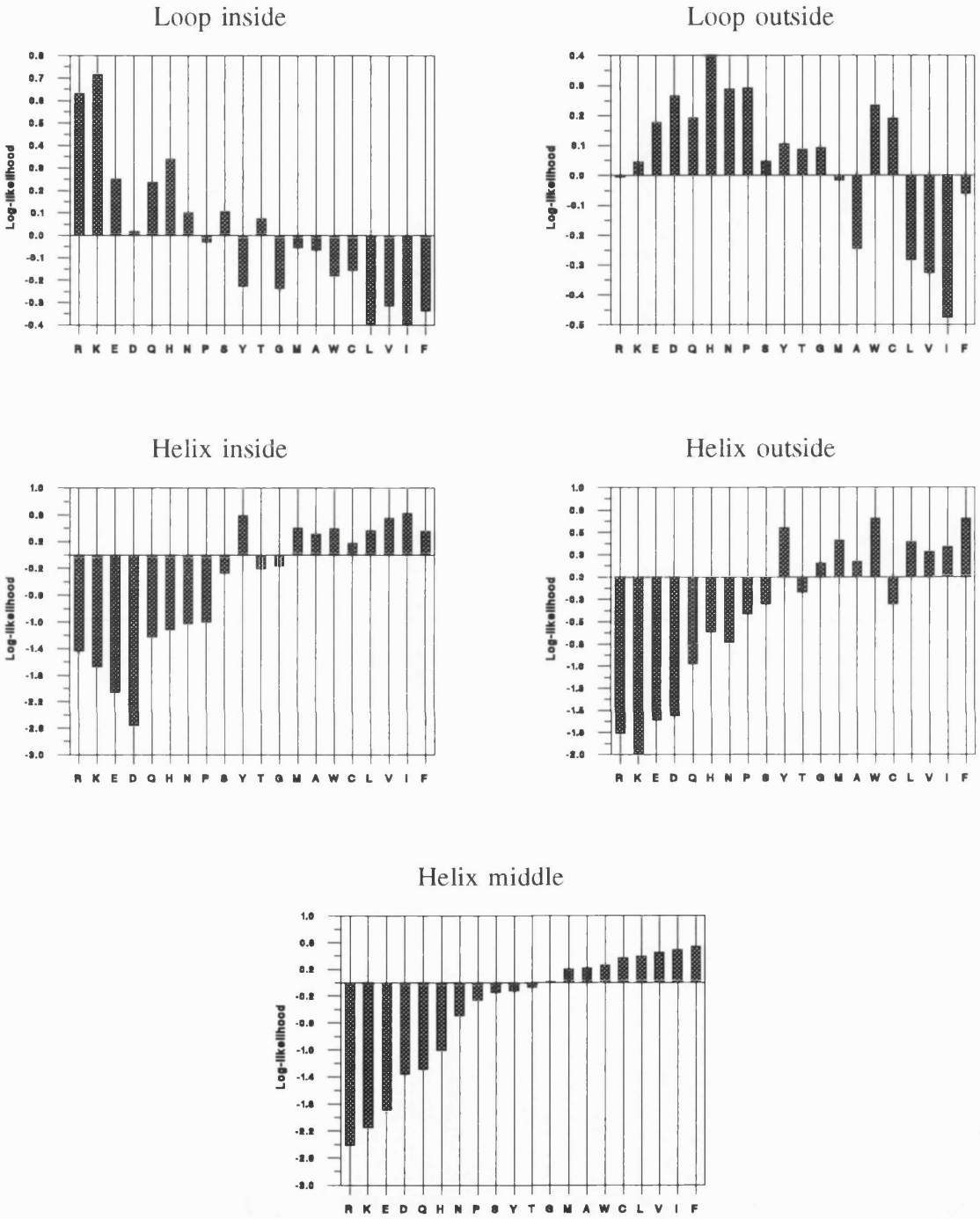


Figure 5.3

Plots of the topogenic parameters for multi-spanning segments. Amino acids are ordered by the helix middle values.

To test the effectiveness of this topogenic scoring system, the ability of these scores to predict the correct location of polar flanking regions was tested. Each documented polar flanking region was extracted from the sequence databank, and the total score calculated using both inside and outside loop log-likelihood values. Where the total inside score exceeded the outside score the region was predicted as being inside, otherwise it was predicted as being outside. The results of predicting the location of the flanking regions of a test set of proteins (see Table 5.7) are shown in Table 5.5 (multi-spanning) and Table 5.6 (single-spanning). In all cases, the protein under test was excluded from the calculation of the topogenic parameters, along with any related sequences (sequence identity > 25%).

The average score for the scheme proposed here is 73% for multi-spanning loops, and 70% for single-spanning loops[7] which compare favourably with the random expected score of 50%. Interestingly, in the case of the single-spanning loops, of the 31 loops shorter < 70 residues in length, the locations of 29 (94%) are correctly predicted, and of the 42 loops ≥ 100 residues in length, only 23 (55%) are correctly predicted. In the case of single-spanning proteins, therefore, loops of 70 residues or more contain little information regarding their location with respect to the membrane. For multi-spanning segments, it is important to note that this scheme is clearly able to predict the location of long flanking-regions as well as short regions. Whilst it is not possible to use the positive-inside rule to directly predict the location of a given flanking region, it should be noted that previous studies have shown no significant positive-bias for regions longer than 70 residues (von Heijne & Gavel, 1988). The results here clearly show little dependence on flanking region length for multi-spanning segments, yet a clear length-dependent effect for single spanning segments. Why should the topogenic parameters be so sharply defined

[7] The term 'loop' is not particularly appropriate for single spanning segments, but the term loop is again used rather loosely, in this case to indicate residues not in *transmembrane* secondary structure. In this definition, therefore, an entire globular domain which happened to be anchored to a membrane would be classfied as either an inside or outside loop. To circumvent this, loops longer than 100 residues are *not* classified as loops, and are ignored in the calculation of the q_i values. These oversized loops are, however, included in the calculation of the overall relative frequencies of occurrence p_i.

for single-spanning segments? The reason for this is no doubt due to the fact that for a single-spanning protein, almost the entire responsibility for correctly orienting the protein in the membrane resides with the short N- or C-terminal flanking segment. For multi-spanning proteins, the location of a loop depends not only on the amino acids in the loop itself, but also those in other loops. The method presented here makes use of this cooperativity, in that it is the score obtained for the whole protein with a given topology that is the basis of the prediction, rather than the score obtained for a segment taken in isolation.

Whilst the analysis by Nakashima & Nishikawa (1992) on the amino acid composition of polar flanking regions produced some similar observations to those presented here, it is hard to directly compare results due to the fact that their analysis was based on compositional preferences, and was therefore inherently limited to regions at least 50 residues in length.

Loop length	Total in data set	Number correct
0-9	34	25
10-19	81	60
20-29	32	25
30-39	29	21
40-49	7	6
50-59	2	0
60-69	3	2
70-79	7	5
80-89	4	2
90-99	0	-
>= 100	12	9

Table 5.5

Results of predicting the location of a set of multi-spanning loop segments using the multi-spanning segment topogenic parameters.

Loop length	Total in data set	Number correct
0-9	2	2
10-19	4	4
20-29	6	6
30-39	8	7
40-49	4	3
50-59	5	5
60-69	2	2
70-79	0	-
80-89	2	0
90-99	1	1
>= 100	42	23

Table 5.6

Results of predicting the location of a set of single-spanning loop segments using the single-spanning segment topogenic parameters.

Using Table 5.3 and Table 5.4 it is possible to calculate a score relating to the compatibility of a given sequence with a given topology and secondary structure. This is analogous to the protein fold recognition approaches described for globular protein folds (Jones *et al.*, 1992; Bowie *et al.*, 1991; Finkelstein & Reva, 1991), though the structural description used here is not at the full tertiary structure level. In common with the globular protein fold recognition methods, an algorithm is required that is capable of finding the optimum match between the given sequence and the given structural model. Given the simplicity of the structural models used here, at first sight it might appear feasible to use a brute-force search to identify the most likely match. For a sequence of length m, and a given transmembrane topology (n,t) there are approximately $9^n.((M - 21n)/n)^n$ possible models. Taking as an example a typical case of a 7-helix transmembranal topology and a sequence of length 250, the total number of different models that could be generated for this sequence $\approx 7 \times 10^{14}$. Clearly a brute-force approach is inappropriate.

Despite the apparent complexity of the problem, it should be noted, however, that the score for a particular residue depends solely on the identity of the residue, and its structural environment (L_i, L_o, H_i, H_m, or H_o). As a result of this single dimensionality, it is straightforward to formulate a dynamic programming solution to the problem, which will ensure that the global optimum model will be found every time.

The overall problem of determining the optimal position and length of n transmembranal helices in a sequence of length m is divided into n subproblems: namely determining the optimal position and length of a single transmembranal helix along with it's associated C-terminal coil segment. Let s^{il} be the score associated with a transmembranal helix of length l at position i in the given sequence. This score is calculated according to the diagram shown in Fig. 1, where the helix is divided into three sections (two caps of length *4*, and a centre region of length *l-8*). Whether the cap and its associated loop is inside or outside depends on the initially specified membrane topology. In order to find the best set of s_j^{il} we use a recursive algorithm almost identical to the algorithms used for pairwise

sequence alignment (Needleman & Wunsch, 1970; Sellers, 1974). A score matrix S_j^i (*i:1..n, j:1..m*) is defined thus:

$$S_j^i = \max_{l=17 \rightarrow 25} \left\{ s_j^{il} + \max_{k=i+l+A \rightarrow n} \left\{ S_{j-1}^k \right\} \right\}$$

where *A* is the minimum length of a loop segment.

Having computed the score matrix *S*, the highest value in the column *j=1* is the score for the best path through the matrix, which represents the optimal lengths and positions of *m* transmembranal helices in the given sequence. It should also be noted that the highest value in column 2 is the optimal path score for *m-1* helices, but with inverted topology, and this can be extended to the other columns. In this way, only two score matrices need to be calculated to evaluate all possible membrane topologies for a given case: one with helix 1 (column 1) defined with the N-terminus on the inside, and the other with the helix 1 N-terminus on the outside. If we calculated two matrices for *m=7*, one matrix would therefore provide optimal paths for topologies +7, -6, +5, -4, +3, -2, +1, and the other would provide paths for -7, +6, -5, +4, -3, +2, -1 (where +ve indicates N-terminus inside). A further point to note is that the raw values in *S* do not include the appropriate score for the N-terminal loop, and this must be added to the appropriate matrix values. For example, if we consider a path starting in column 1, row 5. This initial cell represents the position (j5) of the first helix, and by definition implicitly represents an N-terminal loop of length 4.

To illustrate the matrix representation, Figure 5.4 shows a portion of the final score matrix for a hypothetical protein sequence encoding 3 transmembrane segments (N-terminus inside). Greyed cells indicate cells which cannot be traversed by a valid 3 segment path, with the condition that loops are of length 3 or greater. Unlike, a traditional sequence alignment matrix path, deletions are not permitted in the horizontal 'sequence', which would represent omitted helical segments in the predicted structure. Figure 5.5 shows the interpretation of the indicated path in terms of segment positions and lengths.

5.2 Implementational details

To enable a different scoring table to be used to evaluate single-spanning topologies, these topologies are treated separately. This requires an extra two calculations of the score matrix *S*, though only for the trivial cases of a single helix in both topologies. For both multi- and single-spanning helices, the transition between helix cap and helix middle states was trapezoidally smoothed (von Heijne, 1992) as follows:

$$Score_{\substack{helix \\ cap}} = \sum_{i=1}^{4} \frac{i H_m}{5} + \frac{(5-i) H_{i/o}}{5} \quad .$$

To prevent overprediction, a filter was applied to the final topologies by means of a score cut-off on helical segments. Predicted topologies including helical segments with scores less than a predetermined cut-off were rejected from consideration. A value of 0.1 for this cut-off produced acceptable results, and was used for all the results shown in this chapter.

5.3 Program implementation

The algorithm described has been implemented in ANSI C on a number of Unix workstations. The results shown here were generated both on a Solbourne 5/602, and a Hewlett Packard 9000/710 workstation. For computational efficiency, the scores in Table 1 scaled and converted to integers. Certain aspects of the final predicted structure and topology depend on a number of control parameters, as follows: *Maxnhel* is the maximum number of transmembrane helices that may be predicted, which also denotes the number of columns in the score matrix ***S***. For a sequence of length *n* the maximum expected number of transmembrane helices is taken as *n/32*, with an upper limit of 20 helices. *Minllen* specifies the minimum length of the flanking regions (loops), and a value of 6 is generally used for this parameter. *Minhlen and maxhlen* specify the range of helix lengths that will be considered in the search, and values of 17 and 25, respectively, are used for these parameters.

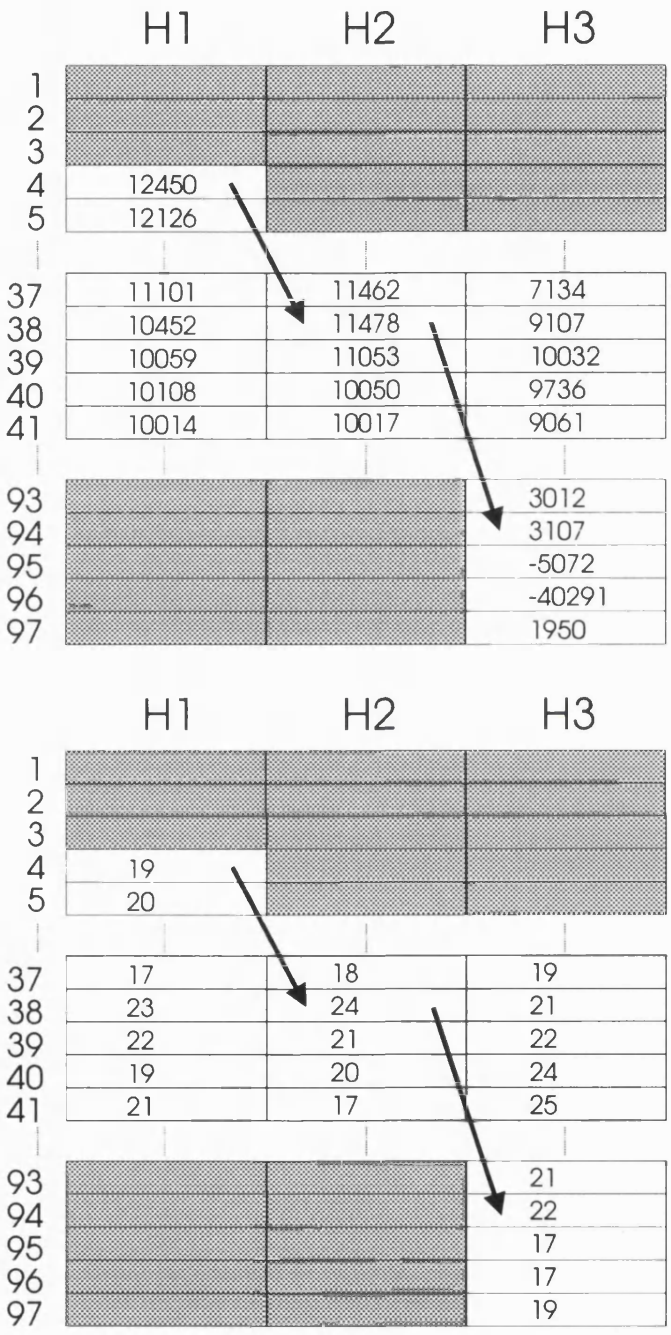


Figure 5.4

A hypothetical score matrix for 3 transmembrane helices. The upper matrix holds the highest achievable path score for each cell, and the lower matrix stores the helix length which permits this score.

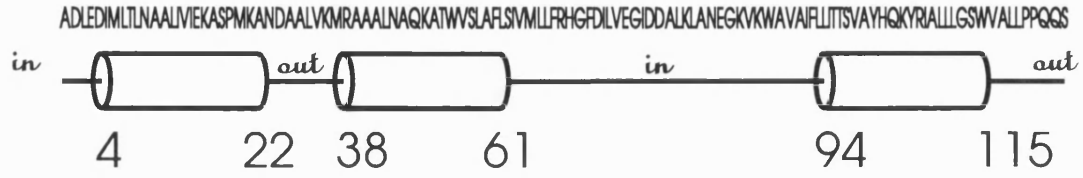


Figure 5.5

The predicted structure and topology relating to the optimal path shown in Figure 5.4.

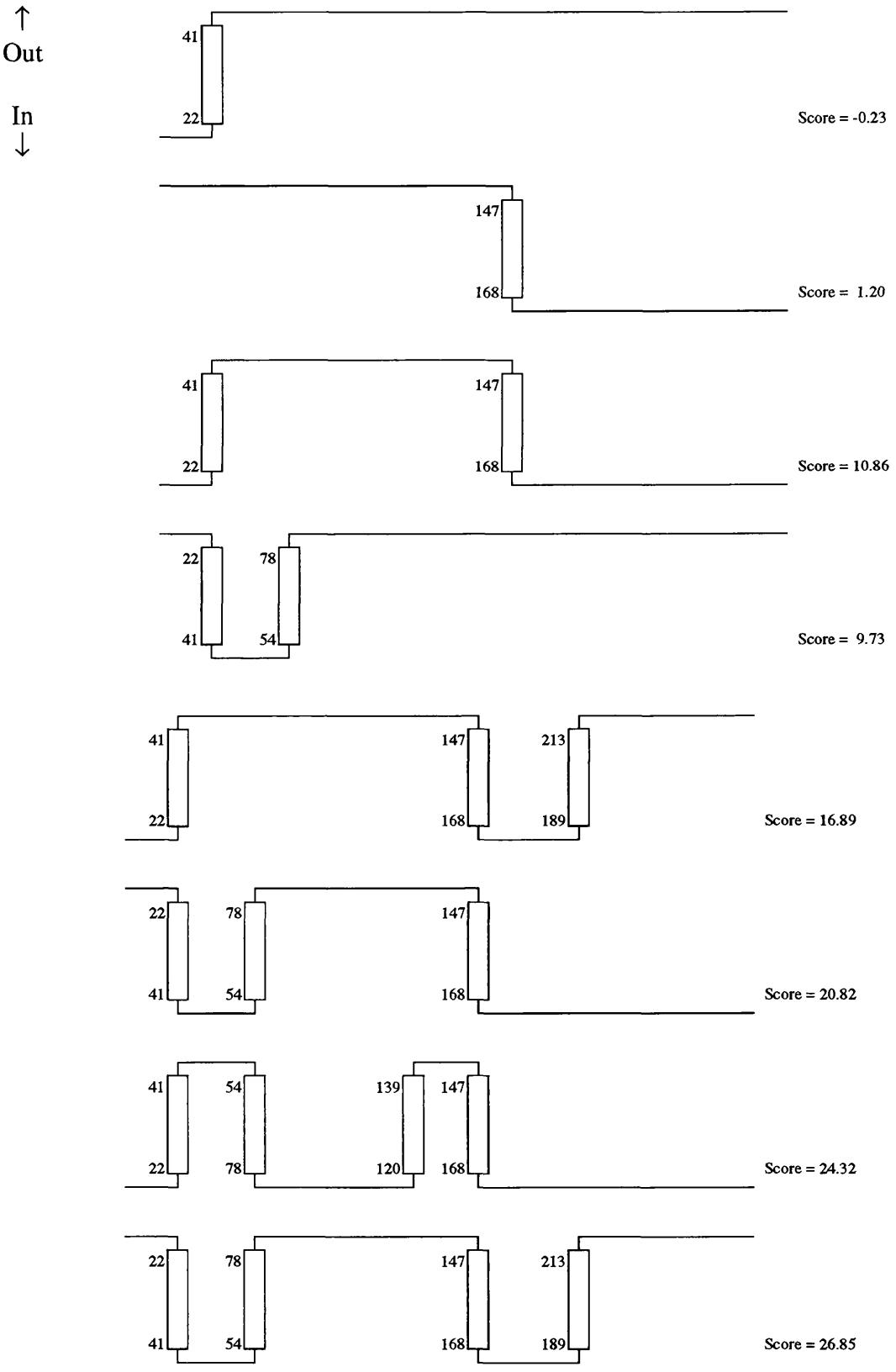
5.4 Results

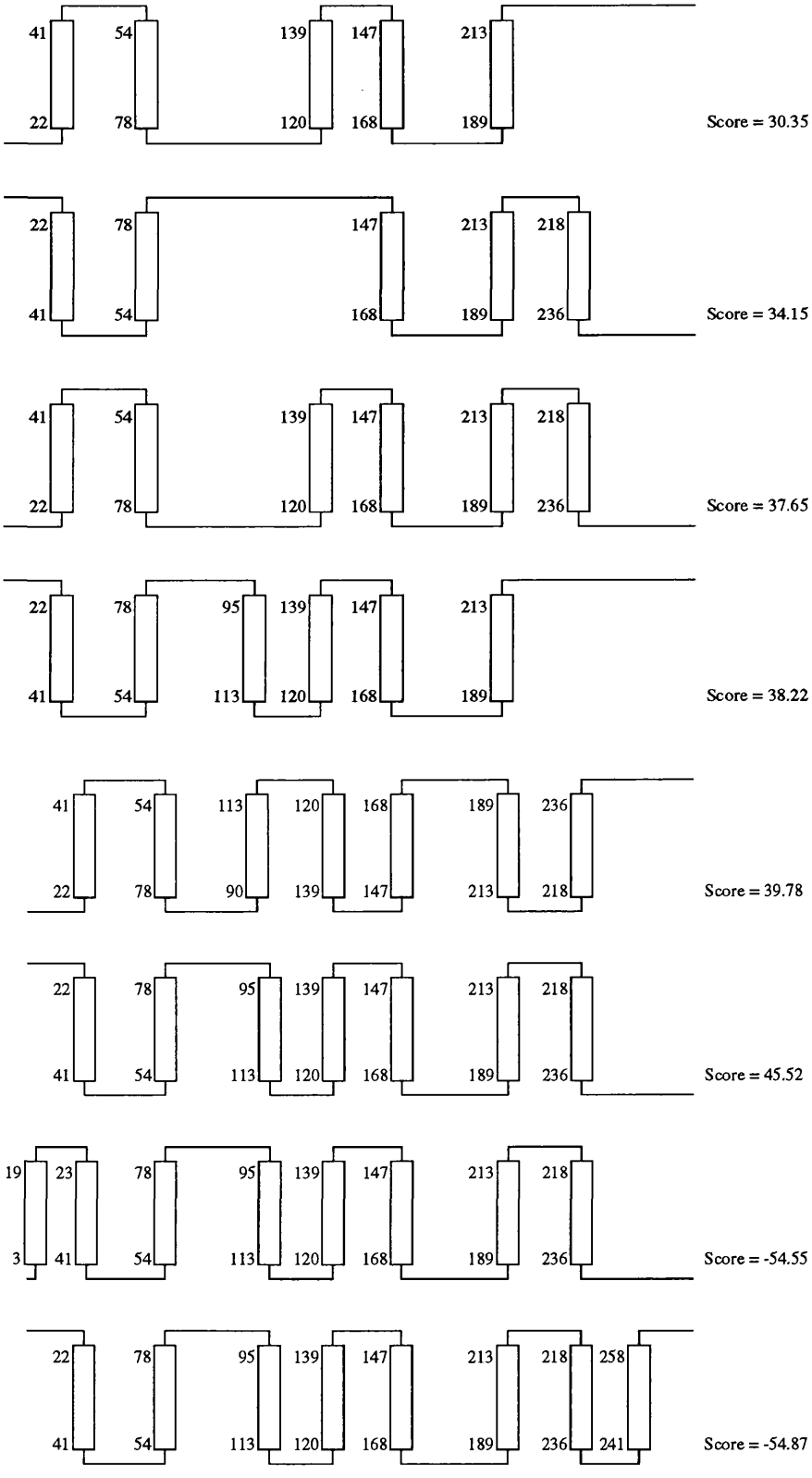
Perhaps the most important aspect of the method described here is that it provides not just a single final prediction, but a list of predictions for all achievable topologies. It is of course reasonable to pick the highest scoring prediction as the final result, but often it is important to take note of predictions whose scores are almost on a par with the optimum. Figure 5.6 shows the optimal models for all the achievable topologies of bacteriorhodopsin from *Halobacterium halobium* using the propensity values in Table 5.3 and Table 5.4. A low resolution structure for this integral membrane protein has been determined by electron microscopic techniques (Henderson *et al.*, 1990), and has been found to comprise 7 transmembrane helices, with the N-terminus located outside. The results in Figure 5.6 show that the native topology (7 helices, N-terminus outside) is clearly favoured over all others. One interesting question that arises from looking at the enumerated segment arrangements is whether the helices which are located early on in the algorithm are the helices which are assembled early on in the folding process itself. Unfortunately there is no experimental evidence available to adequately answer this question, but given the fact that the computer algorithm is keyed primarily on the hydrophobicity of sequence

segments, and that this is the principal contribution to the stability of transmembrane protein structures, then this premise is at the very least tenable.

Given that the correct topology of at least one integral membrane protein with a well-determined structure is favoured over the possible alternatives, the next step is to evaluate the method over other examples. A significant problem again arises here in that very few integral membrane proteins have a known 3-D structure. Fortunately, the membrane topology of transmembrane proteins may be determined with a moderate degree of accuracy, and comparatively quickly by chemical, immunological or genetic means (by the use of fusion proteins). The latter technique, whereby an alkaline phosphatase protein is genetically fused to the C-terminal end of part of the protein under study (Manoil & Beckwith, 1986), has provided topological information on many proteins over the past few years, and is the source of most of the topological information used here.

