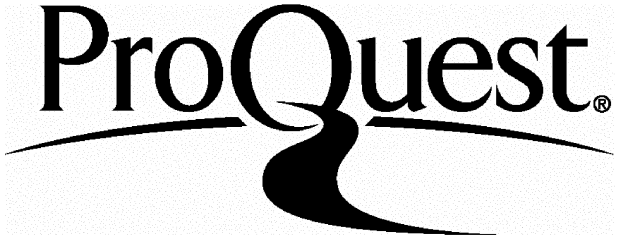
Consensus Templates for Protein Structure Recognition

Ian Sillitoe

Biomolecular Structure and Modelling Unit
Department of Biochemistry and Molecular Biology
University College London

A thesis submitted to the University of London in the
Faculty of Science for the degree of Doctor of Philosophy

December 2002



Abstract

Molecular biology has moved into the new millennium with the human genome sequenced and publicly available. The challenge now facing the bioinformatics field is to assign structure and functional information to protein sequences generated by this and many other genomic projects. To meet this challenge, several structural genomics initiatives are currently underway with the aim of providing, where possible, a protein structure within homology modelling distance for every known sequence. As a result, structure classification databases will need to provide novel methods in order to cope with this high influx of structures.

This thesis presents work on the classification, analysis and recognition of protein structures using the CATH protein structure classification database. Structural similarity is measured by comparing contact maps, or the points of contact between amino acid residues. By examining related structures, it has been possible to identify contacts that have been highly conserved during the process of evolution. Protocols to generate accurate multiple structure alignments and 3D templates based on consensus contact patterns found in these alignments have been developed. Templates have been generated for all homologous superfamilies in CATH to create a library of unique and identifying 'fingerprint' patterns.

These templates were applied to the recognition of models generated at an early stage of *ab initio* protein structure prediction. Scanning these early models against a library of templates describing conserved contacts allowed the most likely superfamily to be identified. An algorithm was also written that performed fold recognition using only a limited set of contacts with the purpose of application to the early stages of experimental NMR structure determination.

Finally, the multiple structural alignments have been used to generate a library of hidden Markov models (HMMs). These structure-based sequence profiles were thoroughly benchmarked using a strict dataset of remote homologues and appear to outperform other commonly used sequence methods.

This work was generously supported by the Biotechnology and Biological Sciences Research Council.

Acknowledgements

I would like to take this opportunity to thank all the people who have contributed both to my academic research and to my sanity throughout the last four years. I owe a great debt of gratitude to my supervisor Christine Orengo who has been so generous with her time, support, and good humour for the duration of my PhD. I would also like to thank Janet Thornton who should take a great deal of credit for her part in making the BSM department at UCL such a friendly and stimulating environment to work in.

There are many people in the department that I would like to acknowledge and thank. The first would have to be the CATH Father himself, Dr. James Bray. James has done his best to keep me organised (a thankless task) and also on my toes (both scientifically and socially), not to mention making himself personally responsible for me finishing this thesis (many people, including myself, are sincerely grateful for this). Also, to Gabby Reeves, who made me porridge when I was writing this thesis. I have absolutely no idea how you put up with me for so long but many thanks, especially for laughing so earnestly at even my most appalling jokes. Many thanks also to Daniel Buchan, who has provided an endless stream of entertainment and thoroughly useless trivia (see also distraction). Stuart Rison and Simon Bergqvist deserve great praise for managing to share an office with me whilst writing this thesis. In particular, Stuart Rison provided me with a means of drinking ridiculously strong coffee at all times of the day and for this I am sincerely grateful.

There are a number of other people who have contributed a great deal to the general atmosphere at UCL over the years. My thanks go to the old crowd: Andreas Brakoulias, Tina Clarke, Jennifer Dawe, Brian Ferguson, Alistair Grant, Andrew Harrison, Thomas "Hot Chocolate" Kabir, Frances Pearl, Mike Plevin, Ollie Redfern, Adrian Shepherd and Annabel Todd. Also to the new crowd: Sarah Addou, Juan Antonio, Chris Bennett, Ilhem Diboun, Mark Dibley, Adrian EdoUkeh, Stefano Lise and Stathis Sideris.

To the people who have now left the lab: I have very much enjoyed my collaboration with Xavier de la Cruz, mainly due to his ceaseless scientific enthusiasm, but also due to the trips to Barcelona; Andrew Martin, Roman Laskowski and William Valdar have contributed to my progress in computer programming; also Gail Bartlett, Richard Jackson, Kevin Murray and Gordon Whamond have all made valuable contributions to both the academic and social ethos of UCL.

To the people behind the scenes: Jahid Ahmed, Donovan Binns, John Bouquiere, Duncan McKenzie and Jesse Oldershaw have displayed commitment above and be-

yond the call of duty in keeping the computers and networks in good working order throughout the last 4 years.

To the people who attempted to keep me sane: Adam Sills, the infamous duo of Neal Houghton and Neil Kerber, Tom Knapp and Ushma Vishram, Jack, Mike Williams and Harriet and Jon Riches, also to James Garvey and the rest of UCLU Jitsu Club who have provided excellent stress-relief. Rob and Sally Cray deserve a special mention for their kindness and support over the last few years.

Many thanks also go to the Palmers: to Carolyn and Bridget for support and encouragement, to Derek for cooking fantastic breakfasts, to Edward for trying to explain the difference between semi-colons and colons (it's not his fault I didn't listen) and to Joseph for trying to teach me how to play golf (see also turf hacking).

My family deserve a huge amount of credit for offering just about every kind of support there is to offer. During the last few years my parents have performed admirably as babysitters, taxi-drivers, bank managers, waiters, chefs and wedding organisers. I am very grateful for all their support. Thanks also to my sister Claire and her husband Faisal for their love and generosity and for accepting the panicked, baby-related early morning phone calls with such good grace. Also, particular thanks go to my Grandparents, Stanhope and Joan Blaikley, for their constant support throughout my PhD.

Finally, I would also like to make a very special thank you to my long suffering fiancée/wife (depending on when I finish) Emma Sillitoe. Em has made writing this thesis possible through her consistent love, support and (gentle) encouragement. The last person to thank is my baby daughter Lauren Emily Sillitoe, who has been of absolutely no help whatsoever in finishing this PhD and I wouldn't have it any other way.

This work is dedicated to my Nan, Phyllis Sillitoe (1917–2002). She would have read it and said "It's lovely, dear." (and meant it).

Ian Sillitoe, December 2002

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Proteins

1.1.1 Background

Proteins form the basis of almost all biological processes (Stryer, 1995). The huge range of functions mediated by these remarkable molecules includes catalysis, transport, mechanical support and molecular recognition. In each case the functional mechanism is closely related to the three-dimensional structure of the protein. Thus, knowledge of the protein structure is essential in order to fully understand the mechanisms by which these functions are achieved at a molecular level. In addition, understanding the structural basis of these functional mechanisms allows rational drug design to specifically target and modify the behaviour of proteins occurring in either a defective or unwanted biochemical pathway.

As a result of this biological importance, protein structure has been the subject of intense academic scrutiny for the majority of the twentieth century. During the early 1930's, W. T. Astbury demonstrated that human hair gave a characteristic X-ray diffraction pattern and that this pattern changed dramatically when the hair was physically stretched (Fundamentals of Fibre Structure, 1933). In 1951, L. Pauling used these diffraction results to make the prediction that proteins form spring-like α-helices with 3.6 amino acid residues per turn. The use of X-ray diffraction as an experimental tool in the field of biophysics continued with the first three-dimensional structure of the protein myoglobin reported in 1958 (Kendrew *et al.*, 1958).

At the end of 2002, the protein data bank (PDB) held the atomic co-ordinates of over 19,000 protein structures determined by experimental techniques such as X-ray crystallography or nuclear magnetic resonance (NMR) spectroscopy. However, this number is almost three orders of magnitude fewer than the number of sequences

in the contemporary sequence databases (the GenBank nucleic acid database Benson *et al.* (1996) contained over 18,000,000 sequence records in November, 2002). Experimental structure determination will certainly struggle to keep up with the explosion of sequence data from various large-scale genome sequence projects, so it is of great importance to accelerate these techniques and automate assignment of structure and function from sequence where possible. The challenge facing biologists is to discover the function of these proteins individually and how they work in concert to form the biochemical machinery of life.

1.1.2 Protein Structure

In order to understand the principles and features of protein structure it can be helpful to dissect a typical structure into its components. Protein structure is often considered in four hierarchical levels of complexity. The first level, called the primary structure, is the sequential chain of amino acids, or residues, in the protein. Local regions of this sequence tend towards distinct geometrical forms such as α-helices, β-strands and random coil, and are known as secondary structure elements (SSE). These local secondary structures can be seen to act as a scaffold as they pack together into a global 3D tertiary structure, known as the protein fold. The fourth level of complexity, termed quaternary structure, describes a collective structure containing more than one separate polypeptide chain.

Despite the almost infinite possibilities for conformations of protein structure from amino acid sequence, only a relatively small number of structural arrangements, or folds, have been observed (less than 800). In 1992, Chothia proposed there would be "no more than 1000 families for the molecular biologist" (Chothia, 1992). One limit on the number of folds available is of a physical nature as there are only a relatively small number of ways to pack a given set of secondary structure elements into a compact, globular form (Finkelstein & Ptitsyn, 1987). An additional reason is the likelihood that all modern proteins have evolved from a small set of common ancestors.

1.1.3 Structural Domains

It is common for a single tertiary structure to fold into two or more structurally distinct regions, known as domains. These structural domains are compact, semi-independent folding units and are often seen recurring in different multidomain proteins (see figure 1.1). It is likely that domains are an important evolutionary unit so for purposes of structure comparison and analysis it is more useful to consider these

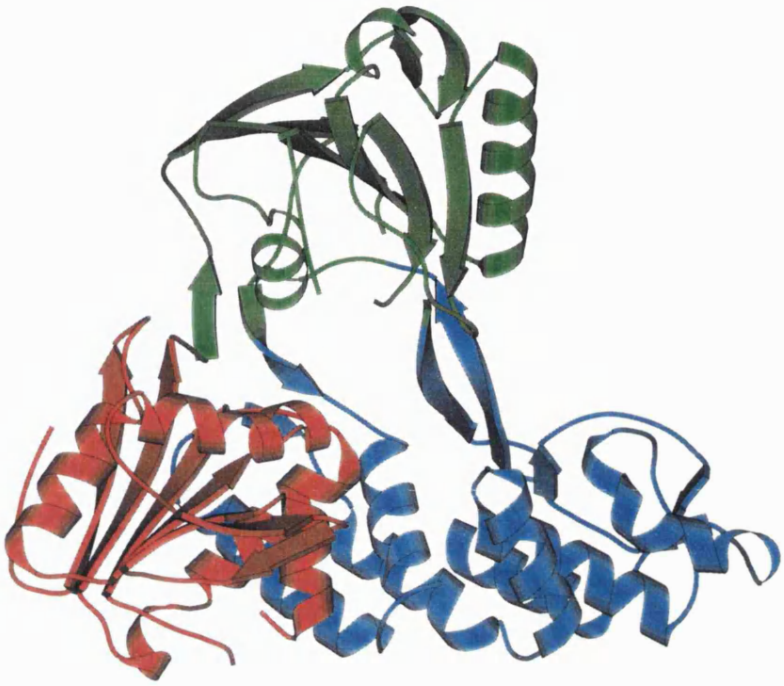


Figure 1.1: MOLSCRIPT (Kraulis, 1991) representations of the three structural domains found in a RNA-helicase protein from Hepititis C virus (PDB structure 1A1V, chain A).

domains individually. Nearly half the known globular structures are multidomain, the majority comprising two domains, though examples of 3, 4, 5, 6 and 7 domain proteins have also been determined. Structural domains do not always occur in sequential order along the protein chain as a domain can consist of two or more non-local sequence fragments.

1.2 Evolutionary Relationships

1.2.1 Identifying Evolutionary Relationships

A common task in the post-genomic bioinformatics era is to extract as much structural and functional information as possible for a target protein sequence in a fast and automated manner. The most efficient method for obtaining this information is through identifying an evolutionary relationship to a protein previously characterised through experimental techniques. Two proteins that are related by evolution, i.e. that share a common ancestor, are termed homologues. Confidently identifying a homologous relationship allows structural, and possibly functional, assignments to

be passed on to sequences for which no annotation exists.

An evolutionary relationship between two proteins is often demonstrated by identifying a significant similarity between the amino acid sequences, tertiary structures or functional mechanisms of the proteins. The measure of significance depends on the confidence required for the assignment of homology, however it is usually based on the likelihood of an equivalent similarity occurring by chance.

1.2.2 Sequence Similarity

A simple measure of similarity between two protein sequences is the number of identical residues one sequence shares with another, i.e. percentage identity. Sequences sharing more than a given threshold value (usually around 30% identity) can be assigned as homologues since it is highly unlikely that this degree of similarity could have occurred by chance. However, when the identity drops below this value it is difficult to assign homology, depending on the size of the protein (see also section 5.2.4.2). This threshold is commonly referred to as the 'Twilight Zone' of sequence similarity (Doolittle, 1986).

Therefore, in order to detect more distantly related proteins, i.e. those having less than 30% idenities in their sequences, sequence alignments are often scored according to the number of aligned residues that share similar properties rather than identities. The probabilities of residue substitutions being accepted and proliferated during protein evolution are summarised in substitution matrices.

1.2.3 Substitution Matrices

1.2.3.1 Based on Amino Acid Properties

All 20 naturally occurring amino acids have distinct chemical structures, however these can be grouped together based on shared chemical and physical properties (see figure 1.2). The difference in physicochemical properties has implications for the evolutionary tolerance of certain amino acid substitutions over others. For example, if a leucine were to replace a valine, the resultant effect on the overall stability of the protein structure would be minimal since both residues are of a similar chemical nature and physical size. However, replacing a valine with a phenylalanine may not be accommodated so easily within the structure due to steric hindrance. Such a drastic residue substitution could destabilise the local packing around the residue and is therefore likely to be selected against.

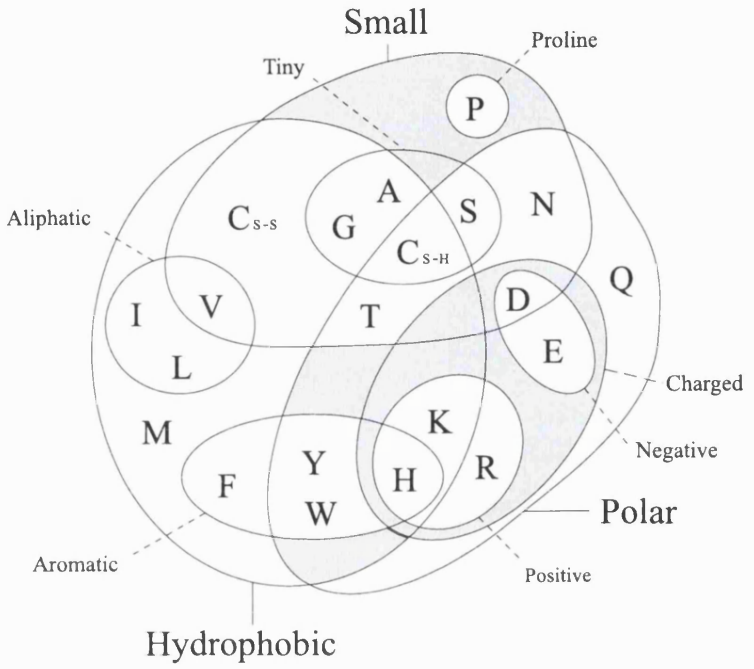


Figure 1.2: A Venn diagram describing the chemical and physical properties of amino acids (Taylor, 1986a). The residues are alanine (A), cysteine (C), aspartic acid (D), glutamic acid (E), phenylalanine (F), glycine (G), histidine (H), isoleucine (I), lysine (K), leucine (L), methionine (M), asparagine (N), proline (P), glutamine (Q), arginine (R), serine (S), threonine (T), valine (V), tryptophan (W) and tyrosine (Y).

1.2.3.2 Based on Observed Mutations

When comparing two proteins, residue substitution probabilities can be used as a more sensitive assessment of similarity than simple amino acid identity. Thus, accounting for the likelihood of specific amino acid substitutions allows distant evolutionary relationships to be modelled more accurately.

The likelihood of a given residue substitution can be quantified in a mutation data matrix (MDM). This 2D matrix describes the probability of each of the 400 potential amino acid substitutions occurring in nature and is based on the observed frequencies of mutations in related sequences. For each amino acid, the 20 substitution probabilities are derived by examining a large number of closely related sequences and counting the occurrence of each type of residue substitution.

Dayhoff or Point Accepted Mutation (PAM) Matrices

The PAM similarity matrices were generated by Dayhoff by initially providing alignments of closely related sequences, i.e. >85% sequence identity (Dayhoff, 1978).

These alignments were global alignment, i.e. encompassing the entire length of the sequences. As a result, both highly conserved regions and more variable regions were included in both the alignments and the subsequent counts of substitution frequency.

The relative substitution frequencies were normalised so that the PAM 1 matrix corresponds to the probability of each residue substitution occurring in an evolutionary period of 1 residue mutation every 100 residues. However, considering substitution values, based on sequences where only 1 in 100 residues has mutated, will not provide much useful information on distant evolutionary relationships since the sequences would be almost identical. So the matrices corresponding to more distant relationships are generated by raising each internal value to the appropriate power, for example giving a PAM 250 matrix. Due to back mutations (A to B to A) and silent mutations (mutations in the genetic code that do not affect the identity of coded amino acid), the PAM 250 matrix corresponds to sequences that are approximately 20% identical.

The Blocks Substitution Matrices (BLOSUM)

The BLOSUM series of matrices (Henikoff & Henikoff, 1992) are generated from local blocks of aligned sequences, rather than full protein sequences, and are taken from the BLOCKS database (Henikoff & Henikoff, 1991). In order to produce matrices reflecting the substitution probabilities of different evolutionary distances, the sequences used in the alignments are first clustered. The sequence identity is calculated for every pair of sequences and any pair with a percentage identity above a given threshold are merged together. A series of BLOSUM matrices have been generated using different clustering thresholds to reflect different evolutionary distances (e.g. BLOSOM50 clusters sequences at 50% identity). These substitution matrices have been seen to outperform the PAM matrices in searching for a defined set of homologous relationships (Henikoff & Henikoff, 1993).

1.2.3.3 Position Specific Score Matrices

Residues that perform crucial roles in the protein structure, such as the residues involved in the folding pathway, important stabilising interactions or catalytic function, would be far less tolerant of dramatically altering substitutions than residues existing in the periphery of the protein structure. Thus, the probabilities for residue substitution will not only be dependent on which two amino acids are being exchanged but also where this exchange occurs in the 3D structure of the protein.

This has implications for identifying the relative importance of different positions in the protein sequence based on the analysis of related protein sequences.

By aligning a series of related protein sequences, i.e. by placing equivalent residues in the same vertical rows of the alignment, conserved patterns of residue identity or physicochemical property can often be identified. If a specific amino acid is seen in a large number of sequences that are otherwise relatively dissimilar, then this residue position is likely to correspond to a key role in the protein structure. Whether this key role is due to an active site, or a crucial interaction in the folding pathway, the highly selective nature of the amino acids identities allowed at such a position provides information specific to the family of proteins being described. By combining the probabilities of residue substitutions observed at each position in an alignment, a position specific score matrix (PSSM) can be generated. This can then be used as a unique 'fingerprint', or profile, that describes the important structural and functional features of a protein family. Profile-based sequence comparisons are discussed in more detail in chapter 5.

1.2.4 Protein Sequence Alignment

1.2.4.1 Insertions and Deletions

Residue substitutions are not the only mechanisms by which proteins evolve. Insertions or deletions (indels) of sequence fragments are also accepted in the protein structure. Usually these indels occur in the variable loops between secondary structure elements. Comparing two protein sequences therefore requires an alignment to be made that allows for the possibilities of these indels. Residues assigned as equivalent from this alignment can then be compared to calculate a score describing the overall sequence similarity.

In order to provide the optimal alignment of two sequences, it is necessary to consider every possible permutation of residues including insertions and deletions. Figure 1.3 shows a 2D matrix, or dot plot (Maizel JV & Lenk, 1981), comparing the identities between residues in two protein sequences (labelled A and B). This provides a simplistic visualisation of the local similarities between the sequences with matching sequence fragments corresponding to diagonal lines in the matrix.

This matrix may provide a useful tool to allow a manual alignment of the two sequences by linking the diagonal segments in the matrix. However as the length of these two sequences grows, manual alignments become far less practical. When searching large databases of sequences it is necessary to use fast and automated procedures.

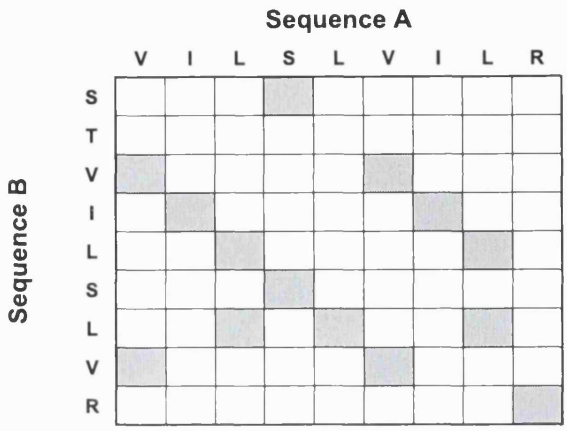


Figure 1.3: Diagram showing a 2D matrix comparing the identity of each residue in sequence A with the identity of each residue in sequence B.

1.2.4.2 Global Alignments

The dynamic programming algorithm is a general mathematical procedure that can be used to find the optimum alignment between two sets of data. This algorithm was first applied to align protein sequences by Needleman and Wunsch in 1970 (Needleman & Wunsch, 1970) and is still widely used in a variety of bioinformatics techniques. The procedure begins by generating a 2D score matrix based on the comparison of residues between two protein sequences, A and B. To illustrate, a simple scoring scheme will be used where identical residues are assigned a score of 5 (see figure 1.4).

The dynamic programming algorithm then accumulates these scores, starting with the bottom-right cell in the matrix. To provide each cell with information of the alignment path up to that point, the maximum score from the previous row or column starting from the cell $(i+1, j+1)$ can be inherited and added to the comparison score of the cell. Since indels occur less frequently than residue substitutions, a penalty is incurred for inheriting the score from any cell other than $(i+1, j+1)$ as this effectively corresponds to opening a gap in the alignment. The rules of inheritance are formalised in equation 1.1 and the accumulation procedure is illustrated by the third step of figure 1.4.

$$S(i,j) = S(i,j) + max \left| \begin{array}{l} S(i+1, j+1) \\ S(i+1, j+2..J) + G \\ S(i+2..I, j+1) + G \end{array} \right| \tag{1.1}$$

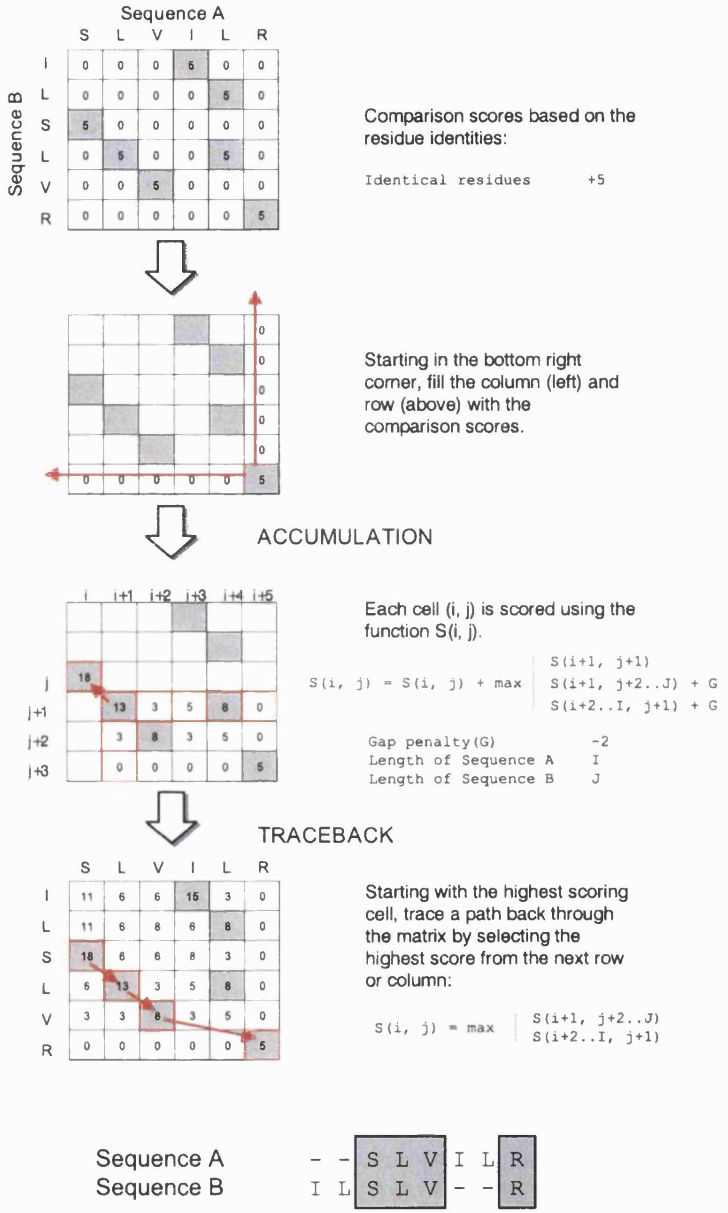


Figure 1.4: Flowchart describing the Needleman-Wunsch dynamic programming algorithm. This compares each residue in sequence A against every residue in sequence B then finds the optimal global alignment between two sequences.

Where $S(i, j)$ is the inheritance score of cell i, j (corresponding to residue i of sequence A and residue j of sequence B) and I and J are the lengths of sequences A and B respectively. The gap penalty, G, is given the value -2 for the example in figure 1.4.

An important part of this algorithm is that the path decisions for each cell,

i.e. the cell from which the inherited score was taken from, are encoded during this accumulation procedure. Thus, when the matrix is fully populated the highest scoring path through the matrix can be quickly identified by starting with the highest score in the first row or column and traversing back through each inherited cell. The alignment between the two sequences can be inferred from the path taken by this traceback procedure using the rules described in table 1.1. The traceback step and final global alignment between the example protein sequences are also shown in figure 1.4.

Traceback from cell (i, j)	
Inherited from cell	*Effect on alignment*
$(i+1, j+1)$	equivalent residues (no gap)
$(i+1, j+N_j)$	open $(N_j - 1)$ gap in sequence A
$(i+N_i, j+1)$	open $(N_i - 1)$ gap in sequence B

Table 1.1: Deriving the alignment from the traceback path. This table shows the effect on the final alignment of the three possible cases of inheritance.

1.2.4.3 Local Alignments

A global alignment algorithm is used to provide the optimal alignment between two protein sequences. However, when searching large sequence databases for putative sequence homologies it is often more important to provide an assessment of homology for each sequence comparison, rather than necessarily produce an accurate alignment. Also, since many proteins contain one or more independent structural domains (see section 1.1.3), it may be more useful to try to match significant fragments of protein sequence, rather than every residue. If the query sequence A, comprising of a single structural domain A_1, was compared against a query sequence B containing two structural domains A_1, B_1, it would only make sense to provide an alignment for the matching domain rather than the whole sequence. In this way, local alignments are used to identify sequence similarity between residue fragments of two proteins.

This implementation of the dynamic programming algorithm was developed by Smith & Waterman (1981) and only differs slightly from the global alignment calculation. Although there are many possible parameters, one usual difference between this local algorithm and the global Needleman and Wunsch algorithm is the introduction of negative scores for non-matching residue comparisons, e.g. matching

residues +5, non-matching residues -1, gap penalty -2.

$$S(i,j) = S(i,j) + max \left| \begin{array}{l} 0 \\ S(i+1, j+1) \\ S(i+1, j+2..J) + G \\ S(i+2..I, j+1) + G \end{array} \right| \tag{1.2}$$

Since this implementation of dynamic programming is designed to align local sequence fragments, it was important to eliminate any penalty for starting a new alignment path. Thus, instead of forcing cells to inherit potentially negative scores from previous paths, an additional choice is included in the score function $S(i,j)$ which allows the value 0 to be inherited (see equation 1.2).

1.2.5 Scoring the Sequence Alignment

1.2.5.1 Assessing Statistical Significance

Having found the optimum alignment between two protein sequences, it is important to assess whether the relationship has occurred as a result of evolution (i.e. descent from a common ancestor) or has arisen by chance. Sequence comparison algorithms are designed to find the optimal alignment between two protein sequences whether they are related or not. Thus every sequence pair will have similarities simply due to the finite number of states each residue can occupy. To assign confidence to a putative homologous match, it is necessary to assess whether the similarity score is statistically significant. This requires knowledge of which scores to expect simply by chance, i.e. the distribution of the random scores from non-related sequence comparisons.

1.2.5.2 Z-scores

The Z-score measures the significance of a putative matching score by assessing the distance between the matching score and the scores for the rest of the (mostly non-related) pairs (see figure 1.5). The Z-score, calculated with respect to a putative match (S), is the number of standard deviations (s.d) from the mean score (m) for the database search. Higher Z-scores denote more significance with regard to the similarity score of the putative match. A Z-score greater than 3 is often deemed significant.

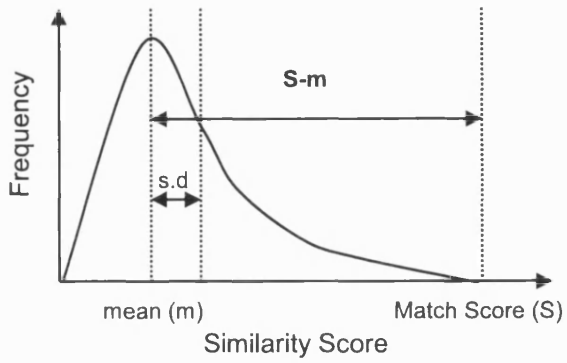


Figure 1.5: The typical distribution observed when searching a database with a query sequence. Z-score is calculated by counting the number of standard deviations (s.d) between the matching score (S) to the mean score (m) of the database search.

1.2.5.3 Expectation Values

It has been observed that the distribution of alignment scores for comparisons between random sequences approximately fits an Extreme Value Distribution (EVD, Dembo *et al.* (1994)). Figure 1.6 illustrates the shape of an EVD and quotes the equation used to model this distribution. This equation describes the frequency of finding a similarity score, S, between two random sequences of lengths n and m. The values k and λ are numerical constants which can be estimated directly from the database used.

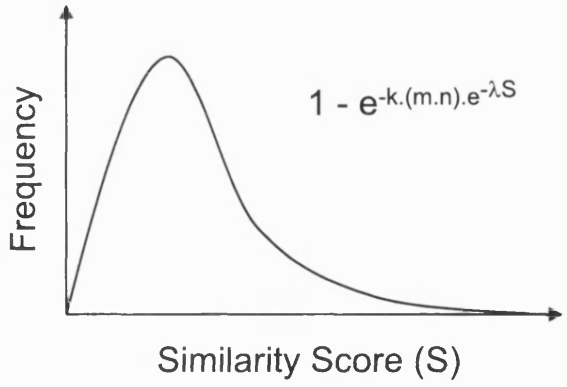


Figure 1.6: Extreme value distribution for pairwise similarity scores (S) between random sequences of length n and m. The values for k and λ are constants derived from the background scores from the database search.

Providing a model of the distribution of random scores allows the probability value, or p-value, to be calculated. This value describes the probability of finding an unrelated protein with a similarity score greater than or equal to the observed

similarity. However, the probability of a random match also depends on the size of the database. In order to take this into account an expectation, or E-value, is calculated. The FASTA sequence comparison algorithm (discussed further in section 5.1.2.3) provides E-values by simply multiplying the p-value by the number of sequences in the database, thus providing the significance of a given similarity score while accounting for database size. For example, if a putative match has an E-value of 0.01, it is expected that an equivalent similarity score could be found in one out of 100 unrelated sequences simply by chance. The CATH database uses a conservative E-value threshold of $1 * 10^{-4}$ to signify homology by sequence similarity, i.e. it is highly unlikely that this degree of sequence similarity could have occurred by chance.

1.2.6 Structural Similarity

1.2.6.1 Conservation of Sequence and Structure

It is well established that protein structure is more conserved that protein sequence during the course of evolution (Chothia & Lesk, 1986). Therefore the protein 3D structure provides a more sensitive probe of evolutionary relationships than the amino acid sequence. This is illustrated in figure 1.7 by making two comparisons of structural similarity and sequence identity (based on the structural alignment) within three globin-like proteins.

1.2.6.2 Methods for Evaluating Structural Similarity

Distance Plots

The earliest and simplest structural comparison methods were based on visually inspecting distance plots between proteins. Distance plots, introduced by Phillips in 1970 (Phillips, 1970), are 2D matrices used to visualise the distances between residues in a protein structure and are often shaded according to this distance. The related contact maps (CM), are used to indicate which residues in a protein structure are in contact, i.e. within an allowed distance threshold, for example <8Å (see figure 1.8 and chapter 2 for more details). This distance threshold is usually based on the distances between C_α or C_β atoms of the amino acid side chains. When these contacts maps are plotted in 2D, by shading the cells associated with contacting residues, patterns arise in the matrix which often prove characteristic of a particular fold.

An example of a typical contact pattern is the thick band along the main diagonal

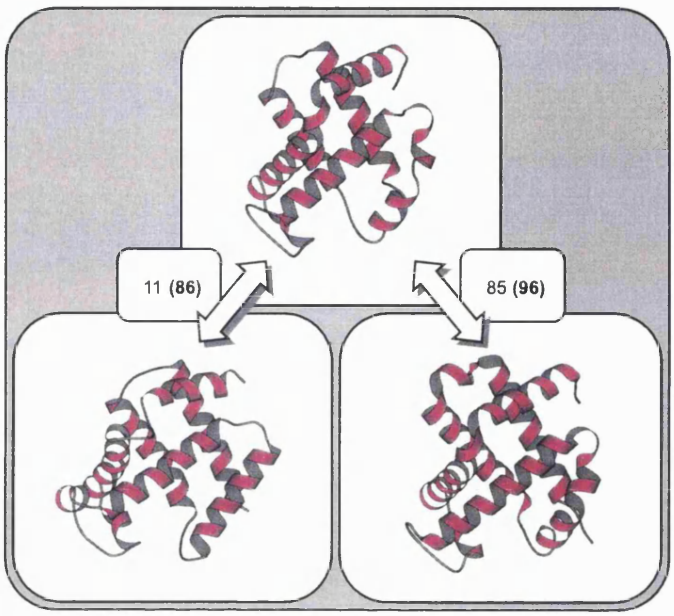


Figure 1.7: MOLSCRIPT (Kraulis, 1991) representations of relatives from the globin superfamily. Pairwise sequence identities are shown for each pair of structures and SSAP structural similarity scores are also given in parentheses. SSAP scores range from 0 up to 100 for identical structures (see section 1.2.7.3).

representing interactions within α-helices which is due to contacts between residues in positions i and i+1 to i+4 (see figure 1.8). Solid lines parallel or anti-parallel to the main diagonal correspond to parallel and anti-parallel β-sheets respectively. Also, any contact involving an α-helix usually can be recognised by a contact pattern repeating every three or four residues, due to residues being brought into contact by the periodicity of the α-helix (see also section 2.2.2.1).

Nearly identical proteins and those with similar lengths can simply be compared by overlaying their distance matrices and searching for similarities or shifts in the observed patterns. Richards & Kundrot (1988) developed a method of assigning secondary and super-secondary, i.e. common motifs of secondary structure, by comparing distance matrices from real structures to distance matrices of idealised secondary structures. The DALI algorithm (Holm & Sander, 1993), described in more detail in section 1.2.7.3, also uses DMs to provide a global alignment and comparison of protein structures.

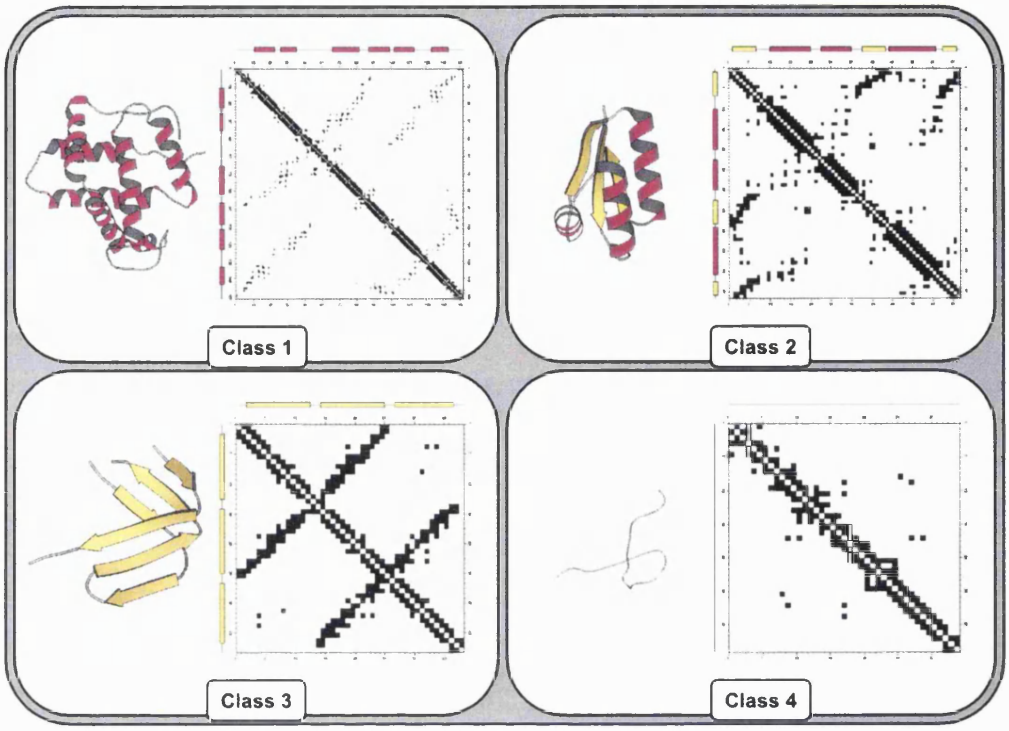


Figure 1.8: Example contact maps for each protein class, mainly-α, mainly-β and mixed-α β. Contacts between parallel secondary structures give rise to diagonal lines parallel to the central diagonal. Contacts between anti-parallel secondary structures give rise to lines perpendicular to the central diagonal. Different folds give rise to characteristic patterns in the matrix.

Root Mean Square Deviation

The root mean square deviation (RMSD) is commonly used as a measure of structural similarity between two sets of 3D co-ordinates. The two structures are first superposed so that they overlap in 3D space as closely as possible (see section 1.2.7.2), then a set of equivalent residues between the two structures is identified. The distance between each pair of equivalent residues, d_i, is then calculated and squared. The sum of these squared distances between equivalent residues is then taken and divided by the number of pairs, N, to give the mean (see equation 1.3).

$$RMSD = \sqrt{\frac{\sum_{i=1}^{N} d_i^2}{N}} \tag{1.3}$$

In this equation N is the number of equivalent residues identified between the structures and d_i is the distance between residue i in the first protein and the equivalent residue in the second protein (see figure 1.9).

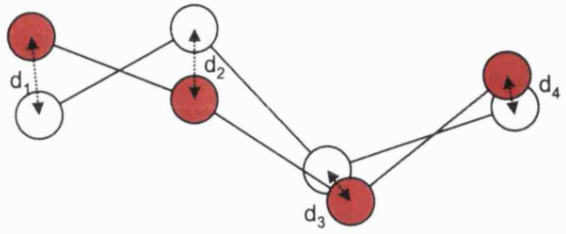


Figure 1.9: Calculating root mean square deviation (RMSD) as a measure of structure similarity following a structural superposition

When comparing identical or highly similar proteins, e.g. when analysing conformational changes on substrate binding, all the residues in the structures may be used for this calculation. However, when considering more distant structural relationships it is more common to only calculate the RMSD with respect to C_α-atoms that are within a specified distance threshold, e.g. <3Å following the structural superposition. When quoting the RMSD it is therefore necessary to provide the number of equivalent residues from which this value was derived. An RMSD of 3Å based on the superposition of 20 residues is far less significant than the same RMSD calculated from 200 residues. As a general rule, proteins of around 150 residues with similar folds would be expected to give a RMSD value of less than 3.5Å over at least 100 residues.

1.2.7 Structure Comparison Algorithms

1.2.7.1 Overview of Comparison Methods

Various algorithms have been proposed to automatically compare and align protein structures, most of which employ dynamic programming at some level. Some of these algorithms use dynamic programming to optimise the alignment of two structures given an initial superimposition. An initial alignment could be made by visual alignment of equivalent positions, however this would prove impractical for the large numbers of structural alignments necessary for large databases searches. Alternatively, this initial alignment could be taken from the sequence alignment of the two proteins. If the two proteins were close homologues, there should be sufficient sequence similarity to provide a reasonable alignment. However, it has already

been mentioned that proteins may share similar structures even when their sequence similarity cannot be distinguished from non-related proteins. Thus, more distant homologues may not share sufficient sequence similarity for this initial alignment to be useful, possibly leading the subsequent optimisation procedure to fail to converge on an optimal alignment. The search for a fast and accurate method able to reliably align and compare protein structures has resulted in many different solutions to this problem. A selection of these algorithms are discussed in more detail below.

As with sequence comparison, methods for comparing protein structures comprise two important components. The first involves techniques for scoring similarities in the structural features of proteins. The second is the use of an optimisation strategy that can identify an alignment which maximises the structural similarities measured.

The majority of methods compare the geometric properties of either the secondary structure elements or residues along the carbon backbone (C_α or C_β atoms are frequently used). The geometric properties of these residues (or secondary structures) are determined from the 3D co-ordinates of the structure deposited in the PDB. Relationship information that can be used to describe the structural environment within a protein includes, for example, distances or vectors between residues.

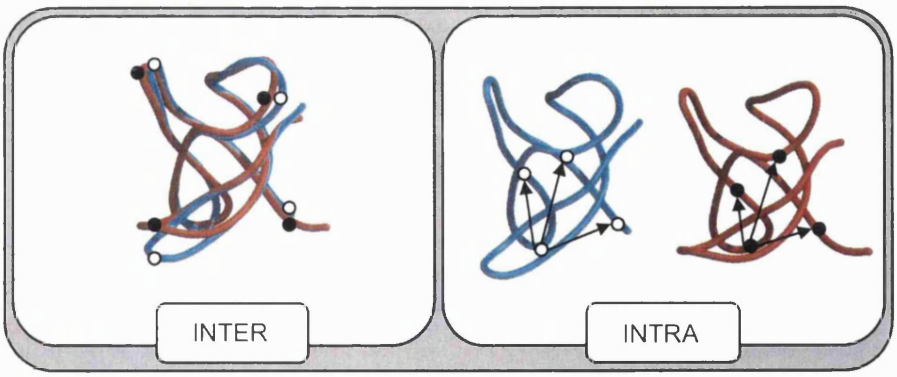


Figure 1.10: Illustration of intermolecular and intramolecular relationships.

Other non-geometric properties are sometimes included in the comparison, e.g. physicochemical properties such as hydrophobicity. Specific bonding networks can also be compared, e.g. hydrogen-bonding patterns, disulfide bonds. However, the contributions of these characteristics need to be very carefully weighted, as they can sometimes increase noise. For example, the hydrogen-bonding properties of pairs of helices or parallel and anti-parallel β-sheets will be similar regardless of the

topology of the protein structures being compared.

Structure comparison algorithms can be separated into two distinct categories; those that compare intermolecular relationships and those that compare intramolecular relationships (see figure 1.10). An example of an intermolecular comparison is a rigid-body superposition method used to calculate RMSD, where co-ordinates of equivalent residues are compared between structures. In contrast, intramolecular comparison involves the identification of similar internal structural features between proteins, e.g. the comparison of inter-residue contacts.

1.2.7.2 Intermolecular Structure Comparison

Rigid Body Superposition

Rossmann and Argos pioneered rigid body superposition methods in the 1970s as the first crystal structures were being deposited in the PDB. Their approaches employed rigid body methods to superpose equivalent C_α atoms between protein structures. The major steps of this method can be described as follows:

- Translate both proteins to a common position in the co-ordinate frame of reference.
- Rotate one protein relative to the other protein, around the three major axes.
- Measure the distances between equivalent positions in the two proteins.

The second and third steps are repeated until there is convergence on a minimum separation between the superposed structures. Usually, the centres of mass of both structures are translated in 3D co-ordinate space towards the origin, an operation performed by the translation vector. Then an optimisation procedure is performed where one structure is rotated around the three orthogonal axes, relative to the second structure (see figure 1.11), so as to minimise the distances between superposed atoms. This operation is described by the rotation matrix. The distance between equivalent positions is generally described by a residual function, which effectively measures the distance between superposed residues, i.e. RMSD. The major difficulty lies in specifying the putative equivalent positions at the start of the optimisation. If this set does not contain a sufficient number of truly equivalent positions the method can fail to converge to a solution. For proteins which are closely related (e.g. >=35% sequence identity), sequence comparison can be used to determine an initial set of equivalent residues. This set can then be refined by the optimisation procedure. For more distantly related proteins, information on equivalent positions

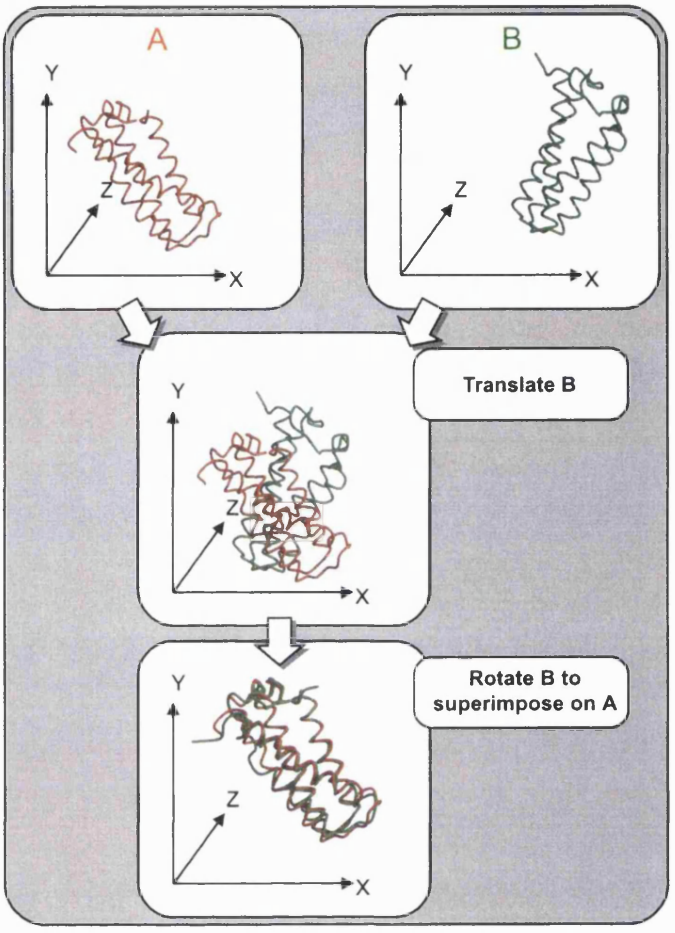


Figure 1.11: Procedure for finding the optimal rigid body superposition between two protein structures.

is often obtained by manual inspections of the structures or by identifying residues binding common ligands in the active sites of the two proteins.

Once the optimal superposition has been determined the difference between the structures is commonly measured by the RMSD. Structures having a similar fold typically give values below 3.5Å, although there is a size dependence to this measure. For example, distant homologues with more than 400 residues may return an RMSD >4.0Å compared to a value of <3.0Å for smaller proteins with less than 100 residues, but of comparable evolutionary distance.

A variety of structure comparison protocols often include a final step to superpose equivalent residues by rigid body techniques, and determine the RMSD value. More recently, statistical approaches have also been developed for assessing the significance of structural similarity detected between protein pairs. These are often more reliable

as they are independent of protein size and are described in more detail below.

STAMP

An elegant approach devised by Russell and Barton in their STAMP method (Russell & Barton, 1992), used dynamic programming to refine the set of equivalent residues given by rigid body superposition (see figure 1.12). An initial set of residue pairs is given by a sequence alignment of the proteins and this is used to guide a superposition of the structures. Intermolecular distances between equivalent residues are then measured and used to score a 2D score or path matrix, which is analysed by dynamic programming. The resulting path gives a new set of possible equivalent residues. Thus a new superposition can be generated using this refined residue pair set and the path matrix re-scored and re-analysed to give a better path and so on. This is repeated until there is no improvement in the RMSD measured over the equivalent pairs. The method can be summarised as follows:

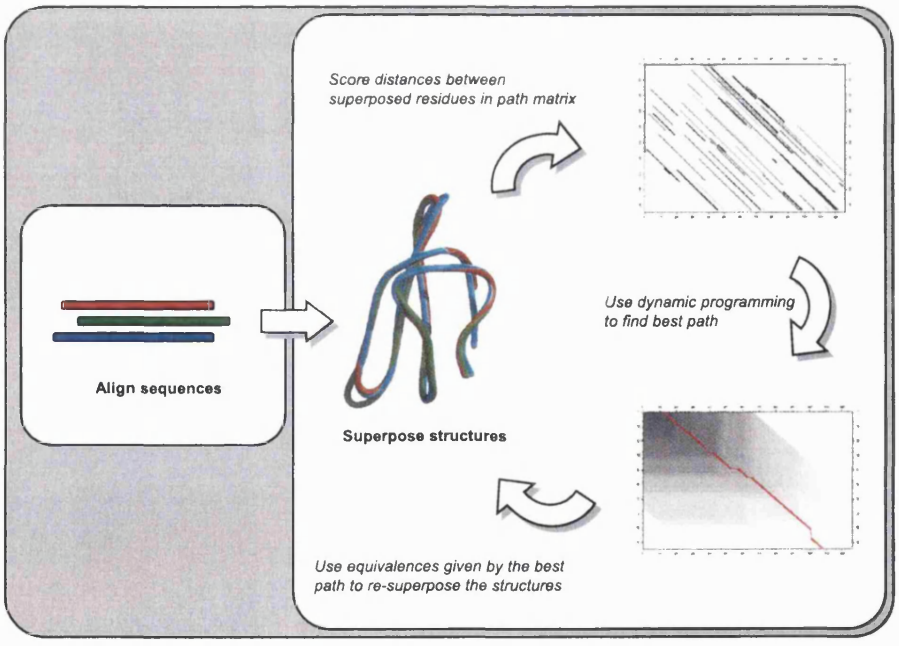


Figure 1.12: Schematic representation of the STAMP method devised by Russell and Barton. Sequence alignment is first used to determine putative equivalent positions in order to superpose the structures. Distances between the superposed positions are then used to score a 2D score matrix which is analysed by dynamic programming to obtain a more accurate set of equivalent positions on which the structures are superposed again. A window is often imposed upon the score matrix to avoid the computational expense of comparing residues that are a long distance apart in the protein sequences.

- Obtain a set of putative equivalent residue positions by aligning the sequences of the two proteins.
- Employ rigid body methods to superimpose the structures using this set of equivalent positions.
- Score the 2D matrix, whose axes correspond to the residue positions in each protein, with values inversely proportional to the distances between the superposed residues.
- Apply dynamic programming to determine the optimal path through this score matrix, giving a new set of putative equivalent positions.

Steps 2 to 4 are repeated until there is no change in the set of equivalent residue positions.

1.2.7.3 Intramolecular Structure Comparison

GRATH

One of the simplest strategies for handling insertions is to discard the loop regions where insertions are more likely to occur. A number of algorithms have been written that employ a mathematical technique, known as graph theory, to identify equivalent secondary structure elements (SSEs) between two structures. Graph theoretical approaches to protein structure comparison were pioneered by Artymiuk, Willett and co-workers in 1993 (Grindley *et al.*, 1993). A more recent approach is the GRATH algorithm (Harrison *et al.*, 2002) which again considers protein structures only in terms of SSEs. Since protein structures contain approximately one order of magnitude fewer SSEs than amino acid residues, this approach drastically reduces the complexity of the problem.

Having broken the protein structures into SSEs, GRATH employs graph theoretical techniques to identify equivalent SSEs, i.e. those possessing similar intramolecular relationships. Graph theory techniques reduce the 3D information embodied in a protein structure, such as SSEs, to a simplified 2D map, or graph (see figure 1.13). In the protein graph, each SSE in the structure is associated with a single node which can be labelled according to the secondary structure type, e.g. α-helix or β-strand. Lines or edges connecting the nodes are labelled with information describing the relationships between the associated secondary structures such as distance, angle and chirality. Graph theory then searches for the most equivalent nodes, or clique, between the two graphs.

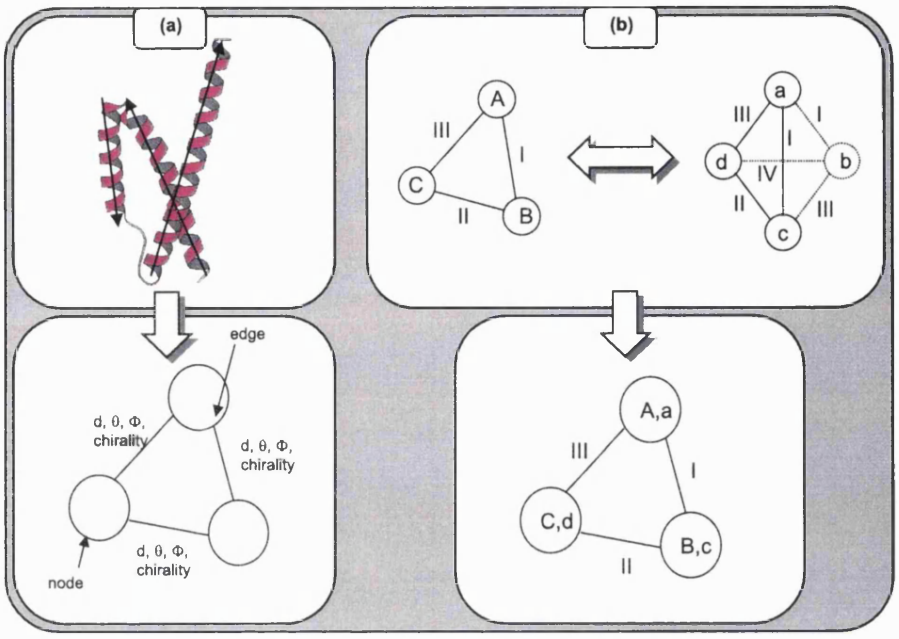


Figure 1.13: The identification of structural similarity with the GRATH algorithm (Harrison *et al.*, 2002). Part (a) describes the procedure for describing a protein 3D structure as a 2D graph based on the relationships between secondary structure elements (SSE). Part (b) illustrates how graph theory is used to identify the maximum sub-graph, or clique, that is common to both graphs representing the proteins being compared.

Since this method aligns SSEs rather than individual residues, it cannot provide an accurate residue-based alignment. However, this algorithm is fast and is able to identify the correct topology within the top 10 matches of a database search 95% of the time. As a result, it can act as perfect filter to generate a list of likely candidates for a more rigorous structural alignment algorithm. The use of GRATH as a pre-filter for the assignment of structural relationships in the CATH server is discussed in more detail in section 3.5.1.