To evaluate the model recognition method, a set of 83 bacterial and eukaryotic integral membrane protein sequences were used. These proteins were deemed to have a reliable experimentally determined topology either from topology records in the databank entry, or from the literature (von Heijne & Gavel, 1988; von Heijne, 1992), though it must be pointed out that without a full 3-D structure, this information is only preliminary. In cases where a sequence in the databank was specified as a precursor, and the location of the leader peptide given, the leader peptide was removed before proceeding with the prediction, else the whole precursor sequence was used.


↑
Out
In
↓
41
22
Score = -0.23
147
168
Score = 1.20
41
22
147
168
Score = 10.86
22
41
78
54
Score = 9.73
41
22
147
168
213
189
Score = 16.89
22
41
78
54
147
168
Score = 20.82
41
22
54
78
139
120
147
168
Score = 24.32
22
41
78
54
147
168
213
189
Score = 26.85


41 54 139 147 213
22 78 120 168 189
Score = 30.35
22 78 147 213 218
41 54 168 189 236
Score = 34.15
41 54 139 147 213 218
22 78 120 168 189 236
Score = 37.65
22 78 95 139 147 213
41 54 113 120 168 189
Score = 38.22
41 54 113 120 168 189 236
22 78 90 139 147 213 218
Score = 39.78
22 78 95 139 147 213 218
41 54 113 120 168 189 236
Score = 45.52
19 23 78 95 139 147 213 218
3 41 54 113 120 168 189 236
Score = -54.55
22 78 95 139 147 213 218 258
41 54 113 120 168 189 236 241
Score = -54.87

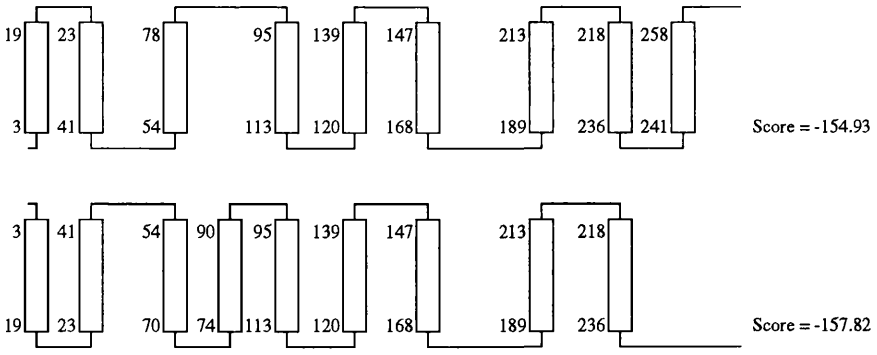


Figure 5.6

Topology schematics showing the optimal achievable topologies for bacteriorhodopsin. The scores (in units of nats) for each topology are shown alongside.

Protein	Predicted Topology	Predicted Segments	Score	Observed Topology	Observed Segments
4F2 Cell-surface antigen heavy chain (human)	✓ in	1: 82-104	5.62	in	1: 82-104
5-Hydroxytryptamine 1A receptor (human)	✓ out	1: 41-63 2: 73-96 3: 111-132 4: 154-174 5: 196-214 6: 344-368 7: 378-401	5.17 3.50 3.06 4.86 4.09 6.22 1.43	out	1: 37-62 2: 74-98 3: 110-132 4: 152-177 5: 191-216 6: 345-366 7: 378-402
5-Hydroxytryptamine 3 receptor precursor (mouse)	✓ out	1: 224-247 2: 259-276 3: 284-307 4: 438-458	4.44 2.52 4.40 3.20	out	1: 223-249 2: 255-273 3: 283-301 4: 442-461
5-Hydroxytryptamine 2 receptor (chinese hamster)	✓ out	1: 76-99 2: 112-136 3: 147-171 4: 196-215 5: 234-258 6: 324-348 7: 357-380	6.24 3.42 1.95 3.59 6.13 5.75 2.91	out	1: 76-99 2: 111-132 3: 148-171 4: 192-215 5: 234-254 6: 325-346 7: 363-384
Adenosine A1 receptor (dog)	✓ out	1: 10-34 2: 45-69 3: 81-103 4: 124-146 5: 185-207 6: 236-259 7: 268-290	3.08 4.64 4.32 5.60 4.39 4.60 1.76	out	1: 11-33 2: 47-69 3: 80-102 4: 124-146 5: 177-201 6: 236-259 7: 268-292
Adenosine A2 receptor (dog)	*x* out	1: 14-33 2: 43-67 3: 78-100 4: 123-143 5: 182-204 6: 235-258	4.46 3.92 4.75 4.89 4.28 5.10	out	1: 8-30 2: 44-66 3: 78-100 4: 121-143 5: 174-198 6: 235-258 7: 267-290
ADP,ATP carrier protein (ricpr)	✓ in	1: 28-45 2: 62-82 3: 93-113 4: 146-168 5: 183-206 6: 219-237 7: 279-297 8: 324-341 9: 349-370 10: 378-396 11: 443-459 12: 466-482	2.45 2.43 5.55 2.27 4.19 5.05 3.88 2.22 4.96 2.98 2.02 5.92	in	1: 35-55 2: 69-89 3: 94-114 4: 149-169 5: 186-206 6: 220-240 7: 281-301 8: 322-342 9: 350-370 10: 381-401 11: 440-460 12: 467-487

Protein	Predicted Topology	Predicted Segments	Score	Observed Topology	Observed Segments
Alpha-1A adrenergic receptor (human)	✓ out	1: 57-81 2: 91-114 3: 129-150 4: 171-193 5: 212-233 6: 308-332 7: 344-360	5.51 2.52 3.45 5.84 4.28 6.35 1.62	out	1: 54-79 2: 92-117 3: 128-150 4: 172-196 5: 210-233 6: 307-331 7: 339-363
Alpha-2A adrenergic receptor (subtype C10 - human)	✓ out	1: 34-58 2: 70-92 3: 108-129 4: 152-172 5: 195-217 6: 372-392 7: 410-429	5.72 3.23 2.31 4.12 6.14 6.00 1.95	out	1: 34-59 2: 71-96 3: 107-131 4: 150-173 5: 193-217 6: 375-399 7: 407-430
Alzheimer's disease amyloid A4 protein precursor (human)	✓ out	1: 683-706	6.39	out	1: 683-706
Archaerhodopsin (hals1)	✓ out	1: 17-35 2: 48-72 3: 91-107 4: 114-133 5: 143-162 6: 180-197 7: 212-230	5.17 4.71 1.71 2.75 5.47 3.92 1.23	out	1: 15-37 2: 49-74 3: 89-106 4: 113-132 5: 143-163 6: 179-198 7: 206-229
Asialoglycoprotein receptor 2 (mouse)	*x* out	1: 59-77 2: 106-125	5.18 0.42	in	1: 59-79
Asialoglycoprotein receptor 1 (human)	✓ in	1: 40-59	5.51	in	1: 40-60
Bacteriorhodopsin (halha) (see Figure 5.6)	✓ out	1: 22-41 2: 54-78 3: 95-113 4: 120-139 5: 147-168 6: 189-213 7: 218-236	4.96 4.57 1.92 2.83 4.97 2.50 1.17	out	1: 24-46 2: 52-76 3: 94-114 4: 122-141 5: 151-171 6: 181-205 7: 217-239
Blue-sensitive opsin (human)	✓ out	1: 38-60 2: 71-94 3: 111-130 4: 150-172 5: 202-220 6: 250-273 7: 284-305	4.17 5.38 3.82 4.82 6.34 4.51 0.12	out	1: 34-57 2: 71-96 3: 111-136 4: 150-173 5: 200-227 6: 250-273 7: 282-306
Cation-dependent mannose-6-phosphate receptor precursor (human)	✓ out	1: 160-184	5.34	out	1: 160-184
Chemotaxis motB protein (E. Coli)	✓ in	1: 33-49	3.47	in	1: 28-49

Protein	Predicted Topology	Predicted Segments	Score	Observed Topology	Observed Segments
Cytochrome O ubiquinol oxidase operon protein CYOE (E. Coli)	✓ in	1 : 13-29 2 : 38-56 3 : 79-103 4 : 111-127 5 : 134-154 6 : 161-183 7 : 209-225 8 : 233-249 9 : 264-281	3.25 4.12 5.23 3.92 2.30 1.87 4.09 3.41 3.82	in	1: 11-31 2: 36-56 3: 85-105 4: 107-127 5: 133-153 6: 160-180 7: 208-228 8: 231-251 9: 264-284
Cytochrome B561 (bovine)	✓ in	1: 37-59 2: 77-94 3: 107-128 4: 148-172 5: 183-201 6: 221-241	4.09 1.58 4.57 3.71 4.43 4.72	in	1: 38-60 2: 75-97 3: 107-129 4: 145-167 5: 185-207 6: 219-241
Cytochrome O ubiquinol oxidase subunit III (E. Coli)	✓ in	1 : 26-50 2 : 68-91 3 : 98-115 4 : 138-162 5 : 177-198	3.88 3.27 3.86 3.48 3.03	in	1: 33-51 2: 68-86 3: 103-121 4: 144-162 5: 186-204
Cytochrome O ubiquinol oxidase operon protein CYOD (E. Coli)	✓ in	1 : 18-37 2 : 44-66 3 : 78-100	5.37 3.77 7.06	in	1: 19-37 2: 47-65 3: 82-100
Cytochrome O ubiquinol oxidase subunit I (E. Coli)	✓ out	1: 16-38 2: 57-79 3: 107-129 4: 136-160 5: 190-213 6: 230-253 7: 287-303 8: 310-332 9: 347-371 10: 380-403 11: 423-440 12: 456-479 13: 494-514 14: 588-604 15: 611-627	6.03 2.61 4.61 3.00 3.56 4.57 2.58 5.25 2.94 4.74 3.69 5.68 6.11 2.76 5.63	out	1: 18-36 2: 59-77 3: 103-122 4: 145-163 5: 196-214 6: 233-251 7: 278-297 8: 321-340 9: 349-367 10: 383-402 11: 411-430 12: 459-477 13: 496-514 14: 589-607 15: 615-633
Cytochrome O ubiquinol oxidase subunit II (E. Coli)	✓ in	1: 10-30 2: 43-67 3: 90-108	2.74 6.33 5.72	in	1: 10-30 2: 47-67 3: 89-109
Epidermal growth factor, kidney (human)	*x* in	1 : 11-28 2 : 1011-1035	0.92 6.02	out	1: 1011-1035
Epidermal growth factor receptor (Drosophila)	*x* in	1 : 26-49 2 : 741-764	0.35 6.82	out	1: 733-764

Protein	Predicted Topology	Predicted Segments	Score	Observed Topology	Observed Segments
Epidermal growth factor receptor (human)	*x* in	1 : 412-429 2 : 619-643 3 : 753-775	0.18 5.33 1.68	out	1: 622-644
Fibronectin receptor alpha subunit (mouse)	✓ out	1: 359-381	8.26	out	1: 356-381
Gap junction beta-1 protein (human)	✓ in	1: 23-40 2: 76-96 3: 131-155 4: 190-214	2.89 2.49 1.78 4.02	in	1: 22-41 2: 75-94 3: 130-149 4: 188-207
Gap junction beta-1 protein (clawed frog)	✓ in	1: 23-40 2: 76-98 3: 132-156 4: 189-213	2.57 2.40 1.54 3.99	in	1: 22-41 2: 75-94 3: 130-149 4: 188-207
Gap junction beta-1 protein (rat)	✓ in	1: 23-40 2: 76-96 3: 131-155 4: 190-214	2.89 2.49 1.78 4.02	in	1: 22-41 2: 75-94 3: 130-149 4: 188-207
Glycophorin (pig)	✓ out	1: 63-85	6.86	out	1: 63-85
Glycophorin A precursor (human)	✓ out	1: 73-95	5.41	out	1: 73-95
Glycophorin C (human)	✓ out	1: 59-81	6.67	out	1: 58-81
Granulocyte-macrophage colony-stimulating factor receptor precursor (human)	✓ out	1: 303-324	6.22	out	1: 299-324
Green-sensitive opsin (human)	✓ out	1: 58-77 2: 90-112 3: 130-149 4: 168-191 5: 219-240 6: 269-293 7: 302-324	4.78 0.71 3.58 4.66 6.41 5.16 0.65	out	1: 53-77 2: 90-115 3: 130-156 4: 169-192 5: 219-246 6: 269-292 7: 300-325
Halorhodopsin (halsp)	✓ out	1: 7-28 2: 40-63 3: 88-106 4: 113-132 5: 141-159 6: 180-203 7: 211-229	4.41 3.98 3.77 2.88 5.31 2.80 1.12	out	1: 7-30 2: 42-65 3: 83-101 4: 112-135 5: 139-163 6: 172-195 7: 208-231
Hemagglutinin-neuraminidase (cdvo)	✓ in	1: 37-58	6.97	in	1: 35-54
Hemagglutinin-neuraminidase (PI4HA)	*x* in	1: 28-46 2: 299-321	5.85 0.90	in	1: 28-47
Hemagglutinin-neuraminidase (measles virus)	✓ in	1: 36-58	6.80	in	1: 36-54

Protein	Predicted Topology	Predicted Segments	Score	Observed Topology	Observed Segments
Hexose phosphate transport protein (E. Coli)	*x* in	1: 28-45 2: 56-80 3: 97-113 4: 120-136 5: 159-183 6: 191-210 7: 260-278 8: 327-344 9: 352-376 10: 422-446	1.53 2.15 3.46 2.77 1.47 5.27 2.64 5.17 4.53 5.38	in	1: 25-45 2: 59-79 3: 97-117 4: 118-138 5: 167-187 6: 190-210 7: 258-278 8: 299-319 9: 326-346 10: 351-371 11: 405-425 12: 427-447
HLA class II histocompatibility antigen, gamma chain precursor (human)	✓ in	1: 47-63	3.15	in	1: 47-72
Immunity protein for colicin A (E. Coli)	✓ in	1: 17-37 2: 72-89 3: 107-123 4: 147-171	7.01 1.29 2.78 4.77	in	1: 17-37 2: 72-92 3: 104-124 4: 143-163
Immunity protein for colicin E1 (E. Coli)	✓ in	1: 9-25 2: 39-57 3: 84-104	2.62 1.87 3.91	in	1: 6-26 2: 37-57 3: 90-110
Immunoglobulin G binding protein precursor (strsp)	✓ out	1: 391-409	3.91	out	1: 390-410
Inner membrane protein secE (E. Coli)	✓ in	1: 19-35 2: 45-61 3: 95-116	4.46 4.24 3.61	in	1: 20-37 2: 46-64 3: 94-112
Insulin-like growth factor I receptor precursor (human)	*x* out	1: 903-927 2: 1157-1173	7.26 1.11	out	1: 906-929
Interleukin-2 receptor alpha chain precursor (human)	✓ out	1: 220-240	5.02	out	1: 220-238
Interleukin-2 receptor beta chain precursor (human)	*x* out	1: 50-66 2: 215-239	0.35 4.54	out	1: 215-239

Protein	Predicted Topology	Predicted Segments	Score	Observed Topology	Observed Segments
Lactose permease (E. Coli)	✓ in	1: 10-34 2: 46-66 3: 75-96 4: 103-125 5: 145-162 6: 169-187 7: 222-239 8: 260-283 9: 291-313 10: 321-337 11: 349-370 12: 385-409	5.67 3.47 6.28 3.17 4.24 5.26 2.42 1.93 1.71 0.48 2.11 4.13	in	1: 8-28 2: 46-66 3: 79-99 4: 103-123 5: 146-166 6: 168-188 7: 220-240 8: 264-284 9: 292-312 10: 316-336 11: 350-370 12: 383-403
Low-affinity nerve growth factor receptor precursor (human)	*x* out	1: 195-211 2: 223-244	1.57 4.85	out	1: 221-242
Low affinity immunoglobulin epsilon FC receptor (human)	✓ in	1: 24-45	6.16	in	1: 22-47
Maltose transport inner membrane protein (E. Coli)	✓ in	1: 17-34 2: 41-58 3: 67-91 4: 282-306 5: 319-336 6: 371-392 7: 417-436 8: 484-504	4.36 3.74 5.35 6.48 3.97 3.29 1.29 6.07	in	1: 17-35 2: 40-58 3: 73-91 4: 277-295 5: 319-337 6: 371-389 7: 418-436 8: 486-504
Matrix (M2) protein (iaann)	✓ out	1: 26-43	4.44	out	1: 25-42
Melibiose carrier protein (E. Coli)	✓ in	1: 6-24 2: 31-49 3: 74-95 4: 102-126 5: 145-162 6: 176-193 7: 228-252 8: 265-281 9: 292-314 10: 321-345 11: 376-392 12: 401-425	0.43 2.75 4.48 2.63 3.02 5.35 3.34 2.48 5.42 4.50 4.51 4.89	in	1: 16-36 2: 42-63 3: 74-95 4: 110-130 5: 146-166 6: 174-194 7: 236-256 8: 263-283 9: 293-314 10: 326-346 11: 374-394 12: 401-422
Myelin-associated glycoprotein, long form precursor (mouse)	✓ out	1: 495-514	5.39	out	1: 498-517
Myelin p0 protein precursor (human)	✓ out	1: 125-149	7.30	out	1: 124-151
NB glycoprotein (inbbe)	✓ out	1: 21-44	4.17	out	1: 19-40
Neprilysin (human)	✓ in	1: 28-49	5.60	in	1: 28-50

Protein	Predicted Topology	Predicted Segments	Score	Observed Topology	Observed Segments
Oligopeptide permease protein OPPB (salty)	✓ in	1: 9-27 2: 100-121 3: 134-158 4: 173-190 5: 228-250 6: 274-298	3.59 5.39 4.93 3.74 3.38 4.41	in	1: 10-29 2: 100-121 3: 138-158 4: 173-190 5: 227-250 6: 272-293
Oligopeptide permease protein OPPC (salty)	✓ in	1: 38-59 2: 105-129 3: 140-157 4: 164-180 5: 214-236 6: 270-290	5.83 5.04 2.32 2.60 3.90 5.04	in	1: 38-59 2: 103-122 3: 140-160 4: 164-180 5: 216-236 6: 268-290
Opsin RH3 (Drosophila)	✓ out	1: 59-83 2: 95-116 3: 132-153 4: 172-189 5: 220-241 6: 285-309 7: 321-340	4.79 1.04 3.26 3.21 5.36 4.72 0.53	out	1: 63-83 2: 96-115 3: 131-151 4: 172-192 5: 220-240 6: 289-309 7: 320-340
Opsin RH2 (Drosophila)	*x* out	1: 59-82 2: 94-116 3: 132-153 4: 177-195 5: 227-250 6: 284-308	5.37 0.58 4.09 4.44 4.37 4.37	out	1: 57-81 2: 94-119 3: 134-160 4: 173-196 5: 221-248 6: 284-307 7: 315-339
Opsin RH4 (Drosophila)	✓ out	1: 55-79 2: 91-110 3: 128-149 4: 168-185 5: 216-237 6: 281-305 7: 317-336	5.26 0.60 1.97 3.12 5.76 4.76 0.64	out	1: 59-79 2: 92-111 3: 127-147 4: 168-188 5: 216-236 6: 285-305 7: 316-336
Opsin RH1 (blow fly)	✓ out	1: 50-73 2: 83-99 3: 123-144 4: 168-186 5: 215-239 6: 275-299 7: 310-329	4.37 1.44 3.47 4.26 5.09 3.50 0.29	out	1: 48-72 2: 85-110 3: 126-151 4: 164-187 5: 212-239 6: 275-298 7: 306-330

Protein	Predicted Topology	Predicted Segments	Score	Observed Topology	Observed Segments
Phosphotransferase enzyme II, mannitol specific (E. Coli)	*x* in	1: 19-43 2: 51-67 3: 79-103 4: 133-154 5: 161-181 6: 243-260 7: 269-292 8: 311-334	4.41 2.74 2.90 5.51 1.17 0.66 5.43 4.74	in	1: 25-44 2: 51-69 3: 135-154 4: 166-184 5: 274-291 6: 314-333
Photosynthetic reaction center protein L chain (rhosh)	✓ in	1: 29-51 2: 85-102 3: 111-134 4: 174-198 5: 232-256	(6.65) (3.11) (4.40) (3.85) (6.15)	in	1: 34-57 2: 86-114 3: 117-142 4: 172-201 5: 227-253
Platelet glycoprotein IB beta chain precursor (human)	*x* in	1: 123-147	4.20	out	1: 122-146
Polymeric-immunoglobulin receptor (human)	*x* in	1: 621-643	4.65	out	1: 621-643
Probable G protein-coupled receptor EDG-1 (human)	✓ out	1 : 47-71 2 : 81-104 3 : 124-140 4 : 160-182 5 : 202-222 6 : 257-277 7 : 294-312	5.00 1.76 4.23 6.13 6.16 6.15 2.41	out	1: 47-71 2: 79-107 3: 122-140 4: 160-185 5: 202-222 6: 256-277 7: 294-314
Reaction center protein M chain (rhosh)	✓ in	1: 50-74 2: 114-130 3: 147-171 4: 203-226 5: 268-291	5.98 4.12 3.62 3.81 4.84	in	1: 54-81 2: 112-141 3: 144-169 4: 199-227 5: 261-287
Reaction center protein H chain (rhosh)	✓ out	1: 12-31	4.44	out	1: 14-32
Red-sensitive opsin (human)	✓ out	1: 58-77 2: 90-112 3: 130-149 4: 168-191 5: 219-240 6: 269-293 7: 302-324	4.22 0.77 3.58 4.40 7.02 4.61 0.12	out	1: 50-74 2: 87-112 3: 127-153 4: 166-189 5: 216-243 6: 266-289 7: 297-322
Rhodopsin (bovine)	✓ out	1: 37-61 2: 74-91 3: 114-133 4: 153-175 5: 203-223 6: 253-276 7: 286-307	5.17 2.88 2.69 4.42 6.00 6.19 0.58	out	1: 37-61 2: 74-98 3: 114-140 4: 153-173 5: 203-230 6: 253-276 7: 285-309
Ribophorin I precursor (rat)	✓ out	1: 416-433	5.42	out	1: 417-435

Protein	Predicted Topology	Predicted Segments	Score	Observed Topology	Observed Segments
secY protein (bacsu)	✓ in	1: 18-39 2: 67-87 3: 115-132 4: 149-166 5: 174-191 6: 210-234 7: 269-291 8: 310-329 9: 367-386 10: 394-410	3.49 2.88 2.87 4.51 2.77 5.60 2.97 4.51 5.41 1.97	in	1: 18-39 2: 59-80 3: 115-132 4: 148-167 5: 174-192 6: 217-234 7: 268-291 8: 310-329 9: 367-386 10: 392-410
Signal peptidase I (E. Coli)	*x* out	1: 7-23 2: 59-76 3: 86-103	2.98 2.09 1.18	out	1: 4-22 2: 58-76
Sucrase-isomaltase, intestinal (human)	*x* in	1: 11-32 2: 622-639	7.15 2.72	in	1: 13-32
T-cell receptor beta chain precursor (rabbit)	✓ out	1: 293-313	2.08	out	1: 292-313
Tetracycline resistance protein tn10 (E. Coli)	✓ in	1: 7-30 2: 43-62 3: 73-95 4: 102-119 5: 130-152 6: 161-179 7: 212-234 8: 244-266 9: 277-295 10: 302-324 11: 336-357 12: 367-388	3.81 2.78 3.70 2.80 4.20 5.00 3.09 2.85 2.92 3.13 3.64 6.82	in	1: 7-27 2: 42-62 3: 81-101 4: 102-122 5: 133-153 6: 159-179 7: 201-221 8: 240-260 9: 276-296 10: 298-318 11: 337-357 12: 369-389
Thrombomodulin precursor (human)	*x* out	1: 162-178 2: 185-202 3: 495-518	1.45 1.10 5.73	out	1: 516-539
Transferrin receptor protein (human)	*x* in	1: 66-86 2: 540-558 3: 735-751	5.96 2.50 0.24	in	1: 63-88
Tyrosine kinase receptor CEK2 precursor (chick)	*x* out	1: 346-370 2: 652-668	6.82 1.28	out	1: 346-370

Protein	Predicted Topology	Predicted Segments	Score	Observed Topology	Observed Segments
UDP-N-acetylglucosamine-dolichyl-phosphate n-acetylglucosaminephosphotransferase (crilo)	✓ out	1: 11-32 2: 59-79 3: 95-114 4: 126-142 5: 157-179 6: 186-208 7: 222-240 8: 248-268 9: 275-297 10: 379-397	3.62 6.06 5.96 2.45 3.06 0.94 4.19 2.69 0.92 3.03	out	1: 8-33 2: 59-80 3: 96-115 4: 127-146 5: 166-185 6: 196-212 7: 223-241 8: 254-270 9: 276-295 10: 380-398

Table 5.7

Results of predicting the structure and topology of 83 proteins from a mixture of organism classes. Again, in all cases, the protein under test was excluded from the calculation of the topogenic parameters, along with any related sequences (sequence identity > 25%). Topology entries indicate the location of the N-terminus; following segments thereafter alternate in/out. Correct predictions are indicated with a ✓, and incorrect predictions with a *x* symbol. The locations of some of the helices in the melibiose carrier protein, including the first two, are not experimentally determined, and it would appear that the locations predicted here are more reasonable. **Key to organisms:** ricpr - *Rickettsia prowazekii*, hals1 - *Halobacterium SP.* (strain Aus-1), halsp - *Halobacterium SP.*, halha - *Halobacterium halobium*, cdvo - Canine distemper virus, pih4a - Human parainfluenza 4A virus, strsp - *Streptococcus SP.*, iaann - Influenza a virus, inbbe - Influenza b virus, salty - *Salmonella typhimurium*, rhosh - *Rhodobacter sphaeroides*, bacsu - *Bacillus subtilis*, crilo - *Cricetulus longicaudatus*.

5.5 Discussion

The results shown in Table 5.7 demonstrate that the proposed method for recognizing membrane topology models by a process of expectation maximization is highly successful, with 65 out of 84 being correctly predicted. Most of the failures were due to overpredictions for large globular (eukaryotic) proteins with single membrane anchoring segments. Typically in these cases, buried β-strands are mistaken for membrane-spanning segments, and this is a recurrent problem in all membrane protein structure prediction methods. A few possible ways of detecting such mispredictions will be discussed later. Taking the human epidermal growth factor receptor prediction as an example, where a single spanning membrane segment is located roughly half-way along the sequence at position 618, two extra helices are predicted. One of these helices can be eliminated on the basis of a marginal helix score (0.18 nats), but the score for the other helix (1.68 nats) is reasonable.

Of the multi spanning proteins, the general trend in misprediction is towards underprediction. For example, the top-scoring topology for *E. Coli* hexose phosphate transport protein includes only 10 of the expected 12 helices. In the case of the 12 helix topologies, helix 11 only achieves a score of -0.164 nats, which is of course below the set cut-off of 0.1 nats. If this cut-off is not applied, then the highest scoring topology is found to correspond with the one which has been experimentally determined. Low scoring helices appear to be common in the larger helix bundles: helix 10 of lactose permease, for example, only has a prediction score of 0.48 nats. This suggests, perhaps, that the helix score cut-off should not be applied to topologies involving more than 6 helices, though this is not yet verified, and is in any case hard to justify. Presumably in large transmembrane helical bundles, fairly hydrophilic helices may be accommodated by means of shielding from the lipid environment by neighbouring helices.

Encouragingly, the proteins of known 3-D structure, or which have relatives of known structure, are correctly predicted by the method. In view of this, it is interesting to observe

that the structure of the opsins is not predicted with great certainty. Despite confident prediction of bacteriorhodopsin, archaerhodopsin, rhodopsin and most of the G-coupled receptors, most of the opsins and the adenosine A2 receptor, which are believed to have a 7-helix structure similar to that observed in bacteriorhodopsin, have weakly predicted final helices, and in some cases the predicted topology misses this helix completely. In the absence of firm experimental data it is quite possible that these proteins only have 6 transmembrane segments, though this is not expected.

The proposed method for membrane protein structure prediction appears to be very powerful. The most important point to note, however, is that it carefully considers all possible predictions in arriving at the final highest-expectation model, and ranks the alternative predictions alongside. As any prediction algorithm will have only a limited degree of accuracy (just over 77% in this case), it is vital that alternatives be considered when making use of the prediction results. Where, say, the top two topologies have almost indistinguishable scores, then the final prediction must be taken as being *either* topology, not just a single topology, which is the form of output from previous membrane structure prediction methods. It is important to realize that the method described here need not only make use of the topogenic parameters calculated here. Indeed any scoring system that can be encoded as a 1-D vector, can be incorporated into the expectation maximization methodology. Despite the effectiveness of the parameters described, it is certain that improvements can be made, if only by an extension to the data set used in their compilation. As more experimental data becomes available, therefore, it is hoped that the predictive power of the parameters will increase.

5.6 Future prospects

As with the globular protein fold recognition method presented in the previous chapter, several improvements to the proposed method can be envisaged. The most immediate problem that needs solution is the problem of overpredicting large globular proteins with single membrane-spanning segments. One possibility here is to calculate *directional* parameters for the scoring of helical segments, similar to those used in the GOR secondary structure prediction method (Garnier *et al.*, 1978), which would allow residue pair information to be incorporated. Including pair information will hopefully allow buried β-strands to be discriminated from membrane-spanning helices, and preliminary experiments seem to bear this out.

Another important development would be to extend the simple in-out helical bundle model to encompass other structural elements observed in membrane-associated proteins. Ion-channel proteins, for example, include highly amphipathic helices in their overall topology, which will not score well with parameters biased towards strictly lipophilic helices. A rather more distant prospect is the consideration of membrane-spanning β-structure. As mentioned earlier, the sole example of an integral membrane protein containing β-structure is porin, and in this case the structure contains only β-strands, formed into a single β-barrel. It is not yet known whether it is possible for an αβ protein to integrate into a membrane. Without more information as to the likelihood of finding β-strands in an integral membrane protein structure, it is not possible to attempt to extend the recognition models along these lines. Clearly if it proves to be the case that both helices and strands can be found in almost any mixture in membrane proteins, then it will be necessary to alter the method such that only *observed* topologies are considered, as is the case for globular protein fold recognition.

The simplest and perhaps most powerful way to enhance the model recognition method is again the use of *multiply aligned sequence families*. Rather than attempting to predict the optimal topology for a single sequence, it is clear that better discrimination will be

achieved by summing the parameters over an aligned block of sequences. Early results along these lines are very promising.

References

Abarbanel, R.M., Wieneke, P.R., Mansfield, E., Jaffe, D.A. & Brutlag, D.L. (1984) *Nucleic Acids Res.* **12**, 263-280.

Adams, M.J., Ford, G.C., Koekoek, R., Lentz, P.J. jr, McPherson, A. jr, Rossmann, M.G., Smiley, I.E., Schevitz, R.W. & Wonacott, A.J. (1970) *Nature* **227**, 1098-1103.

Altschul, S.F., Gish W., Miller W., Myers E.W. & Lipman D.J. (1990) *J. Mol. Biol.* **214**, 403-410.

Anderson, C.M., Stenkamp, R.E. & Steitz, T.A. (1978) *J. Mol. Biol.* **123**, 15-33.

Anfinsen, C.B., Haber, E., Sea, M. & White, F.H. (1961) *Proc. Natl. Acad. Sci. U.S.A.* **47**, 1309-1314.

Argos, P., Rao, J.K.M. & Hargrave, P.A. (1982) *Eur. J. Biochem.* **128**, 565-575.

Arutiunian, E.G., Kuranova, I.P., Vainshtein, B.K. & Steigemann (1980) *W. Sov. Phys. Crystallogr.* **25**, 43.

Bairoch, A. (1991) *Nucleic Acids Res.* **19**, 2241-2245.

Bairoch, A. & Boeckmann, B. (1991) *Nucleic Acids Res.* **19**, 2247-2249.

Banner, D.W., Bloomer, A.C., Petsko, G.A., Phillips, D.C., Pogson, C.I., Wilson, I.A., Corran, P.H., Furth, A.J., Milman, J.D., Offord, R.E., Priddle, J.D. & Waley, S.G. (1975) *Nature* **255**, 609-614.

Barker, W.C., Hunt, L.T., Seibel-Ross, E., Yeh, L.-S. & George, D.G. (1987) *Fed. Proc.* **46**, 2232.

Barker, W.C., Hunt, L.T., George, D.G. (1988) *Protein Seq. Data Anal.* **1**, 363-373.

Barker, W.C., George, D.G., Mewes, H.-W. & Tsugita, A. (1992) *Nucleic Acids Res.* **20**, 2023-2026.

Barton, G.J. (1990) *Meth. Enzymol.* **188**, 403-428.

Bashford, D., Chothia, C. & Lesk, A.M. (1987) *J. Mol. Biol.* **196**, 199-216.

Baumann, G., Frommel, C. & Sander, C. (1989) *Protein Engineering* **2**, 329-334.

Benner, S.A. & Gerloff, D. (1991) *Adv. Enz. Reg.* **31**, 121-181.

Benner, S.A., Cohen, M.A. & Gerloff, D. (1992) *Nature* **359**, 781.

Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, E.F., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977) *J. Mol. Biol.* **112**, 535-542.

Bilofsky, H.S. & Burks, C. (1988) *Nucleic Acids Res.* **16**, 1861-1863.

Bjorkman, P.J., Saper, M.A., Samraoui, B., Bennett, W.S., Strominger, J.L. & Wiley, D.C. (1987) *Nature* **329**, 506-512.

Blundell, T.L., Carney, D., Gardner, S., Hayes, F., Howlin, B., Hubbard, T.J.P., Overington, J., Singh, D.A., Sibanda, B.L. & Sutcliffe, M. (1988) *Eur. J. Biochem.* **172**, 513-520.

Bode, W. & Schwager, P. (1975) *J. Mol. Biol.* **98**, 693-717.

Bolognesi, M., Onesti, S., Gatti, G., Coda, A., Ascenzi, P. & Brunori, M. (1989) *J. Mol. Biol.* **205**, 529-544.

Bork, P., Sander, C. & Valencia, A. (1992) *Proc. Natl. Acad. Sci. U.S.A.* **89**, 7290-7294.

Bouzida, D., Kumar, S. & Swendsen, R.H. (1992) *Phys. Rev. [A]* **45**, 8894-8901.

Bowie, J.U., Clarke, N.D., Pabo, C.O. & Sauer, R.T. (1990) *Proteins* **7**, 257-264.

Bowie, J.U., Lüthy, R. & Eisenberg, D. (1991) *Science* 253, 164-170.

Brooks, B., Bruccoleri, R.E., Olafson, B.D., States, D.J., Swaminathan, S. & Karplus, M. (1983) *J. Comp. Chem.* **4**, 187-217.

Busetta B. & Barrans Y. (1984) *Biochim. Biophys. Acta* **790**, 117-124.

Casari, G. & Sippl, M.J. (1992) *J. Mol. Biol.* **224**, 725-732.

Chan, H.S. & Dill, K.A. (1990) *Proc. Natl. Acad. Sci. U.S.A.* **87**, 6388-6392.

Chothia, C. (1976) *J. Mol. Biol.* **105**, 1-14.

Chothia, C. (1992) *Nature* **357**, 543-544.

Chou, P.Y. & Fasman, G.D. (1974) *Biochemistry* **13**, 212-245.

Cohen, F.E., Sternberg, M.J.E. & Taylor, W.R. (1980) *Nature* **285**, 378-382.

Cohen, F.E., Sternberg, M.J.E. & Taylor, W.R. (1982) *J. Mol. Biol.* **156**, 821-862.

Cohen, F.E., Abarbanel, R.M., Kuntz, I.D. & Fletterick, R.J. (1983) *Biochemistry* **22**, 4894-4904.

Cohen, F.E., Abarbanel, R.M., Kuntz, I.D. & Fletterick, R.J. (1986) *Biochemistry* **25**, 266-275.

Cornette, J.L., Cease, K.B., Margalit, H., Spouge, J.L., Berzofsky, J.A. & DeLisi, C. (1987) *J. Mol. Biol.* **195**, 659-685.

Crippen, G.M. (1991) *Biochemistry* **30**, 4232-4237.

Czelusniak, J., Goodman, M., Moncrief, N.D. & Kehoe S.M. (1990) *Meth. Enzymol.* 183, 601-615.

Dayhoff, M.O. (1968) *Atlas of Protein Sequence and Structure.* Nat. Biomed. Res. Found., Washington D.C.

Dayhoff, M.O., Eck, R.V. & Park, C.M. (1972) In *Atlas of Protein Sequence and Structure.* **5**, 89-99. Nat. Biomed. Res. Found., Washington D.C.

Dayhoff, M.O., Schwartz, R.M. & Orcutt, B.C. (1978) In *Atlas of Protein Sequence and Structure.* **5 (suppl. 3)**, 345-352. Nat. Biomed. Res. Found., Washington D.C.

Deisenhofer, J., Epp, O., Miki, K., Huber, R. & Michel, H. (1985) *Nature* **318**, 618-624.

Dill, K. & Chan, H.S. (1990) *Biochemistry* **29**, 2183.

Doolittle, R.F., Feng, D.F., Johnson, M.S. & McClure, M.A. (1986) *Cold Spring Harbor Symp. Quant. Biol.* **51**, 447-455.

Doolittle, R.F. (1992) *Prot. Sci.* **1**, 191-200.

Drexler, K.E. (1981) *Proc. Natl. Acad. Sci. U.S.A.* **78**, 5275-5278.

Eisenberg, D., Schwarz, E., Komaromy, M. & Wall, R. (1984) *J. Mol. Biol.* **179**, 125-142.

Eisenberg, D. & McLachlan, A.D. (1986) *Nature* **319**, 199-203.

Engelman, D.M., Steitz, T.A. & Goldman, A. (1986) *Annu. Rev. Biophys. Biophys. Chem.* **15**, 321-353.

Farber, G.K. & Petsko, G.A. (1990) *Trends Biochem.* Sci. **15**, 228.

Fasman G.D. (1989) *Prediction of protein structure and the principles of protein conformation*, Plenum.

Fauchere, J.L. & Pliska, V.E. (1983) *Eur. J. Med. Chem.* **18**, 369-375.

Feng, D.-F., Johnson, M.S. & Doolittle, R.F. (1985) *J. Mol. Evol.* 21, 112-125.

Fields, B.A., Guss, J.M. & Freeman, H.C. (1991) *J. Mol. Biol.* **222**, 1053-1065.

Finkelstein, A.V. & Reva, B.A. (1991) *Nature* 351, 497-499.

Finzel, B.C., Clancy, L.L., Holland, D.R., Muchmore, S.W., Watenpaugh, K.D. & Einspahr, H.M. (1989) *J. Mol. Biol.* **209**, 779-791.

Fitch, W.M. & Margoliash, E. (1967) *Science* **155**, 279-284.

Flaherty, K.M., de Luca-Flaherty, C. & Mckay, D.B. (1990) *Nature* **346**, 623-628.

Flaherty, K.M., McKay, D.B., Kabsch, W. & Holmes, K.C. (1991) *Proc. Natl. Acad. Sci. U.S.A.* **88**, 5041-5045.

Flores, T.P., Moss, D.S. & Thornton, J.M. (1993) (in preparation)

French, S. & Robson, B. (1983) *J. Mol. Evol.* **19**, 171-175.

Furey, W. Jr., Wang, B.C., Yoo, C.S. & Sax, M.F. (1983) *J. Mol. Biol.* **167**, 661-692.

Gamblin, S.J., Davies, G.J., Grimes, J.M., Jackson, R.M., Littlechild, J.A. & Watson, H.C. (1991) *J. Mol. Biol.* **219**, 573-576.

Garnier, J., Osguthorpe, D.J. & Robson, B. (1978) *J. Mol. Biol.* **120**, 97-120.

Garrett, T.P.J., Clingeleffer, D.J., Guss, J.M., Rogers, S.J. & Freeman, H.C. (1984) *J. Biol. Chem.* **259**, 2822-2825.

George, D.G., Barker, W.C. & Hunt, L.T. (1990) *Meth. Enzymol.* **188**, 333-351.

Gibrat, J.-F., Garnier, J., Robson, B. (1987) *J. Mol. Biol.* **198**, 425-443.

Godzik, A., Kolinski, A. & Skolnick, J. (1992) J. Mol. Biol. **227**, 227-238.

Goldberg, D.E. (1989) *Genetic Algorithms in Search, Optimization, and Machine Learning*, (Addison-Wesley, Reading, MA).

Gotoh, O. (1982) *J. Mol. Biol.* **162**, 705-708.

Grantham, R. (1974) *Science* **185**, 862-864.

Gregoret, L.M. & Cohen, F.E. (1990) *J. Mol. Biol.* **211**, 959-974.

Gregoret, L.M. & Cohen, F.E. (1991) *J. Mol. Biol.* **219**, 109-122.

Gribskov, M., McLachlan, A.D. & Eisenberg D. (1987) *Proc. Natl. Acad. Sci. U.S.A.* **84**, 4355-4358.

Gribskov, M., Lüthy, R. & Eisenberg, D. (1990) *Meth. Enzymol.* **188**, 146-159.

Hazes, B. & Dijkstra, B.W. (1988) *Prot. Engng.* **2**, 119-125.

Henderson, R., Baldwin, J.M., Ceska, T.A., Zemlin, F., Beckmann, E. & Downing, K.H. (1990) *J. Mol. Biol.* **213**, 899-929.

Hendlich, M., Lackner, P., Weitckus, S., Floeckner, H., Froschauer, R., Gottsbacher, K., Casari, G. & Sippl, M.J. (1990) *J. Mol. Biol.* **216**, 167-180.

Hendrickson, W.A., Klippenstein, G.L. & Ward, K.B. (1975) *Proc. Natl. Acad. Sci. U.S.A.* **72**, 2160-2164.

Henrick, K., Collyer, C.A., Blow, D.M. (1989) *J. Mol. Biol.* **208**, 129-157.

Hill, E., Tsernoglou, D., Webb, L. & Banaszak, L.J. (1972) *J. Mol. Biol.* **72**, 577-591.

Hodgman, T.C. (1989) *Comput. Applic. Biosci.* **5**, 1-13.

Holm, L. & Sander, C. (1992) *J. Mol. Biol.* **225**, 93-105.

Holmes, M.A. & Stenkamp, R.E. (1991) *J. Mol. Biol.* **220**, 723-737.

Holmgren, A. & Branden, C.I. (1989) *Nature* **342**, 248-251.

Hyde, C.C., Ahmed, S.A., Padlan, E.A., Miles, E.W. & Davies, D.R. (1988) *J. Biol. Chem.* **263**, 17857-17871.

Jahnig, F. (1990) *Trends Biochem. Sci.* **15**, 93-95.

Janin, J. & Chothia, C. (1985) *Meth. Enzymol.* **115**, 420-430.

Jones, D.T., Taylor, W.R. & Thornton, J.M. (1992a) *Nature* **358**, 86-89.

Jones, D.T., Taylor, W.R. & Thornton, J.M. (1992b) *Comput. Applic. Biosci.* **8**, 275-282.

Kabsch, W. & Sander, C. (1983) *Biopolymers* **22**, 2577-2637.

Kabsch, W., Mannherz, H.G., Suck, D., Pai, E.F. & Holmes, K.C. (1990) *Nature* **347**, 37-44.

Karlin, S., Altschul, S.F. (1990) *Proc. Natl. Acad. Sci. U.S.A.* **87**, 2264-2268.

King, R.D. & Sternberg, M.J.E. (1990) *J. Mol. Biol.* **216**, 441-457.

Kostrowicki, J. & Scheraga, H.A. (1992) *J. Phys. Chem.* **96**, 7442-7449.

Kraulis, P.J. (1991) *J. Appl. Cryst.* **24**, 946-950.

Kretsinger, R.H. (1980) *Crit. Rev. Biochem.* **8**, 119.

Kruskal, J.B. & Sankoff, D. (1983). *(in) Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison, Sankoff, D., Kruskal, J.B., (eds.), (Addison-Wesley, Reading)*, pp 265-310.

Kyte, J. & Doolittle, R.F. (1982) *J. Mol. Biol.* **157**, 105-132.

Lee, B. & Richards, F.M. (1971) *J. Mol. Biol.* **55**, 379-400.

Lee, C. & Subbiah, S. (1991) *J. Mol. Biol.* **217**, 373-388.

Levinthal, C. (1968) *Chim. Phys.* **65**, 44-45.

Levitt, M. (1976) *J. Mol. Biol.* **104**, 59-107.

Lifson, S. & Sander, C. (1979) *Nature* **282**, 109-111.

Lim, V.I. (1974a) *J. Mol. Biol.* **88**, 857-872.

Lim, V.I. (1974b) *J. Mol. Biol.* **88**, 873-894.

Lindqvist, Y. (1989) *J. Mol. Biol.* **209**, 151-166.

Lipman, D.J. & Pearson, W.R. (1985) *Science* **227**, 1435-1441.

Lüthy, R., McLachlan, A.D. & Eisenberg D. (1991) *Proteins* **10**, 229-239.

Lüthy, R., Bowie, J.U., Eisenberg, D. (1992) *Nature* **356**, 83-85.

Maiorov, V.N. & Crippen, G.M. (1992) *J. Mol. Biol.* **227**, 876-888.

Manoil, C. & Beckwith, J. (1986) *Science* **233**, 1403-1408.

Marquart, M., Deisenhofer, J., Huber, R. & Palm, W. (1980) *J. Mol. Biol.* **141**, 369-391.

Mathews, F.S., Bethge, P.H. & Czerwinski, E.W. (1979) *J. Biol. Chem.* **254**, 1699-1706.

McLachlan, A.D. (1971) *J. Mol. Biol.* **61**, 409-424.

McLachlan, A.D. (1972) *J. Mol. Biol.* **64**, 417-437.

Morris, A.L., MacArthur, M.W., Hutchinson, E.G. & Thornton, J.M. (1992) *Proteins* **12**, 345-364.

Moult, J. & Unger, R. (1991) *Biochemistry* **30**, 3816-3824.

Murzin, A.G. & Finkelstein, A.V. (1983) *Biofizika* **28**, 905-911.

Murzin, A.G. (1992) *Nature* **260**, 635.

Murzin, A.G. & Chothia, C. (1992) *Curr. Opin. Struct. Biol.* **2**, 895-903.

Murzin, A.G., Lesk, A.M. & Chothia, C. (1992) *J. Mol. Biol.* **223**, 531-543.

Nagano, K. (1977) *J. Mol. Biol.* **109**, 251-257.

Nagano, K. (1980) *J. Mol. Biol.* **138**, 797-832.

Nakai, K., Kidera, A. & Kanehisa, M. (1988) *Protein Eng.* **2**, 93-100.

Nakashima, H. & Nishikawa, K. (1992) *Febs Lett.* **303**, 141-146.

Needleman, S.B. & Wunsch, C.D. (1970) *J. Mol. Biol.* **48**, 443-453.

Nishikawa, K. & Ooi, T. (1986) *J. Biochem.* **100**, 1043-1047.

Norris, G.E., Anderson, B.F. & Baker, E.N. (1983) *J. Mol. Biol.* **165**, 501-521.

Novotny J., Bruccoleri R.E. & Karplus M. (1984) *J. Mol. Biol.* **177**, 787-818.

Novotny, J., Rashin, A.A. & Bruccoleri, R.E. (1988) *Proteins* **4,** 19-30.

Ollis, D.L., Cheah, E., Cygler, M., Dijkstra, B., Frolow, F., Franken, S.M., Harel, M., Remington, S.J., Silman, I., Schrag, J., Sussman, J.L., Verschueren, K.H.G., Goldman, A. (1992) *Prot. Engng.* **5**, 197-211.

Onesti, S., Brick, P. & Blow, D.M. (1991) *J. Mol. Biol.* **217**, 153-176.

O'shea, E.K., Klemm, J.D., Kim, P.S. & Alber, T. (1991) *Science* **254**, 539-544.

Orengo, C.A. & Taylor, W.R. (1990) *J. Theor. Biol.* **147**, 517-551.

Orengo, C.A. & Taylor, W.R. (1992) *J. Mol. Biol.* (accepted)

Orengo, C.A., Brown, N.P. & Taylor, W.R. (1992a) *Proteins* **14**, 139-167.

Orengo, C.A., Flores, T.P. & Taylor, W.R., Thornton, J.M. (1993) *Protein Engineering* (accepted)

Overington, J., Johnson, M.S., Šali, A. & Blundell, T.L. (1990) *Proc. R. Soc. Lond. B.* 241, 132-145.

Overington, J., Donnelly, D., Johnson, M.S., Šali, A. & Blundell, T.L. (1992) *Protein Sci.* **1**, 216-226.

Palau, J. & Puigdomenech, P. (1974) *J. Mol. Biol.* **88**, 457-469.

Pastore, A. & Lesk, A.M. (1990) *Proteins* **8**, 133-155.

Pearl, L.H. & Taylor, W.R. (1987) *Nature* **328**, 351-354.

Pearson, W.R. (1990) *Meth. Enzymol.* **188**, 63-98.

Petratos, K., Banner, D.W., Beppu, T., Wilson, K.S. & Tsernoglou, D. (1987) *Febs Lett.* **218**, 209-214.

Phillips, S.E.V. (1980) *J. Mol. Biol.* **142**, 531-554.

Ploegman, J.H., Drent, G., Kalk, K.H. & Hol, W.G.J. (1978) *J. Mol. Biol.* **123**, 557-594.

Ponder, J.W. & Richards, F.M. (1987) *J. Mol. Biol.* **193**, 775-791.

Presnell, S.R. & Cohen, F.E. (1989) *Proc. Natl. Acad. Sci. U.S.A.* **86**, 6592-6596.

Ramachandran, G.N. & Sassiekharan, V. (1968) *Adv. Prot. Chem.* **28**, 283-437.

Rao, J.K.M. & Argos, P. (1986) *Biochim. Biophys. Acta* **869**, 197-214.

Richmond, T.J. & Richards, F.M. (1978) *1978* **119**, 537-555.

Risler, J.L., Delorme, M.O., Delacroix, H. & Henaut, A. (1988) *J. Mol. Biol.* 210, 181-193.

Robson, B. & Osguthorpe, D.J. (1979) *J. Mol. Biol.* **132**, 19-51.

Robson B. & Garnier J. (1993) *Nature* **361**, 506-506.

Rondeau, J.M., TeteFavier, F., Podjarny, A., Reymann, J.M., Barth, P., Biellmann, J.F. & Moras, D. (1992) *Nature* **355**, 469-472.

Rooman, M.J. & Wodak, S.J. (1988) *Nature* **335**, 45-49.

Rooman, M.J., Wodak, S.J. & Thornton, J.M. (1989) *Protein Engineering* **3**, 23-27.

Rooman, M.J., Rodriguez J. & Wodak S.J. (1990) *J. Mol. Biol.* **213**, 337-350.

Rooman, M.J., Kocher, J.P.A. & Wodak, S.J. (1991) *J. Mol. Biol.* **221**, 961-979.