DALI

Another approach for coping with indels is to divide the protein into fragments and compare the contact maps for these fragments. For example, the DALI method of Holm and Sander (Holm & Sander, 1993) separates the proteins into all permutations of hexapeptide fragments (see figure 1.14). Contact maps derived from these fragments can be rapidly compared to identify potentially equivalent hexapeptide matches. These matches are then concatenated using a Monte Carlo optimisation strategy. This type of optimisation strategy effectively explores all combinations

that satisfy the constraints on the overall topology and generate a structural unit with an acceptable RMSD. The method can be summarised as follows:

- Divide the protein into hexapeptides and derive the contact map for each hexapeptide.
- Identify hexapeptides whose contact maps match within an allowed threshold, i.e. where there is a similar pattern of distances between equivalent residues.
- Concatenate matching hexapeptide contact maps to extend the similarity between the proteins.
- Superpose the extended fragments and check the difference between the fragments, as measured by RMSD, is within an allowed threshold, else reject the extension.

Steps 3 and 4 are repeated until no further fragments can be added.

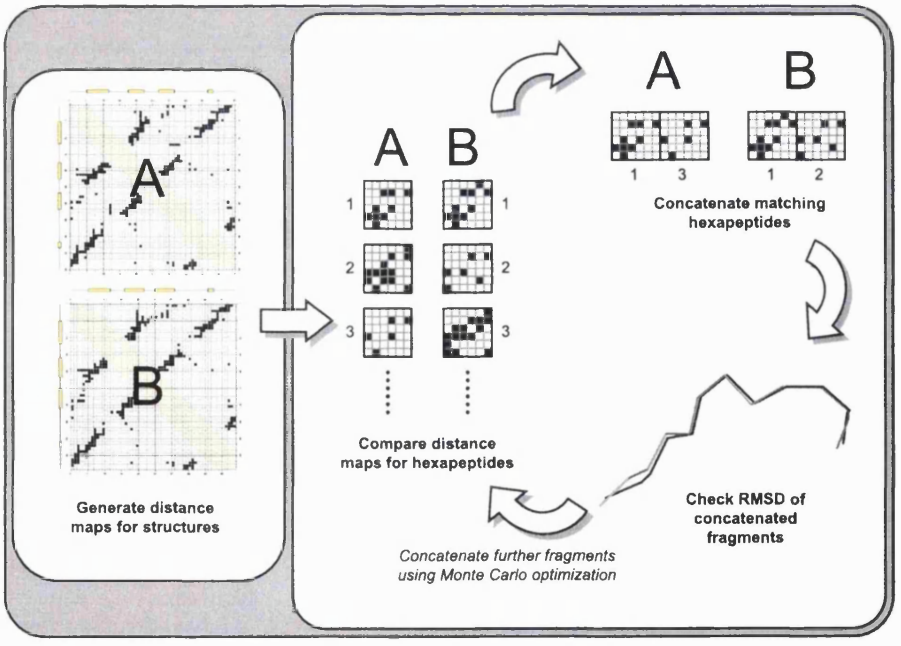


Figure 1.14: Flowchart describing the DALI protocol. Proteins are broken into hexapeptides and contact maps are derived for each fragment. Similar contact map fragments are then concatenated if the resulting RMSD for the resulting structural fragments is within a given threshold.

SSAP

The SSAP method (Taylor & Orengo, 1989) employs dynamic programming at two levels to cope with the extensive indels between homologues. The first level of DP is employed for the comparison of residue structural environments between proteins. The second level of DP is used in a final pass to determine the set of equivalent residues. The structural environment, or view, for a given residue is defined as the set of vectors from the C_β atom of this residue to the C_β atoms of all others residues in the protein (see figure 1.15). In order to compare views between residues in two proteins, vectors must be determined within a common co-ordinate frame. In SSAP, the co-ordinate frame is based on the local geometry of the C_α atom and is therefore independent of the global orientation of the structure.

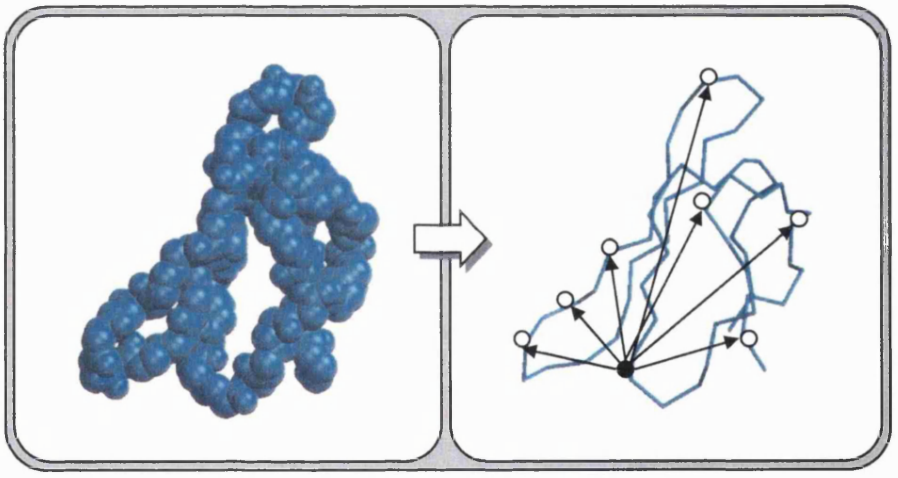


Figure 1.15: Illustration of a residue structural environment, or vector view, employed by the SSAP method (Taylor & Orengo, 1989). The view for a given residue is taken as the set of vectors from the C_β atoms of this residue and all other residues in the protein.

The vector views are first determined for all residue positions in the proteins being compared (protein A and protein B). Vector views for residues selected for being potentially equivalent between the two structures are then compared. The selection criteria identifying potentially equivalent residue pairs are based on the residues having similar physical and chemical properties, such as accessibility, secondary structure state, local conformation (measured by ϕ, ψ angles). Comparisons between two vector views are scored in a 2D score matrix which is referred to as the residue level score matrix. This is analogous to the score matrix used in sequence comparison, except that the horizontal axis is labelled by the vectors associated with

the view for residue i in structure A and the vertical axis is labelled with vectors for the view from residue j in structure B. Cells are then scored depending on the similarity of these vectors. As with sequence alignment, dynamic programming is used to find the optimal path through this matrix giving the best alignment of the residue views (see section 1.2.4.2).

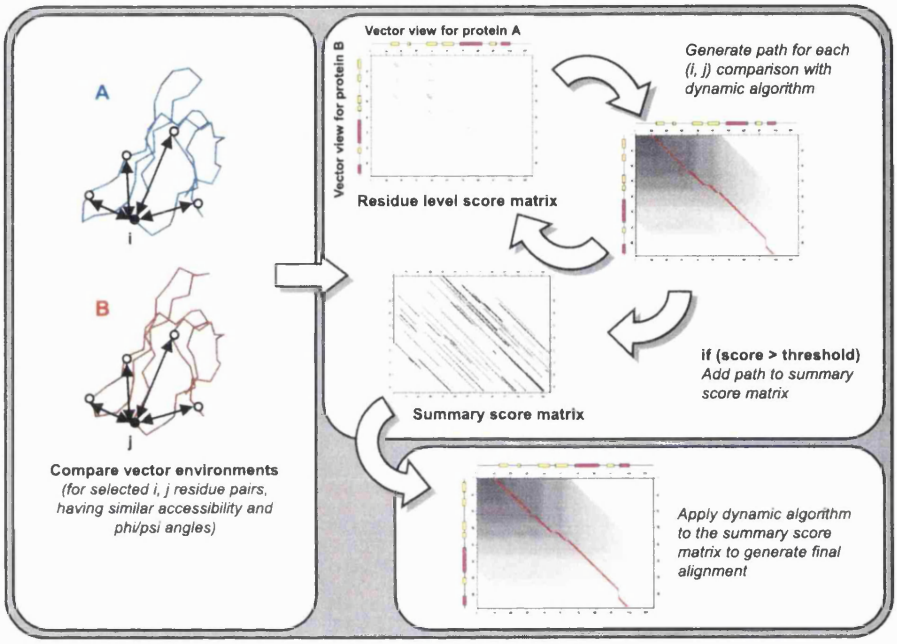


Figure 1.16: Flowchart describing the SSAP protocol. DP is applied on two levels. First to find the optimal alignment based on the comparison of structural environments of two residues (scored in the residue-level score matrix). If the residues are deemed sufficiently similar (i.e. the alignment path scores greater than a given threshold), the scores from this path are added to a summary score matrix. DP is then used again to find the optimal global alignment through the summary score matrix to identify equivalent residues between the two proteins.

The alignment path from this score matrix effectively represents the best alignment between the two structures based on the structural environments of the two residues being compared (residue i in structure A and residue j in structure B). The degree of similarity between residue pairs can be inferred from the relative score of this alignment path. If the residues are very similar, i.e. provide a high scoring alignment path, then the scores from this alignment path are added to a summary score matrix. Once the vector views of all residue pairs have been aligned and high scoring optimal paths added to the summary matrix, the best path through this summary matrix can be found, again using dynamic programming (see figure 1.16).

Since dynamic programming is applied on two levels, to provide an alignment at

the residue level and a final global alignment, the algorithm is often referred to as double dynamic programming (DDP). This double dynamic algorithm allows indels to be modelled accurately based on 3D structure and has been applied to other areas such as fold recognition (see section 1.3.3).

A simple outline of the SSAP algorithm can be summarised as follows:

- Calculate the view for each residue in the two proteins, given by the set of vectors from the residue to all other residues in the protein.
- For each potentially equivalent residue pair between the proteins, e.g. possessing similar torsional angles and accessible areas, compare vector views by using dynamic programming to find the best path through a residue level score matrix, scored according to the similarity of vectors.
- For residue pairs scoring highly, add the scores along the optimal path obtained in step 2, to a 2D summary score matrix.
- Repeat steps 2 and 3 until all potentially equivalent pairs have been compared.
- Use dynamic programming again to determine the optimal path through the 2D summary score matrix, giving the equivalent residues between the proteins.

Various modifications of this approach have been developed over the years but the basic concepts remain the same.

1.3 Predicting 3D Protein Structure from Sequence

1.3.1 Background to Protein Structure Prediction

Although many characteristics of protein molecules are now well documented, such as the chemical and physical attributes of the constituent amino acids, plenty of questions regarding protein structure and function remain unanswered. Perhaps the most sought after answer is how a linear chain of amino acids folds, both quickly and accurately, into the highly specific, functionally active three-dimensional protein structure. This question, also known as the protein folding problem, is often referred to as the 'Holy Grail' of molecular biology due to both the relative importance and the elusive nature of the answer. Until our knowledge extends sufficiently to provide a comprehensive solution to this problem, a more pragmatic approach to

understanding the relationship between protein sequence and protein structure must be taken.

Assigning the structure of a protein directly from sequence usually begins by identifying relationships to known protein structures on the basis of sequence similarity. It is well accepted that proteins sharing at least 30% of their amino acid sequence will adopt the same fold and will often exhibit similar functional mechanisms (Todd *et al.*, 2001). Therefore, if an evolutionary relationship can be identified between a protein sequence and a known protein structure, many structural features can often be assigned. The technique used to assign such structural features depends on the sequence similarity between the protein sequence and known protein structure. If the sequence identity is more than 30%, then homology modelling, or comparative modelling is used. When the sequence identity to a known structure is lower than this threshold, homology modelling does not always prove reliable and fold recognition, or threading techniques are used as an alternative. However, both homology modelling and fold recognition are only applicable when the protein sequence adopts a known folding arrangement. If the sequence presents a novel fold then *ab initio* prediction methods must be used, i.e. methods that predict structure directly from amino acid sequence.

1.3.2 Homology Modelling

Modelling by homology is currently by far the most accurate method of structure prediction if a close sequence relative with known structure can be identified. Once this close relative, known as the template or parent structure, has been identified, an alignment is made between the sequence and template structure. The quality of the predicted model is highly dependent on the quality of this alignment. Thus, identifying a template structure with a high sequence identity to the query sequence is vital to produce a high quality model. Once this parent has been identified, a typical procedure for generating a structural model from the query sequence can be illustrated in the following points.

- Align the query sequence to the parent structure.
- Determine the structurally variable regions (e.g. loops in the periphery of the structure) and structurally conserved regions (e.g. core secondary structure elements) in the parent structure.
- Assign structurally conserved regions directly to the query sequence.

- Model structurally variable regions. This can be approached using manual techniques, knowledge-based methods or from *ab initio* protocols.
- Build the amino acid side chains (e.g. by inheriting torsion angles from the parent where possible).
- Refine the model (e.g. using energy minimisation techniques).
- Evaluating the model (e.g. using RMSD, see section 1.2.6.2).

1.3.3 Fold Recognition

The fold recognition approach to structure prediction 'threads' the protein sequence through a library of structural templates to see which fold would 'fit' the sequence best (Sippl & Weitckus, 1992; Jones *et al.*, 1992). The compatibility between the sequence and a given template is often assessed using knowledge-based pair-potentials which describe the likelihood of the residues in the query sequence being within the distances specified in the template structure. Other structural features, such as accessibility have also been used. As in sequence and structure comparison algorithms, the alignment between the query sequence and template structure must allow for indels. As a result, an implementation of the double dynamic algorithm is often used (see SSAP 1.2.7.3). This provides the optimal alignment between the query sequence and template structure by scoring all the internal residue interactions against pair-potentials derived from known structures.

Threading has allowed structural information to be assigned to proteins where no relatives could be identified by sequence comparison. However, this method is inherently limited to identifying sequences that form known protein folds and cannot be used to predict the structure of novel folds.

1.3.4 *Ab initio* Prediction of Protein Structure

The ultimate aim of structure prediction is to generate the protein structure directly from amino acid sequence. Many methods have been submitted for this *ab initio* approach ranging from those attempting to recreate the physical and chemical forces involved in protein folding to those training neural networks to match predictions of secondary structure and distance constraints from sequence patterns.

Prediction of protein structure directly from amino acid sequence can be broken down into two separate tasks.

- Defining an energy function that gives the native conformation a lower energy than all other conformations.
- Developing an algorithm that can use this energy function to identify the correct, native conformation.

Generating an energy function or potential that can evaluate the 'nativeness' of a given conformation can be approached in two different ways. Potentials using a classical approach attempt to recreate the physical and chemical forces involved in protein folding. An alternative approach is to use an empirical potential based on distributions of distances between pairs of amino acids in known structures, i.e. pair potentials. However, both of these types of potentials have, in some cases, failed to uniquely identify the native conformation. *Ab initio* methods are discussed in more detail both in chapter 2 and chapter 4.

1.4 Protein Structure Classification Databases

1.4.1 Overview of Structure Classification

Over the last five years, several groups have attempted to classify proteins into evolutionary families and fold groups in order to understand the relationship between sequence and structure and to help identify remote evolutionary relatives. Databases such as CATH and SCOP employ hierarchical classifications and are discussed below.

1.4.2 CATH

The CATH classification database (Orengo *et al.*, 1997; Pearl *et al.*, 2001b), developed at UCL, holds over 19,000 structural domains (release 2.4). Boundaries for structural domains are identified by a consensus approach which seeks agreement between three domain assignment algorithms (Jones *et al.*, 1998). Once defined, domains are clustered in a hierarchy according to four major levels, class (C), architecture (A), topology (T) and homologous superfamily (H).

The first level, class, simply reflects the proportion of α-helix or β-strand secondary structures and is split into three major categories, mainly-α, mainly-β and mixed-$\alpha\beta$. Architecture is a description of the general spatial arrangement of the secondary structures and there are 37 distinct architectures in CATH v2.4. The topology, or fold level incorporates additional information on the connectivity or sequential order of these secondary structure elements, giving a total of 775 different

fold groups for CATH v2.4. Proteins are only grouped at the last level of the hierarchy, homologous superfamilies or H-level, if there is sufficient evidence that proteins are related by evolution. Homology is defined when at least two of the following evolutionary similarities can be assigned; high sequence similarity (>35% identity or significant E-value), high structural similarity (SSAP score > 80), or functional similarity.

The entries in each level of the hierarchy are assigned an identifying number and these numbers can be combined to give the CATH classification code. For example, the CATH classification code for the globins superfamily is 1.10.490.10. Therefore this superfamily is in the mainly-α class (C identifier 1), in the orthogonal bundle architecture (C.A identifier 1.10) and globin-like topology (C.A.T identifier 1.10.490).

1.4.3 SCOP

The SCOP database (Murzin *et al.*, 1995; Lo Conte *et al.*, 2000) also classifies domains into a hierarchy of structural relationships although classification is based mainly on visual inspection using Alexei Murzin's encyclopedic knowledge of protein folds. SCOP defines four levels of classification; class, common fold, superfamily and family in descending hierarchical order. The SCOP definition of class is similar to that of CATH apart from the mixed-α β class is split into α/β, where α-helices and β-strands are interspersed, and $\alpha + \beta$, where the regions of α-helices and β-strands are segregated. The common fold is equivalent to the topology (T) level of CATH and the superfamily level is most similar to a homologous superfamily (H) in CATH. Structures are clustered into the same family if they have a sequence identity of at least 30% or if they exhibit close functional and structural similarity.

1.4.4 Other Structure Classification Databases

A variety of structural databases have been constructed with contrasting levels of manual and automated intervention. Table 1.2 summarises some of the more commonly used structure databases. With the exception of SCOP (see section 1.4.3), all the databases discussed here use an automated method for protein structure comparison at some point in the classification procedure. Rather than generate strict hierarchical boundaries, some resources, such as FSSP (Holm & Sander, 1998) and Entrez (Madej *et al.*, 1995), provide lists of domains with similar structures. These lists, also known as nearest neighbour lists, describe a model of protein folding space that resembles a continuum rather than a series of discrete structural clusters.

Database	Location	Structure Comparison Method	Description
3Dee	EBI, Cambridge, UK	STAMP (Russell & Barton, 1992)	Fully automated, multi-hierarchical classification with some class and fold definitions taken from SCOP.
DDD	EBI, Cambridge, UK	DALI (Holm & Sander, 1993)	**D**ali **D**omain **D**ictionary. Fully automated structural classification of recurring protein domains.
ENTREZ/ MMDB	NCBI, Bethesda, MD, USA	VAST (Madej *et al.*, 1995)	Fully automated structural descriptions using pre-calculated nearest neighbour lists.
FSSP	EBI, Cambridge, UK	DALI (Holm & Sander, 1993, 1998)	**F**old classification based on **S**tructure-**S**tructure alignment of **P**roteins. Fully automated structural descriptions using nearest neighbour lists.
HOMSTRAD	Cambridge University, UK	COMPARER (Sali & Blundell, 1990)	**HOM**ologous **STR**ucture **AL**ignment **D**atabase. Manual classification using information from SCOP, and various sequence family databases.

Table 1.2: Summary of structure classification databases.

1.5 Overview of the Thesis

The work presented in this thesis examines the conservation of inter-residue contact patterns between evolutionarily related structures in the CATH database and applies this consensus data to the classification, analysis and fold recognition of both protein structures and protein sequences. Figure 1.17 provides a flowchart summarising the aims and protocols discussed in this thesis.

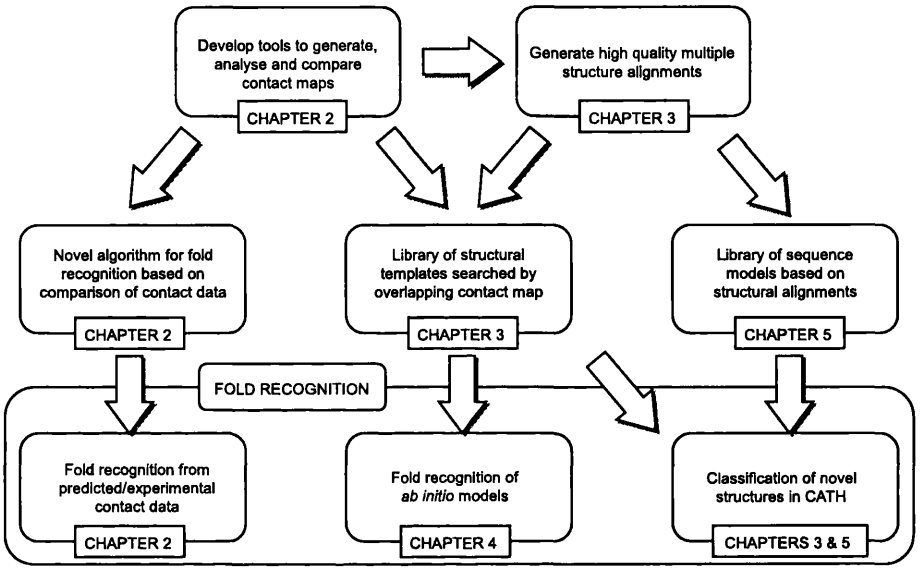


Figure 1.17: Flowchart providing an overview of the work discussed in this thesis.

Chapter 2 describes the computational tools used in this thesis for analysis, comparison and visualisation of contact patterns. The **CO**nserved **CO**ntact **PLOT**ting (COCOPLOT) software (I. Sillitoe, computer program, 2002), was used throughout this thesis to identify inter-residue contacts from single structures and identify conserved contacts from multiple structure alignments, or structural templates. This program was also employed to provide a measure of structural similarity based on structure–structure alignments and structure–template alignments by calculating the degree of overlap of contact patterns. The diagnostic plots, or contact maps, generated by this program allowed these features to be visualised and provided a useful tool for analysing the structural variation within superfamilies. This chapter also introduces a novel fold recognition algorithm that aligns structures based on their inter-residue contacts. This algorithm employs a similar approach to structure comparison to that used by SSAP (Taylor & Orengo, 1989), i.e. a double dynamic programming algorithm (see section 1.2.7.3), however rather than compare internal

vectors, this program restricts itself to comparing internal contacts. This program was intended to be used in a collaborative project with Mitsu Ikura, University of Toronto, as a fold recognition tool using preliminary experimental distance restraints by NMR. The protocol for optimising this algorithm is discussed in depth, in addition to preliminary tests assessing the ability to recognise the native fold from a reduced set of contacts.

In order to identify contacts that have been highly conserved during the process of evolution, it was first necessary to generate multiple structure alignments of proteins known to be related by evolution. The work in chapter 3 discusses a novel clustering protocol to provide structurally coherent protein clusters within homologous superfamilies in the CATH database. Superfamilies containing highly variable structures were represented more accurately by allowing more than one template per superfamily. This clustering procedure was optimised by assessing the ability of the resulting templates to discriminate between related and non-related structures. The optimal clustering parameters were then used to generate a library of structural templates for CATH version 1.7. The performance of this structural template library was then tested by assessing the ability to recognise the correct fold for a structurally validated dataset of remote homologues.

Chapter 4 applies this library of structural templates to recognise the native folding arrangement from approximate models of protein structure. Currently, small protein structures can be modelled to low resolution directly from amino acid sequence using *ab initio* prediction techniques (Jones, 1997; Orengo *et al.*, 1999; Lesk *et al.*, 2001). However, due to the large volume of conformational space that the protein chain can possibly occupy, the time taken to refine these low-resolution models is often prohibitively long. Currently, it is not practical to apply these methods to the large number of genome sequences. To address this, structure comparison methods were employed to recognise the native fold of an approximate protein model at an early stage in the refinement process, thus saving time and identifying constraints from the native fold that could be used to improve the protein model. The fold recognition performance using pairwise SSAP scores and the library of structural templates was then examined with models obtained from a variety of sources. Approximate models of native structures (simulating resolutions of around 3–8Å RMSD from native) were generated and *ab initio* predicted structural models were also obtained both from Simons *et al.* (1997) and from submissions to the *ab initio* category of the third Critical Assessment of Structure Predictions (CASP3). This work was carried out in collaboration with Xavier de la Cruz, University of Barcelona (de la Cruz *et al.*, 2002).

Chapter 5 discusses a novel protocol that incorporates genomic sequence information into multiple structure alignments. The clustering protocol, optimised and tested in chapter 3, provided structurally coherent protein clusters which were subsequently converted into high quality multiple structure alignments. Since structure is more conserved that sequence, structure comparison techniques can be used to validate the alignment between very distantly related protein sequences. Thus, the set of high quality multiple structure alignments allowed the alignments of remote sequence families (identified using SAM-T99 software, see section 5.1.3.3) to be combined in a structurally validated manner. The SAMOSA protocol (**S**equence **A**ugmented **MO**dels of **S**tructural **A**lignments) included very distant sequence relationships and as a result were expected to provide highly sensitive probes of sequence similarity. The increase in performance when including models in the library generated from the SAMOSA protocol was assessed using the same set of structurally validated remote sequences as that used in chapter 3.

Performances measured for all the protocols developed in this thesis confirmed the benefits of using consensus patterns of inter-residue contacts in fold recognition. Methods presented have been employed to facilitate more frequent updating of structures and sequences in the CATH database of protein families.

Chapter 2

Inter-Residue Contacts for Structural Analysis, Comparison and Alignment

2.1 Introduction

2.1.1 Background

Inter-residue contacts are important in protein structure as they help to constrain and define the overall fold. In 1970, Phillips used 2D matrices, or distance plots, to characterise and compare protein structures (Phillips, 1970) (discussed in section 1.2.6.2). These distance plots were 2D matrices displaying the distances between all pairs of residues in the protein structure. A related plot, the contact map (CM), only displays the inter-residue distances which are less than a given threshold value and these contact patterns have also been used to describe and compare protein structures.

Since protein structure is more conserved than sequence, structure comparison methods are often used to identify evolutionary similarities for relationships that are too distant to detect by sequence. However, it has been observed that some protein folds can accept more insertions and deletions than others. Although the buried core of related proteins usually remains conserved, a high degree of embellishments can often be observed in the periphery of the structure (see figure 2.1). As a result, finding a score of structural similarity that can be widely applied across all types of structures in the database is not a trivial task (see section 1.2.6).

Identification of remote evolutionary relationships facilitates the assignment of functional information from sequence or possibly from structure. Defining an au-

1bcn, chain A (domain 2)
Biotin Carboxylase

1iow (domain 2)
D-alanine D-alanine Ligase

Figure 2.1: Structurally diverse relatives from the ATP-dependent carboxylase-amine/thiol ligase superfamily. These distant proteins (18% sequence identity) are functionally similar and share a common buried structural core, however 1bcn, chain A (domain 2) has a high degree of structural embellishment.

tomated method that can reliably distinguish remote evolutionary relationships requires knowledge of features only found to be conserved in related proteins. Russell and Barton examined this problem by looking at the structural features conserved between 607 pairs of proteins with similar 3D structure (Russell & Barton, 1994). Three types of relationship were defined based on the degree of structural and sequence similarity (see table 2.1).

Similarity type	**Conserved features**		
	Sequence	*Function*	*Structure*
A	yes	yes	yes
B	no	yes	yes
C	no	no	yes

Table 2.1: Three types of protein relationships as described by Russell & Barton (1994).

Homologues were defined as pairs of proteins with similarity types A and B, and analogues as pairs of proteins with type C similarity (i.e. sharing similar structures by chance rather than as a result of divergent evolution). After structural alignment using the STAMP package (Russell & Barton, 1992) (see section 1.2.7.2), the structural similarity score and root mean square deviation (RMSD) of C_α atoms

found little or no separation between type B and type C similarities. Also, analysing these structural alignments demonstrated that accessibility and secondary structure assignment have a level of conservation similar to that expected by chance.

In a single structure the number of residue-residue interactions, or contacts, was found to have an approximately linear relationship with the protein length. Of these contacts only about half were considered favourable, i.e. the interacting residues had attractive physicochemical properties. This suggests that only around half the contacts found in a protein help to stabilise the fold by favourable interaction (i.e. help to cause the fold). The other interactions are therefore just a result of residues being brought into close contact due to other features of the protein structure (i.e. effect of the fold).

Points of contact common to both structures after a structural alignment were also examined for any conserved features. It was found that similar structures could have as little as 20% of residue-residue interactions in common. Although there was little separation between type B and C similarities, there was good separation between similar and dissimilar structures, especially when using the contact definition of C_β-C_β distances closer than 8Å. For many distantly related proteins (type B similarities), the number of interactions that were both shared and favourable was found to be the same as that expected by chance. This suggested that many proteins with structural and functional similarities have completely different stabilising interactions.

This work highlighted the value of using inter-residue contacts to differentiate between related and non-related structures. Also, that the most discriminatory definition of an inter-residue contact was using a C_β-C_β distance of less than 8Å. However, this work was limited by the relatively small size of the structural databases in 1994. Now the structure databases are more populated, there are many protein families that contain a large number of related structures. The structures within these protein families can provide insights into evolutionary processes and a great deal of evolutionary information can be extracted from analysing these structural families as a whole. Analysing and comparing families of related proteins, rather than pairs, allows the identification of structural features that have been highly conserved during evolution. The identification of these highly conserved structural features, especially the conservation of inter-residue contacts, forms the basis of the work described throughout this thesis.

Various other structure comparison and analysis methods have also used inter-residue contacts to discriminate between related and non-related structures (Holm & Sander, 1993; Russell & Barton, 1994; Orengo, 1999). Inter-residue contacts

can be calculated using any given side-chain atom and distance threshold. However, it is generally accepted that the optimal definition of an inter-residue contact, i.e. the most descriptive, is where the C_β atoms of each residue are within 8Å (Lesk *et al.*, 2001). It is likely that an increase in discrimination when using C_β distances (e.g. rather than C_α) is due to the fact that C_β atoms, being further away from the peptide backbone, contain information regarding the orientation of the side-chains and this increases their description of the local environment. Since it is generally the side-chains that are involved in the important stabilising inter-residue interactions, it is intuitive to use a measure that incorporates a more descriptive account of the side-chain orientation. When analysing contacts involving glycine residues, which do not contain a C_β atom, a dummy atom was created based on the typical stereochemical features of the C–C bond.

Since contacts act to constrain the protein fold, knowledge of which residues are in contact can be helpful in predicting the structure of a protein from sequence or when determining the structure of a protein. The experimental technique of Nuclear Magnetic Spectroscopy (NMR) identifies such distance constraints and allows models of the protein structure to be built that are consistent with this data (see section 2.1.2). Contacts have also been predicted directly from amino acid sequences by analysing contact propensities for given residues (see section 2.1.3.1 below) and identifying sites of correlated mutations from large sequence alignments (see section 2.1.3.2 below). Although these methods have not yet proved reliable enough to build accurate models of protein structures (Orengo *et al.*, 1999; Lesk *et al.*, 2001), there has been progress in this field in recent years (Lesk *et al.*, 2001).

2.1.2 Deriving Contacts from Experimental Methods

Nuclear magnetic resonance (NMR) is an experimental technique which allows inter-residue distances to be measured for residues in relatively close spatial proximity (approximately <5Å) through a phenomenon known as Nuclear Overhauser Enhancement (NOE). The assignment and measurement of these close inter-residue contacts reveals a great deal of structural information about the protein fold. However, obtaining such quantitative data for an entire protein from a complex NMR spectrum requires an unambiguous assignment of NMR peaks associated with given residues in the protein. This is currently the rate limiting factor of full protein structure determination by NMR methods.

In order to overcome the problems of overcrowding in NMR spectra, a method has been proposed for visualising only the backbone interactions, thus enabling a faster

assignment (Standley *et al.*, 1999). A fully deuterated, C^{13} and N^{15} labelled sample of the protein is first prepared then the main chain amide deuterium atoms are exchanged for protons. The C, N and H NMR spectra are obtained and assignments made using 3D heteronuclear methods. Since the deuterium atoms in the side-chains are 'invisible' to NMR (deuterium nuclei do not possess spin), this approach produces far less crowded NMR spectra. The longer relaxation times also cause much sharper peaks, allowing a reasonable set of long-range distance constraints to be assigned. Although it would be impossible to derive a complete solution for the full protein structure, this method may provide sufficient constraints for the general folding arrangement to be established (Standley *et al.*, 1999).

2.1.3 Deriving Contacts from Theoretical Methods

2.1.3.1 Prediction of Contacts from Pair Potentials

Several methods have been submitted for predicting protein inter-residue contacts directly from sequence. The general approach is to use statistically derived probabilities of inter-residue contacts, i.e. pair potentials (based on the pairwise distances between amino acids in known structures), to assess the likelihood of residues in the protein sequence with unknown structure being in contact. In order to take into account the effects of local sequence around these amino acid pairs, small sequence windows around the amino acids in contact are often used as input for neural networks (Lund *et al.*, 1997). This 'black box' methodology constructs a model that 'learns' to predict relationships between sequence patterns and contact propensities from the patterns observed in sequences with known structure. This has allowed residue contacts to be predicted directly from sequence. (Lund *et al.*, 1997) compared contact prediction methods and found that neural networks were more accurate at predicting inter-residue distances than using only pair potentials. The results from the neural network were also improved by using sequence profiles from multiple sequence alignments (see section 1.2.3.3) rather than individual sequences.

2.1.3.2 Prediction of Contacts from Correlated Mutations

Another approach for the prediction of contacts is based on the identification of correlated mutation sites, i.e. where the mutation of one residue is influenced by the mutation of another. This method makes the assumption that correlated mutations usually occur as a compensatory effect between residues in physical contact. The challenge is to identify structurally constrained sequence changes from the back-

ground noise of neutral mutational drift. Complications arise where more than one residue is mutated as a compensatory response or when a domino-effect of interactions enables compensatory mutations to occur at spatially distant sites in the protein.

Gobel *et al.* (1994) addressed the problem of identifying sites involved in correlated mutation by generating multiple alignments of homologous sequences, i.e. sequences related by evolution, to describe a common structure. The similarity of amino acid residues at each position in the alignment was then described in a mutation matrix (see section 1.2.3). These matrices used statistically derived similarity constants to compare the residues in the alignment position. All these individual matrices were then compared to identify any correlated changes in the amino acid properties between sequences in the two alignment positions. When the correlation between mutation matrices was above a given threshold, the two residues were predicted to be in contact.

2.1.4 Using a Limited Set of Distance Constraints to Predict 3D Structure

In order to simplify this task, several groups have attempted to assign the protein fold using a minimal set of distance constraints. This additional information is intended to reflect the limited amount of data available from the early stages of techniques such as NMR structure refinement. Smith-Brown *et al.* (1993) used both pre-assigned (DSSP) and predicted secondary structure assignment and introduced a minimum of three distance constraints between each secondary structure unit. These secondary structure units were then added to each other sequentially using a randomly chosen set of distance constraints. At each stage of addition a Monte Carlo molecular dynamics simulation (IMPACT), based on an empirically derived all-atom force field, was used to optimise the global structural arrangement. This procedure was continued until either the whole protein had been modelled or until it became impossible to satisfy the newly added constraints. Four proteins were folded using this method and they gave backbone RMSD of 3–5Å from the native structures, but as many as 147 distance constraints were required to correctly fold the flavodoxin protein which contained only 138 residues.

The problem of folding proteins using a small set of distance constraints has also been addressed in the MONSSTER algorithm (Skolnick *et al.*, 1997). In this method the folding space for the C_α backbone was restricted to discrete positions on a high-resolution lattice, which was able to represent all structures in the PDB

to an accuracy of 0.6 to 0.7Å RMSD. Both intrinsic and knowledge-based potentials were used to provide 'protein-like' behaviour for model structures. This method dramatically reduced the number of distance constraints required to find the correct fold, e.g. flavodoxin is folded with an accuracy of 4.3Å RMSD using only 35 constraints. The authors attributed this considerable improvement to be due to the involvement of knowledge-based potentials, as these possibly compensated for native physicochemical forces not included in the force field potentials.

2.1.5 Aims of this Chapter

This chapter aims to provide a brief description of some of the computer programs used throughout the body of work presented in this thesis. Since many groups have demonstrated the value of using inter-residue contact information for structure analysis and comparison, the work presented throughout this thesis has been largely based on developing novel approaches for using inter-residue contacts in the analysis, classification and fold recognition of protein structures. In order to investigate these areas in an automated manner, it was therefore necessary to write a computer program that could identify, manipulate and display information on inter-residue contacts in a variety of ways.

The COCOPLOT toolkit was written to provide a series of functions that would facilitate the fast and automated handling of inter-residue contact information. This chapter provides an overview of the analytical tools and scoring mechanisms incorporated into this program in addition to a discussion of the research carried out in optimising the identification of conserved contacts from multiple structural alignments.

For structural analysis and diagnostics, the ability to generate contact maps based on single structures was incorporated into the toolkit in postscript format. In addition to contact maps for single structures, consensus contact maps based on multiple structural alignments were also included. The selection criteria for obtaining these consensus contacts from a multiple structural alignment is discussed in greater detail. The procedure and scoring function for the comparison of two protein structures based on a pairwise alignment is presented. This is then extended to allow the comparison between a single structure and a template of consensus contact information.

This chapter also presents a brief summary of an algorithm written to align protein structures only using information of inter-residue contacts. This algorithm was written in order to test the hypothesis that related structures could be reli-

ably aligned and compared even when using only a small fraction of the distance constraints observed in the native structure. This would have implications for the application this approach either using distance constraints taken from NMR experiments or inter-residue contacts predicted directly from protein sequence.

This alignment method was originally intended to facilitate structure determination by NMR in collaboration with Professor Mitsu Ikura in Toronto. However, during this project it became apparent that there were problems identifying sufficient distance constraints from the early stages of the NMR analysis. It was therefore decided to explore other applications for this algorithm, such as structure comparison and fold recognition based on inter-residue contacts predicted directly from sequence. The parameters for this algorithm were optimised and the performance was assessed by attempting to find the minimum percentage of native contacts that were required to recognise the correct fold

2.2 Analysis and Comparison of Contact Maps: COCOPLOT

2.2.1 Overview

The first part of this chapter discusses the COCOPLOT program which has been used for analysis and assessment of structural similarity throughout the body of work covered in this thesis. Examples of contact maps (CMs) representing single protein structures and consensus contact maps (CCMs) that represent the conserved contact patterns observed in multiple structure alignments, or templates, are presented. The variability of consensus contacts within these multiple structure alignments is also considered by generating plots that take into account factors such as alignment gaps and standard deviation.

The contact overlap score is also introduced which provides a measure of the structural similarity based on a structure-structure alignment or a structure-template alignment. This effectively superimposes two contact maps according to a structural alignment then calculates the percentage of overlapping contacts with respect to the structure with the larger number of contacts.

2.2.2 Structural Analysis

2.2.2.1 Contact Maps for Single Structures

The contact definition used in this thesis is where C_β atoms of a pair of residues are within 8Å. Given this definition, a visual representation of all the contacts within a protein structure can be generated. This representation is known as a contact map (CM) (see figure 2.2).

This type of plot allows many structural features to be identified, such as secondary structure and super-secondary structural motifs. Examples of the contact patterns for typical secondary structure interactions are shown in figure 2.2. The distinctive thickening along the diagonal axis corresponds to the close proximity of residues involved in α-helices. Parallel β-strand interactions usually appear as thick lines parallel to the diagonal and anti-parallel β-strand interactions appear perpendicular to the diagonal. Non-local interactions involving α-helices typically have a regular scattered pattern with contacts occurring every three or four residues. This characteristic pattern is due to residues being brought into contact as a result of the periodicity of the α-helix.

Residues that are close in sequence will also be close in 3D proximity and as

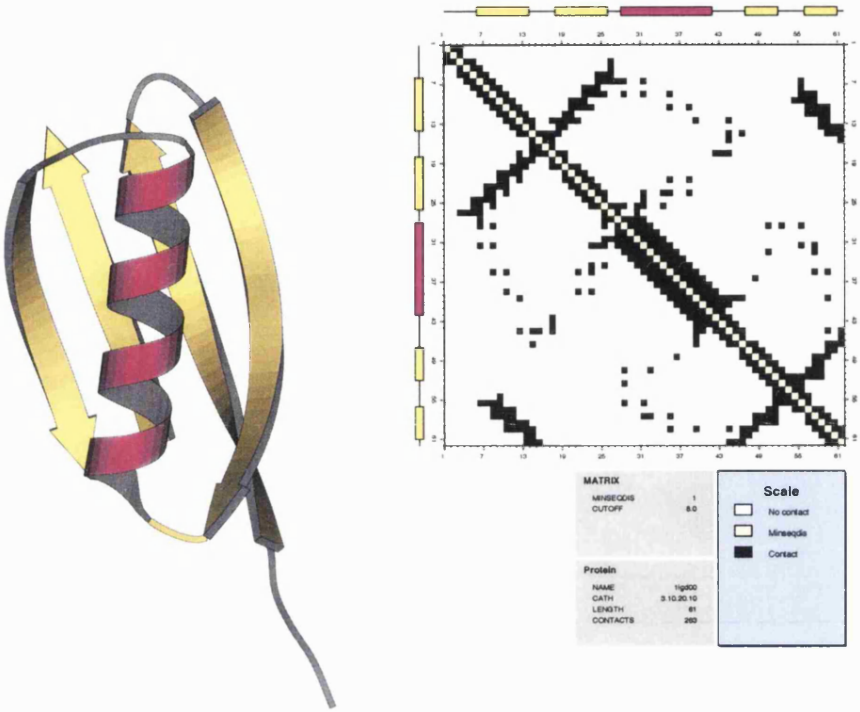


Figure 2.2: Example contact map (I. Sillitoe, COCOPLOT) for a single protein structure (PDB code 1igd) with MOLSCRIPT representation of the 3D structure. A schematic description of the secondary structure is plotted along both axes with α-helices in magenta and β-strands in yellow. The heavy lines parallel to the diagonal are the parallel β-strand interactions, the lines perpendicular to the diagonal correspond to anti-parallel β-strand interactions. α-Helix interactions usually involve a regular contact pattern every three or four residues which corresponds to the periodicity of the α-helix.

a result a large number of contacts are solely due to local interactions within secondary structures. However, these local contacts contain very little information on the overall tertiary structure of the protein. In fact, since these local contacts are common to all proteins, this information often distorts the signal when trying to differentiate between the contact maps of two structures. In order to avoid this, a minimum sequence distance of eight residues was introduced, so only the contacts that occur between residues at least eight residues apart in sequence are considered.

2.2.2.2 Consensus Contact Maps for Multiple Structural Alignments

A large section of this thesis deals with identifying contacts that are highly conserved across families of related structures and using this consensus information for analysis

and classification. Since protein structure is well conserved during evolution, and contacts can be seen to describe the overall fold it would be expected that contacts would also be conserved between related proteins. In addition, inter-residue contacts form stabilising interactions that can play a crucial role in the folding pathway or act as a positive adhesive effect in maintaining the global structure. Contacts having such an active role in the protein structure would be expected to be highly conserved during evolution. By including these highly conserved contacts and ignoring the more variable data from unique embellishments of individual proteins, a structural "fingerprint" can be constructed that encapsulates the most important identifying features of the family.

Identifying these highly conserved contacts can prove useful when analysing the structural features of individual proteins as these contacts represent key interactions in the structure. However, a more important application of these consensus contact maps, in the context of this thesis, is to use these fingerprints to recognise structures that may be too remote to identify by simple pairwise methods.

In order to identify contacts which are conserved throughout related proteins, it was necessary to generate a multiple structure alignment. The CORA algorithm (Orengo, 1999) was used to generate a multiple alignment containing a series of related structural domains. This method uses a double dynamic programming algorithm similar to the SSAP pairwise structure comparison algorithm Taylor & Orengo (1989) and is discussed in further detail in chapter 3.

When generating contact maps for single structures, each residue-residue interaction was assigned as either being in contact or not in contact. When analysing a multiple structural alignment, it was possible to calculate the degree that a contact was conserved by taking two positions in the alignment (i, j) and examining the inter-residue distances for each of the structures at those positions. In this way a consensus contact map was generated describing not just whether a contact was present, but the degree to which a contact is conserved throughout the alignment. Figure 2.3 provides an example of a multiple structure alignment for a family of related proteins, the associated consensus contact map and a superposition of these structures. An example of a consensus contact is highlighted in the contact map and is the result of the comparison of alignment positions 21 and 37. In this case all the structures in the alignment have a contact between these two alignment positions so the contact is deemed to be 100% conserved.

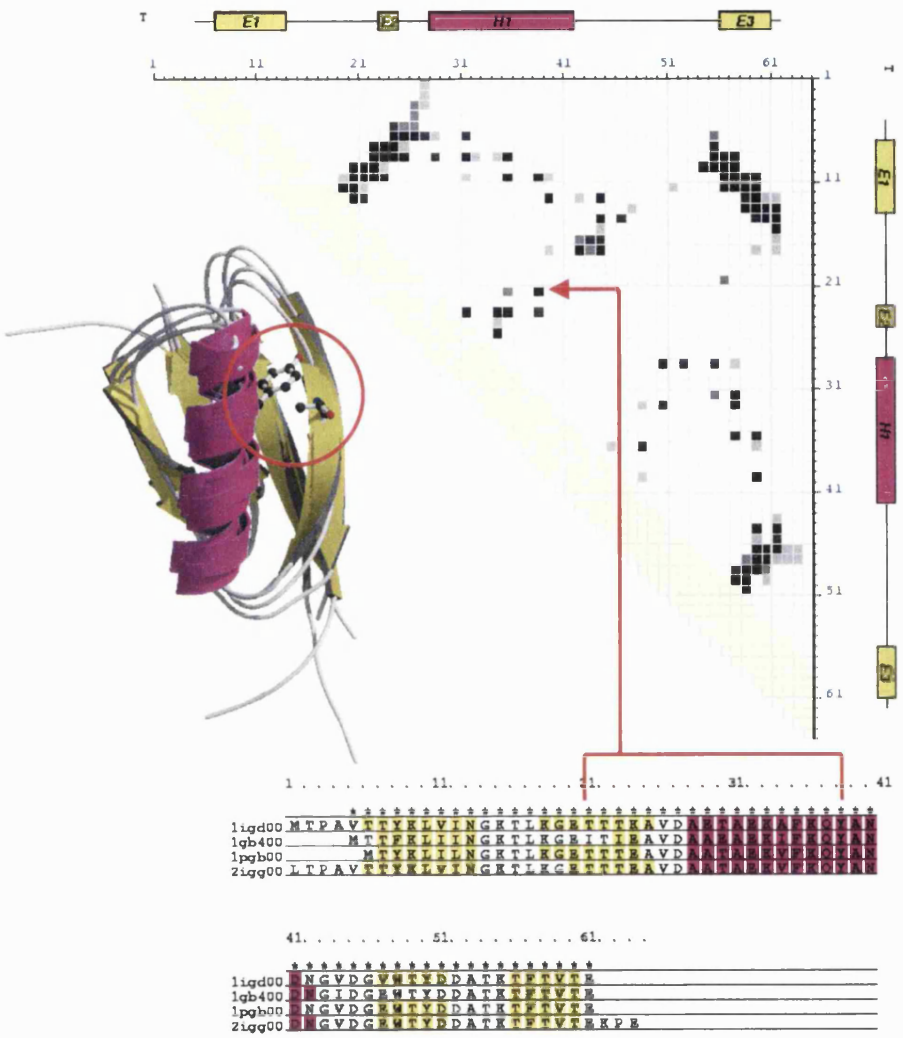


Figure 2.3: Defining a consensus contact by examining interactions between positions in the alignment. The contacts are shown as grey dots with the intensity of each dot indicating the degree of conservation of the consensus contact (black is 100% conserved).

2.2.2.3 Extending the Consensus Plots

Conservation of Alignment Position

When assessing whether a contact is conserved between two alignment positions, it is necessary to take into account gaps that have been included in the multiple structure alignment. A conserved contact is the percentage of all the contacts observed between two alignment positions. However, it is not possible to assess the contact status between residues corresponding to gaps in the alignment. Therefore, a residue-residue comparison for a given structure is considered void if either one of

the residues equates to a gap in the alignment. The degree of contact conservation (CC) for a given alignment position (i, j) can therefore be defined in equation 2.1.

$$CC_{(i,j)} = \frac{\text{Number of structures with residues in contact}}{\text{Number of structures with residues present in both alignment positions}} \quad (2.1)$$

However, if this measure is to be employed as an assessment of contact conservation, it is also necessary to take into account the number of structures used for the calculation. Otherwise, if only one structure had residues present in both alignment positions (i, j) then a single contact in this structure would be considered as 100% conserved despite only being derived from a single structure. Therefore an additional measure of alignment conservation (AC) was also calculated for each alignment position (i, j) (see equation 2.2).

$$AC_{(i,j)} = \frac{\text{Number of structures with residues present in both alignment positions}}{\text{Number of structures in alignment}} \quad (2.2)$$

Since contact maps are symmetrical about the diagonal axis, the two halves of the plots could be used to display a different statistical analysis of the contacts observed in the same multiple structure alignment. In the case of the conserved contact map, the degree of plot intensity reflects the contact conservation in the top-right half and alignment conservation in the bottom-left half. An example of the plot generated by COCOPLOT is shown in figure 2.4. This is derived from the CORA multiple structure alignment of three representative structures in the oxidoreductase superfamily found in the $\alpha\ \beta$-roll architecture of CATH (classification code 3.10.30.70).

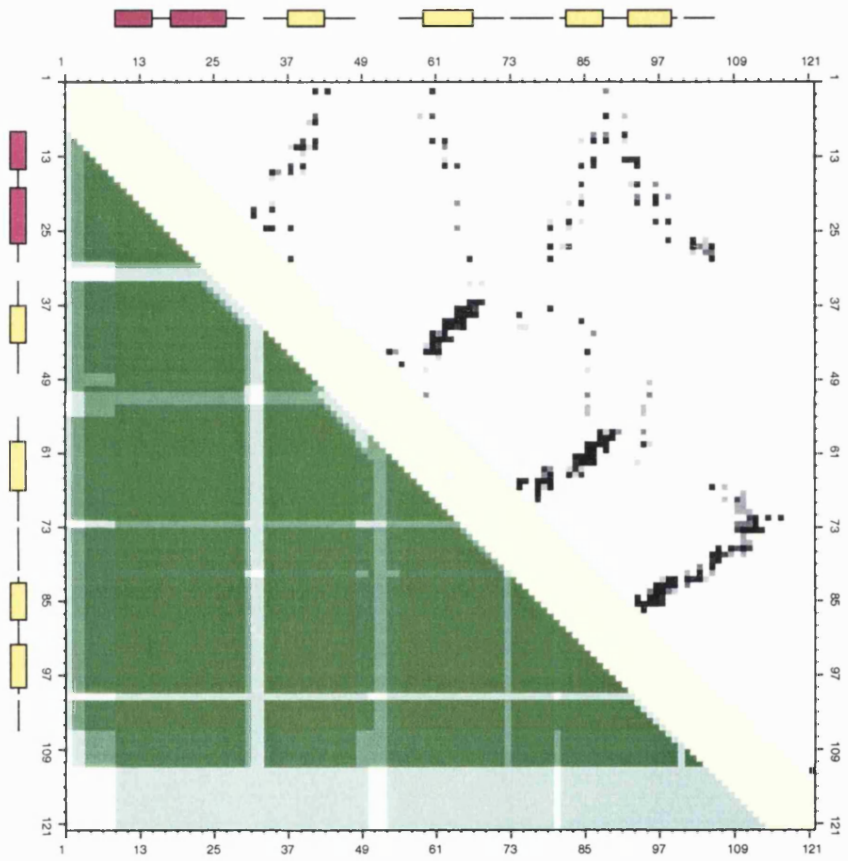


Figure 2.4: A consensus contact map (COCOPLOT, I. Sillitoe) for a multiple structure alignment from the oxidoreductase superfamily in the $\alpha\beta$-roll architecture (3.10.30.70). The top-right half of the plot indicates the degree of conservation of contacts with plot intensity increasing for a higher proportion of contacts between positions in the alignment. The bottom-left half indicates the degree of conservation between the alignment positions regardless of inter-residue distance. The plot intensity increases with a higher proportion of structures included in those alignment positions, thus gaps in the alignment are shown as pale bands and intersecting gaps as white blocks.

Average Distance and Standard Deviation

In addition to generating plots solely based on contacts, COCOPLOT was also written to generate distance plots, where all inter-residue distances are calculated and the magnitude of the distance is indicated by the plot intensity. In the case of the consensus plots from the multiple structural alignments, the distance was taken to be the average inter-residue distance between the two positions in the alignment. Again this distance was only calculated for structures with residues present in both alignment positions (i, j) being considered.

To complement the measure of average distance, the standard deviation for these values was also calculated and displayed in the opposing half of the plot. An example of this plot is illustrated in figure 2.5 using the same multiple structure alignment from figure 2.4. The standard deviation also provides a useful measure of the conservation of inter-residue distances between alignment positions, regardless of the distance itself. Thus, insights can be gained into the conservation of sections of structure that may not necessarily be in contact. As expected, a correlation between low average distance and low standard deviation is observed in most cases, which is especially noticeable between secondary structure interactions. This presents more evidence for the use of contacts as a measure of structural homology since the most conserved regions of structure are usually found close in 3D space.

Interestingly, there are examples where the distances between regions of the multiple structure alignment are well conserved (low standard deviation) despite being spatially distant (high average distance). Since the representative proteins used in this alignment all display less than 35% sequence identity, conserved features are likely to be the result of genuine structural or functional constraints rather than lack of evolutionary distance. This type of highly conserved long-range constraint is unusual within such evolutionarily distant proteins and could therefore be of use when assessing a new structure for evolutionary relationships. However, due to the small number of structural relatives in many superfamilies, the majority of these multiple structural alignments only contain between two and four structures (see section 2.7). This relatively low number of structures negates the use of the standard deviation measure for anything other than a qualitative investigative tool.

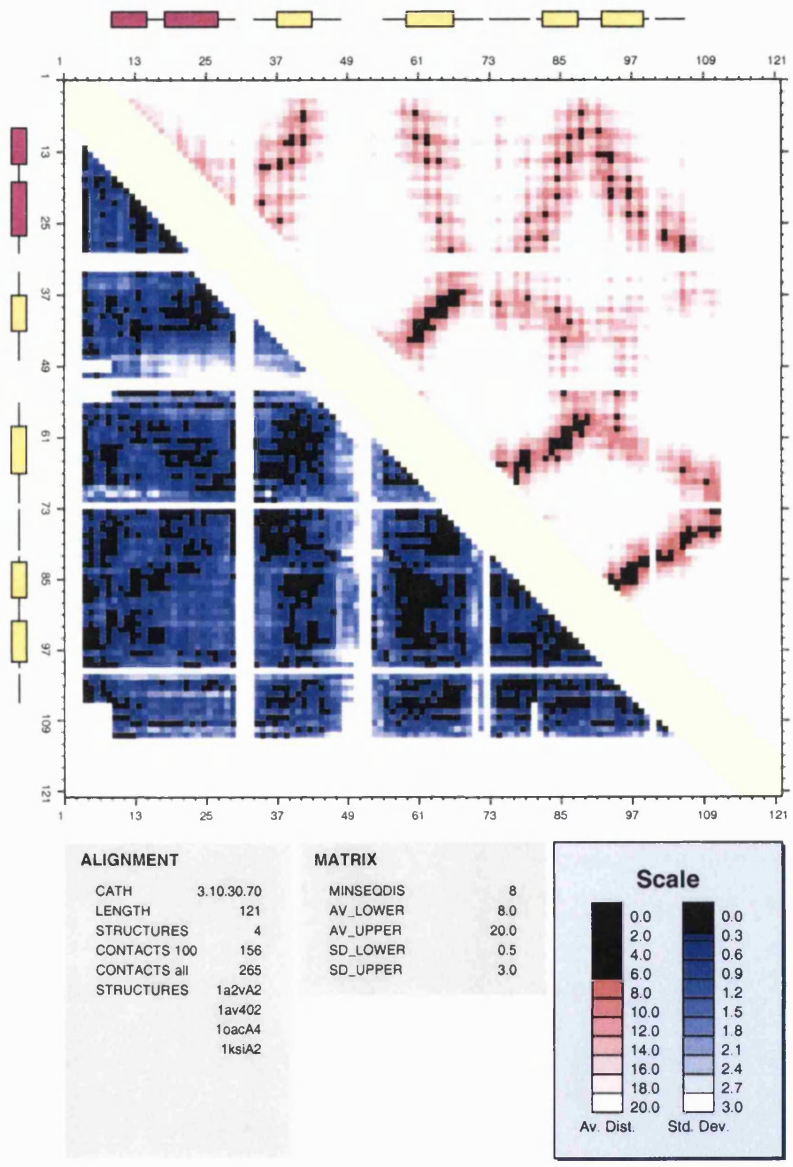


Figure 2.5: A consensus distance plot (I. Sillitoe, COCOPLOT) for the oxidoreductase superfamily in the α β-roll architecture of CATH (classification code 3.10.30.70). The top-right half of the plot illustrates the average distance between alignment positions with plot intensity increasing with shorter distances. The bottom-left half of the plot illustrates the standard deviation for these alignment positions as the plot intensity increases where the structural variability decreases.