Rose, G.D. (1985) *Meth. Enzymol.* **115**, 430-440.

Rossmann, M.G., Moras, D. & Olsen, K.W. (1974) *Nature* **250**, 194-199.

Ryu, S.-E., Kwong, P.D., Truneh, A., Porter, T.G., Arthos, J., Rosenberg, M., Dai, X., Xuong, N.-H., Axel, R., Sweet, R.W. & Hendrickson, W.A. (1990) *Nature* **348**, 419-426.

Sack, J.S., Saper, M.A. & Quiocho, F.A. (1989a) *J. Mol. Biol.* **206**, 171-191.

Sack, J.S., Trakhanov, S.D., Tsigannik, I.H. & Quiocho, F.A. (1989b) *J. Mol. Biol.* **206**, 193-207.

Šali, A. & Blundell, T.L. (1990) *J. Mol. Biol.* **212**, 403-428.

Sander, C. & Schneider, R. (1991) *Proteins* **9**, 56-68.

Sankoff, D. (1972) *Proc. Natl. Acad. Sci. U.S.A.* **69**, 4-6.

Schiffer, M. & Edmundson, A.B. (1968) *Biophys. J.* **8**, 29-39.

Schiffer, M. & Edmunson, A.B. (1967) *Biophys. J.* **7**, 121-135.

Schiffer, M., Chang, C.-H., Stevens, F.J. (1992) *Prot. Engng.* **5**, 213-214.

Schirmer, T., Bode, W., Huber, R., Sidler, W. & Zuber, H. (1985) *J. Mol. Biol.* **184**, 257-277.

Sellers, P.H. (1974) *J. Combinator. Theor.* **16**, 253-258.

Sellers, P.H. (1984) *Bull. Math. Biol.* **46**, 501-514.

Shrake, A. & Rupley, J.A. (1973) *J. Mol. Biol.* **79**, 351-371.

Sielecki, A.R., Hendrickson, W.A., Broughton, C.G., Delbaere, L.T.J., Brayer, G.D., James, M.N.G. (1979) *J. Mol. Biol.* **134**, 781-804.

Sippl, M.J. (1990) *J. Mol. Biol.* **213**, 859-883.

Smith, R.F. & Smith, T.F. (1990) *Proc. Natl. Acad. Sci. U.S.A.* **87**, 118-122.

Smith, T.F. & Waterman, M.S. (1981) *Adv. Appl. Math.* **2**, 482-489.

Staden, R. (1982) *Nucleic Acids Res.* **14**, 363-374.

Stirk, H.J., Thornton, J.M. & Howard, C.R. (1992) *Intervirology* **33**, 148-158.

Swindells, M.B. (1992) *Science* **258**, 1160-1161.

Taylor, W.R. & Thornton, J.M. (1983) *Nature* **301**, 540-542.

Taylor, W.R. (1986a) *J. Mol. Biol.* **188**, 233-258.

Taylor, W.R. (1986b) *J. Theor. Biol.* **119**, 205-218.

Taylor, W.R. (1987) (In) *Nucleic acid and protein sequence analysis a practical approach*, Bishop M.J., Rawlings C.J., Eds., pp359-385, IRL Press, Oxford.

Taylor, W.R. (1988a) *J. Mol. Evol.* **28**, 161-169.

Taylor, W.R. (1988b) *Protein Engineering* **2**, 77-86.

Taylor, W.R. & Orengo, C.A. (1989) *J. Mol. Biol.* **208**, 1-22.

Taylor, W.R. (1990) *Meth. Enzymol.* **188**, 456-474.

Taylor, W.R. (1991) Protein Engineering **4**, 853-870.

Taylor, W.R. & Jones, D.T. (1991) *Curr. Opin. Struct. Biol.* **1**, 327-333.

Taylor, W.R. & Jones, D.T. (1993) *J. Theor. Biol.* (accepted).

Thornton, J.M., Sibanda, B.L., Edwards, M.S. & Barlow, D.J. (1988) *BioEssays* **8**, 63-69.

Thornton, J.M., Flores, T.P., Jones, D.T. & Swindells, M.B. (1991) *Nature* **354**, 105-106.

von Heijne, G. (1981) *Eur. J. Biochem.* **120**, 275-278.

von Heijne, G. & Gavel, Y (1988) *Eur. J. Biochem.* **174**, 671-678.

von Heijne, G. (1991) *J. Mol. Biol.* **218**, 499-503.

von Heijne, G. (1992) *J. Mol. Biol.* **225**, 487-494.

Vonderviszt, F. & Simon, I. (1986) *Biochem. Biophys. Res. Commun.* **139**, 11-17.

Vyas, N.K., Vyas, M.N. & Quiocho, F.A. (1988) *Science* **242**, 1290-1295.

Walker, J.E., Saraste, M., Runswick, W.J. & Gay, N.J. (1982) *EMBO J.* **1**, 945-951.

Wang, J., Yan, Y., Garrett, T.P., Liu, J., Rodgers, D.W., Garlick, R.L., Tarr, G.E., Husain, Y., Reinherz, E.L. & Harrison, S.C. (1990) *Nature* **348**, 411-418.

Waterman, M.S. (1986) *Nucleic Acids Res.* **14**, 9095-9102.

Weber, E., Steigemann, W., Jones, T.A., Huber, R. (1978) *J. Mol. Biol.* **120**, 327-336.

Weiss, M.S. & Schulz, G.E. (1992) *J. Mol. Biol.* **227**, 493-509.

Wierenga, R.K. & Hol, W.G.J. (1983) *Nature* **302**, 842-844.

Wilbur, W.J. & Lipman, D.J. (1983) *Proc. Natl. Acad. Sci. U.S.A.* **80**, 726-730.

Wilmanns, M. & Eisenberg, D. (1993) *Proc. Natl. Acad. Sci. U.S.A.* **90**, 1379-1383.

Wilson, S.A. & Drew, R.E. (1991) *J. Bacteriol.* **173**, 4914-4921.

Wilson, S.A., Chayen, N.E., Hemmings, A.M., Drew, R.E. & Pearl, L.H. (1992) *J. Mol. Biol.* **222**, 869-871.

Wilson, S.A., Wachira, S.J., Drew, R.E., Jones, D.T. & Pearl, L.H. (1993) (submitted)

Wolfenden, R.V., Cullis, P.M. & Southgate, C.C.F. (1979) *Science* **206**, 575-577.

Woolfson, D.N., Mortishiresmith, R.J. & Williams, D.H. (1991) *Biochem. Biophys. Res. Commun.* **175**, 733-737.

Wootton, J.C. (1974) *Nature* **252**, 542-546.

Appendix A - Publications arising during the course of the project

Jones, D.T. (1991) The application of fractal clustering to efficient molecular ray-tracing on low-cost computers. *J. Mol. Graphics.* **9**, 249-253.

Taylor, W.R. & Jones, D.T. (1991) Templates, consensus patterns and motifs. *Curr. Opin. Struct. Biol.* **1**, 327-333.

Thornton, J.M., Flores, T.P., Jones, D.T. & Swindells, M.B. (1991) Prediction of progress at last. *Nature (News and Views)* **353**, 388-389.

Jones, D.T. (1992) A brief review of protein sequence pattern matching. (In) *Patterns in proteins sequence and structure.* Taylor, W.R. Ed., pp11-28, Springer-Verlag, Heidelberg.

Jones, D.T., Taylor, W.R. & Thornton, J.M. (1992) The rapid generation of mutation data atrices from protein sequences. *Comput. Applic. Biosci.* **8**, 275-282.

Jones, D.T., Taylor, W.R. & Thornton, J.M. (1992) A new approach to protein fold recognition. *Nature* **358**, 86-89.

Jones, D.T. & Thornton, J.M. (1993) Protein fold recognition. *J. Comput. Aided Mol. Des.* (in press).

Taylor, W.R. & Jones, D.T. (1993) Deriving an amino acid distance matrix. *J. Theor. Biol.* (accepted).

Jones, D.T., Taylor, W.R. & Thornton, J.M. (1993) A Mutation Data Matrix for Transmembrane Proteins. *J. Mol. Biol.* (submitted).

Jones, D.T., Taylor, W.R. & Thornton, J.M. (1993) A model recognition approach to membrane protein topology prediction. (in preparation).

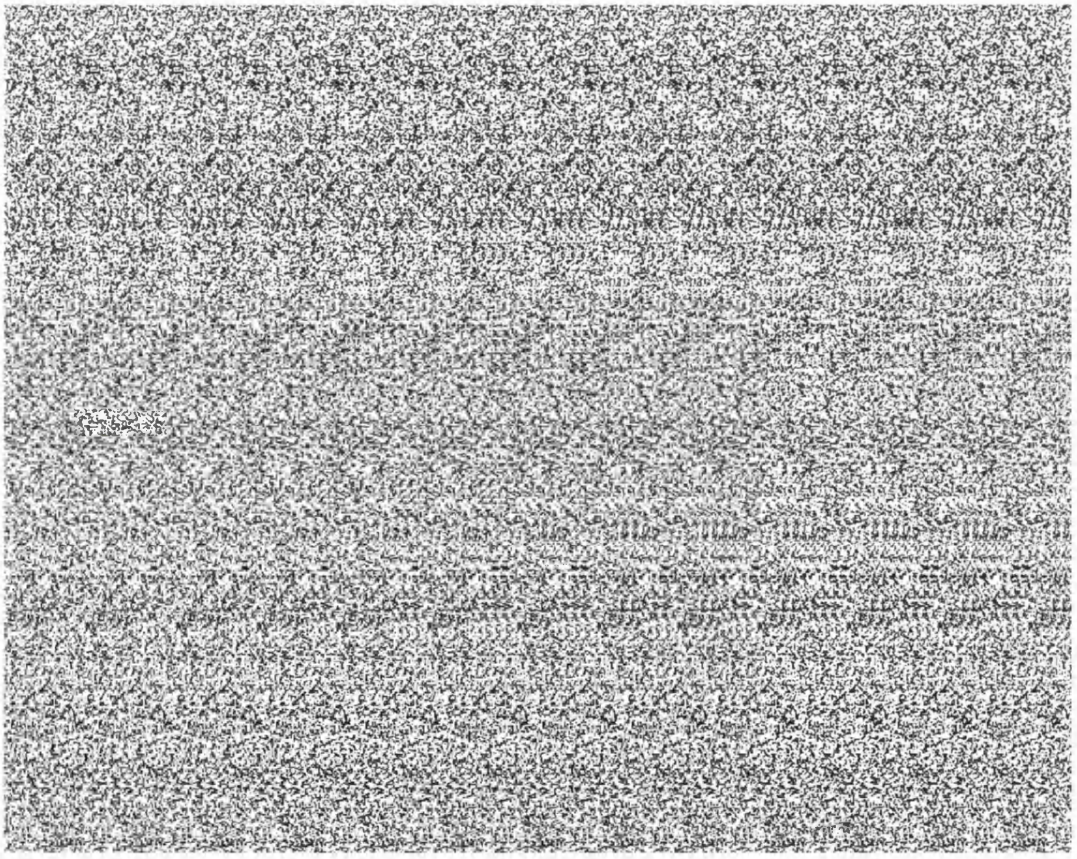
Appendix B - Computer Programs Used

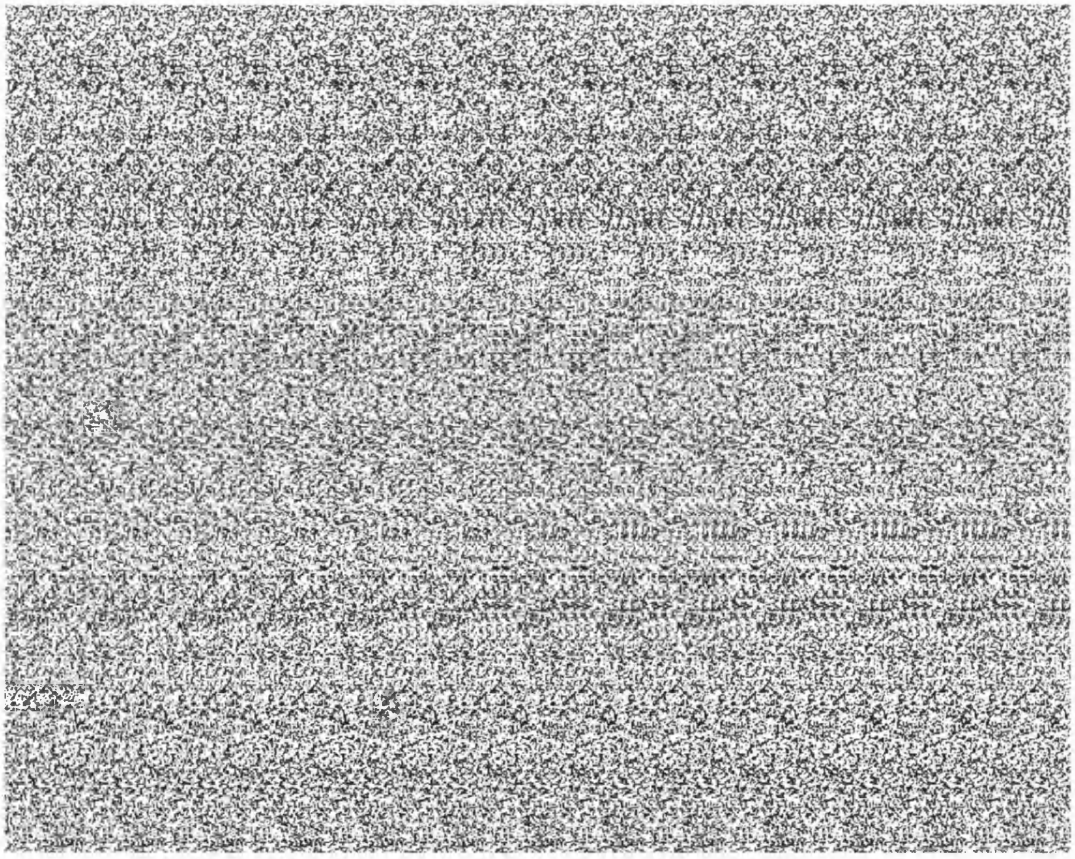
Name	Author(s)	Source	Description
ANNTHRD	D. Jones	1	Simulated annealing threading program
DSSP	W. Kabsch & C. Sander	2	Calculates protein secondary structure
GENSTATS	D. Jones	1	Compiles potentials of mean force
MAKEPET	D. Jones	1	Computes mutation data matrices
MEMSAT	D. Jones	1	Membrane structure and topology prediction
MEMSTATS	D. Jones	1	Compiles topogenic parameters
MOLSCRIPT	P. Kraulis	3	Draws protein ribbon cartoons
RANDTHRD	D. Jones	1	Exhaustive threading program
SEQGREP	D. Jones	1	Compiles sequence lists by keyword search
SSAP	W. Taylor & C. Orengo	4	Structurally aligns proteins
SUPRMS	F. Rippmann	4	Superposes protein structures
THREAD	D. Jones	1	Double dynamic programming threading
TOPS	T. Flores	5	Draws protein topology schematics

Sources

1. Source code for these programs is freely available to academic users from the author via e-mail (mail jones@bsm.bioc.ucl.ac.uk).
2. Available freely to academic users via e-mail from Dr Chris Sander (mail sander@embl-heidelberg.de).
3. Available by post from Dr Per Kraulis. For details, write to Dr Per Kraulis, Department of Biochemistry, Cambridge University, Tennis Court Road, Cambridge, U.K.
4. These programs are freely available to academic users from Dr Christine Orengo via e-mail (mail orengo@bsm.bioc.ucl.ac.uk).
5. E-mail Dr Tom Flores for details on availability (mail t-flores@nimr.mrc.ac.uk).

Index





A new approach to protein fold recognition

D. T. Jones*†, W. R. Taylor† & J. M. Thornton*

* Biomolecular Structure and Modelling Unit,
Department of Biochemistry and Molecular Biology,
University College, Gower Street,
London WC1E 6BT, UK
† Laboratory of Mathematical Biology, National Institute for Medical Research,
The Ridgeway, Mill Hill, London, NW7 1AA, UK

THE prediction of protein tertiary structure from sequence using molecular energy calculations has not yet been successful; an alternative strategy of recognizing known motifs[1] or folds[2–4] in sequences looks more promising. We present here a new approach to fold recognition, whereby sequences are fitted directly onto the backbone coordinates of known protein structures. Our method for protein fold recognition involves automatic modelling of protein structures using a given sequence, and is based on the frameworks of known protein folds. The plausibility of each model, and hence the degree of compatibility between the sequence and the proposed structure, is evaluated by means of a set of empirical potentials derived from proteins of known structure. The novel aspect of our approach is that the matching of sequences to backbone coordinates is performed in full three-dimensional space, incorporating specific pair interactions explicitly.

In outline our method is simple. A library of different protein folds is derived from the database of protein structures. In our case, the library contained all the unique, moderately well resolved chains (sequence identity < 30%, resolution ≤ 2.8 Å) in the July 1991 release of the Brookhaven database[5], totalling 102 chains. Each fold is considered as a chain tracing through space; the original sequence being ignored completely. The test sequence is then optimally fitted to each library fold (allowing for relative insertions and deletions in loop regions), with the 'energy' of each possible fit (or threading) being calculated by summing the proposed pairwise interactions. The library of folds is then ranked in ascending order of total energy, with the lowest energy fold being taken as the most probable match.

In previous work, the difficult problem of optimizing the threading of the test sequence onto a structure with respect to the detailed pairwise interactions has been avoided by matching at the sequence level. At its most basic, this involves matching the sequence of the given fold with the test sequence, scoring each residue-residue match by means of a score matrix such as the Dayhoff matrix[6]. But there are now many examples of proteins exhibiting high structural similarity yet little or no similarity in their sequences (sequence identity < 15%). In view of this, several groups have attempted to match sequences to folds by describing the fold not in terms of its amino-acid sequence, but in terms of the environment of each residue location in the structure[2–4]. The environment (for example, the local secondary structure and solvent accessibility) of a particular residue tends to be more highly conserved than the identity of the residue itself, and so methods that match each residue in the test sequence to the environments of each residue in a protein fold are able to detect more distant sequence-structure relationships than purely sequence-based methods. Finkelstein and Reva attempted to take pairwise interactions into account by addressing the problem of fitting a sequence onto idealized lattice models of 8-stranded β-sandwich folds using an iterative procedure[3,7]. We have used a dynamic programming-based algorithm[8,9] capable of optimizing pairwise interactions, which was originally applied to the problem of structural comparison. This algorithm uses a standard sequence alignment method to optimize the threading of the sequence onto the structure around each residue in turn, finally computing the best threading through the whole structure by means of a shortest-path algorithm.

To evaluate the energy of a sequence in a particular conformation we need a set of potentials for residue interactions that do not require explicit modelling of all side-chain atoms. Previous work[10] has shown that classical potentials (for example, CHARMm[11]) cannot identify proteins that have been folded into non-native conformations. For these reasons we use a set of knowledge-based potentials and explicitly consider the degree of residue solvation, both of which do in fact identify such misfolded proteins[12,13]. In particular, we use a set of pairwise potentials similar to those described by Sippl[14] which are derived from a statistical analysis of known protein structures (see Fig. 1 legend for details). For a given pair of atoms, a given residue sequence separation and a given interaction distance, these potentials provide a measure of pseudo-energy, which relates to the probability of observing the proposed interaction in native protein structures. By dividing the empirical 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 α-helical conformation and evaluating the short-range potential terms, the probability of the sequence folding into an α-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. Some sample potentials are shown in Fig. 1*a–d*.

Our medium and long-range pairwise potentials differ from those proposed in ref. 14 in that interactions beyond 10 Å are ignored. These interactions are not residue-specific and are determined simply by solvation effects. In place of these long-distance terms, we substitute a 'solvation potential'. This potential measures the propensity of each amino-acid type for a certain degree of solvation, approximated by the residue solvent-accessible surface area.

Many studies have shown that only the cores of distantly related structures are conserved, therefore in calculating the energy of a given threading we ignore all pairwise terms involving loop residues. Loop positions are evaluated by the solvation potential alone, which takes into account the tendency for loop regions to be exposed to solvent.

An obvious first test of these potentials was to attempt to thread a sequence onto its own native structure. Taking a number of small structures, for which it was practical to evaluate every possible threading (disallowing gaps in regions of regular secondary structure), we have found that the native threading of a sequence onto its own structure is usually found to be the lowest energy threading. As an example, the native threading histogram for the C-terminal ribosomal protein fragment (CTF) is shown in Fig. 1*e*. One small protein for which the native threading does not have the lowest evaluated energy is crambin, but this is attributable to the fact that this protein is not soluble in water, and consequently the solvation effects are not correctly modelled by our solvation potential.

To demonstrate the capability of our method for recognizing protein folds and generating an accurate sequence-structure alignment, we consider here the example of C-phycocyanin. The striking feature of the chain fold of C-phycocyanin is that the globular portion (helices A–H) closely resembles the globin fold[15]. Despite the similarity in fold, the sequence homology between the globins and C-phycocyanin is very low, with only 14 identities between the 174 β-chain residues of C-phycocyanin and sperm whale myoglobin. So far, sequence analysis methods have proved unable to detect the globin fold in C-phycocyanin. For example, despite success in constructing templates to select almost every available globin sequence, it has not been possible to match these templates against the phycocyanins[16]. Using the optimal threading algorithm, the C-phycocyanin sequence was threaded on each member of the library of protein folds to find its most compatible fold. The two lowest-energy folds were

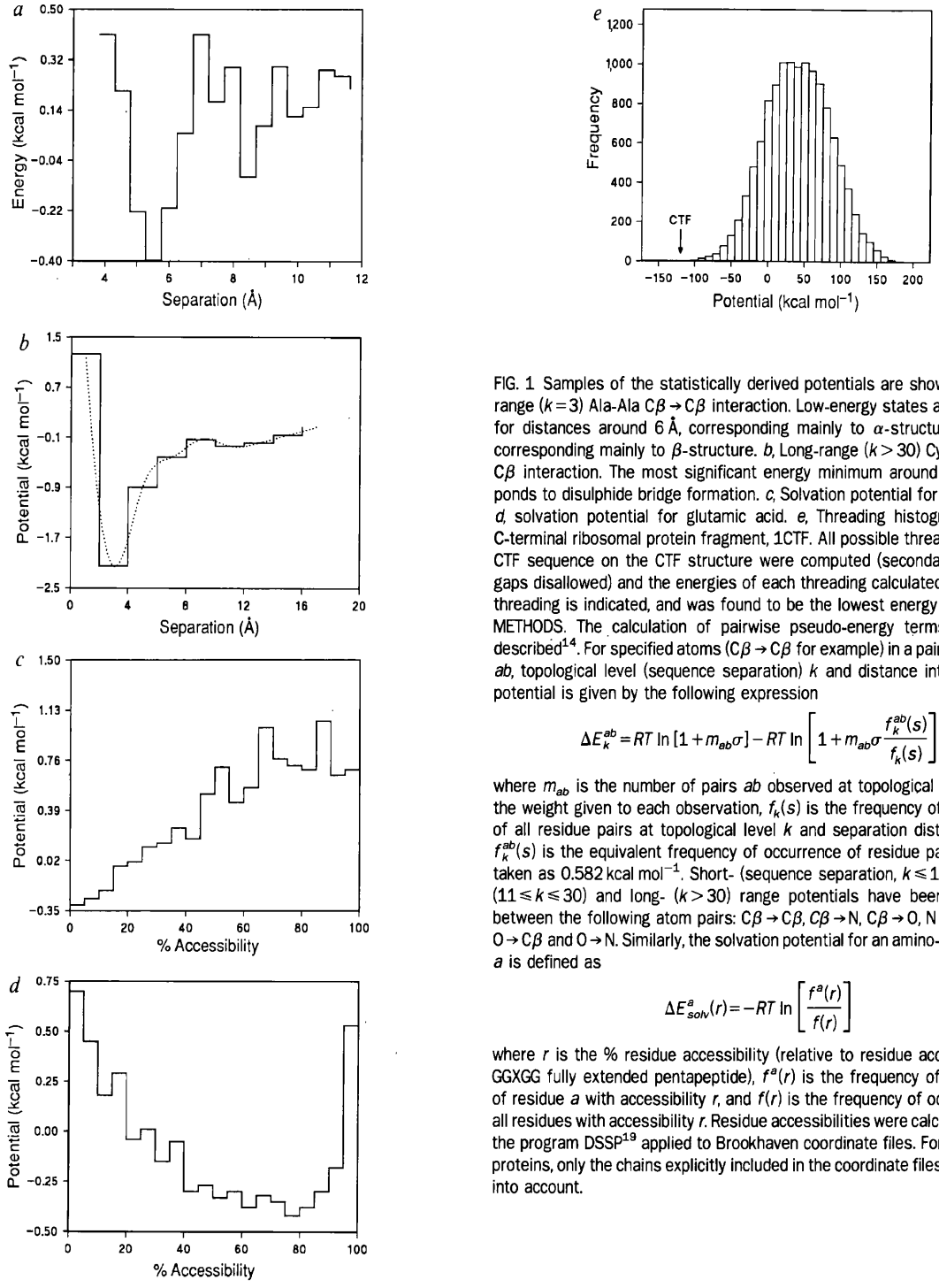


FIG. 1 Samples of the statistically derived potentials are shown. *a*, Short-range ($k=3$) Ala-Ala C$\beta \rightarrow$Cβ interaction. Low-energy states are observed for distances around 6 Å, corresponding mainly to α-structure, and 9 Å, corresponding mainly to β-structure. *b*, Long-range ($k>30$) Cys-Cys C$\beta \rightarrow$ Cβ interaction. The most significant energy minimum around 4 Å corresponds to disulphide bridge formation. *c*, Solvation potential for leucine, and *d*, solvation potential for glutamic acid. *e*, Threading histogram for the C-terminal ribosomal protein fragment, 1CTF. All possible threadings of the CTF sequence on the CTF structure were computed (secondary structure gaps disallowed) and the energies of each threading calculated. The native threading is indicated, and was found to be the lowest energy threading.