2.2.3 Structural Comparison

2.2.3.1 Structure-Structure Comparison

The COCOPLOT toolkit generates a measure of structural similarity based on the number of equivalent, i.e. overlapping, contacts between two protein structures. Inter-residue contacts buried in the protein core not only provide a description of tertiary protein structure but also contribute to the stability of the overall fold. As a result, these contacts are often highly conserved during evolution.

During evolution a structure can be affected not only by point mutations, where one residue changes identity to another, but also by indels of sometimes large fragments of amino acid sequence (see section 1.2.4.1). Therefore, any method that examines these distant structural relationships must take these indels into account by allowing gaps to be introduced into the alignment. For structure-structure comparisons the SSAP algorithm (Orengo & Taylor, 1996) was used to generate a structural alignment between the two proteins. To generate the pairwise contact overlap score, the contact maps of each structure were first generated then the SSAP alignment used to identify equivalent positions in the pairwise alignment where both structures have an inter-residue contact. The contact overlap score, $S_{structure-structure}$, is given by the overlapping contacts as a percentage of the larger number of contacts between the two structures (see equation 2.3).

$$S_{structure-structure} = \frac{C_{overlap}}{C_{max}} * 100 \tag{2.3}$$

Where

$C_{overlap}$ = Number of overlapping contacts between structures (I) and (J)

C_{max} = Max (Contacts_I, Contacts_J)

Figure 2.6 shows the comparison contact map between two actin-binding proteins (PDB codes 2vik and 1svq) based on a SSAP structural alignment. These two proteins are structurally similar (SSAP score of 83), yet evolutionarily distant (17% sequence identity). In the comparison contact map, the contacts in the first structure (2vik) are shown as grey dots, the contacts in the second structure (1svq) are shown as black dots and the overlapping contacts are shown as red dots. The minimum sequence distance of 8 residues can be seen as the yellow band on the main diagonal. This is imposed to avoid including frequently occurring contact patterns between residues close in sequence, since these patterns (typical of secondary structures) are common to both related and unrelated structures.

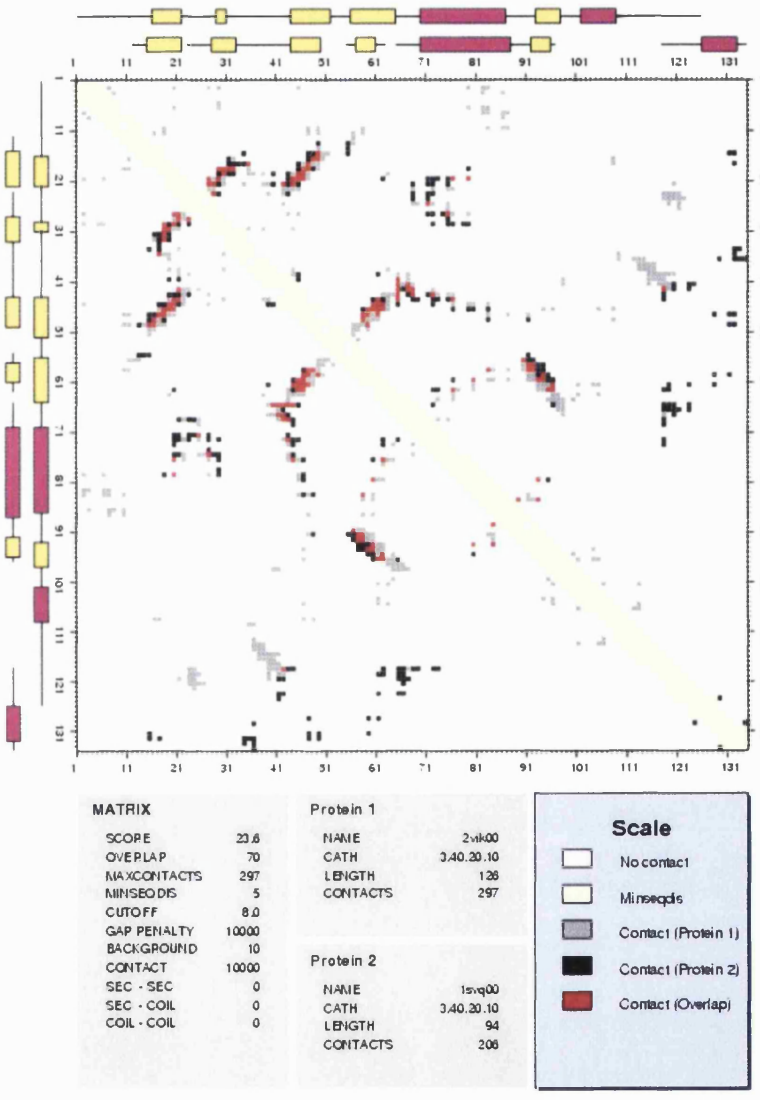


Figure 2.6: Pairwise structure-structure comparison by overlapping contact maps. The alignment between the structures is shown by a secondary structure schematic (alpha-helices in magenta, beta-strands in yellow) with the contacts of the first structure shown as black dots, the contacts of the second structure as grey dots and overlapping contacts as red dots. The values seen in the bottom-left box relate to CONALIGN parameters (see section 2.3)

2.2.3.2 Structure-Template Comparison

To compare the similarity in contact maps between a query structure and 3D template, a structural alignment must again be performed to identify the set of equivalent positions. This structural alignment was achieved using the CORALIGN program from the CORA suite (Orengo, 1999). The CORALIGN program is used to align a single protein structure to the consensus structural template generated from a CORA multiple structure alignment (Orengo, 1999). The COCOPLOT program is then used to generate a CCM for the template and a contact map for the single structure. The contact overlap score, $S_{structure-template}$, is then calculated as the number of contacts that occur between equivalent positions in the alignment (see equation 2.4)

$$S_{structure-template} = \frac{C_{overlap}}{C_{max}} * 100 \tag{2.4}$$

Where

$C_{overlap}$ = Number of overlapping contacts between contacts in structure (I) and consensus contacts in structural template (J)

C_{max} = Max (Contacts$_I$, Consensus Contacts$_J$)

2.2.4 Extending the Definition of a Consensus Contact

The definition of a conserved contact could be given by the criterion that every structure in the alignment is required to have a residue at both alignment positions (i, j) and every one of these residue-residue pairs must be a contact. However, if this was the case it would only take one structure in the alignment to have a gap, or one residue-residue pair not to be in contact, to result in the consensus contact being dismissed on the grounds of being insufficiently conserved. This would result in a bias against alignments with a large number of structures as every extra structure provides an extra possibility for the consensus contact to be dismissed.

This can be highlighted by showing the distribution of the percentage of conserved contacts for different numbers of structures in the alignments (see figure 2.7). The percentage of conserved contacts was calculated by dividing the number of consensus contacts in the multiple alignment by the average number of contacts for the single structures. The plot clearly shows that as the number of structures increases, the percentage of conserved contacts decreases.

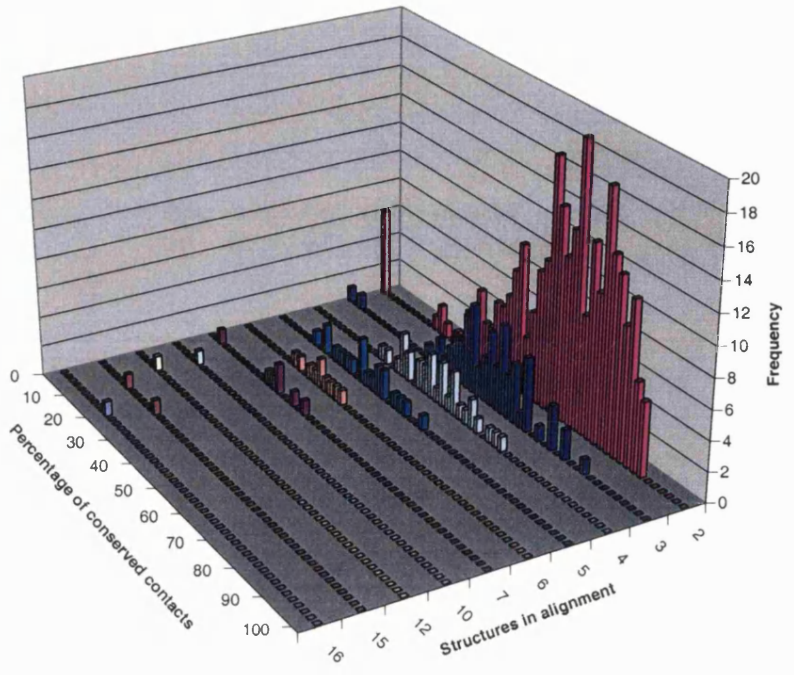


Figure 2.7: Distribution of percentage of conserved contacts with a harsh conservation criteria demonstrates dependence on the number of structures in the alignment

One solution to this problem is to use a ratio rather than an absolute true or false value for these two variables: the proportion of structures that have residues present at both positions in the alignment (minimum alignment ratio, MAR), and the proportion of these residue-residue distances that are actually contacts (minimum contact ratio, MCR). Setting the minimum values for these ratios too low results in large numbers of consensus contacts being classified as highly conserved, thus producing contact maps that are too general to be useful in differentiation between superfamilies. Conversely, setting these ratios too high restricts the consensus criteria to the point that no contacts are deemed to be conserved, making any comparison to the consensus contact map void.

$$\text{Minimum Alignment Ratio (MAR)} = \frac{\text{Structures with residues aligned at } (i,j)}{\text{Total structures}} \tag{2.5}$$

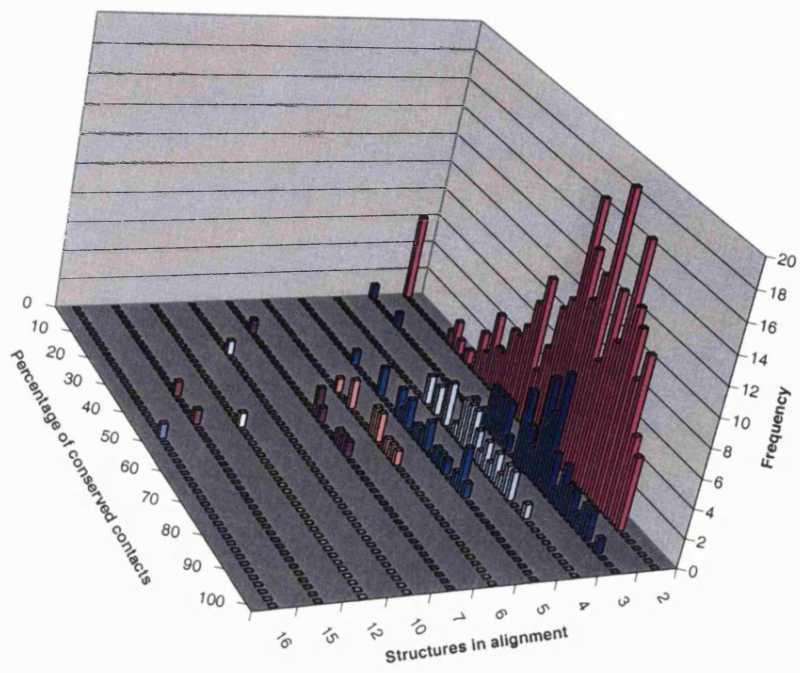
$$\text{Minimum Contact Ratio (MCR)} = \frac{\text{Contacts in alignment positions } (i,j)}{\text{Structures with residues aligned at } (i,j)} \tag{2.6}$$

Figure 2.8(a) shows the effect of relaxing the criteria for selecting a conserved contact. In this graph a consensus contact is regarded as conserved if all the structures have residues present at both positions in the alignment (MAR = 1.0), but

only more than half of the residue-residue interactions between these positions are required to be in contact (MCR > 0.5). The distribution for alignments with larger numbers of structures now reflects more closely the distribution for alignments with fewer structures.

Figure 2.8(b) relaxes this criteria further by adding that a consensus contact is regarded as conserved if more than half of the structures between two alignment positions have residues present (MAR > 0.5) and more than half of these interactions are in contact (MCR > 0.5). Although there are far fewer points in the distributions for alignments with fewer structures, it can be seen that the centre of all these distributions are now within a similar range.

(a)



(b)

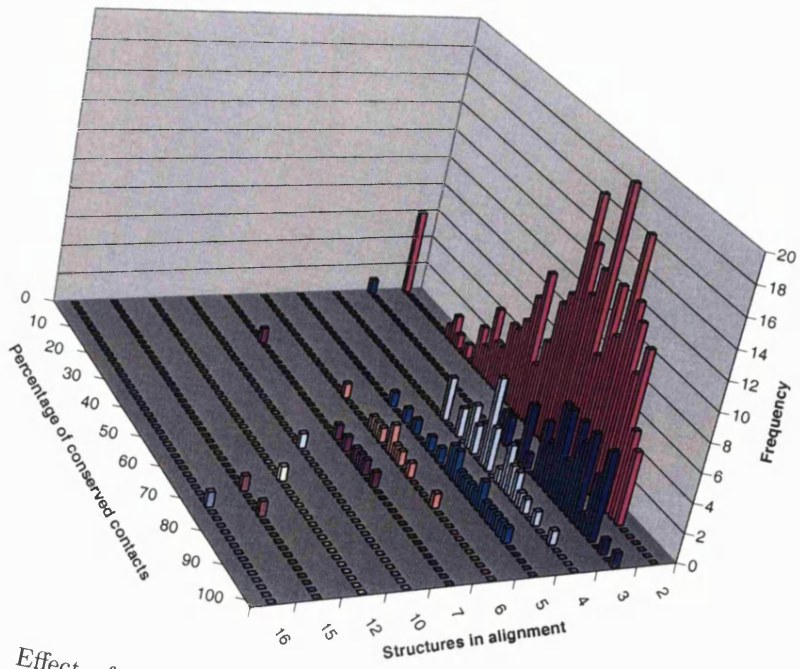


Figure 2.8: Effect of relaxing the criteria of conservation on the distribution of percentage of conserved contacts. (a) has the conservation criteria ($MAR = 1.0$, $MCR > 0.5$), (b) has the conservation criteria ($MAR > 0.5$, $MCR > 0.5$)

2.3 Protein Structure Alignment from Contact Data: CONALIGN

2.3.1 Overview

The second part of this chapter describes the double dynamic programming algorithm written to align protein structures using only contact data, CONALIGN (I. Sillitoe, computer program). The aim was to devise a method that could correctly identify the fold of a query protein for which only limited structural data was available (e.g. a subset of all inter-residue contacts). Such a method would facilitate structure determination by NMR (e.g. using contact data derived from NOE interactions). Alternatively, inter-residue contacts predicted directly from sequence could also be used (e.g. from correlated mutations). This algorithm used three main parameters to align two sets of contact data; background score (BG), contact score (CN) and gap penalty (GP). The protocol and results from the optimisation of these parameters are described below. The algorithm was then tested using three example structures from the CATH database.

2.3.2 Implementing the Double Dynamic Programming Algorithm

2.3.2.1 Dynamic Programming

The dynamic programming (DP) algorithm (Needleman & Wunsch, 1970) is described at length in section 1.2.4.2. In summary, the algorithm provides the optimal alignment between two sets of data (e.g. amino acid sequences for protein A and protein B) based on a pre-defined scoring scheme (e.g. matching residues score 5, non-matching residues score 0). The score matrix, a 2D matrix with the residues of protein A on one axis (length N_A) and the residues of protein B on the opposing axis (length N_B), is filled with comparison scores for each combination of residues between the two proteins. The score matrix is then accumulated, starting from the bottom-right corner, by adding the comparison score for each cell (i, j) in the matrix to the highest score from the previous row or column starting from $(i + 1, j + 1)$. Since insertions and deletions generally occur less frequently than residue substitutions (in terms of events accepted in protein evolution), a gap penalty is incurred when opening a gap in the alignment (e.g. gap penalty, $GP = -2$). Thus, if a score is inherited from any cell other than the diagonal $(i+1, j+1)$, e.g. from the previous column $(i+1, j+2..N_B)$ or row $(i+2..N_A, j+1)$, the gap penalty is included in the

inheritance score. The inheritance path, i.e. the cell containing the inherited score, is also stored in each cell. The highest scoring alignment path is then identified in the traceback step by starting with the highest score in the matrix (always from the left-most column or top-most row) and traversing through the inheritance path.

2.3.2.2 Double Dynamic Programming

The SSAP algorithm uses DP on two levels to compare protein structures and as such, has been called double dynamic programming (DDP) (see section 1.2.7.3). The first level of DP compares the residue environments, or vector views, between all potentially equivalent residues in the two structures (residues are deemed potentially equivalent if they share similar accessibility and torsional angles). That is, for a given residue comparison (i, j), rather than comparing residue identities, each cell in the residue-level score matrix (x, y) now compares the C_β–C_β vector between atoms i and x ($i\vec{.}x$) to the C_β–C_β vector between atoms j and y ($j\vec{.}y$). If the score from the optimal alignment path exceeds a given threshold, the structural environments for residues i (structure A) and j (structure B) are deemed similar and all the scores from this alignment path are added to the summary score matrix. After DP has been applied to identify the alignment paths of all these equivalent residues, DP is again used in a final pass of the scores in the summary score matrix. This provides the optimal alignment through all the residue-level paths and represents the best alignment between the two structures based on the scoring scheme provided.

Since it is rare to find extremely large indels, a window is imposed upon both the residue-level score matrix and the summary matrix for efficiency. This is calculated by taking the differences in the protein length and adding 50 (Taylor & Orengo, 1989). Residue comparisons are therefore only made if they fall within this window, thus reducing the algorithm search time considerably.

2.3.2.3 CONALIGN

The CONALIGN program was written using the C programming language and employs a similar DDP algorithm to that used in SSAP. However, rather than comparing all the internal C_β–C_β vectors between two protein structures, CONALIGN assesses the similarity of the C_β–C_β contact environments. In order to illustrate the procedure for this DDP algorithm, CONALIGN is used to compare two actin-binding proteins sharing 17% sequence identity and a SSAP structural similarity score of 83 (PDB codes 2vik and 1svq). Figure 2.9a illustrates a typical CONALIGN summary matrix resulting from the comparison of these two structures. This matrix contains

the scores from all the high scoring residue-level paths, i.e. the best alignment of the structure from the perspective of residues (i, j). The accumulation and traceback steps to find the optimal global algorithm from the summary matrix is illustrated in figure 2.9b. Figure 2.10 demonstrates the comparison contact map of these two structures based on the CONALIGN alignment.

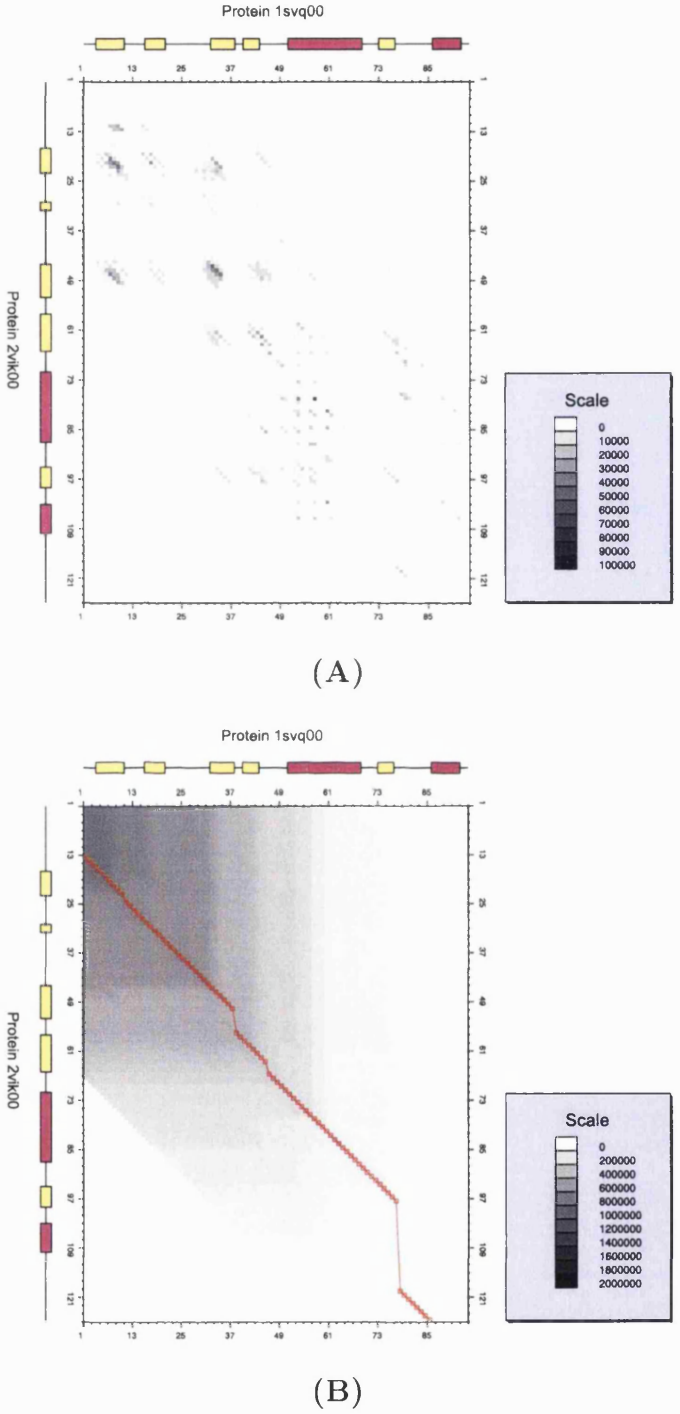
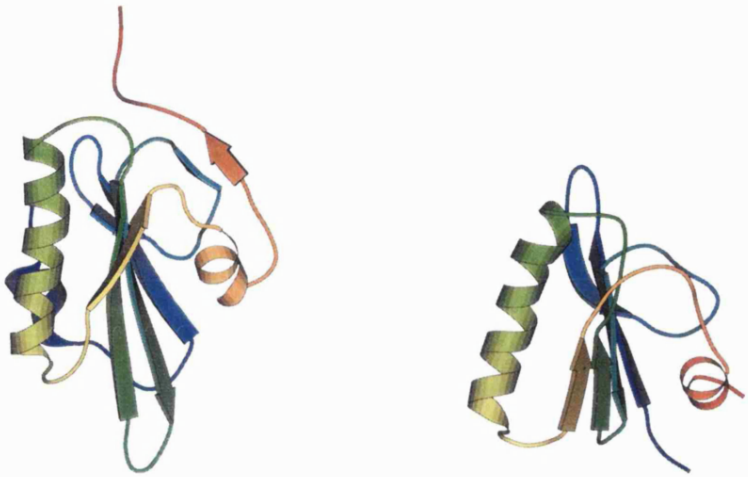


Figure 2.9: Illustration of double dynamic programming. (A) Demonstrates the summary score matrix for the comparison of two actin-binding proteins (PDB codes 2vik and 1svq). (B) Demonstrates the accumulation and traceback steps based on the summary score matrix. The red path shows the optimal alignment between the two sets of contact data as identified by the DDP.



2vik
Actin-severing domain Villin 14T

1svq
Severin domain 2

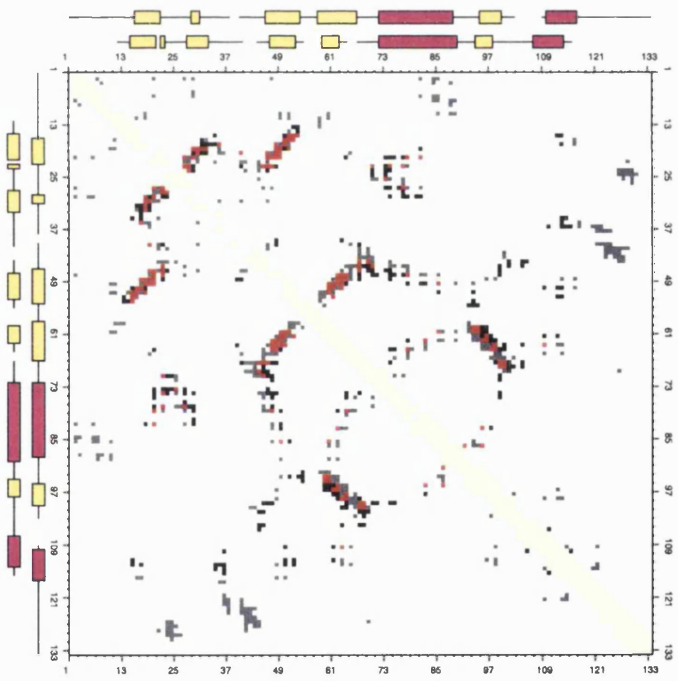


Figure 2.10: Comparison contact map based on the CONALIGN alignment for the two actin-binding proteins (PDB structures 2vik and 1svq). MOLSCRIPT (Kraulis, 1991) representations of the two structures are illustrated (coloured red to blue from N-terminus to C-terminus respectively) The contact maps are superposed and compared using the CONALIGN alignment to assign equivalent residues.

2.3.3 Optimisation Protocol

2.3.3.1 Overview

The basic premise of the optimisation procedure was to identify CONALIGN input parameters that would provide the greatest discrimination between related and non-related structures (see figure 2.11). Three structures were selected for this optimisation procedure spanning the three main classes in CATH. The fumerase protein (PDB structure 1fup, chain A, domain 2) was used to represent the mainly-α class, the oxidoreductase protein (PDB structure 1cbj, chain A) used to represent the mainly-β class and the DNA-binding bovine papillomavirus (PDB structure 2bop, chain A) used to represent the mixed-α β class (see table 2.2).

Test Protein	**PDB code**	**Class**	**Length**	**Related structure(s)**
Fumarase	1fup, chain A (domain 2)	mainly-α	269	1
Oxidoreductase	1cbj, chain A	mainly-β	151	17
DNA-binding bovine papillomavirus	2bop, chain A	mixed-α β	85	27

Table 2.2: Dataset of structures used to optimise CONALIGN parameters. The number of related structures is the number of homologous superfamily representatives sharing the same topology as the test protein (after structures with $\geq$25% sequence identity have been removed.

The three main CONALIGN parameters requiring optimisation can be described as follows:

- **Contact Score** (CN)

  This parameter was used to signify a match on the residue-level score matrix. For example, when comparing residues (i, j), a contact between $(i.x, j.y)$ would result in the the cell (x, y) having a comparison score of CN.

- **Gap Penalty** (GP)

  The gap penalty: a negative value introduced when opening up a gap in the alignment, i.e. inheriting scores from cells other than $(i + 1, j + 1)$ during the accumulation phase would incur a penalty of $-GP$.

- **Background Score** (*BG*)

  Since a protein structure typically only displays around 3 contacts per residue, there is often little data to direct the alignment path of the residue-level score matrix. To address this, residues sharing the same secondary structure state were given a positive background score (*BG*). When applying this algorithm to situations with only limited 3D data (e.g. NOE distance constraints, predicted contacts), the secondary structure assignments could be taken from automated secondary structure prediction servers (e.g. PSI-PRED, Jones (1999b)).

Rather than attempting to independently vary three parameters, two ratios were employed to implicitly describe all three parameters (i.e. GP/CN and BG/GP). From an early stage in the optimisation procedure, it became obvious that the GP/CN parameter ratio was insensitive to small changes. Therefore a logarithmic scale was used to test the range for this value. The parameter space for this value ranged from 1.0 to 0.0002 (corresponding to $-\log(\text{GP/CN})$ values of 0 to 3.70), whereas the ratio (BG/GP) ranged from 0.05 to 0.5.

To assess the discrimination power for each of these parameter settings, it was necessary to generate a validated dataset containing related (i.e. same topology classification code in CATH) and non-related proteins for each test structure. Since the main focus of this procedure was to maximise the differentiation between related structures and the first non-related structures, it was not necessary to use a full database search for every change in the parameters. Thus, each test structure was first searched against the 693 superfamily representatives in CATH v2.0 using the SSAP structure comparison algorithm. Structures sharing $\geq 25\%$ sequence identity were removed from the list to avoid the possibility of recognising close structural homologues. The top 100 structures from these database scans were then used, providing a subset of the closest, unrelated superfamily representatives for each test structure. The test structures were then searched against their specific structural libraries by applying the CONALIGN algorithm with all possible combinations of parameters ratios within the accepted ranges (see above). In addition to the contact overlap score calculated from the CONALIGN alignment, the SSAP score for the same comparison was calculated in addition to the Z-score (based on the contact overlap score). The Z-score (described in section 1.2.5.2) assesses the significance of a result by calculating the number of standard deviation units a given score is from the mean of the distribution.

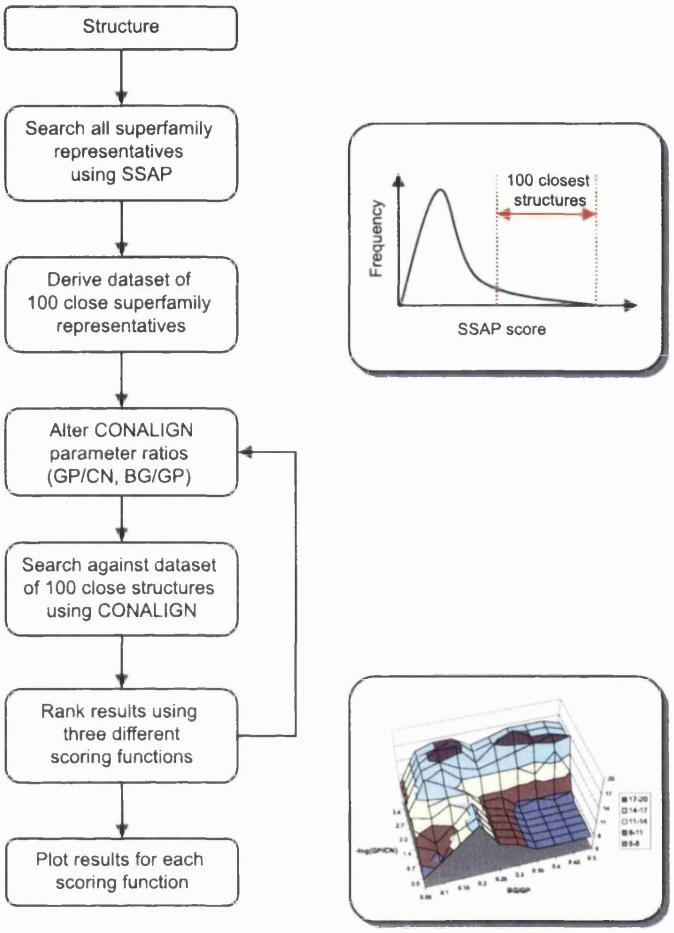


Figure 2.11: Summary flowchart of the CONALIGN optimisation procedure.

2.3.3.2 Scoring Schemes

A global scoring scheme was created to quantify the effect of each parameter setting on the ability of the CONALIGN algorithm to discriminate between related and non-related structures. The results from each database scan were ranked according to decreasing contact overlap score and each comparison was assigned a related or non-related status depending on whether the two structural domains were found in the same CATH topology (i.e. sharing the first three classification numbers). Two global scores were investigated that attempted to provide an overall description of the differentiation between scores for related and non-related structures.

Global Score (1): Coverage at Zero Error

The first score was simply the number of related structures found before the first non-related score, i.e. coverage at zero error. Obviously the maximum value for this score was heavily dependent on the total number of related structures in the dataset and scores were not compared between test structures. However, within the results for a given test structure, this score provided a reasonably sensitive account of the effect of the parameter settings on introducing errors into highly ranked positions in the database search. This scoring scheme provided a similar set of results for all three test structures so only one example is shown based on the results from the mixed-α β protein 2bop, chain A (see figure 2.12). This figure plots the values of the parameter ratio on the x and y axes and plots the value of the global scoring scheme on the z-axis. The plot provides a 3D landscape that describes the parameter space with peaks corresponding to parameters performing well and troughs corresponding to parameters performing poorly.

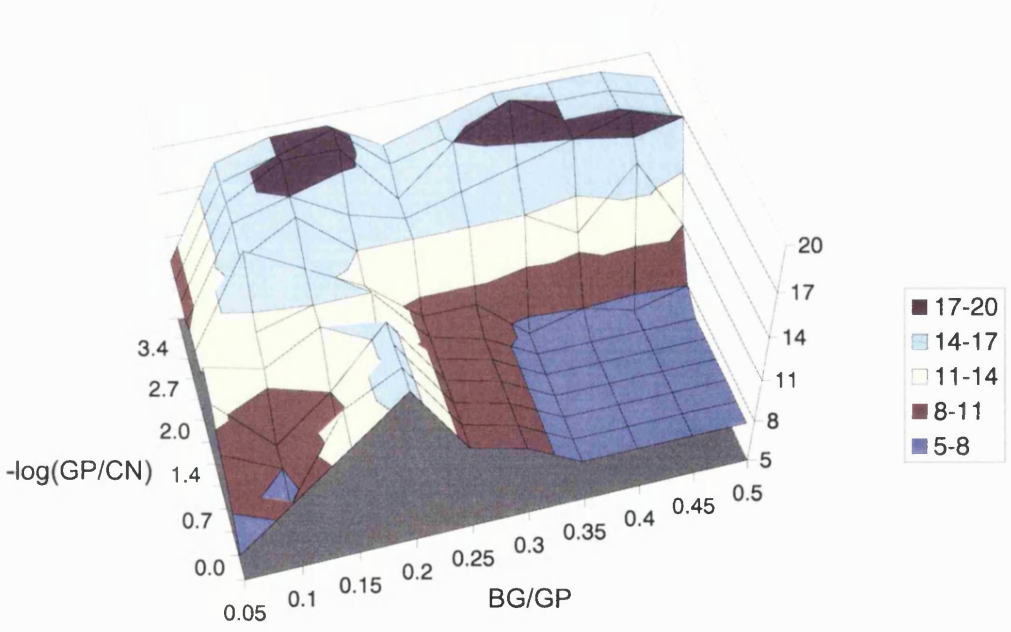


Figure 2.12: Results of optimisation score (1) for the mixed-α β structure 2bop, chain A. This score was based on the number of related matches observed before the first non-related structure (shown as the height of the surface and the legend on the right). The peaks in the parameter landscape correspond to fewer errors, i.e. non-related structures, appearing before related structures.

Global Score (2): Distance between the distribution of related and non-related scores

The second score provided a description of the difference in the distributions between the distribution of contact overlap scores for related and non-related structures (see figure 2.13). Z-scores, based the contact overlap score from the CONALIGN alignment, were used to describe the significance of each comparison score, then average Z-scores were calculated for both the related and non-related structures. The global score (2) was given as the difference between these two averages. Again, the parameter surfaces for all three structures were similar so the results for 2bop, chain A are highlighted as an example (see figure 2.14).

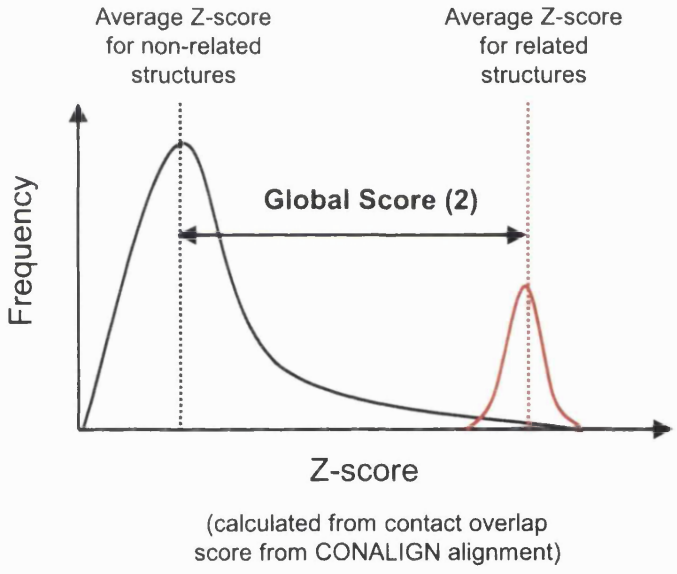


Figure 2.13: Illustration of the calculation of global optimisation score (2) describing the difference in distributions for contact overlap scores of related and non-related structures.

2.3.3.3 Summary of Optimisation Results

The results from the two scoring schemes display similar 3D surfaces for all three test structures. Figures 2.12 and 2.14 both contain a ridge corresponding to a (BG/GP) ratio of around 0.2 and these surfaces also both plateau at a $-\log(GP/CN)$ ratio of around 2.4. Since the two scoring schemes are based on quite different approaches, the apparent agreement between the two methods suggests that they provide a consistent measure of the discrimination between related and non-related structures.

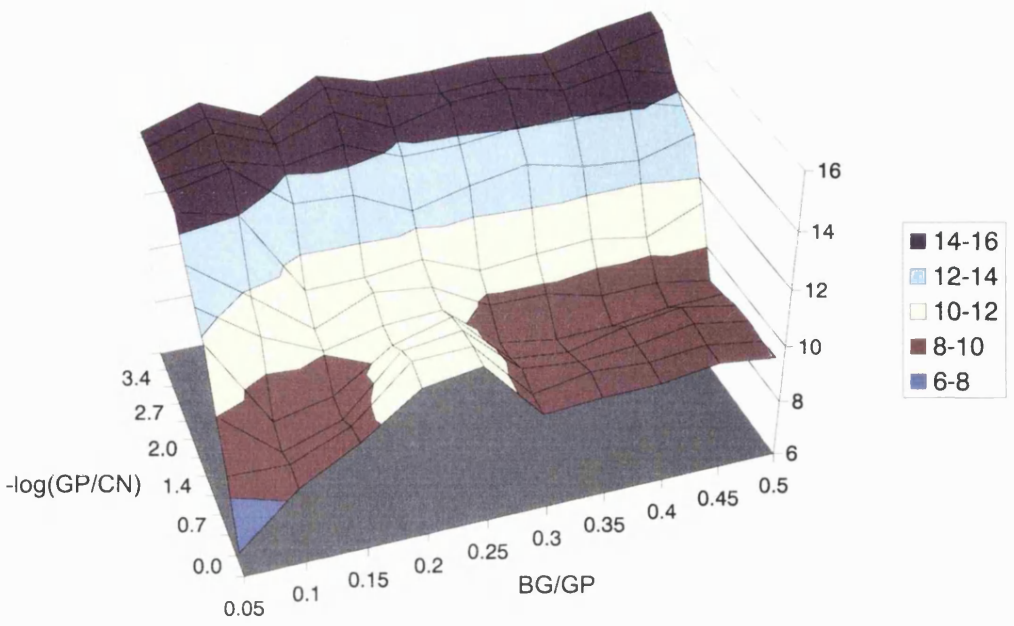


Figure 2.14: Results of optimisation score (2) for the mixed-α β structure 2bop, chain A. This score was calculated from the difference between the average Z-scores for related structures and the average Z-scores for non-related structures

The stable region of parameter space, corresponding to large values of CN, i.e. high values of $-\log(GP/CN)$, is likely to be the result of a lack of information at the residue level when only considering inter-residue contacts. That is, in contrast to comparing the vector environments of residues where the vectors to all neighbouring residues can be compared, this algorithm has to provide the alignment based on far fewer guides.

2.3.4 Testing the algorithm

The ability of the algorithm to recognise the native fold of a protein from partial contact data was examined by randomly removing contacts from the three test structures and identifying the point at which the correct fold is no longer recognised. The test structures were first searched against the database of structures using CONALIGN, then the resulting scores were ranked by decreasing contact overlap score. A consensus view of the folds seen to occur in the top five positions was then taken from these results, i.e. a running total considering the frequency of the first three CATH classification digits (C.A.T) for the top five contact overlap scores. If the correct topology appears more frequently than any other topology in these top 5 positions, the CONALIGN algorithm is deemed successful as the fold has been correctly recognised. This scheme is also described in chapter 4.

The reduced sets of contact data (models) were generated for each of the three test structures by randomly removing a given percentage (between 10% and 90%) of the contacts observed in the native structure. To account for possible irregularities when randomly removing contacts (i.e. some contacts may prove more important than others when attempting to recognise the protein fold), 20 different sets of randomly selected contact data were generated for each test structure at each percentage threshold (0, 10, 20...90%).

The results of this experiment were then plotted for each test structure by counting the number of models able to recognise the correct fold within each percentage bin (out of the maximum number of 20 models). Figure 2.15 provides a histogram summarising the results for all three test structures. This can be illustrated by highlighting a particular set of results. For the fumarase protein (PDB code 1fup, chain A, domain 2), out of the 20 models generated by randomly selecting a set of 20% of the native contacts, only one model failed to recognise the correct fold. Indeed, even when using as few as 10% of the contacts observed in the native structure, the fold recognition protocol identified the correct fold in every model, for all three test structures.

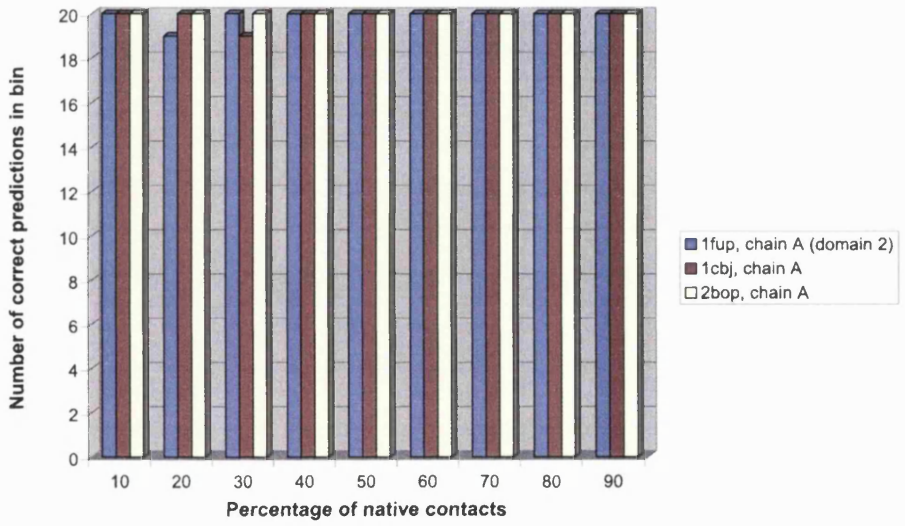


Figure 2.15: Results from the sets of reduced contact data. For each of the three test structures, 20 models of randomly selected contacts were generated from a series of thresholds for the percentage of native contacts (e.g. 10, 20, ...90% of the contacts observed in the native structure). The y-axis simply counts the number of models that could be assigned the correct fold (from the maximum number of 20 models in each percentage bin.

2.4 Discussion

This chapter has described methods for displaying inter-residue contacts and has introduced simple scoring schemes for comparing contacts between 3D structures and identifying those contacts that are highly conserved across a family of related structures.

A proposal for future work is to invert this problem by assessing the number of correlated mutations occurring in spatially close residues. This would involve comparing mutation matrices for positions in a structural alignment that are known to be in contact. The CATH database (version 2.0) holds structural alignments for 362 homologous families and a procedure for identifying residue contacts within these structural alignments has already been presented in this chapter. This work would provide useful information on the occurrence and characteristics of correlated mutations between both individual and conserved contacts within a structural family. For example, what is the total number of correlated mutations observed in a given homologous family? Are contacts conserved across a structural family more likely to exhibit correlated mutation behaviour? Are correlated mutations more likely to occur in residues near an active site? Answers to these questions certainly have implications for contact prediction and may also help to increase our understanding of the mechanisms involved in the evolution of protein structure.

An algorithm has been presented which allows a protein with limited structural data (e.g. inter-residue distance constraints from NMR data) to be scanned against a library of 3D structures in order to identify the correct fold group. The three CONALIGN parameters have been optimised in a comprehensive manner through the introduction of cross-validated scoring schemes.

The initial testing protocol, discussed in section 2.3.4, provided an interesting set of results as it suggested that the native fold could be found from a small number (as low as 10%) of native contacts for the three structures tested. However, these initial results cannot be viewed as an exhaustive examination of this structure comparison algorithm.

If this algorithm were to be considered as a possible application of fold recognition from contacts predicted by sequence, it would be necessary to answer further questions. For example, how well does the method perform when incorrect contact data is included in the reduced set of native contacts? Could the information used to compare residues be expanded to include predicted accessibility or residue similarity scores from sequence substitution matrices? Is there any increase in performance when using highly conserved contact data from multiple structure alignments?

At the time of developing this algorithm, the accuracy of *ab initio* prediction of contacts from correlated mutations was deemed to be very poor. The contact analysis tools developed in this chapter were used to assess submissions to the contact prediction category of CASP3 (Orengo *et al.*, 1999) (see attached paper). The accuracy of these sets of predicted contact data were scored by simply shifting the alignment between the predicted contacts and native structure one residue at a time and calculating the overlap score at each step. This effectively provided a distribution of random contact overlap scores and allowed a significance score (Z-score) to be calculated for the original alignment. However, the most significant prediction only gave a Z-score of 3.4 (Casadio group, Fariselli & Casadio (1999)) and all other predictions gave Z-scores in the range 0.2–2.8.

Furthermore, the method being developed by our collaborator Prof. Ikura was also unable to rapidly obtain NOE contact data. This reduced the value of using CONALIGN to speed up structure determination by NMR since the most time consuming step was still the assignment of peaks in the spectra (which precedes the assignment of NOE data). Therefore, rather than further optimising and testing CONALIGN as suggested above, it was decided to pursue other related research themes based on comparing inter-residue contacts for structure prediction and classification. However, this algorithm has demonstrated sufficient promise to warrant future research, especially since the accuracy of predicting contacts from sequence continues to improve (Lesk *et al.*, 2001).

Chapter 3

Generation and Application of Representative Structural Templates for Homologous Superfamilies in CATH

3.1 Introduction

3.1.1 Background

At the end of 2002, the Protein Data Bank (Berman *et al.*, 2000) contained the 3D co-ordinates for more than 19,000 protein chains. One common approach in rationalising this amount of data is to group together proteins based on similarities in protein sequence, structure or function. Many classification schemes have been proposed in order to provide reliable clusters of related protein structures, using varying degrees of manual intervention such as CATH (Orengo *et al.*, 1997; Pearl *et al.*, 2001b), SCOP (Murzin *et al.*, 1995; Lo Conte *et al.*, 2000), FSSP (Holm *et al.*, 1992) and 3Dee (Dengler *et al.*, 2001) (see section 1.4.1 for descriptions of each). The number of structures available to these databases is expected to expand dramatically with the advent of several large-scale structural genomics initiatives (Pennisi, 1998). As a result the challenge now facing structure databases is to provide methods to cope with this influx of structures and also to fully utilise this wealth of data.

Grouping related proteins together into evolutionary clusters with similar structures provides two main benefits. First, the enormous amount of redundancy present in the database can be reduced by selecting a single structure to represent a whole

cluster of related proteins rather than considering each structure individually. Also, once these evolutionary clusters have been defined, the common structural features and highly conserved amino acid positions can be identified to help to provide insights into evolutionary relationships which may not be apparent from analysis of the separate structures.

The technique of using the consensus information from a series of related proteins to examine constraints on protein evolution is well established in the field of protein sequence analysis. When aligning a series of related sequences, it is possible to identify recurring patterns using either the identities or chemicophysical properties of amino acids at each position in the alignment. If a particular amino acid or amino acid property is seen to appear in a large number of non-redundant sequences then it is likely that this residue feature has been conserved due to a functional or energetic constraints (Mirny & Shakhnovich, 1999). This constraint may be structural in nature as it could represent an important interaction in the folding pathway, or it could be functional as it could represent an active site residue which is vital for the biological function of the protein. Either way, the accumulation of this consensus information from a set of related proteins can be used as an identifying fingerprint that describes important evolutionary features.

The concept of gathering a consensus of information from related proteins can also be applied to the field of protein structure alignments. However, using structure rather than sequence information enables this concept to be extended to even more distant evolutionarily relationships since protein structure is more conserved than sequence during evolution (Sander & Schneider, 1991; Flores *et al.*, 1993; Orengo *et al.*, 1993). This is illustrated in figure 3.1 by comparing haemoglobin, α-chain from pig (1QPW, chain A) with haemoglobin, domain 1 from pig roundworm (1ASH) and haemoglobin, α-chain from horse (1IBE, chain A). The proteins involved in both of these comparisons have highly similar structures (SSAP scores greater than 80) and similar functional characteristics (oxygen-binding proteins), yet the sequence similarity between 1IBE and 1ASH is low (11% sequence identity).

For this reason, and also because tertiary structure contains so much more information than amino acid sequence, alignments from structural comparisons usually prove to be far more robust than those based on sequence for detecting distant evolutionary relationships. As a result, structural alignments have often been used to validate sequence alignments of distant proteins (Gotoh, 1996). Increasing the accuracy of the alignment in turn increases the ability to recognise conserved features, especially when aligning large numbers of distant structures. Thus, multiple structural alignments provide a powerful tool for identifying residues with functional

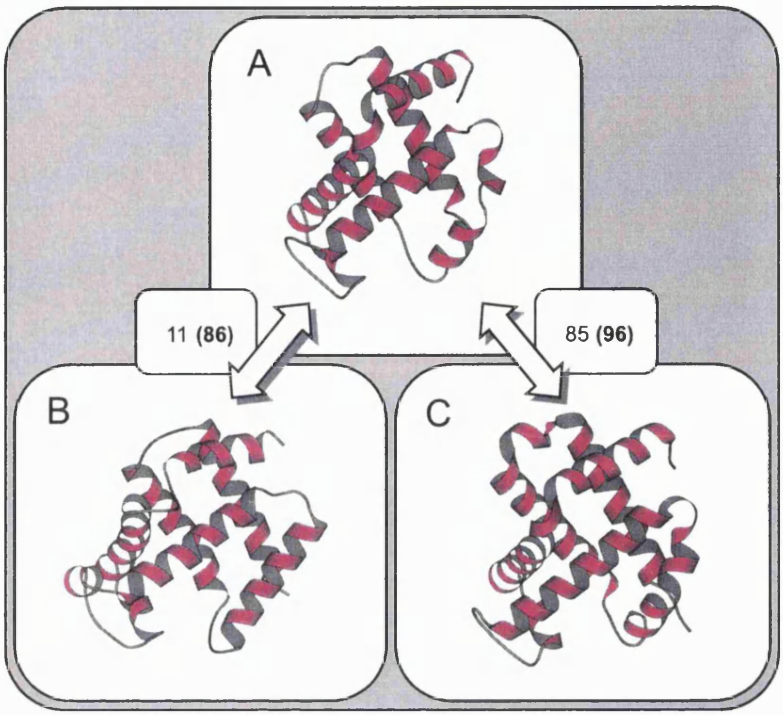


Figure 3.1: Comparison of the structures in the globin-like superfamily illustrating that protein structure can be conserved even at low sequence similarity. Sequence identity is shown as the first number with the SSAP structural comparison score in parentheses. (A) haemoglobin, α-chain from pig (1QPW, chain A), (B) haemoglobin, domain 1 from pig roundworm (1ASH), (C) haemoglobin, α-chain from horse (1IBE, chain A). For ease of reference this figure is repeated from figure 1.7

importance and can therefore be used to assist the putative assignment of biological function.

The evolutionary information found in a multiple structural alignment is often encoded into a structural template containing the conserved structural features at each position in the alignment. This is analogous to sequence 'profiles' generated from sequence alignment protocols such as PSI-BLAST (Altschul *et al.*, 1997), a topic discussed in more detail in chapter 5. A representative family template can often prove more powerful than a series of single structures when identifying distant relatives, as structural features that are highly conserved during the process of evolution can be identified. Gathering all the structural information within a family can also identify the degree of conservation, i.e. the relative importance, of these conserved features. Thus, features such as secondary structure elements buried in the core and motifs integral to the function of the protein family would be given higher weighting in the template. Conversely, highly variable regions (e.g. peripheral coils) can be recognised as noise and removed from the signal. These consensus

features act as a fingerprint for the whole family rather than the individual members and can provide a fast and sensitive probe for finding structural relationships and homologies.

3.1.2 Multiple Structure Alignment Algorithms

A large number of methods have been proposed to compare and align two protein structures (see section 1.2.7). However there have been fewer developments in the comparison of more than two protein structures, possibly due to the sparse nature of the available structural data. Many structural families contain a high number of very similar structures, but only a relatively small subset of these families display sufficient structural diversity to encourage multiple structural alignments rather than simple pairwise alignments. However as mentioned previously, this situation is likely to change with the results of the structural genomics initiatives.

Possibly the most simple approach of generating a multiple structural alignment is to successively chain together pairwise alignments, usually starting with the most closely related pair of proteins and ending with the most distantly related. This resembles protocols developed for generating multiple sequence alignments, e.g. CLUSTALW (Thompson *et al.*, 1994). A different approach is to build a template after each alignment, containing a description of the highly conserved structural features, then use this template to align successive structures.

Two multiple structure alignment methods are discussed in more detail; the STAMP method (Russell & Barton, 1992) proposed by Russell and Barton (see section 3.1.2.1) and the CORA algorithm (Orengo, 1999) proposed by Orengo (see section 3.1.2.2).

3.1.2.1 STAMP

The STAMP algorithm can be used for pairwise and multiple structure alignments, however the alignment algorithm is similar in both cases. The structural alignment is achieved by successive refinement of an initial set of equivalent residues. This initial set of residues is identified from a simple sequence alignment and the equivalent pairs of residues used to superpose the two structures. A matrix of scores is then generated based on the intermolecular distances between residues in the superposed structures and a dynamic algorithm used to find the best path, or alignment, from these scores (see chapter 1.2.4.2). This alignment then provides a new set of equivalent residues from which a more accurate superposition of the two structures can be generated. This process continues until there in no observed improvement in RMSD (see section

1.2.6.2) for the two structures.

The first step in generating a multiple structure alignment is a pairwise structural alignment for each pair of proteins in the alignment, thus providing a phylogenetic tree based on structural similarity scores. This phylogenetic tree is then traversed from leaves to root, calculating a structural alignment at each branch point, thus clustering the most similar proteins first. The growing alignments are merged by transforming the coordinates of all the structures in one node to the coordinates of the second node, and repeating the structural superposition refinement process (see figure 3.2).