METHODS. The calculation of pairwise pseudo-energy terms has been described[14]. For specified atoms (C$\beta \rightarrow$Cβ for example) in a pair of residues *ab*, topological level (sequence separation) k and distance interval s, the potential is given by the following expression

$$\Delta E_k^{ab} = RT \ln [1 + m_{ab}\sigma] - RT \ln \left[1 + m_{ab}\sigma \frac{f_k^{ab}(s)}{f_k(s)}\right]$$

where m_{ab} is the number of pairs *ab* observed at topological level k, σ is the weight given to each observation, $f_k(s)$ is the frequency of occurrence of all residue pairs at topological level k and separation distance s, and $f_k^{ab}(s)$ is the equivalent frequency of occurrence of residue pair *ab*. RT is taken as 0.582 kcal mol^{-1}. Short- (sequence separation, $k \leq 10$), medium- ($11 \leq k \leq 30$) and long- ($k>30$) range potentials have been calculated 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. Similarly, the solvation potential for an amino-acid residue *a* is defined as

$$\Delta E_{solv}^{a}(r) = -RT \ln \left[\frac{f^a(r)}{f(r)}\right]$$

where r is the % residue accessibility (relative to residue accessibility in GGXGG fully extended pentapeptide), $f^a(r)$ is the frequency of occurrence of residue *a* with accessibility r, and $f(r)$ is the frequency of occurrence of all residues with accessibility r. Residue accessibilities were calculated using the program DSSP[19] applied to Brookhaven coordinate files. For multimeric proteins, only the chains explicitly included in the coordinate files were taken into account.

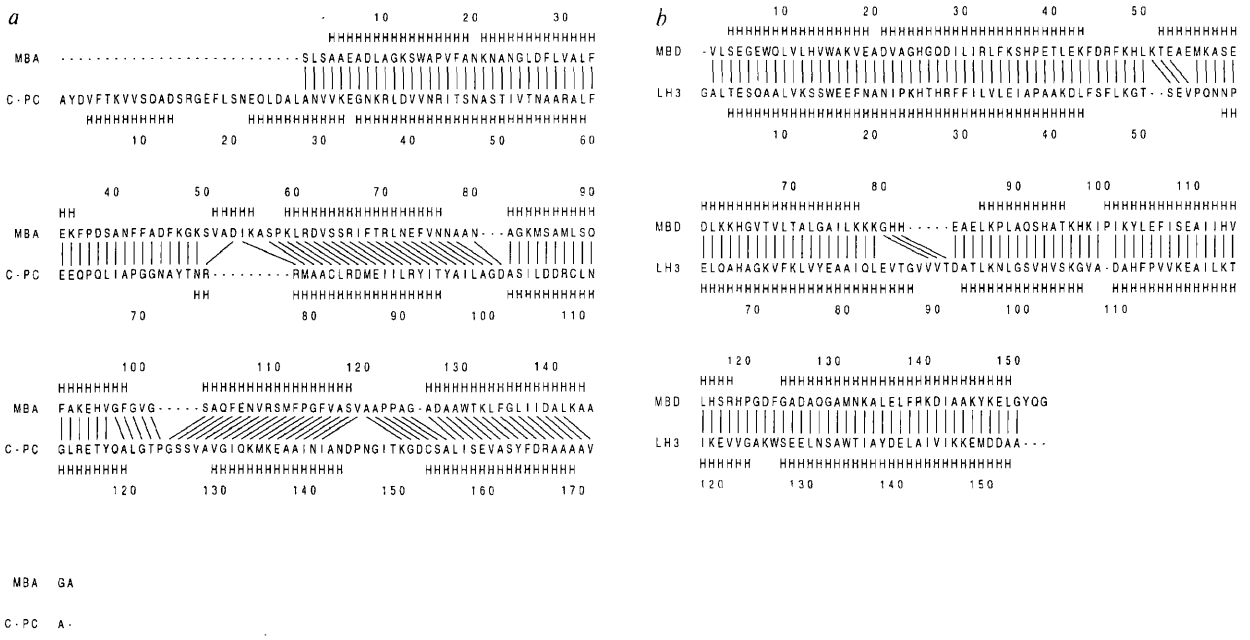


FIG. 2 *a*, Alignment of sea hare myoglobin (1MBA) with C-phycocyanin (C-PC) β-chain from *Mastigocladus laminosus* (SWISSPROT code PHCB$MASLA), found by optimal threading. Author-assigned secondary-structure codes are shown. The alignment is compared to the structurally determined alignment by Pastore and Lesk, where lines are drawn between structurally equivalent residue pairs as determined in the reference alignment[17]. *b*, Optimal threading of yellow lupin leghaemoglobin (LH3) on the structure of sperm whale myoglobin (Brookhaven code 1MBD). Alignment of protein structures is compared with the structural alignment obtained using the program SSAP[9].

found to be sea hare myoglobin (-451 kcal mol^{-1}) and midge erythrocruorin (-356 kcal mol^{-1}) followed by several other all-α protein folds. Figure 2*a* shows the alignment corresponding to the optimal threading of the C-phycocyanin β-chain sequence onto the best matching fold (sea hare myoglobin). For comparison, the optimal threading alignment of myoglobin and leghaemoglobin is shown in Fig. 2*b*. In terms of sequence, myoglobin and leghaemoglobin are only distantly related (17% sequence identity), but their structural similarity is much higher than in the case of phycocyanin and myoglobin, leading to a relatively unambiguous alignment. It should be borne in mind that even the structural alignment of phycocyanin and myoglobin is uncertain[17]. The fact that the optimal threading algorithm finds sea hare myoglobin to be the best model for C-phycocyanin is in accordance with a report[17] in which the

TABLE 1 Summary of trial fold-recognition searches

Test protein	Source	Fold	Best match	ΔE	% Sequence identity	Matches
C-phycocyanin β (C-PC)	Red algae	Globin	1MBA	101	7	1, 2, 9, 18, 25
Glycolate oxidase (GOX)	Spinach	TIM barrel	1WSY(A)	52	10	1, 3, 49
Muscle aldolase (ALD)	Human	TIM barrel	4XIA(A)	80	6	1, 2, 3
Lactate dehydrogenase (LDH)	Dogfish	Rossmann	4MDH(A)	87	15	1*
Elastase (EST)	Pig	Trypsin	4PTP	110	35	1, 14
CD4	Human	Ig	2FB4(H)	87	10	1, 2, 31
Stellacyanin (STEL)	Varnish tree	Cu binding	2AZA(A)	18	14	1, 6, 20
Cytochrome B562 (B562)	*E. coli*	4-helix bundle	2MHR	78	6	1
Trypsin inhibitor DE-3 (ETI)	Kaffir tree	Interleukin 1β	1I1B	14	5	1
PapD-chaperonin	*E. coli*	Ig	2FB4(L)	64	15	1, 5, 9, 35
70K, Heat-shock cognate (HSC)	Cow	Actin	1ATN(A)	94	9	1
Hexokinase B (HEX)	Yeast	Actin	1ATN(A)	0	12	1

In each case the database included 102 protein chains, except where the test protein was itself in the database, in which case it was excluded. Templates for each chain were constructed as described in the text, with residues not in helices or strands (as calculated by DSSP[19]) assigned as loop residues. For the 70K heat-shock cognate protein and hexokinase searches, the coordinates for actin were also included (coordinates deposited under the code 1ATN, but not yet released). Proteins with >25% sequence identity to the test protein were also excluded from the calculation of potentials. The pairwise and solvation terms were summed and stored separately, and standard deviations ($s.d._{pair}$ and $s.d._{solv}$) for the two contributing factors calculated over the set of 102 folds. To balance the contributions of the pairwise and solvation terms, the final energy was taken as $E = E_{pair} + WE_{solv}$, where $W = (s.d._{pair}/s.d._{solv})$. The 'confidence' of the match (ΔE) is given in terms of the absolute energy difference between the top scoring fold and the next highest scoring, different, fold. The 'best match' column gives the Brookhaven ID of the best matching chain fold (including chain identity where appropriate), along with the sequence identity between the best matching chain and the test protein. Positions in the sorted list of threading energies of similar folds are also shown. A constant set of alignment parameters (gap penalty for example) was used for all databank searches shown. Typical execution times for a single search of 102 chains are around 100 minutes on a Unix workstation (Solbourne 5/602). The 102 chains used were as follows: 351C, A256B, 2AAT, 1ABP, A5ACN, 8ADH, 3ADK, A8ATC, B8ATC, A2AZA, 3BLM, 1BP2, 2CA2, A7CAT, 1CC5, 1CCR, A2CCY, 1CD4, 2CDV, 3CLA, 2CNA, I4CPA, 5CPA, 2CPP, 4CPV, 1CRN, 2CRO, E1CSE, I1CSE, 1CTF, 1CY3, 2CYP, 3DFR, A4DFR, A1DHF, 1ECA, E2ER7, H2FB4, L2FB4, 1FD2, 1FX1, 3FXC, 4FXN, A3GAP, 2GBP, 1GCR, O1GD1, 3GRS, A3HHB, 1HIP, A2HLA, B2HLA, 1HOE, 1I1B, 3ICB, 3ICD, 1L01, 2LBP, 6LDH, 1LH1, 31LRD, A2LTN, 1LZ1, 1MBA, 1MBD, A4MDH, 2MHR, 2OVO, A2PAB, 9PAP, 1PAZ, 1PCY, A1PFK, 3PGK, 3PGM, 1PHH, 5PTI, 4PTP, 1RHD, 2RHE, 2RNT, 7RSA, 4RXN, 2SGA, I4SGB, 1SN3, 2SNS, O2SOD, 2SSI, 2STV, I1TGS, E2TMN, 4TNC, A1TNF, 1UBQ, 1UTG, A9WGA, R2WRP, A1WSY, B1WSY, A4XIA, A1YPI.

* Other topologically similar (yet structurally different) parallel αβ folds were positioned at 3, 7, 11, 12, 13, 17, 19, 31, 34, 82.

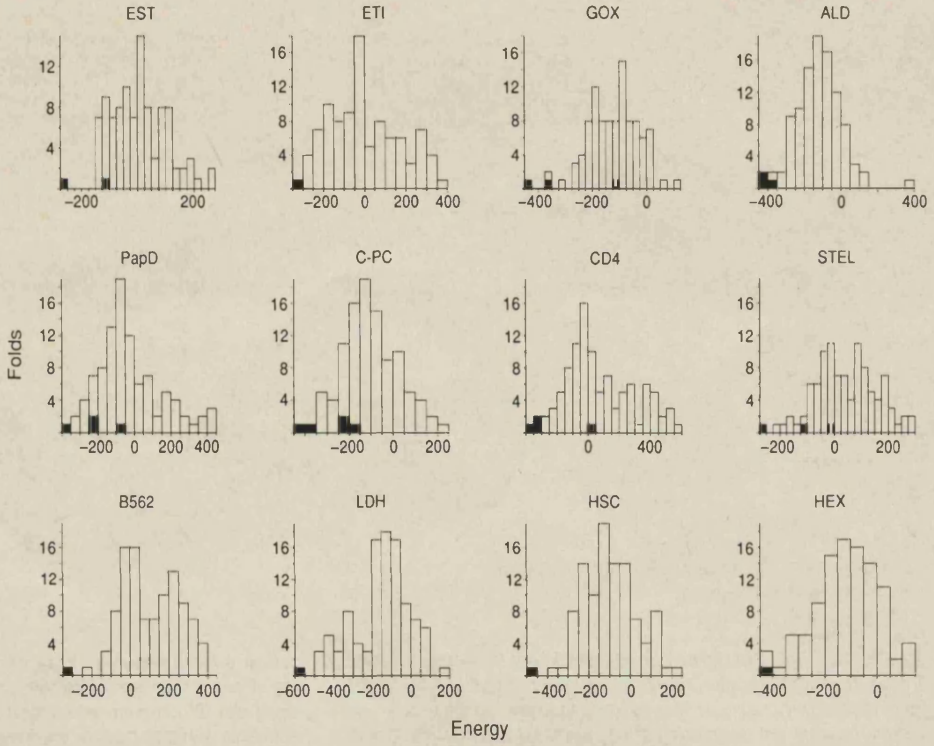


FIG. 3 For a number of test cases (see Table 1) the histogram of energies for optimally threading onto each of the 102 folds is given. In each histogram, the positions of folds expected to match the given sequence (that is, those folds similar to the known fold of the test sequence) are shown as filled bars. For example, in the case of LDH (lactate dehydrogenase), the expected match in the database of folds is MDH (malate dehydrogenase). This match is shown as a single filled bar representing an energy of −577 kcal mol^{-1}, an energy which is lower than that achieved by any other fold. As noted in the text, in some cases expected folds are apparently not detected. This occurs for two reasons: either the expected structures are not sufficiently similar to the native fold, or the optimization method does not succeed in producing a satisfactory alignment. The C-PC results demonstrate the former case. A number of unrelated highly helical proteins (carp parvalbumin and T4 lysozyme, for example) score better than the low-scoring globins. The worst-scoring globin is in fact human haemoglobin, in which case the poor score is due not only to the substantial secondary structural shifts relative to C-PC, but also to the fact that the calculated accessibilities are for the complete tetramer. The second situation arises in the results for GOX, where the algorithm fails to find an optimal threading of GOX onto XIA (xylose isomerase), resulting in an unexpectedly poor score. Of particular note in the results shown are the $(\alpha\beta)_8$ (TIM) barrel and trypsin inhibitor DE-3 examples. The degree of sequence homology between different $(\alpha\beta)_8$ barrel enzyme families and between trypsin inhibitor DE-3 and interleukin-1β is extremely low (5–10%). As a consequence of this, again, sequence template methods have been unable to detect these folds. Also of note are the results for the 70K heat-shock cognate protein (HSC70), and yeast hexokinase B. The N-terminal ATPase fragment of the heat-shock cognate protein has an almost identical structure to that of actin, but the similarity between hexokinase and actin is more topological than at the level of specific structural interactions. The two degrees of similarity are borne out by the threading results for these proteins, in that although actin is the lowest energy fold for hexokinase, the separation between the actin fold and the next-best-matching fold (aspartate transcarbamylase, ATC) is almost zero (0.1 kcal mol^{-1}); the rather weak structural similarity between hexokinase and actin would therefore appear to be just at the limits of our method. In contrast, the match between HSC70 and actin is clearly significant.

helix geometry of this globin was found to be closest to that of C-phycocyanin. Not only has the method correctly identified its globin fold, but has accurately located it in the C-phycocyanin sequence and has generated an alignment close to that obtained by careful structural alignment. It is clear that the method has identified the related folds in the database. It should be emphasized that no specific sequence information was used in the threading process: the structure was considered only as a chain of anonymous placeholders onto which the given sequence is threaded.

The results of other trial searches using the method of optimal sequence threading are shown in Table 1 and Fig. 3. From Fig. 3 it is apparent that in some cases expected matches are far from the top of the list. On inspection it was found that in these cases the threading algorithm had clearly misaligned the proteins, and had failed to find a reasonable optimum of the objective function, although this could generally be corrected by adjusting the alignment gap penalty.

The method described here shows promise as a new means for sensitively recognizing protein folds, and it is evident from the results that new information beyond sequence similarity is being exploited here. We are now exploring the generation of model folds[3,18], to escape from the limitation of only being able to predict previously observed folds, and the incorporation of multiple sequence data (from aligned sequence families) in the recognition process. □

Received 5 February; accepted 21 May 1992.

1. Taylor, W. R. & Thornton, J. M. *J. molec. Biol.* **173,** 487–514 (1984).
2. Overington, J., Johnson, M. S., Šali, A. & Blundell, T. L. *Proc. R. Soc.* B**241,** 132–145 (1990).
3. Finkelstein, A. V. & Reva, B. A. *Nature* **351,** 497–499 (1991).
4. Bowie, J. U., Lüthy, R. & Eisenberg, D. *Science* **253,** 164–170 (1991).
5. Bernstein, F. C. *et al. J. molec. Biol.* **112,** 535–542 (1977).
6. Dayhoff, M. O., Schwartz, R. M. & Orcutt, B. C. in *Atlas of Protein Sequence and Structure* Vol. 5 suppl. 3 345–352 (Natn. Biomed. Res. Fnd, Washington DC, 1978).
7. Thornton, J. M., Flores, T. P., Jones, D. T. & Swindells, M. B. *Nature* **354,** 105–106 (1991).
8. Taylor, W. R. & Orengo, C. A. *J. molec. Biol.* **208,** 1–22 (1989).
9. Orengo, C. A. & Taylor, W. R. *J. theor. Biol.* **147,** 517–551 (1990).
10. Novotny, J., Bruccoleri, R. E. & Karplus, M. *J. molec. Biol.* **177,** 787–818 (1984).
11. Brooks, B. *et al. J. comp. Chem.* **4,** 187–217 (1983).
12. Hendlich, M. *et al. J. molec. Biol.* **216,** 167–180 (1990).
13. Eisenberg, D. & McLachlan, A. D. *Nature* **319,** 199–203 (1986).
14. Sippl, M. J. *J. molec. Biol.* **213,** 859–883 (1990).
15. Schirmer, T., Bode, W., Huber, R., Sidler, W. & Zuber, H. *J. molec. Biol.* **184,** 257–277 (1985).
16. Bashford, D., Chothia, C. & Lesk, A. M. *J. molec. Biol.* **196,** 199–216 (1987).
17. Pastore, A. & Lesk, A. M. *Proteins* **8,** 133–155 (1990).
18. Taylor, W. R. *Prot. Engng* **4,** 853–870 (1991).
19. Kabsch W. & Sander C. *Biopolymers* **22,** 2577–2637 (1983).

ACKNOWLEDGEMENTS. We thank T. P. Flores, S. J. Hubbard, C. A. Orengo and M. B. Swindells for discussion, and K. C. Holmes for permission to use the coordinates for actin. D.T.J. acknowledges receipt of an SERC CASE studentship with the MRC.

CABIOS

Vol.8. no.3. 1992
Pages 275–282

The rapid generation of mutation data matrices from protein sequences

David T.Jones[1,2], William R.Taylor[2] and Janet M.Thornton[1]

Abstract

An efficient means for generating mutation data matrices from large numbers of protein sequences is presented here. By means of an approximate peptide-based sequence comparison algorithm, the set sequences are clustered at the 85% identity level. The closest relating pairs of sequences are aligned, and observed amino acid exchanges tallied in a matrix. The raw mutation frequency matrix is processed in a similar way to that described by Dayhoff et al. (1978), and so the resulting matrices may be easily used in current sequence analysis applications, in place of the standard mutation data matrices, which have not been updated for 13 years. The method is fast enough to process the entire SWISS-PROT databank in 20 h on a Sun SPARCstation 1, and is fast enough to generate a matrix from a specific family or class of proteins in minutes. Differences observed between our 250 PAM mutation data matrix and the matrix calculated by Dayhoff et al. are briefly discussed.

Introduction

Despite the great diversiy of methods devised for the alignment and comparison of protein sequences, all of these depend at some point on the simple comparison of two amino acid residues. The most popular method for measuring the similarity between amino acids is to use a scoring matrix of some form. At its simplest, a typical scoring matrix comprises 20 × 20 elements, each element representing some metric that relates two residues.

The least sophisticated matrix is the 'Unitary Protein Matrix' (UPM), also known as the 'identity matrix'. The UPM scores a 1 for exactly matching residues and a 0 for every other combination. Obviously this matrix lacks sensitivity, as it is unable to detect the possibility of phenotypically silent mutational events between two sequences. One advantage of the UPM is that it is wholly unbiased, providing a very easily understood alignment metric. The 'percentage identity' between two sequences is often offered as a universal means of describing the mutual degree of 'homology' between them. Although a low identity score can in no way prove or disprove the existence of homology, it has proved easier to provide rules of thumb for identity scoring than for any other scheme. In general, for two sequences of reasonable length (say 50 residues or more), a percentage identity of >25% points to a significant structural homology between them. Feng and Doolittle have described a fuzzy region around 20% identity which they call the 'Twilight Zone'. Within this zone and below, it is not possible to tell the difference between real sequence similarity implying a common structural framework, and accidental similarity providing no useful structural information.

Probably the next simplest amino acid scoring matrix is the 'Genetic Code Matrix' (GCM). This matrix scores amino acid similarity by the maximum number of common nucleotide bases between their closest matching representative codons. Identical residues of course share a maximum of 3 bases, whereas non-identical residues may have only 0, 1 or 2 bases in common. This matrix has a pleasantly 'genetic flavour' to it, but it must be realized that the bulk of the selection pressure is on the protein sequence and not on the underlying DNA sequence. Although there does seem to be a reasonable correlation between the nucleotide codons associated with amino acids and the degree of chemical similarity between them (Woese, 1969, for example), the rather limited range of match-scores puts the GCM somewhat in the shade. To detect weak homologies between sequences a more accurate amino acid comparison table is required.

McLachlan (1972) published a scoring matrix that attempted to quantify explicitly the degree of chemical similarity between amino acids. This matrix, known as the 'Structure-Genetic Matrix' (SGM), incorporated two sources of information in evaluating the similarities of amino acids. The first source was a statistical analysis of observed amino acid exchanges in available families of proteins, the second was from the assignment of transition values for each pair of amino acids depending on the number of overlapping physico-chemical properties between them. These data were used to 'bias' the UPM in such a way that only 20 of the 190 possible interchanges were significantly preferred (Feng *et al.*, 1985). The problem with the SGM and other matrices that attempt to incorporate 'real' amino acid similarities is that the groupings used are artificial, there is no guarantee that an arbitrary common amino acid property is at all important for strucural and functional conservation between proteins. A better approach is to concentrate on the observed exchanges between amino acids

[1]*Biomolecular Structure and Modelling Unit, Department of Biochemistry and Molecular Biology, University College, Gower Street, London WC1E 6BT and* [2]*Laboratory of Mathematical Biology, National Institute for Medical Research, The Ridgeway, Mill Hill, London, NW7 1AA, UK*

in very similar aligned sequences. Evidently amino acids that share the appropriate properties will exchange more frequently than ones that do not. McLachlan's earlier attempt to compare amino acids (McLachlan, 1971) was based entirely on such a statistical approach.

Recently, matrices based on the principles of structural comparison have been described (Risler *et al.*, 1988; Overington *et al.*, 1990). These matrices essentially contain statistics on the pairwise substitutions observed at structurally equivalent positions in aligned families of protein structures. In the case of Overington *et al.*, a range of matrices are calculated, one from each class of structural environment, an example of one such class being 'buried coil' for example. These matrices show great promise in increasing the accuracy of sequence-to-sequence, and sequence-to-structure alignments, though the sparsity of structural data presently available is a significant disadvantage of this approach.

The most widely used comparison matrix today is the 'Log-Odds Matrix' and the very closely related 'Mutation Data Matrix' (MDM) published by Dayhoff *et al.* (1978). The MDM was calculated from a study of the exchange probabilities (or odds) derived from an analysis of the evolutionary changes seen in groups of very similar proteins. A strictly Markovian model (i.e. the current probabilities are independent of previous events) of amino acid exchange is assumed in the Dayhoff model. This model has been criticized (see George *et al.*, 1990, for a review), but comparisons of different scoring schemes have tended hesitantly to recommend the MDM over other matrices (Feng *et al.*, 1985).

In this paper we show a straightforward and automatic procedure for generating mutation data matrices, in order that very large sets of sequences can be processed without using inordinate amounts of computing resources. In particular we are able to improve the generality of the MDM, in that we now have access to a much greater variety of protein sequences than were available to Dayhoff and her workers in 1978, and it is our hope that the matrices presented here will more clearly express the general nature of the underlying amino acid similarities.

The original mutation data matrix (MDM68) was presented in the original *Atlas of Protein Sequence and Structure* (1968), and the method (outlined below) remained virtually unaltered through each of the subsequent updates. There are five main steps required for the creation of a mutation data matrix:

1. Construction of the raw PAM matrix

The basic unit of molecular evolution expressed in a MDM is the 'accepted point mutation', or with a little license to ease pronunciation: PAM. One PAM is simply the mutation of a single amino acid in a sequence such that the new amino acid may be accommodated in the structure and function of the protein. In general, therefore, amino acid residues that are frequently seen to exchange in a PAM matrix typically have similar physico-chemical properties.

The raw PAM substitution matrix is created by considering the possible mutational events that could have occurred between two closely related sequences. Ideally we would like to compare every present-day sequence with its own immediate predecessor and thus accurately map the evolutionary history of each sequence position. Of course this is impossible, and so two main courses of action may be taken to approximate this information. The method used by Dayhoff was the 'common ancestor' method. Here closely homologous pairs of present-day sequences are taken and a common ancestral sequence inferred. Given only a pair of present-day sequences, an unambiguous inferred common ancestor cannot be generated. A complete phylogenetic tree is required in this case to allow the most probable common ancestors to be inferred for each tree node. The important thing to realize is that the inference of common ancestors must consider the overall topology of the tree. Every suggested common ancestor must be traced back to higher level nodes and evaluated in order to determine whether or not that ancestral sequence is the most probable for the tree as a whole.