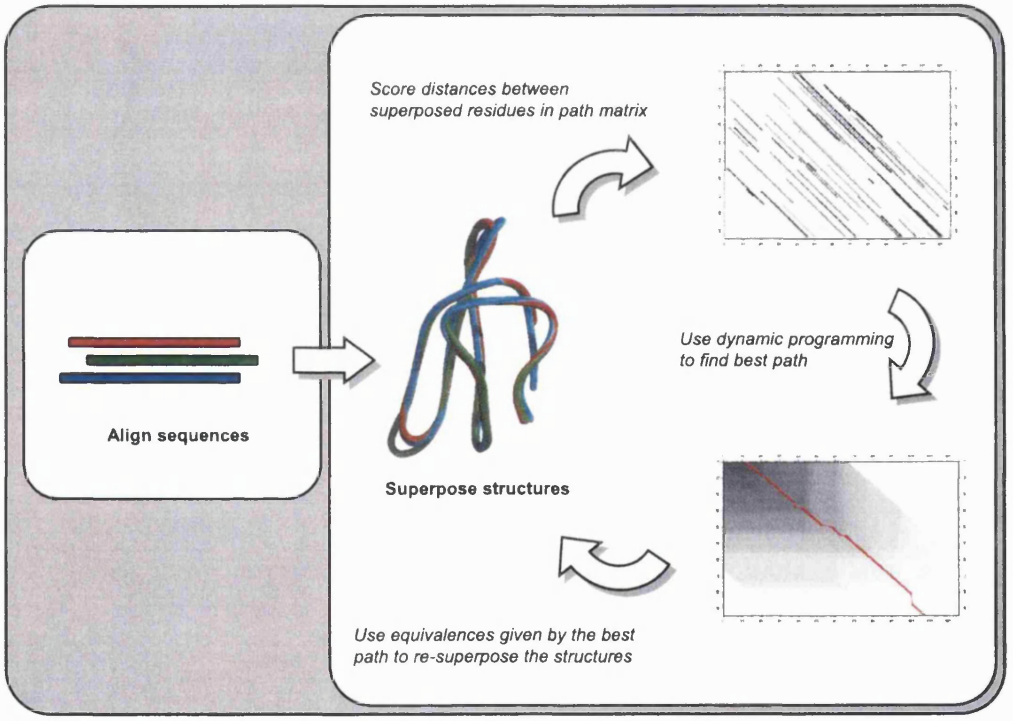


Figure 3.2: A flowchart describing the STAMP structure comparison protocol (Russell & Barton, 1992). Sequence alignment is first used to determine putative equivalent positions in order to superpose the structures. Distances between the superposed positions are then used to score a 2D score matrix which is analysed by dynamic programming to obtain a better set of equivalent positions on which the structures are re-superposed. The dark diagonal lines in the score matrix represent high scoring paths based on intermolecular residue distances from the two structures. The red path through the score matrix is the highest scoring path identified from the dynamic programming algorithm. For ease of reference this figure is repeated from figure 1.12

3.1.2.2 CORA

The CORA algorithm (Orengo, 1999) generates multiple structure alignments through the use of a double dynamic programming (DDP) algorithm. The use of DDP was initially developed for the pairwise comparison of protein structures and is based on the comparison of intramolecular C_β vectors between two structures (SSAP, Taylor & Orengo (1989); Orengo & Taylor (1996)). Comparing the structures based on intramolecular rather than intermolecular structural features negates the need for an initial alignment since intramolecular, or internal, features are independent of the frame of reference (see section 1.2.4.2).

Briefly, dynamic programming (DP) is first applied on the residue level in order to compare the structural environments of all residues judged potentially equivalent. This potential equivalence is based on the properties of the residue such as secondary structure state, accessibility and torsional angle. Removing the residue comparisons that are deemed unlikely to be equivalent not only accelerates the algorithm considerably, but also helps to remove noise from the overall alignment. The structural environments of these residues are then compared by filling a 2D matrix with scores based on the comparison of the average internal C_β vectors in the structure template to the C_β vectors seen in the structure being aligned. The DP algorithm is then used to pick the highest scoring alignment path from these scores while allowing for gaps in the alignment, in order to account for insertions and deletions during the process of evolution. If the overall score from this path is above a given threshold, i.e. the structural environments of the two residues are sufficiently similar, then the scores from this alignment path are added to the summary score matrix (see figure 3.3).

After all these residue comparisons have been made, DP is used on a second level to find the highest scoring alignment path through the summary score matrix. Given this alignment CORA then analyses the structures and calculates average structural properties which are then encoded into a consensus template. This template incorporates critical core properties such as internal vectors, residue accessibility and torsional angles together with information on the variability of the property within the alignment. As the algorithm is only used to align similar structures, a reliable multiple alignment can be generated by successively aligning proteins to the evolving consensus template in the order of decreasing pairwise structural similarity, or SSAP, scores. After each new protein is aligned, the template is calculated again, thus recapturing and enriching the consensus structural environment at each residue position.

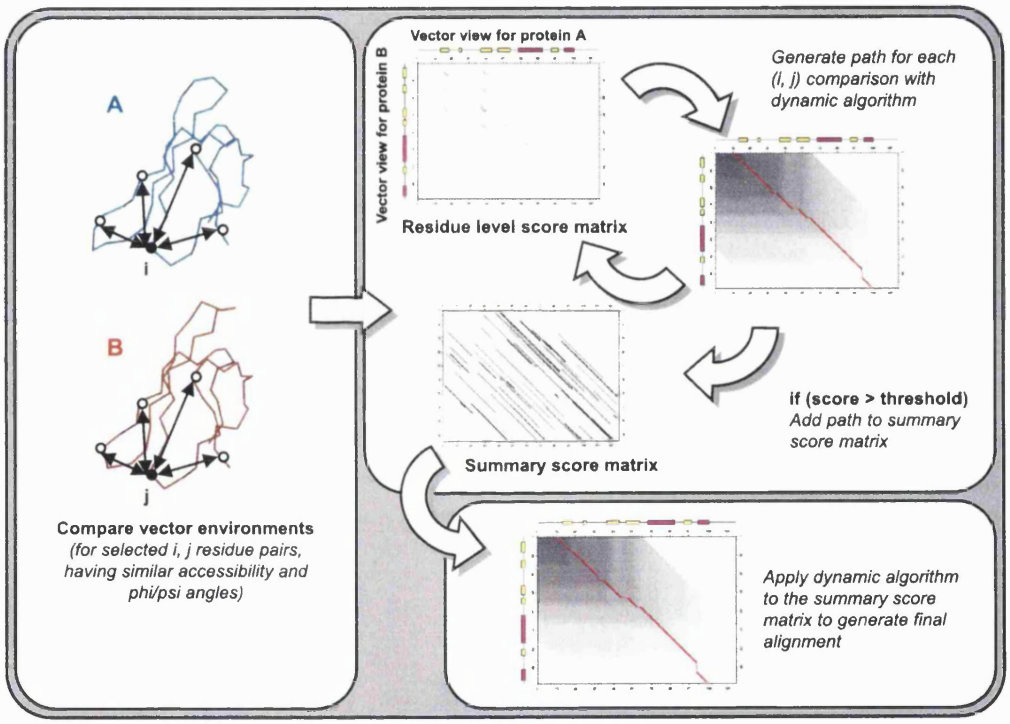


Figure 3.3: Flowchart describing the double dynamic programming algorithm for the comparison of two structures. CORA extends this pairwise method by constructing a consensus template encoding average structural properties and information on variability after each structure is aligned. The consensus structural information from this template is then used to align the next structure. For ease of reference this figure is duplicated from figure 1.16

3.1.3 Representing Structurally Diverse Superfamilies

In the January 2002 release of the CATH database (release 2.4), only 3% (34/1221) of superfamilies have 10 or more sequence families with sequences clustered at $\geq$35% sequence identity. However, this small number of superfamilies represents 26% (687/2679) of all the sequence families in the CATH database. With the increasing numbers of structures being experimentally determined, especially with the structural genomics initiatives, many more superfamilies are likely to be expanded in this way.

For large and structurally diverse superfamilies, generating a single template that accurately models all the identifying structural features within the superfamily is not a trivial task. Effectively this task involves reaching a compromise between including a small, highly conserved representative set of structures and including structural information from all the representatives within a superfamily. The former case of restricting the templates only to highly similar structures would result in highly selective templates, however the templates will no longer represent the full structural diversity present in the superfamily and will provide poor coverage as a result. The alternative of including all the structural diversity present in a superfamily may result in a poor quality multiple structure alignment due to a large number of gaps. The poor quality alignment would result in a poor quality template as the important, descriptive features would be misaligned and lost in the noise.

An alternative to this compromise is to allow more than one template for a given homologous superfamily (see figure 3.4). In this way, both the selectivity and coverage is maintained by using a series of highly descriptive templates. The only disadvantage is the speed penalty of allowing more templates, however when compared with a pairwise structure comparison procedure this time penalty is negligible.

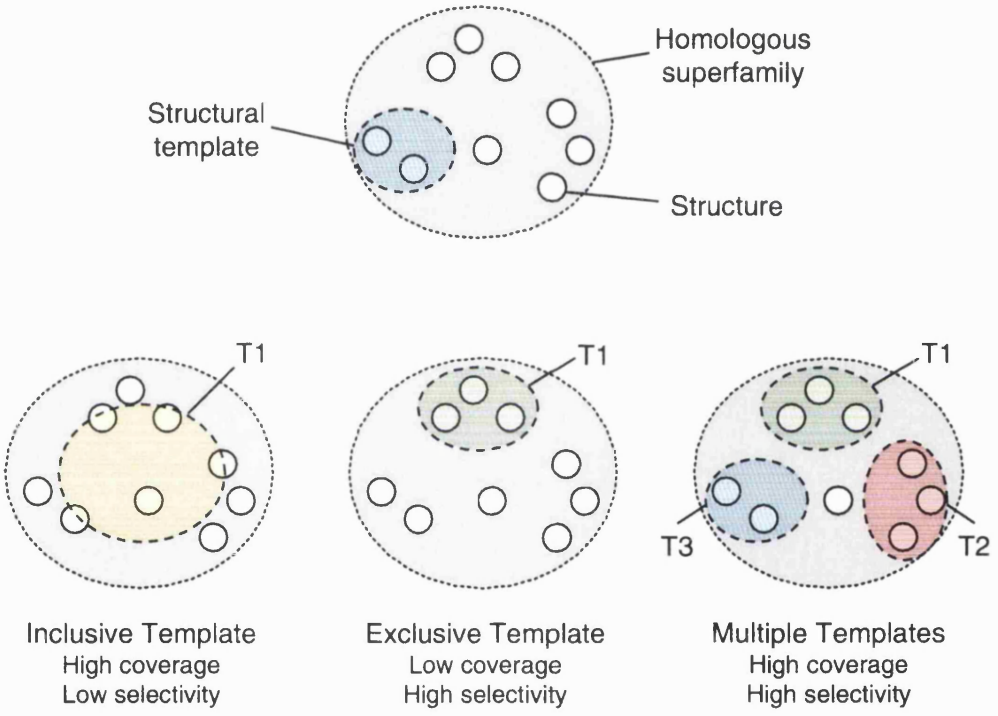


Figure 3.4: Selecting representative proteins for the structural templates. The figure gives a schematic representation of the structural space occupied by proteins within an homologous superfamily with proteins displaying high structural similarity located in close proximity. Three different template selection criteria are shown; inclusive, exclusive and multiple. The inclusive template displays high coverage but low selectivity. The exclusive template displays low coverage but high selectivity. The multiple templates show high coverage and high selectivity. It should also be noted that even when using the multiple templates, some structures still may not be included in any template if they form single structure clusters.

3.1.4 Aims

This chapter describes the process of generating, optimising and testing a library of multiple structure templates that best represent homologous superfamilies in the CATH database. The flowchart in figure 3.5 presents an outline of the work presented in this chapter.

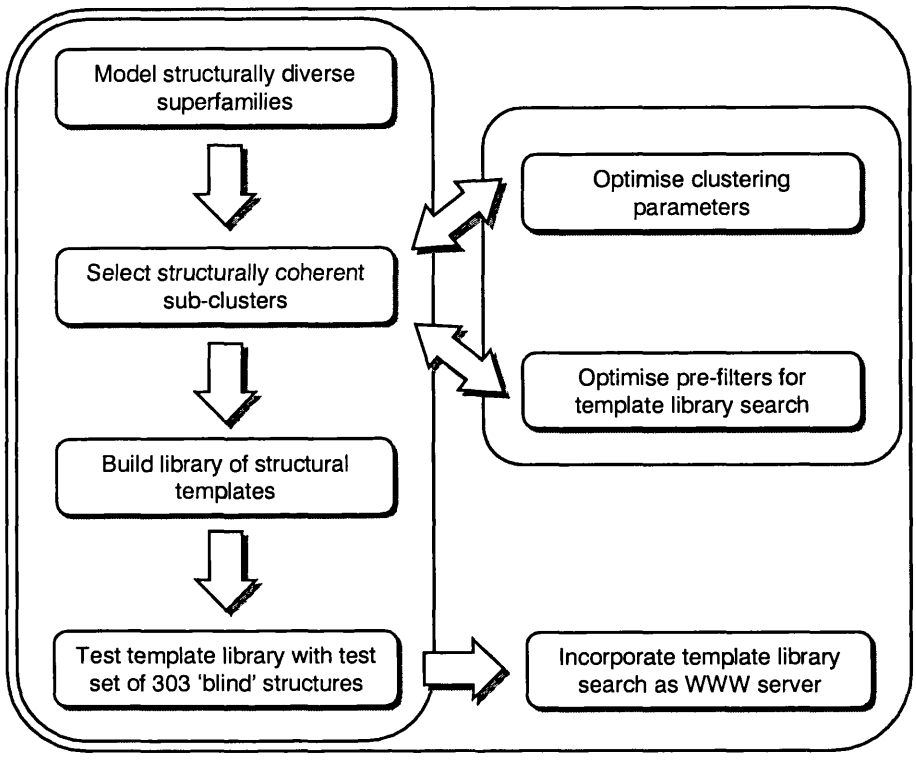


Figure 3.5: Flowchart outlining work presented in this chapter, based on generating, optimising and testing structural templates.

Since a single structural template is designed to represent a large number of individual structures, a library of structural templates would be far smaller than an equivalent library of individual structures. Thus, the time and computational expense of the current classification protocol of searching the structural database in a pairwise manner could be dramatically reduced if classifications could be made with the structural templates. Also, the templates take advantage of all the evolutionary information available in an homologous superfamily by encompassing the highly conserved features in the structural core and ignoring the random noise caused by embellishments in individual structures. As a result, the structural templates may be able to identify more distant evolutionary relationships to aid classification in the CATH database.

An analysis of the optimal clustering criteria and parameters is presented that aims to select the most descriptive set of representative structures for inclusion in the structural templates. The clustering methodology was extended to allow highly populated and structurally diverse superfamilies to be appropriately represented in the library, a potential problem in the CATH database that had not been addressed previously.

The accuracy of the template library was also tested using a 'blind' set of independently classified structures, sharing no detectable sequence similarity to structures in the library. This blind dataset provided an effective assessment of the performance when using the structural template library for the homologous superfamily assignment of novel structures. Finally, the structural templates were incorporated into the CATH server thus making the classification tool available for external use via the Internet.

3.2 Methods

3.2.1 Methods Overview

This section describes the procedures involved in generating and optimising the searchable library of structural templates. As mentioned in section 3.1.3, the structural templates use a representative subset of structures rather than all the proteins within a superfamily. Section 3.2.4 discusses the protocol used to optimise the criteria responsible for selecting structures in order to select structures that would represent the structural diversity within an homologous superfamily yet retain the integrity of the structural alignment. To achieve this, a clustering algorithm was developed and optimised that would provide structurally coherent subgroups within homologous superfamilies (see section 3.2.3.2).

This optimisation procedure involved scanning a set of single structures against templates constructed under a variety of conditions for four test superfamilies. The novel scoring function used in these database scans, based on the comparison of conserved contact patterns, is also described. The results of these database scans are presented in the form of coverage-versus-contact plots which allowed the results from various parameter sets to be compared. Once the optimal parameters had been identified, these contact-versus-coverage plots were used to demonstrate the overall performance of these templates for all the superfamilies in CATH.

3.2.2 Definitions of Evolutionary Relationships

The purpose of the structural templates is to recognise distant evolutionary relationships. The optimisation and testing precedures discussed in further detail in this section are based on maximising the differentiation between the scores for related structures and non-related structures. For clarity the terms used later in this chapter to describe specific evolutionary and structural relationships in CATH are discussed below.

In the CATH database (Pearl *et al.*, 2001b), a significant evolutionary relationship, i.e. homology, is defined by the presence of at least two of the following three criteria between two proteins.

- High sequence similarity (>35% sequence identity, or significant E-value using PSI-BLAST).
- High structural similarity (SSAP score over 70).

- Evidence of functional similarity.

If two of these criteria are met then proteins are clustered into the same homologous superfamily in CATH. Since CATH is a hierarchical classification database, the superfamilies themselves are clustered into fold groups, or topologies (T-level), that share a similar spatial and sequential arrangement of secondary structures. Proteins that share the same topology but are clustered into different homologous superfamilies within CATH are given the term analogues. Proteins that do not have similar structures, that is they are not in the same homologous superfamily or fold group, are termed non-relatives.

Analogues generally share a similar folding arrangement which could occur either as an example of convergent evolution, where proteins arrive at the same fold through an independent evolutionary pathway, or by divergent evolution, where the proteins share a common ancestor albeit so distant that the only evidence of this relationship is the structural similarity. However, this lack of evolutionary evidence, in terms of sequence or functional similarity, does not necessarily mean that they are unrelated, only that no evidence of the relationship is currently available. Often a protein sequence or structure will be found that has evolutionary relationships to more than one superfamily and this provides a 'missing link'. This allows superfamilies to be merged resulting in analogous relationships being redefined as homologous relationships. As a result, it is useful to include analogues when examining fold recognition methods. However, it is also useful to discriminate between analogous and homologous relationships for purposes of classification.

3.2.3 Generating Structural Templates

3.2.3.1 Selecting Representative Structures

To distill the greatest amount of information from a large number of structures, a careful screening process is required in order to identify suitable representatives to include in the final template. Ensuring that a template identifies all the important structural features that have been conserved during the process of evolution necessitates including proteins that are highly divergent (i.e. have low sequence identity). That is, a structural feature shared by a series of very distantly related proteins is likely to be the result of important structural or functional constraints. It is more difficult to identify highly conserved structural features when comparing closely related structures as the similarity could be simply due to the lack of time for evolutionary divergence. However, if the structures within the template are too

divergent then important consensus features can be hidden or missed altogether by the poor quality of the resulting alignment.

Also, it is often the case that some gene sequences and therefore protein structures are more thoroughly researched than others. This leads to some families containing large numbers of proteins that have highly similar sequences. It follows that these proteins will have near identical structures purely because there has not been time to adapt and evolve rather than from any specific structural constraint. If proteins involved in these highly populated areas of structural space were included with the same weighting alongside more distant structures, the template would be unfairly biased.

3.2.3.2 Selecting Structurally Coherent Sub-Groups

A multiple-linkage algorithm was written in order to group the structures within each of the superfamilies into structurally coherent clusters. The algorithm first reads a matrix of pairwise structural similarity scores generated by the SSAP structural comparison algorithm for all proteins being clustered. The algorithm then selects the highest resolution structure for each sequence family clustered at 35% identity, i.e. no two representative structures have sequence identities greater than 35%. This helps to remove redundancy in the structural templates as sequences that are more than 35% identical will nearly always have highly similar 3D structures. Starting with the highest SSAP score, these representative structures are then clustered on the basis that a structure can only join a cluster if it has a structural similarity above a given threshold to all the existing members of that cluster.

The multiple-linkage approach was chosen over single-linkage as the objective was to define structural clusters that were internally consistent. Single-linkage cannot guarantee this consistency as it joins clusters on the requirement that only one structure from each needs to be similar. This allows clusters to be chained together and can contain very remote structures which, in turn, can result in poor quality structural alignments. Figure 3.6 illustrates the differences between single-linkage and multiple-linkage clustering and highlights a single-linkage chain that results in two dissimilar structures being clustered together.

A more robust method might be to introduce a weighting scheme that would allow all structures to be included in the structural template but would downweight the contribution from proteins that have similar sequences. This is a common feature of sequence alignment methods when dealing with large numbers of sequences, e.g. CLUSTALW, Thompson *et al.* (1994). However, in practice, the majority of

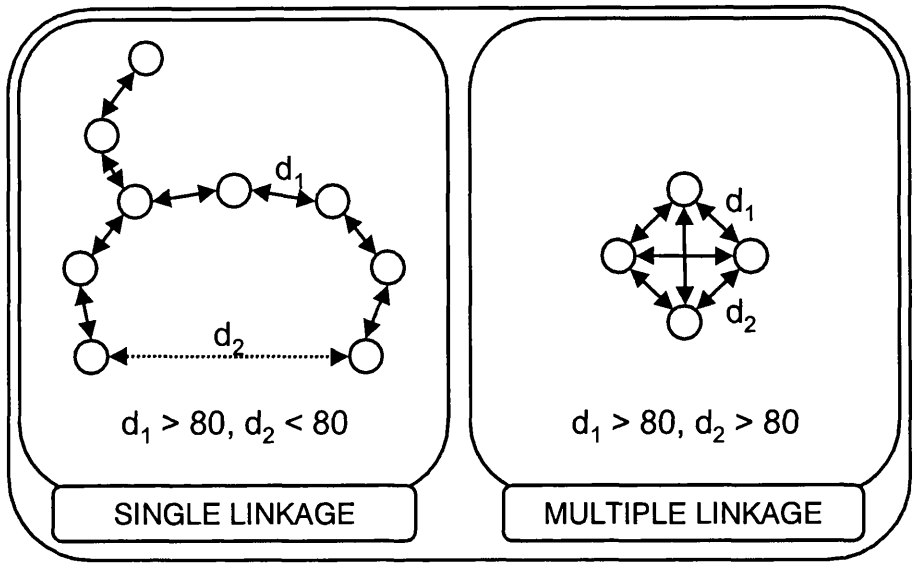


Figure 3.6: Single and multiple linkage clustering. Single-linkage clustering only requires one comparison to meet the clustering criteria (e.g. SSAP score $d_1 > 80$) for a structure to be included in a cluster. This allows structures to be chained together and can result in clusters containing very diverse structures (e.g. SSAP score $d_2 < 80$). Multiple-linkage will only allow a structure to join a cluster if the clustering criteria is met with all members of the cluster (e.g. SSAP score for $d_1, d_2, ..., d_n > 80$).

structural families are small and this simple approach of removing redundancy has been seen to work well (Orengo, 1999). As the structural genomics initiatives expand the population and diversity of these homologous superfamilies a weighting scheme may well prove more appropriate in order to incorporate the maximum amount of evolutionary information.

3.2.3.3 Building the Structural Templates

Once the structurally coherent sub-clusters had been selected within an homologous superfamily, the CORA algorithm was used to generate a multiple structural alignment. As discussed in section 3.1.2.2, CORA analyses protein structural families then identifies the consensus structural features and variability at each alignment position. The conserved structural characteristics of the cluster are then stored as a consensus structural template which can be used to align further structures.

3.2.4 Optimising the Clustering Procedure

In order to find the optimal cutoff for the multiple-linkage clustering, i.e. one that provides templates with the highest degree of discrimination, a series of structural

similarity cutoffs was selected (SSAP score 70, 75, 80 and 85). The multiple-linkage clustering algorithm would therefore guarantee that all structures within a given cluster would have a structural similarity SSAP score of greater than this cutoff. The low structural similarity threshold generally results in a smaller number of large, structurally diverse clusters. Using a higher structural similarity threshold typically results in a larger number of small, structurally specific clusters. However, if this threshold is set too high then many of the generated clusters only contain a single structure which cannot be converted into a structural template and is therefore left unrepresented in the template library.

The efficacy of each of these four structural similarity thresholds was then tested by comparing the ability of templates from a given superfamily to differentiate between related and non-related structures. Due to the high computational cost of running many structure comparisons, a dataset of four superfamilies was selected to optimise the clustering parameters. This dataset consisted of the cytokine-like superfamily from the mainly-α class (CATH classification code 1.20.160.30, see section 1.4.2 for more details on CATH codes), the mainly-β cupredoxin superfamily (CATH code 2.60.40.420) and two mixed-$\alpha\beta$superfamilies; the $\alpha\beta$-plaits (CATH code 3.30.70.330) and the thioesterase superfamily from the Rossmann fold (CATH code 3.40.50.950). These four superfamilies were chosen as their variation in population and structural diversity would provide a range of different numbers of templates within a superfamily and they would also cover four completely different architectures across the three main classes in CATH.

Superfamily	**Sreps S35**	**Nreps S95**	**Cluster cutoff** *Number of Templates (sequence families)*			
			70	**75**	**80**	**85**
1.20.160.30	13	19	2(12)	1(9)	3(9)	1(3)
2.60.40.420	18	46	2(18)	2(18)	5(17)	6(14)
3.30.70.330	6	14	1(6)	1(6)	1(4)	1(3)
3.40.50.950	19	31	2(17)	2(16)	3(12)	1(2)

Table 3.1: Summary of the superfamilies within the test set containing the number of sequence and non-identical representatives (Sreps and Nreps respectively), the number of templates generated for each cluster cutoff with the number of sequence families represented within these clusters in parentheses. Sreps are the representatives for structures clustered at $\geq$35% sequence identity (S35). Nreps are the representatives for structures clustered at $\geq$95% sequence identity (S95).

Table 3.1 gives a summary of the number of sequence representatives (Sreps) and non-identical representatives (Nreps) contained in each superfamily and the number of templates generated at each of the cluster cutoff values (see section 3.2.3.3). The

number of sequence families that are represented by the structural templates for each of these threshold values is also provided. Although the representation of sequence families cannot be directly related to the overall structural representation of the clusters, it does give some indication of the degree of diversity seen in the structural templates and the number of structures left unrepresented by these clustering restrictions.

3.2.4.1 Scanning the Template Database

In order to assess the structural similarity between a query protein and a structural template, a structural alignment is first generated using the CORA program. The resulting alignment is then used to superpose the contact map of the query structure with the consensus contact map for the structural template. A structural similarity score between the template and the query structure is then calculated from the percentage of contacts that are overlapping between the two contact maps (see chapter 2 for more details). If a superfamily contains more than one structural template then only the highest scoring template match is included for this superfamily.

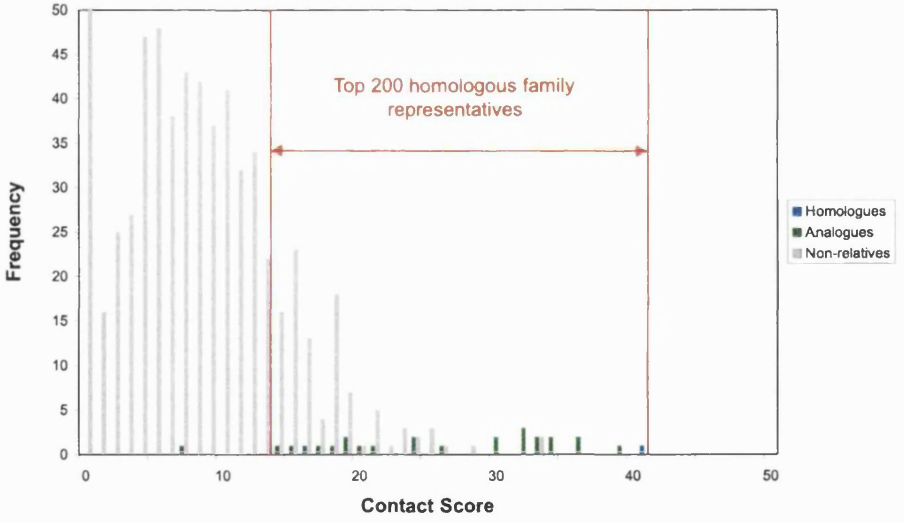


Figure 3.7: A distribution graph showing the results of a database scan of a superfamily template against the 900 superfamily representatives from CATH v1.7, scored by contact overlap. A reduced testset was generated for each superfamily by taking all the sequence representatives from the closest 200 superfamilies.