An alternative to the common ancestor method is to relate present-day sequences by their pairwise alignment distances, estimating a possible phylogenetic tree from this distance matrix. This method was first described by Fitch and Margoliash (1967). Although construction of the distance matrix is a trivial exercise, the generation of an optimal phylogenetic tree from this data again requires an exhaustive iterative analysis such that the total number of mutations required to produce the present day set of sequences is minimized. Although both of the above methods have advantages and disadvantages, matrix methods are now most widely used.

No matter which method is finally used to infer the phylogenetic tree, construction of the PAM matrix is the same. The raw matrix is generated by taking pairs of sequences, either a present-day sequence and its inferred ancestor, or two present-day sequences, and tallying the amino acid exchanges that have apparently occurred. Given the following alignment:

```
ACDEFL
AGDEAL
```

we count four PAMs (C → G, G → C, F → A and A → F). The raw PAM matrix is obviously symmetric given the fact that we cannot know whether for example C mutated to G or G mutated to C; there is no harm in this as we are interested in discerning the extent of similarity between amino acids here, and 'similarity' is generally thought of as being symmetric. Treatment of gaps/insertions in an alignment is arbitrary: one possibility is to count gap characters as another type of amino acid; another possibility that is probably the safer of the two is simply to ignore gaps. We are after all only interested in

the exchange of amino acids, the deletion of a particular amino acid tells us nothing of its relative similarity to other amino acids, though it does provide information as to the amino acid's characteristic 'mutability'.

2. Calculation of relative mutabilities

Evidently if we are to estimate the probability of a given mutation event, we must consider two pieces of information. Firstly how likely is it that a given amino acid A changes at all, secondly how likely is it that the given amino acid changes to amino acid B given that A does change? We are therefore interested in the conditional probability that amino acid A changes to amino acid B given that A is seen to change. The probability of amino acid A changing at all in a given unit of time is usually expressed as the 'relative mutability' of A. Relative mutability is simply calculated as the number of observed changes of an amino acid divided by its frequency of occurrence in the aligned sequences. From the alignment shown earlier, A is seen to change once, but occurs three times in the alignment. The relative mutability of A from this alignment alone is therefore calculated as ⅓. An overall measure of relative mutability must allow for the different evolutionary distances and different sequence lengths found in a non-specific collection of sequences. Mutability is normalized by defining the basic unit of evolutionary distance as being a single accepted point mutation in a sequence of length 100. The average relative mutability of an amino acid given this definition is therefore the total number of changes observed for this amino acid in all the families of proteins considered, divided by the total sum of all local frequencies of occurrence of the amino acid multiplied by the numbers of mutations per 100 residues in each of the branches of all the family trees.

3. Calculation of the mutation probability matrix

The basic matrix in the generation of MDM type matrices is the 'mutation probability matrix'. Elements of this matrix give the probability that a residue in column j will mutate to the residue in row i in a specified unit of evolutionary time. Evidently a diagonal element of this matrix represents the probability of residue $i = j$ remaining unchanged, and hence being easily calculated according to the following formula:

$$M_{ji} = 1 - \lambda m_j \qquad (1)$$

where m_j is the average relative mutability of residue j, and λ is a proportionality constant.

Non-diagonal elements are given by:

$$M_{ij} = \frac{\lambda m_j A_{ij}}{\sum_i A_{ij}} \qquad (2)$$

where A_{ij} is a (non-diagonal) element of the raw PAM matrix.

The value of λ relates to the evolutionary distance represented by the probability matrix, accordingly:

$$\sum_i f_i M_{ii} = 1 - \frac{P}{100} \qquad (3)$$

where f_i is the normalized frequency of occurrence of residue i, and P approximates the evolutionary distance (in PAMs) represented by the matrix. This relationship breaks down for $P >> 5$.

P is usually given the value 1 so that the basic mutation probability matrix represents a distance of 1 PAM. Matrices representing larger evolutionary distances may be derived from the 1 PAM matrix by matrix multiplication. Squaring the 1 PAM matrix gives a 2 PAM matrix, cubing it a 3 PAM matrix and so forth.

4. Calculating the log-odds matrix

Of more use than the mutation probability matrix in the alignment of protein sequences is the 'relatedness odds matrix'. This symmetric matrix represents the probability of residue j being replaced by residue i per occurrence of i, and is derived from the mutation probability matrix simply by dividing each element M_{ij} by the normalized frequency of occurrence of i, f_i. For the purposes of sequence comparison the relatedness odds for each alignment position should be multiplied together in order to arrive at a total 'alignment odds' value. To avoid slow floating-point multiplications, the relatedness odds matrix is usually converted to the log odds-matrix (also known as the mutation data matrix) thus:

$$MDM_{ij} = 10 \log_{10} R_{ij} \qquad (4)$$

where R_{ij} are elements of the relatedness odds matrix (MDM_{ij} values are rounded to the nearest integer).

Automating the procedure

Although computational tools were used in constructing the original MDMs, in particular for the inference of common ancestral sequences and the generation of phylogenetic trees, the whole process was only partially automated. This was hardly of consequence considering the small number of available sequences in the 1970s, but as at the time of writing some 23 000 protein sequences are available for analysis, it is evident that a more streamlined approach is now required.

Our method for generating MDMs is in fact very similar in essence to that described by Dayhoff *et al.* (1978). The method involves three steps: (i) clustering the sequences into homologous families, (ii) tallying the observed mutations between highly similar sequences and (iii) relating the observed mutation frequencies to those expected by pure chance. The main difference here is in our use of an approximate method (a

pairwise present-day ancestor scheme) for inferring the phylogenetic relationships among the sequences in the data set. A program was written to compute all the relevant data automatically from a file of protein sequences.

In view of the relative inefficiency of standard methods for inferring maximum parsimony phylogenetic trees it was found to be necessary to implement an approximate method to find the reasonable family trees by means of cluster analysis of the sequence data. Although the limitations of using such simple means alone for the inference of phylogenetic trees are well known (Czelusniak *et al.*, 1990), and the large-scale structure of such crude phylogenetic trees tends to be somewhat incorrect, the relationships between closely related sequences are inferred correctly. To verify our methodology, we attempted to re-create the set of sequences used to construct MDM78. Using these sequences we found our mutation data closely approximated those in the original work with 164 of the 400 mutation frequencies (number of mutations occurring per 10000 observations) being identical, and 350 differing by five or less. It should be pointed out that though our results very closely match those of Dayhoff *et al.*, our matrices are not derived from the same explicit evolutionary model outlined in the original work. The practical significance of this fact depends on the intended application of the matrices. In terms of sequence analysis applications, a derivation independent of the choice of evolutionary model might well be preferred due to the reduced possibility of bias (in particular, maximum parsimony nucleotide substitution methods will tend to produce results biased towards the exchanges expected from the genetic code rather than generally observed amino acid similarities). A further justification for determining relationships via a pairwise scheme is that of the 2621 families of proteins in the current release of SWISSPROT, 79% contain fewer than five sequences. With such small families the results of simple clustering and those of rigorous maximum parsimony analysis are indistinguishable with respect to the present application.

In generating the initial distance matrix, we do not assume that the input sequences are in any way pre-clustered into family groups, and are therefore forced to calculate the entire distance matrix to sort the sequences into families, and thereafter produce trees for each family. Evidently the vast majority of pairwise comparisons are unnecessary, so some simple (and quick) means is needed to filter out sequence pairs that have no chance of producing alignment identity scores >85%. We propose here a simple approximate algorithm for 'estimating' the percentage identity between two protein sequences without prior alignment. Our algorithm considers the distribution of residue triplets (or 3-tuples) between the two sequences. If there are sufficient identical triplets between both sequences we assume that the sequences show a potential homology. The longest sequence is taken and a hash table constructed containing the frequencies of occurrence of the constituent triplets. The triplet frequencies

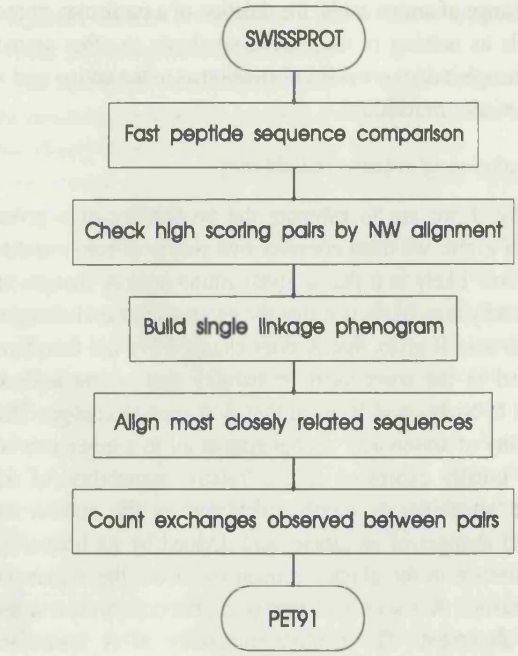


Fig. 1. An outline of the described method for generating mutation data matrices.

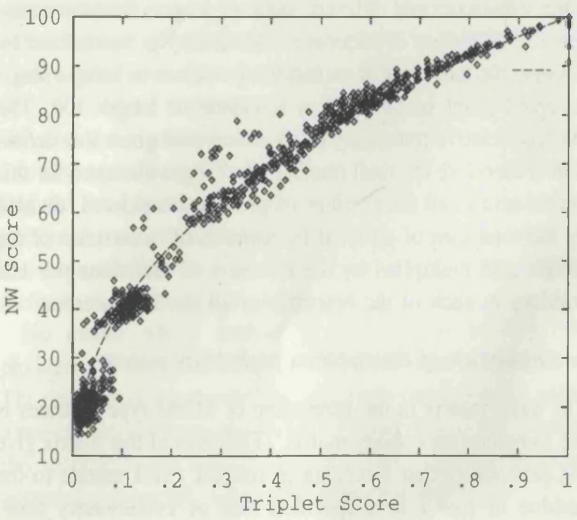


Fig. 2. Relationship between triplet scores and per cent identity after Needleman–Wunsch alignment with constant gap penalty.

of the shorter sequence are then compared with those of the longer. A comparison score is calculated thus:

$$S = \frac{\sum_{pqr=AAA}^{VVV} \min(f_a^{pqr}, f_b^{pqr})}{\min(n_a, n_b) - 2} \quad (5)$$

where f_a^{pqr} and f_b^{pqr} are the frequencies of occurrence of triplet pqr in sequences a and b, and n_a and n_b are the respective sequence lengths.

This normalized score (S) is effectively the fractional area

Table I. The 250 PAM PET91 matrix ($\log_{10}$ relatedness odds), based on 59 190 accepted point mutations found in 16 130 protein sequences

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	2	247	216	386	106	208	600	1183	46	173	257	200	100	51	901	2413	2440	11	41	1766
R	-1	5	116	48	125	750	119	614	446	76	205	2348	61	16	217	413	230	109	46	69
N	0	0	3	1433	32	159	180	291	466	130	63	758	39	15	31	1738	693	2	114	55
D	0	-1	2	5	13	130	2914	577	144	37	34	102	27	8	39	244	151	5	89	127
C	-1	-1	-1	-3	11	9	8	98	40	19	36	7	23	66	15	353	66	38	164	99
Q	-1	2	0	1	-3	5	1027	84	635	20	314	858	52	9	395	182	149	12	40	58
E	-1	0	1	4	-4	2	5	610	41	43	65	754	30	13	71	156	142	12	15	226
G	1	0	0	1	-1	-1	0	5	41	25	56	142	27	18	93	1131	164	69	15	276
H	-2	2	1	0	0	2	0	-2	6	26	134	85	21	50	157	138	76	5	514	22
I	0	-3	-2	-3	-2	-3	-3	-3	-3	4	1324	75	704	196	31	172	930	12	61	3938
L	-1	-3	-3	-4	-3	-2	-4	-4	-2	2	5	94	974	1093	578	436	172	82	84	1261
K	-1	4	1	0	-3	2	1	-1	1	-3	-3	5	103	7	77	228	398	9	20	58
M	-1	-2	-2	-3	-2	-2	-3	-3	-2	3	3	-2	6	49	23	54	343	8	17	559
F	-3	-4	-3	-5	0	-4	-5	-5	0	0	2	-5	0	8	36	309	39	37	850	189
P	1	-1	-1	-2	-2	0	-2	-1	0	-2	0	-2	-2	-3	6	1138	412	6	22	84
S	1	-1	1	0	1	-1	-1	1	-1	-1	-2	-1	-1	-2	1	2	2258	36	164	219
T	2	-1	1	-1	-1	-1	-1	-1	-1	1	-1	-1	0	-2	1	1	2	8	45	526
W	-4	0	-5	-5	1	-3	-5	-2	-3	-4	-2	-3	-3	-1	-4	-3	-4	15	41	27
Y	-3	-2	-1	-2	2	-2	-4	-4	4	-2	-1	-3	-2	5	-3	-1	-3	0	9	42
V	1	-3	-2	-2	-2	-3	-2	-2	-3	4	2	-3	2	0	-1	-1	0	-3	-3	4

Values have been multiplied by 10 and rounded to the nearest integer. The upper half of the matrix shows the actual numbers of exchanges observed.

Table II. Mutation probability matrix for an evolutionary distance of 1 PAM. Values are scaled by a factor of 10^5

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	98759	27	24	42	12	23	66	129	5	19	28	22	11	6	99	264	267	1	4	193
R	41	98962	19	8	21	125	20	102	74	13	34	390	10	3	36	69	38	18	8	11
N	43	23	98707	284	6	31	36	58	92	26	12	150	8	3	6	344	137	0	23	11
D	63	8	235	98932	2	21	478	95	24	6	6	17	4	1	6	40	25	1	15	21
C	44	52	13	5	99450	4	3	41	17	8	15	3	10	28	6	147	28	16	68	41
Q	43	154	33	27	2	98955	211	17	130	4	64	176	11	2	81	37	31	2	8	12
E	82	16	25	398	1	140	99042	83	6	6	9	103	4	2	10	21	19	2	2	31
G	135	70	33	66	11	10	70	99369	5	3	6	16	3	2	11	129	19	8	2	32
H	17	164	171	53	15	233	15	15	98867	10	49	31	8	18	58	51	28	2	189	8
I	28	12	21	6	3	3	7	4	4	98722	212	12	113	31	5	28	149	2	10	630
L	24	19	6	3	3	29	6	5	12	122	99328	9	90	101	53	40	16	8	8	117
K	28	334	108	14	1	122	107	20	12	11	13	99101	15	1	11	32	57	1	3	8
M	36	22	14	10	8	19	11	10	8	253	350	37	98845	18	8	19	123	3	6	201
F	11	3	3	2	14	2	3	4	11	41	230	1	10	99357	8	65	8	8	179	40
P	150	36	5	7	3	66	12	16	26	5	97	13	4	6	99278	190	69	1	4	14
S	297	51	214	30	44	22	19	139	17	21	54	28	7	38	140	98548	278	4	20	27
T	351	33	100	22	9	21	20	24	11	134	25	57	49	6	59	325	98670	1	6	76
W	7	65	1	3	23	7	7	41	3	7	49	5	5	22	4	21	5	99684	24	16
Y	11	12	30	23	43	10	4	4	134	16	22	5	4	222	6	43	12	11	99377	11
V	226	9	7	16	13	7	29	35	3	504	161	7	71	24	11	28	67	3	5	98772

of overlap between the two triplet histograms. Scatter plots based on all possible pairwise alignment scores in a set of 200 protein sequences (containing a mixture of related and unrelated sequences) plotted against our scoring metric were produced (a subset of this data is shown in Figure 2). The raw triplet scores were thus compared with Needleman–Wunsch scores

(>40% ID), and the following relationship (correlation coefficient 0.986) was observed:

$$I \approx 100S^{0.3912}$$

where S is the normalized triplet frequency score, and the result I is in units of percentage identity.

By aligning only those sequence pairs with corrected triplet scores indicating sequence identity ≥45% and subsequently excluding sequence pairs with alignment scores of ≤85% identity we were able rapidly to generate a sparse distance matrix complete enough for our purposes. By combining this very rapid heuristic measure of identity with an efficiently coded dynamic programming algorithm as a 'second level filter' we were able to construct the distance matrix at an average rate of over 1000 similarity score calculations per second on a Sun SPARCstation 1 (standard Sun C compiler). Out of the 130 million pairwise alignments that would normally be required, only 559 692 passed the initial similarity filter, speeding up the process nearly 200-fold.

Using this matrix of identity scores, the sequences were subjected to an efficient single-linkage clustering algorithm, with mutation statistics being generated for each sequence by aligning it with the sequence that offers the highest pairwise alignment score. For each sequence pair, amino acid substitutions are tallied with alignment positions containing at least one non-standard residue code (B, Z, X or 'Gap') being ignored.

Implementation

The matrix generation program MAKEPET is coded in standard Sun C, and should be portable to most platforms supporting a C compiler. The required matrix PAM distance and other control parameters are specified as command line arguments. MAKEPET takes as input a single file of sequences in 'compact PIR' format, where each sequence is preceded by two description lines and terminated by a '*' character. A simple keyword searching program SEQGREP allows specific sets of sequences to be compiled from the complete sequence databank, permitting the easy generation of matrices biased towards particular structural or functional classes (membrane-bound proteins for example).

Results

The upper half of Table I shows how many of each of the possible 190 exchanges were observed, with the lower half of Table I showing our equivalent of the widely used MDM78 matrix ($\log_{10}$ relatedness-odds matrix for 250 PAMs), which we call PET91 (Pairwise Exchange Table 1991). The 1 PAM mutation probability matrix required to generate mutation data matrices for evolutionary distances other than 250 PAMs is shown in Table II. PET91 was generated from Release 15.0 of the SWISS-PROT protein sequence database (Bairoch, 1990),

Table III. Relative mutabilities and normalized frequencies of occurrence for the 20 amino acid residues, calculated from the PET91 data set, compared with the values from Dayhoff *et al.* (1978)

	Relative Mutability* (1991)	Relative Mutability* (1978)	Relative Frequency of Occurrence (1991)	Relative Frequency of Occurrence (1978)
Ala (A)	100	100	0.077	0.087
Arg (R)	83	65	0.051	0.041
Asn (N)	104	134	0.043	0.040
Asp (D)	86	106	0.052	0.047
Cys (C)	44	20	0.020	0.033
Gln (Q)	84	93	0.041	0.038
Glu (E)	77	102	0.062	0.050
Gly (G)	50	49	0.074	0.089
His (H)	91	66	0.023	0.034
Ile (I)	103	96	0.053	0.037
Leu (L)	54	40	0.091	0.085
Lys (K)	72	56	0.059	0.081
Met (M)	93	94	0.024	0.015
Phe (F)	51	41	0.040	0.040
Pro (P)	58	56	0.051	0.051
Ser (S)	117	120	0.069	0.070
Thr (T)	107	97	0.059	0.058
Trp (W)	25	18	0.014	0.010
Tyr (Y)	50	41	0.032	0.030
Val (V)	98	74	0.066	0.065

* Relative to Ala which is arbitrarily assigned a mutability of 100.

containing 16 941 sequences, though sequences <20 residues were excluded to avoid insignificant alignments. It should be noted that the 250 PAM matrix is shown here for reasons of comparison with the most common variant of the original matrix, and that matrices calculated for evolutionary distances other than 250 PAMs are often found to perform better for some sequence comparisons. The recently described sequence databank search program, BLAST (Altschul *et al.*, 1990), for example, uses a 120 PAM Dayhoff matrix by default.

Of particular interest here are the differences between these results and those of the original work, a rough impression of which may be gained from a comparison of the relative mutabilities shown in Table III with those observed by Dayhoff (1978). A value of 0.76 is obtained for the Spearman rank correlation coefficient between the old and new relative mutabilities, indicating little overall change. Ser (serine) and Thr (threonine) are found to be the most mutable residues in this work, as opposed to asparagine and serine in the 1978 table. Trp (tryptophan) and Cys (cysteine) are found to be least mutable here, which agrees with the earlier findings, though the mutability of Cys found here is double the original value. The frequencies of occurrence of the amino acid residues (Table I) show no significant differences from the earlier values.

Table IV. The difference matrix ($PET91_{ij}$ − $MDM78_{ij}$) between the 250 PAM PET91 matrix and the MDM78 matrix

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	0	+1	0	0	+1	-1	-1	0	-1	+1	+1	0	0	+1	0	0	+1	+2	0	+1
R	+1	-1	0	0	+3	+1	+1	+3	0	-1	0	+1	-2	0	-1	-1	0	-2	+2	-1
N	0	0	+1	0	+3	-1	0	0	-1	0	0	0	0	+1	0	0	+1	-1	+1	0
D	0	0	0	+1	+2	-1	+1	0	-1	-1	0	0	0	+1	-1	0	-1	+2	+2	0
C	+1	+3	+3	+2	-1	+2	+1	+2	+3	0	+3	+2	+3	+4	+1	+1	+1	+9	+2	0
Q	-1	+1	-1	-1	+2	+1	0	0	-1	-1	0	+1	-1	+1	0	0	0	+2	+2	-1
E	-1	+1	0	+1	+1	0	+1	0	-1	-1	-1	+1	-1	0	-1	-1	-1	+2	0	0
G	0	+3	0	0	+2	0	0	0	0	0	0	+1	0	0	0	0	-1	+5	+1	-1
H	-1	0	-1	-1	+3	-1	-1	0	0	-1	0	+1	0	+2	0	0	0	0	+4	-1
I	+1	-1	0	-1	0	-1	-1	0	-1	-1	0	-1	+1	-1	0	0	+1	+1	-1	0
L	+1	0	0	0	+3	0	-1	0	0	0	-1	0	-1	0	+3	+1	+1	0	0	0
K	0	+1	0	0	+2	+1	+1	+1	+1	-1	0	0	-2	0	-1	-1	-1	0	+1	-1
M	0	-2	0	0	+3	-1	-1	0	0	+1	-1	-2	0	0	0	+1	+1	+1	0	0
F	+1	0	+1	+1	+4	+1	0	0	+2	-1	0	0	0	-1	+2	+1	+1	-1	-2	+1
P	0	-1	0	-1	+1	0	-1	0	0	0	+3	-1	0	+2	0	0	+1	+2	+2	0
S	0	-1	0	0	+1	0	-1	0	0	0	+1	-1	+1	+1	0	0	0	-1	+2	0
T	+1	0	+1	-1	+1	0	-1	-1	0	+1	+1	-1	+1	+1	+1	0	-1	+1	0	0
W	+2	-2	-1	+2	+9	+2	+2	+5	0	+1	0	0	+1	-1	+2	-1	+1	-2	0	+3
Y	0	+2	+1	+2	+2	+2	0	+1	+4	-1	0	+1	0	+2	+2	+2	0	0	-1	-1
V	+1	-1	0	0	0	-1	0	-1	-1	0	0	-1	0	+1	0	0	0	+3	-1	0

A positive matrix element indicates that the PET91 value is higher than the related value in MDM78. Absolute differences ≥2 are shown shaded.

Table IV shows the pattern of changes between the MDM78 and the PET91 matrices. Both Cys and Trp show very different patterns of mutability, both now showing a much greater tendency to exchange with other amino acid residues than in the previous study. This can be attributed mainly to the paucity of mutational events involving Cys and Trp in the original data set. Overall, in Dayhoff's data 35 amino acid exchanges were never observed at all (e.g. Cys and Trp); here, however, all possible exchanges have been observed (Cys and Trp exchanging 38 times in the current data set). PET91 incorporates 442 Trp exchanges and 1292 Cys exchanges, where only 7 Trp exchanges and 28 Cys exchanges were recorded for the MDM78 matrix. Interestingly, however, the average absolute change of the Cys matrix elements is higher than that of Trp, even though the Cys sample was larger than that of Trp in the 1978 data set. This anomaly is attributable to the fact that Cys residues occur in three very different chemical roles in proteins: as free sulphydryl groups (−S−H), in disulphide bridges (−S−S−), and as ligands for metals (−S..X). The number of observed cys exchanges in the original work would have been insufficient to sample these three situations effectively. In addition, the Cys residue exchanges observed in the original work were mostly from the metallothionein sequences included in the data set.

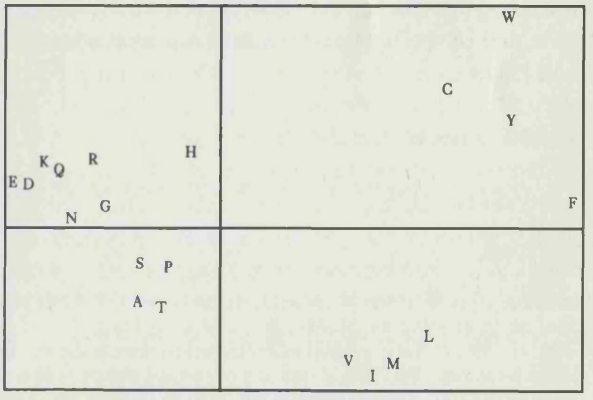


Fig. 3. The general trends in amino acid residue similarity shown in the PET91 relatedness odds matrix, visualized by means of multidimensional scaling.

It is also interesting to note that even with the very large amount of data collected here, some amino acid exchanges are still very seldom observed: Trp and Asn (asparagine), for example, were only seen to exchange twice. Indeed it is hard to be certain whether these highly infrequent exchanges are real observations or artefacts caused by errors in the sequence database.

A common method for interpreting the complex trends in a similarity matrix is to project the 20 × 20 = 400-dimensional

pattern onto a plane via multidimensional scaling (French and Robson, 1983). The plot in Figure 3 shows such a projection, which clearly delineates the relationships between the 20 amino acids found in PET91. The general trends shown in the PET matrix are essentially those found in the original Dayhoff matrix: hydrophobicity and size being the most significant factors.

Discussion

In general, the most significant differences (PET91 matrix elements differing from MDM78 elements by ±2 or more) correspond almost exactly to exchanges that were observed no more than once in Dayhoff's sequence alignments. Despite these few anomalous differences, however, it is interesting to see how little the bulk of PET91 differs from MDM78. The fundamental amino acid similarities remain unchanged, and given that we have now collected enough data to iron out the residual sampling errors in the mutation data matrix, we feel confident that PET91 represents a relatively unbiased measure of amino acid similarity in sequence data and should be used in preference to the MDM78 in sequence analysis applications. Investigation is currently under way as to the performance of our matrices compared to others with regard to sequence alignment and databank searching. We are also developing matrices biased to particular protein classes and residue environments, and a dipeptide mutability matrix (400 × 400 elements) which has enabled us to investigate short-range sequence neighbourhood effects on residue mutability.

The matrix generation programs and the complete data, including all intermediate matrices and tables required for constructing matrices for evolutionary distances other than 250 PAMs, may be obtained from the authors in printed or machine-readable form.

Acknowledgements

D.T.J. acknowledges receipt of a SERC CASE studentship with the MRC.

References

Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **214**, 403–410.

Bairoch,A. (1990) PC/Gene: a protein and nucleic acid sequence analysis microcomputer package, PROSITE: a dictionary of sites and patterns in proteins and SWISS-PROT: a protein sequence data bank. Ph.D. thesis, University of Geneva.

Czelusniak,J., Goodman,M., Moncrief,N.D. and Kehoe,S.M. (1990) Maximum parsimony approach to construction of evolutionary trees from aligned homologous sequences. *Methods Enzymol.*, **183**, 601–615.

Dayhoff,M.O., Schwartz,R.M and Orcutt,B.C. (1978) In *Atlas of Protein Sequence and Structure*. National Biomedical Research Foundation, Washington, DC, Vol. 5 Suppl. 3, pp. 345–352.

Feng,D.-F., Johnson,M.S. and Doolittle,R.F. (1985) Aligning amino acids sequences: comparison of commonly used methods. *J. Mol. Evol.*, **21**, 112–125.

Fitch,W.M. and Margoliash,E. (1967) Construction of phylogenetic trees. *Science*, **115**, 279–284.

French,S. and Robson,B. (1985) What is a conservative substitution? *J. Mol. Evol.*, **19**, 171–175.

George,D.G., Barker,W.C. and Hunt,L.T. (1990) Mutation data matrix and its uses. *Methods Enzymol.*, **188**, 333–351.

McLachlan,A.D. Test for comparing related amino acid sequences. Cytochrome c and cytochrome c551. (1971) *J. Mol. Biol.*, **61**, 409–424.

McLachlan,A.D. Repeating sequences and gene duplication in proteins. (1972) *J. Mol. Biol.*, **64**, 417–437.

Needleman,S.B. and Wunsch,C.D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.*, **48**, 443–453.

Overington,J., Johnson,M.S., Šali,A. and Blundell,T.L. (1990) Tertiary structural constraints on protein evolutionary diversity: templates, key residues and structure prediction. *Proc. R. Soc. Lond. B*, **241**, 132–145.

Risler,J.L., Delorme,M.O., Delacroix,H. and Henaut,A. (1988) Amino acid substitutions in structurally related proteins. A pattern recognition approach. *J. Mol. Biol.*, **210**, 181–193.

Woese,C.R. (1969) Models for the evolution of codon assignments. *J. Mol. Biol.*, **43**, 235–240.

Received on October 21, 1991; accepted on December 6, 1991

Current Opinion in STRUCTURAL BIOLOGY

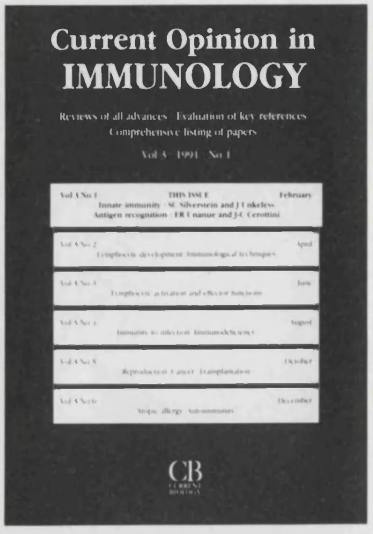
Reprinted from Volume 1 1991

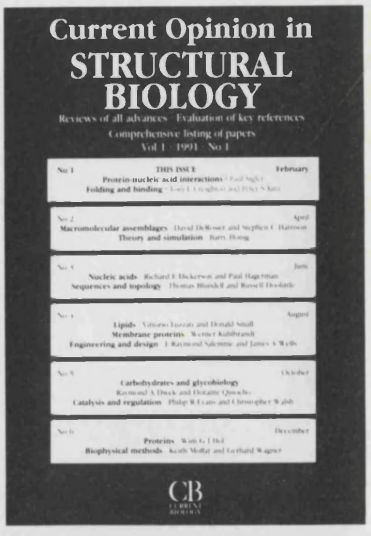
CB
CURRENT
BIOLOGY

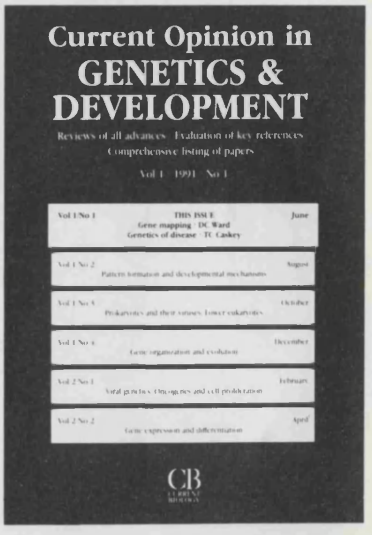
AN EXCELLENT WAY

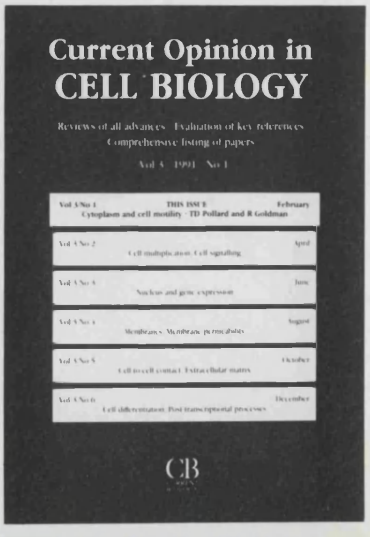
TO STAY INFORMED

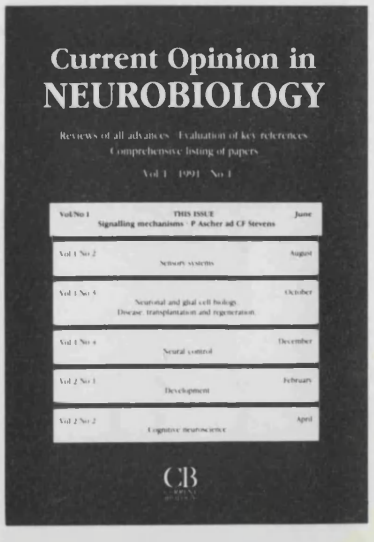
PROFESSOR J. SCHELL, DIRECTOR
MAX PLANCK INST., COLOGNE

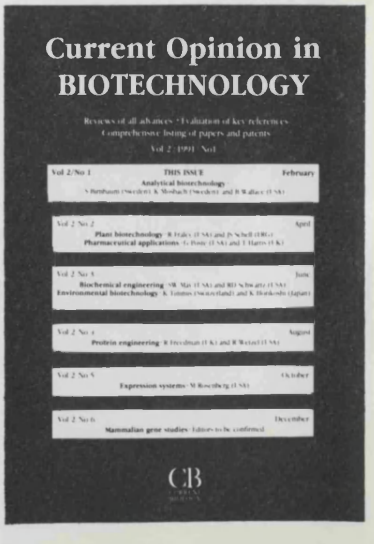












Templates, consensus patterns and motifs

William R. Taylor and David T. Jones

The National Institute for Medical Research, London, UK and University College, London, UK

Current methods in pattern and consensus-sequence matching are reviewed. Attention is focused on those studies in which these methods have been applied to either known structures or structure prediction, including some applications that use machine learning and artificial intelligence. Rather than attempt to cover the wide range of known sequence motifs, examples of Ca^{2+}-binding and DNA-binding motifs are selected.

Current Opinion in Structural Biology 1991, **1**:327–333

Introduction

If a new sequence shares a clear similarity with a protein of known function (and perhaps even structure), then much can be learnt very rapidly by simply recognizing the homology. All too often, however, a search of the sequence databases reveals no significant match, or perhaps only a match to an equally uncharacterized protein. Faced with this situation, two lines of investigation can be pursued: one is to look for fragmentary similarities with other proteins rather than search for a similarity over the whole of the new sequence; the other is to attempt to predict the structure of the new protein. Both approaches rely on identifying characteristic sequence patterns and, where possible, relating these to known structures.

Given several aligned members of a family of proteins, it is possible to construct an average, or consensus, sequence. Instead of a single amino acid code at each alignment position, a histogram is constructed where the numbers of each of the 20 common amino acids occurring at that position are tallied. This allows each sequence to 'vote' for the appropriate residue at any given alignment position. If the consensus sequence is short and well conserved, it may be reasonable to ignore the possibility of inserting gaps in it in order to obtain a good alignment. This greatly simplifies the matching process and the consensus fragment is then usually referred to as a pattern, fingerprint or template. Such patterns are often characterized by a simple scoring scheme; however, many variants can be found and the nomenclature is confused (see Table 1, for some attempt at classification).

Methods

Consensus alignment

Many alignment and consensus-alignment methods have been reviewed in a recent volume of *Methods in Enzymology*. Most of these are established methods that have been developed over the past several years and range from simple tree-based alignments [1] through trees of consensus alignments [2] to simple consensus alignments [3]. Since a flurry of activity a few years ago, there has been little development of consensus-alignment methods and, today, most workers seem to have settled on some form of tree-based condensation of sequences based on a pairwise score. The form of this score is variable, but each method generally has some intrinsic measure that is convenient to use. A more fundamental

Table 1. A classification of pattern-matching methods.

	Matching method		
Position descriptor	Any gap allowed (dynamic programming)	Restricted gaps	No gaps allowed
Histogram	Consensus alignment of profile	Flexible pattern and linked templates	Weight matrix
Match set	Simple consensus or extended pairwise	Linked templates and regular expression	Regular expression fingerprint template
Identity	Standard pairwise alignment	Linked templates and regular expression	As above and left, or peptide or tuple

Abbreviation

AI–artificial intelligence.

 ISSN 0959-440X

problem is how to avoid bias in the composition of the consensus and a recent solution to this old problem has been provided by Sibbald and Argos [4].

Advances in basic pairwise alignment methods are fundamental to the development of methods of consensus- and multiple-sequence alignment. For example, a useful algorithm which may well find application to multiple alignments has been devised by Vingron and Argos [5] to calculate the reliability at each point along a pairwise alignment. Some workers have attempted to increase the reliability of pair alignments still further by incorporating ancillary data, such as predicted secondary structure (giving marginal improvement) [6] or hydrophobicity and solvent-accessibility data [7••]. The latter is very effective but, of course, relies on having a three-dimensional structure for the area data. Using more conventional statistical approaches, others have attempted to reassess the significance of alignment scoring schemes [8,9•,10•]. The latter work has paved the way for a very fast method of searching databanks. The resulting program, BLAST (basic local alignment search tool) [11••], searches around 500 000 amino acid residues per second on a SUN 4/280. Being based on pairwise comparisons, BLAST still falls short of the best pattern or consensus methods with respect to sensitivity, although some steps have been taken to improve this by considering sequence triplets [12].

Pattern matching

Although innovation in the 'classic' consensus-sequence field appears to have slowed slightly, there is clearly still scope for experimentation in the less constrained world of pattern matching. Some developments have combined aspects of the dynamic-programming approach with pattern-matching methods. The resulting method has been referred to by Barton and Sternberg [13••,14] as flexible pattern matching. Sibbald and Argos [15••] have developed a fast multi-faceted pattern-matching tool called SCRUTINEER, and the older 'template' method of Taylor [16] has been fully described and extended to match against multiple alignments [17•]. The latter two methods incorporate a wide variety of features, including predicted secondary structure, that are intended principally to aid searches for tenuous motifs. The method of Taylor also introduces the idea of matching match-sets between a probe (template) and an aligned sequence family. This has the useful property of equating the degree of conservation between two positions as well as the type of conservation. Extending some older work related to the prediction of β/α-barrel structures, Niermann and Kirschner [18] have also combined the various physicochemical properties of these aligned sequences, including secondary-structure propensities and hydrophobicity, to produce an effective signature for this class of structure.

A simpler type of protein-sequence pattern that has recently become popular is based on regular expressions. A regular expression is simply a linear sequence pattern that permits the use of wildcards (matching any residue), set closures (matching residues in a particular set) and gaps (matching a number of residues or none at all). Many sequence motifs have been expressed in this or a similar formalism but they all tend to be relatively simple (i.e. well defined). Using such patterns, Smith and Smith [19••] have developed an automatic pattern maker and search program (PLSEARCH). In their method, a strictly hierarchical grouping of amino acids is used, subdividing the amino acids into small mutually exclusive property sets. They then attempt to boot-strap the pattern creation by aligning each sequence in a clustered family of sequences against a rigid pattern string, according to a binary tree ordered by the pairwise similarity score between each sequence and all others. The examples given of applications of the PLSEARCH pattern-generation algorithm involve only clear alignments, and it is uncertain that the method would cope well with more divergent families of proteins.

Structural patterns

Building on some earlier work, Rooman, Rodriguez and Wodak [20•,21•] analysed the predictive power of sequence patterns in an attempt to gain some insight into the current limitations in structure-prediction accuracy. They correlated simple sequence patterns with the structures of the corresponding peptides and, by extracting such motifs from differently sized data-sets, they surmised that the current lack of success in pattern-based structural prediction is simply due to the limited size of the current structural database. Despite some good correlations of individual patterns with structure, the overall results obtained were no better than average when all motifs were combined into a method for secondary-structure prediction [22].

Following a similar line, others have considered the correlation of larger fragments of structure with sequence. Sternberg and Islam [23] analysed the results of local alignments of protein sequences with known structure, and found that many 'significant' sequence alignments did not have similar structure. Adopting a slightly different viewpoint, Sander and Schneider [24•] compared all structure fragments. They plotted the correspondence of secondary structures for each pair on a graph of percentage identity of the corresponding sequences against different fragment lengths, giving a convenient visual guide to the secondary structural significance of any local alignment (Fig. 1).

Such comparisons are potential sources of new motifs, but also permit the reassessment of familiar motifs, such as the α/β-structure found in the β/α-barrel proteins [25]. As the structures become larger, however, their comparison becomes more difficult and increasingly labour-intensive, tending as a result to incorporate subjective judgments on the definition of the motif. In reanalysing the four-α-helix bundle, Presnell and Cohen [26] examined how past analyses have been selective in this way and provided a good example of how to proceed more rigorously. Remaining elements of subjectivity in such analyses might be minimized by new automatic methods of structure comparison, which are variously

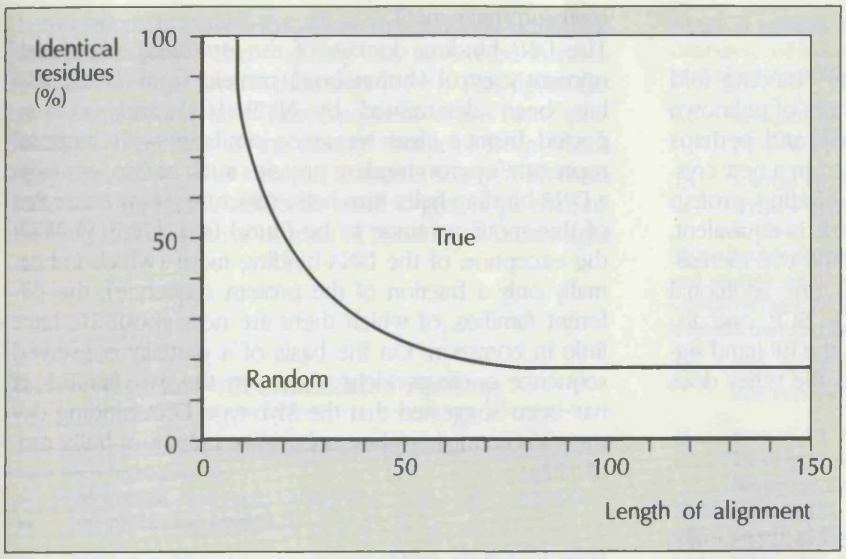


Fig. 1. Plot of the probable structural correspondence of sequence comparison. The structural correspondence of pairs of peptide fragments when plotted according to their measured sequence identity and length, divides the field into two regions, simplistically labelled 'true' and 'random'. In the true region, the secondary structures of the peptide pair correspond over more than 70% of their length, indicating a probable true structural relationship. Adapted from [24•].

based on graph theory [27•,28•], simulated annealing [29•] and recursive dynamic programming [30,31,32•].

Artificial intelligence

Pattern matching is central to fields other than molecular biology; in particular, developments in the field of artificial intelligence (AI) have often been exploited in biocomputing applications. AI methods split neatly into two categories: supervised, and unsupervised methods. Supervised methods require a 'teacher' to provide both questions and answers, whereas unsupervised methods attempt to discover interesting patterns without external influence.

Originally developed as mathematical models of the brain, neural networks have evolved into a general method for complex pattern detection. By presenting many examples of different patterns to a neural network (supervised learning), it may be trained to recognize a particular class of pattern. Neural networks have been applied to several problems in protein-sequence analysis; recent examples are secondary-structure prediction [33•], the identification of accessible residues in proteins [34], and the sensitive detection of immunoglobulin domains [35•].

One problem with neural networks is that they provide no explanation for their conclusions. In contrast, machine-learning techniques (unsupervised learning) attempt to both detect patterns and to explain their significance. Though underused, machine learning has been recently applied to secondary-structure prediction [36•], though the prediction accuracy was no higher than that of traditional statistical methods. Pursuing the automatic definition of motifs, Smith and co-workers [37•] have extended their regular-expression pattern matcher to correlate sequence features with associated biochemical knowledge, and found that the acid-helix motif is associated with transcription activation.

A fundamental problem with automatic pattern searching arises in that, with most protein sequences now being derived from the translation of a nucleic acid sequence only, protein sequences are often functionally classified only by patterns in their sequence. Thus, the criteria of truth (about function) and the patterns that we hope to refine have become interdependent.

Motifs

The number of motifs based only on sequence data is very large and expanding rapidly, and little attempt will be made to provide any guide herein. Readers who wish to pursue this aspect would be best advised to do so electronically and, for this purpose, the best source is the PROSITE database [38]. This well documented library of semi-automatically generated protein-sequence patterns (of the regular expression class) has now been integrated into the SWISS-PROT protein-sequence database, and is distributed by the European Molecular Biology Laboratory (Heidelberg, Germany). Pattern libraries such as PROSITE will almost certainly play an important role in the identification and analysis of the vast number of sequences that will appear as the various large-scale sequencing efforts begin to bear fruit. However, the maintenance of such a library will become increasingly onerous, and may require more automatic methods, such as those based on global alignment [2], templates [17•], regular expressions [19••] or fragments [39•].

For those without convenient access to the computer networks, a paper-based collection of motifs has been produced by Aitken [40]. Some selected, but more detailed, reports can also be found in the 'Sequence Motif' series of the journal *Trends in Biochemical Sciences*, which features reviews on phosphate-binding loops [41,42], and some targeting motifs [43,44].

In the following sections, the current states of knowledge of a few of the more popular motifs, in particular those that bind Ca^{2+} and DNA, are reviewed in the light of recent structural studies.

Calcium-binding motifs

EF-hand

The original motif — the EF-hand Ca^{2+}-binding fold [45] — continues to be found in sequences of unknown structure (e.g. inositol phospholipase [46]) and, perhaps more interestingly, has also been identified in a new crystal structure for a sarcoplasmic Ca^{2+}-binding protein (SCP) [47]. Like parvalbumin, to which it is equivalent, this structure contains two clear motifs and two increasingly distorted motifs (parvalbumin has one additional helix pair which may be a motif relic). In SCP, one distorted motif still binds Ca^{2+} and retains the EF-hand signature in the binding sequence whereas the other does not.

Annexins

The growing annexin (lipocortin or calpactin) family, which was once postulated to have a calmodulin-like (four-helix bundle) fold [48], has been shown by X-ray analysis to have a quite distinct fold of five helices arranged as a super-helix [49]. The structure can also be interpreted as a four-helix bundle in which the two loops in equivalent positions to the two EF-hand motifs in calmodulin run in parallel instead of antiparallel. Even recently, the predicted Ca^{2+}-binding loops were thought to have some sequence similarity with the EF-hand loop [50], but a second crystal structure has shown that although the binding sites are located in the loops (as predicted) they differ from the EF-hand loops in detail [51].

DNA-binding motifs

As a result of the growing application of NMR to protein structure determination, small DNA-binding motifs, recognized for many years from their sequence patterns, now have a structural solution.

Zinc finger

The Cys/His finger motif has (as was predicted in [52,53]) Zn^{2+} held between a β-hairpin and an α-helix structure [54,55]. The motif and many variants continue to be found in a wide variety of proteins (e.g. [56–59]). However, not all finger motifs, such as the $(Cys/Cys)_2$ type found in the steroid receptor [60], have the same structure as the Cys/His motif originally observed.

Leucine zipper

Like the zinc fingers, new leucine zippers abound (e.g. [61,62]) and candidates are now turning up in prokaryotes [63,64] and even in proteins that have no apparent connection with DNA binding [65]. The NMR-determined structures of synthetic fragments [66] and site-directed mutagenesis experiments [67] support the suggested role of these leucine zippers in dimer formation *via* the tight hydrophobic packing of α-helices. Indeed, the motif may simply be a slightly specialized coiled-coil heptad repeat [68].

Helix-turn-helix motif

The DNA-binding domain of the *Antennapedia* development control (homeobox) protein from *Drosophila* has been determined by NMR [69] and, as suspected from a clear sequence similarity with bacterial repressor/operon-binding proteins such as Cro, contains a DNA-binding helix-turn-helix structure. Many examples of this motif continue to be found (e.g. [70,71]). With the exception of the DNA-binding motif (which is normally only a fraction of the protein sequence), the different families, of which there are now about 10, have little in common. On the basis of a partially conserved sequence correspondence between the two families, it has been suggested that the Myb-type DNA-binding domain also contains a homeobox-like helix-turn-helix motif [72].

Helix-loop-helix motif

Not to be confused with the helix-turn-helix motif described above, the helix-loop-helix motif [73] is often typified by its occurrence in the *myc* oncogene product. Although the structure of this motif is unknown, the motif is predicted to be helical and shares similarities with the leucine zippers, having conserved periodic hydrophobic residues (commonly leucine).

Multi-motifs

Like the blood proteins (clotting Factors, etc.), the motifs described above are now being found, in combinations such as the homeobox and finger [74] and EF-hand and finger [75] combinations.

RNA and DNA enzymes

Among the many new motifs describe in the literature, of greatest interest are those that draw together a wide group of proteins, giving rise to a new superfamily. Such an exercise has been performed on the polymerases, unifying all four classes of RNA and DNA dependency [76].

Conclusions

As motifs of ever greater diversity are unified, either by the elucidation of 'missing links' or new computational techniques, a question that recurs is where will it all end — will everything be found to be derived from a few basic structures? We can imagine an ancestral nucleotide-phosphate-binding fold common to kinases, dehydrogenases and other proteins, or an ancient sugar-phosphate-binding fold in the form of the β/α/β structure found in triosephosphate isomerase and at least a dozen other distinct proteins. Similarly, an ancient haem-binding protein has been postulated as the common precursor of both cytochromes and globins. Given the mechanism of evolution, it might be reasonably supposed that the structures we see today ultimately derive from a limited number of basic precursor structures.

It has been proposed that these fundamental units correspond to exons and that they may be limited in number to as few as several thousand [77••]. This figure derives from an analysis of how frequently exons recur, providing an estimate of the size of the pool from which they are drawn. The analysis contains many difficulties but the number is sufficiently small to give hope that, ultimately, the structures of each of these basic units might be known.

References and recommended reading

Papers of special interest, published within the annual period of review, have been highlighted as:

- • of interest
- •• of outstanding interest

1. FENG D-F, DOOLITTLE RF: **Progressive Alignment and Phylogenetic Tree Construction of Protein Sequences.** *Methods Enzymol* 1990, 188:375–402.

2. TAYLOR WR: **Hierarchical Method to Align Large Numbers of Biological Sequences.** *Methods Enzymol* 1990, 188:456–474.

3. GRIBSKOV M, LUTHY R, EISENBERG D: **Profile Analysis.** *Methods Enzymol* 1990, 188:146–159.

4. SIBBALD PR, ARGOS P: **Weighting Aligned Protein or Nucleic Acid Sequences to Correct Unequal Representation.** *J Mol Biol* 1990 216:813–818.

5. VINGRON M, ARGOS P: **Determination of Reliable Regions in Protein Sequence Alignments.** *Protein Eng* 1990, 3:565–569.

6. FISCHEL-GHODSIAN F, MATHIOWITZ G, SMITH TL: **Alignment of Protein Sequences Using Secondary Structure: a Modified Dynamic Programming Method.** *Protein Eng* 1990, 3:577–581.

7. •• BOWIE JU, CLARKE ND, PABO CO, SAUER RT: **Identification of Protein Folds: Matching Hydrophobicity Patterns of Sequence Sets with Solvent Accessibility Patterns of Known Structures.** *Proteins* 1990, 7:257–264.

A novel application of the prediction of residue solvent accessibility from sequence data. Predicted accessibilities from sequences are matched against real accessibilities calculated from protein structures. This could well be a powerful means of assigning the correct fold to a protein sequence family of unknown structure.

8. MOTT RF, KIRKWOOD TBL, CURNOW RN: **Tests for the Statistical Significance of Protein Sequence Similarities in Databank Searches.** *Protein Eng* 1990, 4:149–154.

9. • MOTT RF, KIRKWOOD TBL: **STATSEARCH: a GCG-Compatible Program for Assessing Statistical Significance During DNA and Protein Databases Searches.** *Comput Appl Biosci* 1990, 6:293–295.

Many programs for searching sequence databases make very little attempt at determining the statistical significance of the resulting hits. In contrast, by using parametric methods, STATSEARCH attempts to model the distribution of random search scores to provide a significance guideline.

10. • KARLIN S, ALTSCHUL SF: **Methods for Assessing the Statistical Significance of Molecular Sequence Features by Using General Scoring Schemes.** *Proc Natl Acad Sci USA* 1990, 87:2264–2268.

This statistical analysis forms the core of the BLAST algorithm. The results are valid for alignments that incorporate no insertions or deletions.

11. •• ALTSCHUL SF, GISH W, MILLER W, MYERS EW, LIPMAN DJ: **Basic Local Alignment Search Tool.** *J Mol Biol* 1990, 215:403–410.

BLAST is currently the fastest method for sensitively searching large databases, and is likely to become the method of choice for performing preliminary homology sweeps.

12. ALTSCHUL SF, LIPMAN DJ: **Protein Database Searches for Multiple Alignments.** *Proc Natl Acad Sci USA* 1990, 87:5509–5513.

13. •• BARTON GJ, STERNBERG MJE: **Flexible Protein Sequence Patterns: a Sensitive Method to Detect Weak Structural Similarities.** *J Mol Biol* 1990 212:389–402.

These flexible sequence patterns are very similar to consensus patterns and multiple-sequence profiles. An interesting variation is that several complementary information sources (such as data from biochemical experiments) may be weighted into the final pattern.

14. BARTON GJ: **Protein Multiple Sequence Alignment and Flexible Pattern Matching.** *Methods Enzymol* 1990, 188:403–428.

15. •• SIBBALD PR, ARGOS P: **Scrutineer: a Computer Program that Flexibly Seeks and Describes Motifs and Profiles in Protein Sequence Databases.** *Comput Appl Biosci* 1990, 6:279–288.

An implementation of several useful sequence pattern-matching methods is described.This is one of the first software packages to provide an integrated suite of sequence pattern-matching tools.

16. TAYLOR WR: **Identification of Protein Sequence Homology by Consensus Sequence Alignment.** *J Mol Biol* 1986, 188:233–258.

17. • TAYLOR WR: **A Template Based Method of Pattern Matching in Protein Sequences.** *Prog Biophys Mol Biol* 1989, 54:159–252.

A full description of an older (1986) method, including new extensions to multiple sequences using set/set matching and 'warped' templates.

18. NIERMANN T, KIRSCHNER K: **Improving the Prediction of Secondary Structure of 'TIM-barrel' enzymes.** *Protein Eng* 1990, 4:137–147.

19. •• SMITH RF, SMITH TS: **Automatic Generation of Primary Sequence Patterns from Sets of Related Protein Sequences.** *Proc Natl Acad Sci USA* 1990, 87:118–122.

A method for automatically deriving regular expression patterns from multi-aligned protein sequences is outlined. In place of the commonly used Venn diagram method for classifying amino acids, a more limited hierarchical (tree) approach is used. An important aspect of this paper is the objective reliability assessment of the generated patterns.

20. • ROOMAN MJ, RODRIGUEZ J, WODAK SJ: **Automatic Definition of Recurrent Local Structure Motifs in Proteins.** *J Mol Biol* 19901 213:327–336.

Commonly, amino acid residues are classified into four secondary structural states: α, β, coil and turn. This paper investigates a more flexible way of classifying local protein conformation by means of inter-C_α distances.

21. • ROOMAN MJ, RODRIGUEZ J, WODAK SJ: **Relations Between Protein Sequence and Structure and Their Significance.** *J Mol Biol* 1990, 213:337–350.

An attempt is made to find simple patterns that almost invariably relate to specific local protein conformations (described by inter-C_α distances). Some reliable relationships are detected, though a much larger structure database would be required to provide enough patterns for accurate structure prediction.

22. ROOMAN MJ, WODAK SJ: **Weak Correlation Between Predictive Power of Individual Sequence Patterns and Overall Prediction Accuracy in Proteins.** *Proteins* 1991, 9:69–78.

23. STERNBERG MJE, ISLAM SA: **Local Protein Sequence Similarity Does Not Imply a Structural Relationship.** *Protein Eng* 1990, 4:125–131.

24. • SANDER C, SCHNEIDER R: **Database of Homology-Derived Protein Structures and the Structural Meaning of Sequence Alignment.** *Proteins* 1991, 9:56–68.

An attempt is made to quantify the relationship between local sequence similarity and structural similarity. A database of sequence-structure relationships is described, providing a partial link between entries

in the Brookhaven structure database and the SWISS-PROT sequence database.

25. RICE PA, GOLDMAN A, STEITZ TA: **A Helix-Turn-Strand Structural Motif Common in αβ Proteins.** *Proteins* 1990, 8:334–340.

26. PRESNELL SR, COHEN FE: **Topological Distribution of 4-Alpha-Helix Bundles.** *Proc Natl Acad Sci USA* 1989, 86:6592–6596.

27. • MITCHELL EM, ARTYMIUK PJ, RICE DW, WILLETT P: **Use of Techniques Derived from Graph Theory to Compare Secondary Structure Motifs in Proteins.** *J Mol Biol* 1990, 212:151–166.
Interactions between secondary structures are represented in graph form, allowing existing algorithms for graph/graph matching to be applied. The method may have difficulties with remotely related structures (see [28•]).

28. • ARTYMIUK PJ, RICE DW, MITCHELL EM, WILLET P: **Structural Resemblance Between the Families of Bacterial Signal-Transduction Proteins and of G Proteins Revealed by Graph Theoretical Techniques.** *Protein Eng* 1990, 4:39–43.
Application of the graph-comparison method finds that the Che-Y chemotaxis protein best matches the structure of elongation factor Tu. Other groups have found Che-Y to be more like flavodoxin, both in sequence [7••] and structure as, unlike elongation factor Tu, flavodoxin and Che-Y share almost identical topology.

29. • OVERINGTON J, JOHNSON MS, ŠALI A, BLUNDELL TL: **Tertiary Structural Constraints on Protein Evolutionary Diversity: Templates Key Residues and Structure Prediction.** *Proc R Soc Lond [B]* 1990, 241:132–145.
Mutability matrices derived from sequence data alone have been used for some time in protein-sequence analysis. This paper describes a set of matrices derived from structural data, in which each matrix relates to a particular structural environment, such as an 'inaccessible β-strand'. Classifying the pattern of mutability in aligned sequences using these matrices may provide a means of predicting the structural roles of residues.

30. TAYLOR WR, ORENGO CA: **Protein Structure Alignment.** *J Mol Biol* 1989, 208:1–22.

31. TAYLOR WR, ORENGO CA: **A Holistic Approach to Protein Structure Alignment.** *Protein Eng* 1989, 2:505–519.

32. • ORENGO CA, TAYLOR WR: **A Rapid Method of Protein Structure Alignment.** *J Theor Biol* 1990, 147:517–551.
The latest and (100-fold) faster variant of this method, based on the simple dynamic-programming algorithm used to align sequences. Speed is increased by selecting subsets of potentially equivalent residues on which to base the initial comparison. These are then refined by iterative application.

33. • KNELLER DG, COHEN FE, LANGRIDGE R: **Improvements in Protein Secondary Structure Prediction by an Enhanced Neural Network.** *J Mol Biol* 1990, 214:171–182.
A new twist to the neural-net approach by biasing the network to recognize periodic patterns — clearly, an addition that would be expected to, and does, help predict secondary structure. The best results are obtained when the structural class of the protein is known.

34. HOLBROOK SR, MUSKAL SM, KIM S-H: **Predicting Surface Exposure of Amino Acids from Protein Sequence.** *Protein Eng* 1990, 3:659–666.

35. • BENGIO Y, POULIOT Y: **Efficient Recognition of Immunoglobulin Domains from Amino Acid Sequences Using a Neural Network.** *Comput Appl Biol* 1990, 6:319–324.
The authors describe an interesting hybrid method that uses both a neural network and dynamic programming to sensitively detect immunoglobulin domains. The network detects patterns of residues specific to selected β-strands, with dynamic programming being used to find optimal combinations of these pattern matches.

36. • KING RD, STERNBERG MJE: **Machine Learning Approach for the Prediction of Protein Secondary Structure.** *J Mol Biol* 1990, 216:441–457.
By means of a machine-learning program (PROMIS), rules are generated that enable secondary structure to be predicted from amino acid sequence. Some of the emerging rules are familiar whereas others are novel. An overall prediction of 60% is achieved.

37. • ZHU Q-L, SMITH TF, LATHROP RH, FIGGE J: **Acid Helix-Turn Activator Motif.** *Proteins* 1990, 8:156–163.
A pattern matcher is 'seeded' with potential motifs for transcriptional activator sequences and these are refined automatically. The paper contains a useful evaluation of pattern sensitivity *versus* specificity.

38. BAIROCH A: *PC/Gene: a Protein and Nucleic Acid Sequence Analysis Microcomputer Package, PROSITE: a Dictionary of Sites and Patterns in Proteins, and SWISS-PROT: a Protein Sequence Data Bank [dissertation].* Geneva: University of Geneva, 1990.

39. • SETO Y, IKEUCHI Y, KANEHISA M: **Fragment Peptide Library for Classification and Functional Prediction of Proteins.** *Proteins* 1990, 8:341–351.
This paper describes an automatically generated library of sequence fragments that appear to be unique to different protein families.

40. AITKEN A: *Identification of Protein Consensus Sequences — Active Site Motifs, Phosphorylation and Other Post-Translational Modifications.* New York: Ellis Horwood, 1990.

41. SARASTE M, SIBBALD PR, WITTINGHOFER A: **The P-loop — a Common Motif in ATP and GTP Binding.** *Trends Biochem Sci* 1990, 15:430–434.

42. KEMPT BE, PEARSON RB: **Protein Kinase Recognition Sequence Motifs.** *Trends Biochem Sci* 1990, 15:342–346.

43. PELHAM HRB: **The Retention Signal for Soluble Proteins of the Endoplasmic Reticulum.** *Trends Biochem Sci* 1990, 15:483–486.

44. DICE JF: **Peptide Sequences that Target Cytosolic Proteins for Lysosomal Proteolysis.** *Trends Biochem Sci* 1990, 15:305–309.

45. MONCRIEF ND, KRETSINGER RH, GOODMAN M: **Evolution of EF-Hand Calcium-Modulated Proteins: Relationships Based on Amino Acid Sequences.** *J Mol Evol* 1990, 30:522–562.

46. BAIROCH A, COX JA: **EF-Hands Motifs in Inositol Phospholipid-Specific Phospholipase C.** *FEBS Lett* 1990, 269:454–456.

47. COOK WJ, EALICK SE, BABU YS, COX JA, VIJAY-KUMAR S: **Three Dimensional Structure of a Sarcoplasmic Calcium-Binding Protein from *Nereis diversicolor.*** *J Biol Chem* 1991, 266:652–656.

48. TAYLOR WR, GEISOW MJ: **Predicted Structure for the Calcium-Dependent Membrane-Binding Proteins p35, p36, and p32.** *Protein Eng* 1987, 1:183–187.

49. HUBER R, ROMISCH J, PAQUES E-P: **The Crystal and Molecular Structure of Human Annexin V, an Anticoagulant Protein that Binds to Calcium and Membranes.** *EMBO J* 1990, 9:3867–3874.

50. MOSS SE, CRUMPTON MJ: **The Lipocortins and the EF-Hand Proteins: Calcium-Binding Sites and Evolution.** *Trends Biochem Sci* 1990, 15:11–12.

51. HUBER R, SCHNEIDER M, MARY I, ROMISCH J, PAQUES E-P: **The Calcium Binding Sites in Human Annexin V by Crystal Structure Analysis at 2.0 Å Resolution.** *FEBS Lett* 1990, 275:15–21.

52. BERG J: **Proposed Structure for the Zn-Binding Domains from Transcriptional Factor IIIA and Related Proteins.** *Proc Natl Acad Sci USA* 1988, 85:99–102.

53. GIBSON TJ, POSTMA JPM, BROWN RS, ARGOS P: **A Model for the Tertiary Structure of the 28 Residue DNA-Binding Motif ('Zinc-Finger') Common to Many Eukaryotic Transcriptional Regulatory Proteins.** *Protein Eng* 1988, 2:209–218.

54. LEE MS, GIPPERT GP, SOMAN KV, CASE DA, WRIGHT PE: **Three-Dimensional Structure of a Single Zinc Finger DNA-Binding Domain.** *Science* 1990, 245:635–637.

55. KLEVIT RE, HERRIOTT JR, HORVATH SJ: **Solution Structure of a Zinc Finger Domain of Yeast ADR1.** *Proteins* 1990, 7:215–226.

56. WEISS MA, MASON KA, DAHL CE, KEUTMANN HT: **Alternating Zinc-Finger Motifs in the Human Male-Associated Protein Zfy.** *Biochemistry* 1990, 29:5660–5664.

57. LEGRAIN P, CHOULIKA A: **The Molecular Characterization of Prp6 and Prp9 Yeast Genes Reveals a New Cystine Histidine Motif Common to Several Splicing Factors.** *EMBO J* 1990, 9:2775–2781.

58. OPIPARI AW, BOGUSKI MS, DIXIT VM: **The A20 cDNA Induced by Tumor Necrosis Factor α Encodes a Novel Type of Zinc Finger Protein.** *J Biol Chem* 1990, 265:14705–14708.

59. LIEBHABER SA, EMERY JG, URBANEK M, WANG X, COOKE NE: **Characterization of a Human cDNA Encoding a Widely Expressed and Highly Conserved Cysteine-Rich Protein with an Unusual Zinc-Finger Motif.** *Nucleic Acids Res* 1990, 18:3871–3879.

60. HARD T, KELLEBACH E, BOELENS R, MALER BA, DAHLMAN K, FREEDMAN LP, CARLSTEDT-DUKE J, YAMAMOTO KR, GUSTAFSSON J-A, KAPTEIN R: **Solution Structure of the Glucocorticoid Receptor DNA-Binding Domain.** *Science* 1990, 249:157–160.

61. GUILTIMAN MJ, MARCOTTE WR, QUATRANO RS: **A Plant Leucine Zipper Protein that Recognises an Abscisic Acid Response Element.** *Science* 1990, 250:267–271.

62. LIOU HC, BOOTHBY MR, FINN PW, DAVIDSON R, NABAVI N, ZELEZNIK-LE NJ, TING JPY, GLIMCHER LH: **A New Member of the Leucine Zipper Class of Proteins that Bind to the HLA DRα Promoter.** *Science* 1990, 247:1581–1584.

63. GIRALDO R, NIETO C, FERNANDEZ-TRESGUERRES M-E, DIAZ R: **Bacterial Zipper.** *Nature* 1989, 342:866.

64. MAXON ME, WIGBOLDUS J, BROT N, WEISSENBACH H: **Structure Function Studies on *E. coli* MetR Protein, a Putative Prokaryotic Leucine Zipper Protein.** *Proc Natl Acad Sci USA* 1990, 87:7076–7079.

65. WEBBER AN, MALKIN R: **Photosystem-I Reaction Centre Proteins Contain Leucine Zipper Motifs: a Proposed Role in Dimer Formation.** *FEBS Lett* 1990, 264:1–4.

66. SAUDEK V, PASTORE A, GASTIGLIONE-MORELLI MA, FRANK R, GAUSEPOHL H, GIBSON T, WEIH F, ROESCH P: **Solution Structure of the DNA-Binding Domain of the Yeast Transcriptional Activator Protein GCN4.** *Protein Eng* 1990, 4:3–10.

67. HU JC, O'SHEA EK, KIM PS, SAUER RT: **Sequence Requirements for Coiled-Coils: Analysis with λ Repressor–GCN4 Leucine Zipper Fusions.** *Science* 1990, 250:1400–1403.

68. COHEN CP, PARRY DAD: **Alpha-Helical Coiled Coils and Bundles: How to Design an α-Helical Protein.** *Proteins* 1990, 7:1–15.

69. BILLETER M, QIAN YQ, OTTING G, MULLER M, GEHERING WJ, WUTHRICH K: **Determination of the Three-Dimensional Structure of the *Antennapedia* Homeodomain from *Drosophila* in Solution by (1)H-Nuclear Magnetic Resonance Spectroscopy.** *J Mol Biol* 1990, 214:183–187.

70. KIM Y, MIRENBERG M: ***Drosophila* NK-homeobox Genes.** *Proc Natl Acad Sci USA* 1989, 86:7716–7720.

71. WEDEEN CJ, KOSTRIKEN RG, MATSUMURA I, WEISBLAT DA: **Evidence for a New Family of Evolutionary Conserved Homeobox Genes.** *Nucleic Acids Res* 1990, 18:1908.

72. FRAMPTON J, LEUTZ A, GIBSON TJ, GRAFF T: **DNA-Binding Domain Ancestry.** *Nature* 1989, 343:134.

73. MURRE C, SCHONLEBER-MCCAW P, BALTIMORE D: **A New DNA-Binding and Dimerization Motif in the Immunoglobulin Enhancer Binding 'Daughterless' MyoD and Myc Proteins.** *Cell* 1989, 56:777–783.

74. KARLSSON O, THOR S, NORBERG T, OHLSSON H, EDLUND T: **Insulin Gene Enhancer Binding Protein Isl-1 is a Member of a Novel Class of Proteins Containing Both a Homeo and a Cys-His Domain.** *Nature* 1990, 344:879–881.

75. SAKANE F, YAMADA K, KANOH H, YOKOYAMA C, TANBE T: **Porcine Diacylglycerol Kinase Sequence has Zinc Finger and EF-Hand Motifs.** *Nature* 1990, 344:345–347.

76. DELARUE M, POCH O, TORDO N, MORAS D, ARGOS P: **An Attempt to Unify the Structure of Polymerases.** *Protein Eng* 1990, 3:416–417.

77. •• DORIT RL, SCHOENBACH L, GILBERT W: **How Big is the Universe of Exons?** *Science* 1990, 250:1377–1382.

The frequency with which exons recur leads to an estimate of the size of the pool from which they are drawn. From the comparison of sequences within a database of exons, the number of significant alignments found suggests an underlying pool of 1000–7000 members. Despite the great care taken with the statistics, some difficulties remain.

WR Taylor, Laboratory of Mathematical Biology, The National Institute for Medical Research, The Ridgeway, Mill Hill, London NW7 1AA, UK.

DT Jones, Biomolecular Structure and Modelling Unit, Biochemistry and Molecular Biology Department, University College, Gower Street, London WC1E 6BT, UK.

Comparisons of protein structures

Mark S. Johnson

Imperial Cancer Research Fund and Birkbeck College, London, UK

The structures of proteins related by evolution are remarkably alike even when the observed sequence similarities are statistically marginal or seemingly non-existent. Similar protein substructures are found in proteins for which there is no evidence of common ancestry and no similarity in their global topology. Recent advances in the comparison of whole proteins, together with the comparison and analysis of their parts, have paved the way for the use of structural information in prediction and modelling, protein engineering, structure and sequence alignments, and investigations of protein evolution, among a host of other applications.

Current Opinion in Structural Biology 1991, **1**:334–344

Introduction

It has been said time and time again that the tertiary structures of proteins are more conserved than their corresponding amino acid sequences. This can be seen if one makes a comparative analysis of these two sets of data from a family of related proteins (Figs. 1 and 2). Long ago, Rossmann and colleagues (see [1] for a review) showed that only a few aligned sequence positions were invariant among the known dehydrogenase and kinase structures, although the structures of the nucleotide-binding regions clearly had similar tertiary structures. A recent example of this phenomenon was reported by Pastore and Lesk [2•] who examined globin and phycocyanin structures of which the structural similarity had previously been noted. They have made a case for a distant evolutionary relationship and have based this on structural arguments: the topologies, and especially the local interactions, are indeed very similar yet a relationship could not be established on the basis of the amino acid sequences alone.

In this review, I shall present some of the more recent developments in the automated comparison of three-dimensional structures of proteins. It has been common to compare similar proteins (i.e. structures whose pairwise sequence identity is greater than about 40 %) by the superposition of their structures as rigid bodies, which will usually align a majority of each structure. Some portions of protein structures will not closely superimpose, however, and this usually occurs in regions where more of the differences in amino acid sequence are apparent after sequence alignment, as well as where insertions/deletions (the 'gaps') appear. The residues concerned are mostly located at the surface of proteins, exposed to solvent and frequently part of more mobile loop structures, and naturally have more freedom to accept mutations than residues located within the protein interior (Fig. 1) [3,4].

When larger differences are found between structures, such as when movements have altered the relative positions of stretches of secondary structure or entire domains, the protein folds can still be recognized [5,6], yet direct superposition techniques may produce results with high uncertainty or fail altogether (Fig. 2) [7••]. These disadvantages that are associated with the technique of rigid-body superposition have been recognized and have led to alternative procedures that rely on techniques such as dynamic programming and graph theory, not forgetting the usefulness of segmental comparisons and the direct visualization of structures on computer graphics devices.

The automated procedures discussed in this review can be used to compare both substructures and the coordinates of entire proteins, and also search collections of protein structures themselves. The publically available assemblage of coordinates of protein three-dimensional structures, the Protein Data Bank, Brookhaven National Laboratory, New York, USA, contains a wealth of information, and the analysis of earlier comparisons made for related and unrelated sets of these structures has already revealed features important to understanding the complex nature and variety of protein structures (for general reviews on protein structures, substructures and the evolution of tertiary structures, see [8–10]). In this review, I shall report on a number of recent studies in which both automated methods and the laborious visualization of structures *via* graphics display devices have been employed. The valuable knowledge base gathered so far from these analyses has many potential applications, some of which I will also highlight.

THE CURRENT OPINION SERIES

CLINICAL MEDICINE

Current Opinion in:	Personal (circle one)	Institutional (circle one)
☐ Anaesthesiology (Vol 4, 6 issues)	£75/$145	£150/$255
☐ Cardiology (Vol 6, 6 issues)	£60/$100	£100/$170
☐ Dentistry (Vol 1, 6 issues)	£60/$100	£100/$170
☐ Gastroenterology (Vol 7, 6 issues)	£60/$100	£100/$170
☐ Infectious Diseases (Vol 4, 6 issues)	£60/$100	£100/$170
☐ Lipidology (Vol 2, 6 issues)	£65/$125	£115/$195
☐ Neurology & Neurosurgery (Vol 4, 6 issues)	£85/$145	£150/$255
☐ Obstetrics & Gynecology (Vol 3, 6 issues)	£70/$125	£125/$215
☐ Oncology (Vol 3, 6 issues)	£85/$145	£160/$270
☐ Ophthalmology (Vol 2, 6 issues)	£120/$195	£200/$340
☐ Orthopaedics (Vol 2, 6 issues)	£90/$145	£180/$290
☐ Pediatrics (Vol 3, 6 issues)	£60/$100	£100/$170
☐ Psychiatry (Vol 4, 6 issues)	£85/$145	£130/$230
☐ Radiology (Vol 3, 6 issues)	£85/$145	£170/$290
☐ Rheumatology (Vol 3, 6 issues)	£60/$100	£100/$170

BIOLOGICAL SCIENCES

☐ Biotechnology (Vol 2, 6 issues)	£200/$350	£300/$500
☐ Cell Biology (Vol 3, 6 issues)	£65/$120	£150/$275
☐ Genetics & Development (Vol 1, 4issues)	£45/$80	£100/$185
☐ Immunology (Vol 3, 6 issues)	£65/$120	£150/$275
☐ Neurobiology (Vol 1, 4issues)	£45/$80	£100/$185
☐ Structural Biology (Vol 1, 6 issues)	£90/$175	£190/$350

☐ 10% discount if you subscribe to more than one title

(excluding Students/ Housemen/ Residents - see below)

Method of Payment

Total order value: .....

(including any discount for which you are eligible)

☐ Cheque enclosed, payable to Current Biology Ltd

☐ VISA ☐ MasterCard ☐ Am Ex

Card No ____ Exp Date ____

Name:

Address:

Postcode: Country:

Signature ..... Date .....

No risk guarantee

Your money will be refunded in full if you write cancelling your journal subscription within 30 days of receiving the first issue.

Send to:

USA & Canada
Current Biology
20 North 3rd Street,
Philadelphia, PA 19106-2113,
USA Tel: (215) 574 2266

Rest of world
Current Biology Ltd
34-42 Cleveland Street,
London, W1P 5FB
UK Tel: (071) 323 0323

US$ prices applicable to subscribers in USA & Canada only
Subscription rates include airspeed delivery.
Subscriptions run on a calendar year basis.

☐ Please send me an institutional invoice

☐ Please send me a sample copy of the title(s) circled above

☐ 30% discount for Students/ Housemen/ Residents. Deduct 30% from the prices above and provide proof of status

☐ Bank transfer (details available upon request

☐ Please send me information on your electronic databases - available for all titles