Due to the fact that each of the template-structure comparisons requires a structural alignment, generating the database scans for this optimisation procedure is highly computationally expensive. Since the objective of this optimisation was

to maximise the differentiation between homologues and non-relatives, a reduced dataset was generated for each superfamily, just containing structures from superfamilies with structural similarities to the template. To achieve this, a full database scan was carried out for each of the four superfamilies, just searching the structural templates against a representative structure from each of the homologous superfamilies in CATH v1.7. The 200 structures with the highest contact overlap scores were then selected as the most structurally similar superfamilies to the templates. This dataset was then expanded by including structures from all the sequence family representatives within these 200 superfamilies. This provided a test set of around 400 structurally similar examples for each of the four test superfamilies (see figure 3.7).

3.2.4.2 Coverage-Versus-Contact Plots

The results from a database scan of query structures against structural templates can be visualised by generating a coverage-versus-contact plot. This plot describes the coverage, i.e. observed number of matches over expected number of matches, above a given contact score threshold. Thus, as the contact score threshold is reduced, a larger proportion of the expected matches is observed until full coverage at the minimum contact score of 0. By splitting the matches from these database scans into homologues, analogues and non-relatives and plotting the coverage on the same axes, this plot can be used to investigate the selectivity and sensitivity of structural templates. An example of a coverage-versus-contact plot can be seen in figure 3.8 with homologous matches shown in red, analogous matches in green and non-relatives in blue.

The aim of these plots is to examine the conditions that allow a structural template to recognise the maximal proportion of the homologues at a contact overlap score high enough to preclude all non-relatives. The results for homologues shows how well the templates represent all the structures in their superfamily and are therefore useful when evaluating the selection procedure of which representative structures to use in building the template. The ability to differentiate between true evolutionary relationships and general structural fold similarities is seen by the difference in coverage between the results for homologues and analogues. Although it is preferable to be able to identify the correct homologous superfamily for a given structure, identifying the correct fold group is also useful for classification purposes and these matches should therefore be treated separately from non-relatives. The selectivity of the templates is seen by the difference in the coverage of homologues and non-relatives.

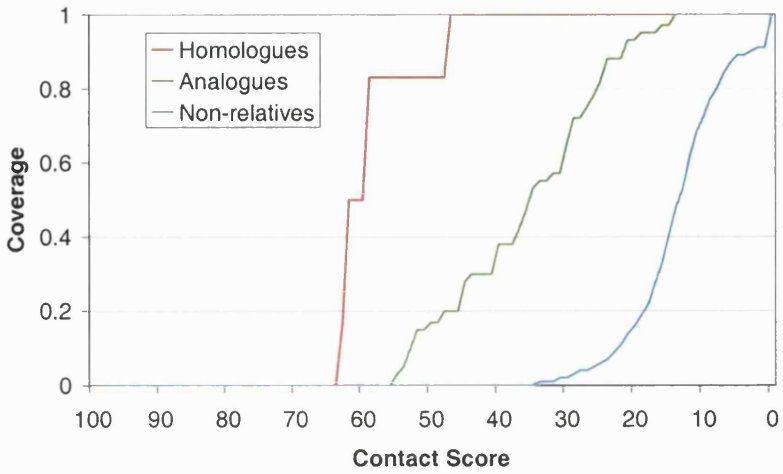


Figure 3.8: An example of a coverage-versus-contact plot. The templates from one superfamily are scanned with a dataset of structurally similar proteins and the results ranked by decreasing contact overlap score. The matches are then separated based on the relationship between the structural template and protein in the dataset (i.e. homologues, analogues and non-relatives). For a particular relationship, e.g. homologues, the plot shows the coverage, or number of observed matches over expected matches, with contact scores above a given threshold.

3.2.5 Searching Novel Structures Against the Template Library

In order to evaluate how this procedure would perform when applied to a real situation, a blind test was constructed that would attempt to classify novel structures containing no significant sequence relationship to structures present in the templates. Currently, if novel structures have no detectable sequence similarity, homology is assigned through automated structural similarity scores and manual functional validation. Assignment of structural similarity is achieved by performing an exhaustive pairwise search of the non-redundant library of structures using the SSAP double dynamic algorithm. For CATH v1.7 this would require 3,851 structural comparisons for each novel structure and this can prove a very time consuming process.

A structural template can represent a large number of related proteins. Thus far fewer structural templates are required to represent the equivalent amount of structural space seen by single structures (database composition is discussed in more detail in section 3.4.1.2). However, the underlying double dynamic algorithm is similar when searching against structural templates (CORA) or single structures (SSAP). As a result, searching a novel structure against a library of structural templates takes far less time than a comprehensive pairwise search of single structures. If any novel structure assignment could be made through the faster method of searching

against the library of structural templates then incorporating this method into the general assignment protocol would help the CATH database keep pace with the explosion of new structures from the various structural genomics projects.

3.2.5.1 Generating the Library of Structural Templates

A library of structural templates was generated from CATH v1.7 using the cluster threshold of 80, which proved the optimal clustering cutoff (see section 3.3.4). This version of CATH contained 3,581 non-identical representative structural domains which cluster into 1,798 sequence families (within 35% sequence similarity) and 903 homologous superfamilies. Of these 903 superfamilies, only 340 resulted in clusters containing more than one structure and could therefore be represented by a structural template. Of these 340 superfamilies, 35 displayed sufficient structural diversity to allow more than one cluster and therefore provide more than one template, giving a total number of 407 structural templates in the library. Although this figure of 340 homologous superfamilies only represents 37% of the superfamilies in the structural database, many of the remaining superfamilies contain very few structures. As a result, the structures within the library of structural templates represent 55% of the total sequence families in CATH, which equates to 66% of the non-identical representatives within the database. This database composition is discussed in more detail in section 3.4.1.2.

3.2.5.2 Generating the Dataset of Remote Structures

The set of remote structures was generated from a protocol described by Bray (Bray, 2001) and is discussed in more detail in section 5.2.4.2. In summary, this protocol involved comparing two versions of the CATH database (CATH v1.7 and v2.0) and extracting structures from CATH v2.0 that had been classified into superfamilies existing in CATH v1.7. Knowledge of this classification therefore provides a reliable validation for the putative homologous assignment given by structural templates built from the earlier version of the CATH database.

To ensure that these novel structure assignments could not be similarly achieved by faster sequence methods, all structures with significant sequence relationships to proteins in CATH v1.7 were also removed from the test set. To achieve this, the SSEARCH algorithm (Smith & Waterman, 1981; Pearson, 1991) was used to scan each structure against a sequence library generated from CATH v1.7. The best scoring match, i.e. lowest E-value, for every superfamily in CATH v1.7 was selected for each structure and plotted on a graph of sequence identity against number of

aligned residues. The HSSP/Rost equation (Rost, 1999), shown in equation 3.1 was then employed to provide a threshold that differentiated between close homologues and remote homologues. This procedure is discussed in more detail in chapter 5.

$$pI(N) = N + 480.L^{-0.32.(1+e^{-L/1000})} \tag{3.1}$$

In this equation, pI is the percentage identity required for the proteins in an alignment to be be considered homologous given the number of aligned residues, L. The variable N is the Rost Threshold and is defined as the number of percentage points away from the baseline curve described by $N = 0$.

A value for this Rost Threshold was empirically determined using the CATH-PFDB resource, which is a database of genomic sequences that have been reliably identified as homologous to protein structures in CATH. The results from a pairwise sequence scan of CATH-PFDB release 1.7 were plotted on a graph of sequence identity versus number of aligned residues. The variable N in equation 3.1 was then increased until all non-homologous matches were found below the HSSP/Rost curve. True homologous matches falling above this curve were seen to be close homologues and were discarded from the data set. The true matches falling below the curve were then left as the remote homologues. Of the 816 examples of homologous relationships identified by the initial database scan, 303 of these structures fell below the HSSP/Rost curve and could be considered remote homologues.

3.2.5.3 Coverage-Versus-Error Plots

One measure of the performance of a search algorithm in the discrimination between homologues and non-homologues can be made by calculating the percentage of true positive matches, or coverage, within a certain degree of error. The 'coverage-versus-error' plot, introduced by Brenner *et al.* (1998), allows the performance of different search methods to be compared by analysing the coverage of true matches for a specific error rate.

Table 3.2 provides a summary of the definitions of true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN). Coverage is defined as the fraction of true positives (TP) that have scores above a given threshold (see equation 3.2). The error rate is defined by the fraction of false positives (FP) above the score threshold with respect to the total possible number of errors (see equation 3.3).

$$\text{Coverage (score)} = \frac{\text{TP (score)}}{\text{TP (score)} + \text{FN (score)}} \tag{3.2}$$

	Superfamily (Homologues)	**Outside Superfamily (Non-homologues)**
Match	TP	FP
Non-Match	FN	TN

Table 3.2: Definitions for measuring performance with database searching. The four categories are true positive (TP), false positive (FP), true negative (TN) and false negative (FN)

$$\text{Error Rate (score)} = \frac{\text{FP (score)}}{\text{FP (score)} + \text{TN (score)}} \tag{3.3}$$

3.3 Results

3.3.1 Overview of Results

The first set of results discussed in section 3.3.2 is from the optimisation of the protocol involved in generating the structural templates. When selecting representative structures to include in a given template, a fine balance must be reached. The template must sample a diverse range of structures in order to encompass important information on evolutionary features, however it cannot include structures that are too diverse as such evolutionary features would be lost in the noise of an inaccurate multiple alignment. The multiple-linkage clustering algorithm employed to select these representative structures was tested with four different thresholds on a dataset of four superfamilies. Each clustering threshold was then analysed by scanning the resulting structural templates against a dataset of structurally similar proteins. The optimal clustering threshold was identified by generating coverage-versus-contact plots and maximising the differentiation between the coverage for homologues and non-relatives. Simple scoring schemes, based on size and contact overlap, were also investigated that could act as a pre-search filter in order to avoid comparisons between structures that were clearly incompatible (see section 3.3.3).

The second set of results, discussed in section 3.3.5, shows the performance of a library of structural templates generated using the optimised protocol. The performance is assessed by attempting to recognise a dataset of structures that have no detectable sequence similarity to structures in the templates.

3.3.2 Optimising the Structural Templates

3.3.2.1 Cytokine superfamily (1.20.160.30)

This is the largest superfamily within the four-helix bundle fold, containing 19 non-identical representatives from 13 sequence families in CATH v1.7. Figure 3.9 shows the structure and associated contact map for a typical member of this superfamily (1RCB). The 'up-down' nature of the α-helix bundle is reflected in the contact map as scattered lines perpendicular to the central diagonal. The scattering in these contact patterns is a common feature of α-helix interactions and is due to the periodic nature of the helix which brings residues into contact every three or four residues.

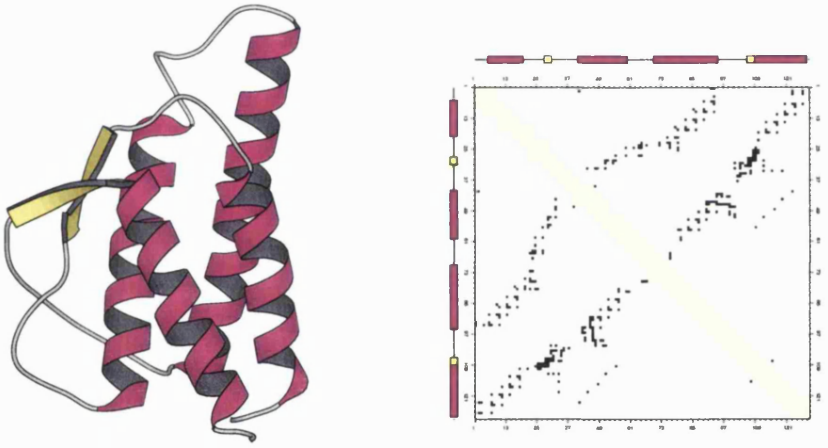


Figure 3.9: A MOLSCRIPT (Kraulis, 1991) description and contact map (generated using the program COCOPLOT, I. Sillitoe) from a representative structure from the cytokine superfamily (1RCB). Short-range contacts (within 8 sequential residues) are omitted from the contact map since they do not provide a discriminatory measure of structural similarity (seen as the yellow band along the diagonal of the contact map).

The structural diversity within this superfamily can be examined from a plot of sequence identity against structural similarity (SSAP) scores for all pairwise comparisons (see figure 3.10). This plot demonstrates the high degree of structural diversity in this superfamily as some structure comparisons have SSAP scores of less than 60. Again, this highlights the need to incorporate multiple templates in order to fully represent all the structures in large and structurally diverse superfamilies.

The coverage-versus-contact plots in figure 3.11 show that clustering the structures using a SSAP score of 80 (Cluster80) gave reasonable differentiation between

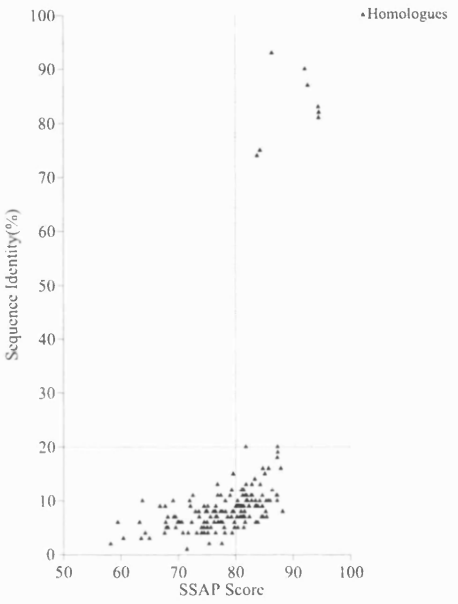


Figure 3.10: Sequence identity versus structural similarity plot for the cytokine superfamily (1.20.160.30).

homologues and non-relatives. It can be seen from these plots that Cluster70 and Cluster75 both give poor coverage, even for homologues, as all the homologues are only recognised when considering matches at low contact overlap scores (<10% contact overlap). On the surface, this would appear to pose a significant problem since, when aligning structures to their own template, one would expect far greater contact overlap than with non-relatives. However, these surprisingly low contact overlap scores can be explained by examining the clusters in question and the subsequent quality of the structural templates.

In the case of the Cluster70 group, two templates were generated covering 12 of the 13 sequence families (see table 3.1). A result of this relaxed cutoff was to generate a cluster which contained 10 structures, each from separate sequence families. This high structural diversity produced a structural alignment where only a small number of the contacts were sufficiently conserved to be included in the consensus contact map. If the structural template contains a small number of conserved contacts then the maximum number of overlapping contacts between the template and query structure would also be small. Therefore, all comparisons with such templates will give low contact overlap scores as the number of overlapping contacts will always be low with respect to the number of contacts observed in the query structure (the contact score is given by the number of overlapping contacts as a percentage of the

maximum number of contacts between template and query structure, see section 2.2.3.2).

Conversely, changing the contact overlap score to be the overlapping contacts as a percentage of the maximum number of contacts in the template would present a strong bias towards the small templates with few conserved contacts. If a small structural template contained only a few conserved contacts, e.g. describing an interaction common to many structures in the database, then it is likely that these contacts can be matched by a large number of mainly-α structures purely by chance. Since the score, in this case, would be calculated as a percentage of the contacts observed in the template (which would now be small), these small templates would tend towards artificially high contact overlap scores. This would result in the recognition ability of the structural library diminishing.

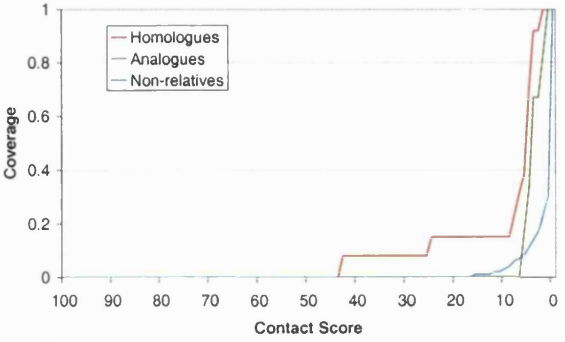
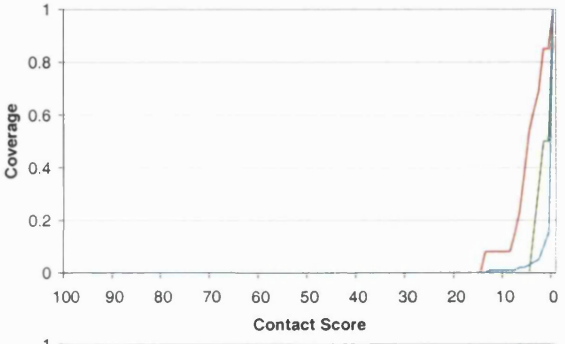
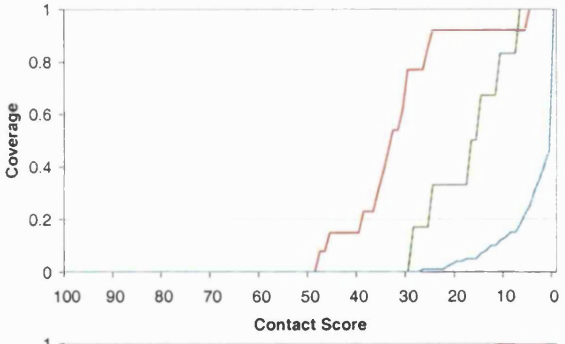
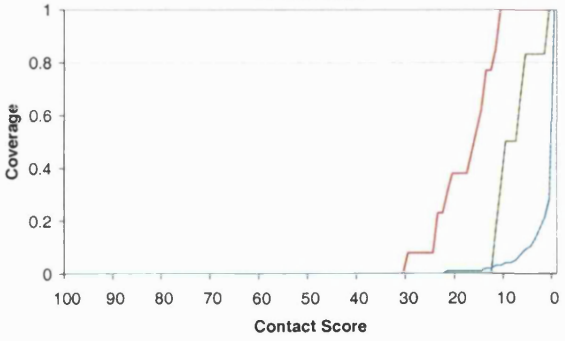
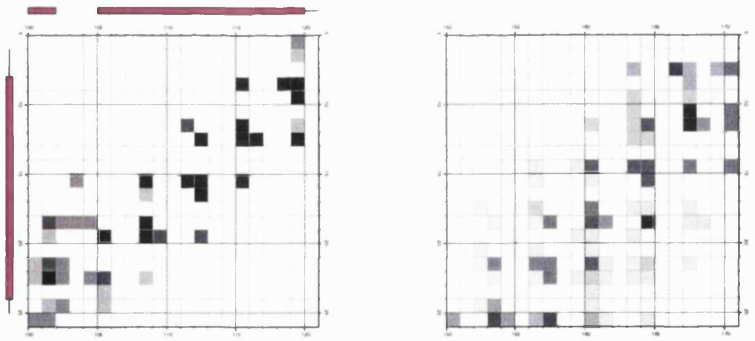
Cluster	Number of Templates (sequence families)	Coverage-Versus-Contact Plot
70	2 (2, 10)	
75	2 (9)	
80	3 (2, 3, 4)	
85	1 (3)	

Figure 3.11: Comparing the coverage-versus-contact plots for homologues, fold-relatives and non-relatives using the multiple templates from the superfamily 1.20.160.30, generated from four different cluster cutoffs (70, 75, 80, 85). The numbers in parentheses indicate the number of proteins in each sequence family.

This lack of consensus agreement for templates with large numbers of diverse structures seems most problematic in the mainly-α superfamilies. It has been observed that interactions between α-helices display a high degree of flexibility and can shift considerably during the process of evolution in order to accommodated residue mutations in the helix-helix interface (Chothia & Lesk, 1985). As a result, or possibly as an artefact of misalignment, it is easily conceivable that the structural comparison between two mainly-α proteins could produce an alignment where equivalent helices are shifted by one or two residues. The effect of this shift on the degree of overlap between contact maps can be quite dramatic as shifting the alignment of a helix by just one residue also shifts the periodically repeating contact pattern for one of the structures in relation to the other. This results in the contact pattern of one structure moving from overlapping precisely with the contact pattern of the second structure, to fitting precisely in the gaps left by the periodic nature of the helix-helix interaction. When generating the consensus contact map from a structural template, this effect dilutes the consensus information by adding noise to the discrete consensus contact patterns (see figure 3.12). When aligning structures to a template, the helix shift effect can produce an artificially low contact overlap score despite obvious structural homology.



(a) (b)

Figure 3.12: Effect of helix shift on consensus contact pattern. **(a)** shows a helix-helix consensus contact pattern for a template containing three highly similar structures. **(b)** shows a helix-helix consensus contact pattern for a template containing 10 diverse structures.

A similar effect is also caused by rotation of the α-helices which is also a common structural feature in protein evolution. Since the angle of the contact pattern of two secondary structures directly reflects the angle of interaction in the structure, rotating this angle by just a small amount can have a serious impact on the degree

of contact overlap. However, in this case the contacts from one structure are less likely to fall exactly in the gaps left by the other structure as the diagonal lines of the contact patterns still intersect rather than just being translated as seen in an alignment shift.

Figure 3.13 shows consensus contact maps for the multiple structure template generated from Cluster70, containing 10 diverse structures and the multiple structure template from Cluster80 containing 3 structures. The vertical and horizontal white bands in the plot correspond to gaps in the alignment, or more specifically those alignment positions where the minimum alignment ratio (MAR) is less than the cutoff value of 0.5 (see section 2.2.4 for more details). The fact that there are many more white spaces in the Cluster70 alignment reflects that the alignment is of lower quality as it contains a larger number of gaps.

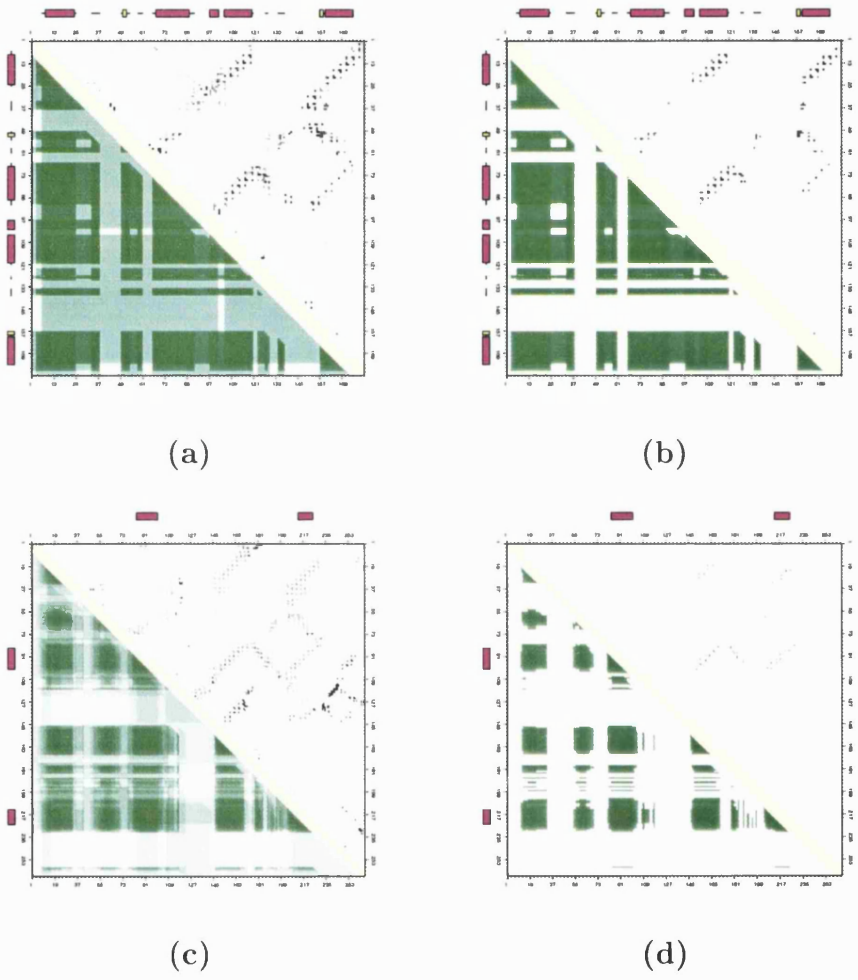


Figure 3.13: Effect of high structural diversity on the consensus contact map (CCM). **(a)** shows the CCM for a template from the Cluster80 cutoff which contains 3 structures, with 133 conserved contacts. **(b)** shows the same CCM with only the conserved contacts and alignment positions highlighted. **(c)** shows the CCM for a template from the Cluster70 cutoff containing 10 structures, with just 40 conserved contacts. Again, **(d)** shows this CCM with just the highly conserved features highlighted. All plots generated by COCOPLOT (I. Sillitoe, computer program).

3.3.2.2 Cupredoxin Superfamily (2.60.40.420)

This superfamily of cupredoxins is found within the highly populated immunoglobin-like fold and consists of 46 non-identical representatives from 18 sequence families in CATH v1.7. A typical structure from this superfamily comprises of a two layer β-sandwich with some structural embellishments seen in the form of small α-helices on the periphery of the structure (see figure 3.14).

In contrast to the scattered nature of the contact map for the mainly-α cytokine superfamily (figure 3.9), the contact map for an example structure from this superfamily can be seen to consist mainly of thick lines. These lines either lie parallel or perpendicular to the diagonal, corresponding to interactions of parallel and anti-parallel β-strands respectively.

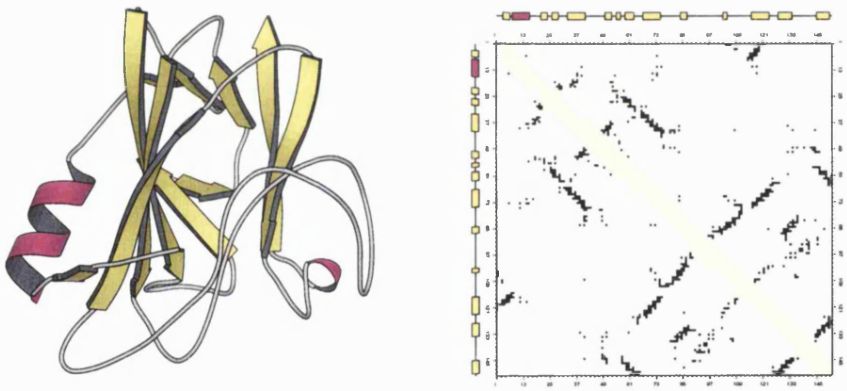


Figure 3.14: A MOLSCRIPT (Kraulis, 1991) description and contact map generated using COCOPLOT (I. Sillitoe, computer program) of a representative structure from the cupredoxin superfamily (1RCY).

There is still a high degree of structural diversity within this superfamily with some pairwise SSAP scores as low as 54. However, despite containing far more sequence families than the mainly-α cytokine superfamily, the vast majority of these pairwise SSAP scores are higher (figure 3.15). Also, as a result of the larger number of non-identical structures, the clustering process gave an equal or greater number of structural templates in each cluster cutoff.

The coverage-versus-contact plots for this superfamily show good coverage and good discrimination between homologues and analogues in each of the cluster bins.

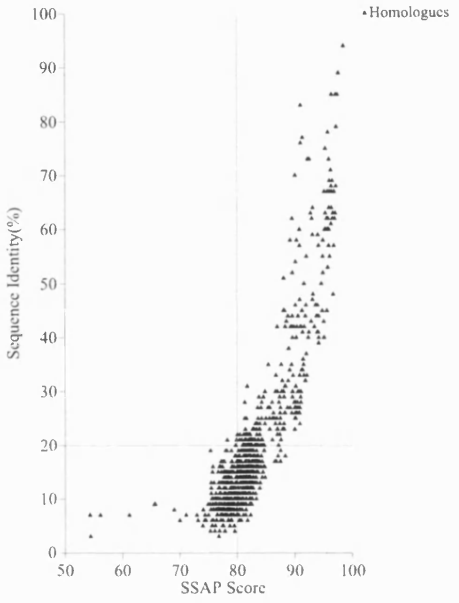


Figure 3.15: Sequence identity versus structural similarity plot for the cupredoxin superfamily (2.60.40.420).

The results for the Cluster70 and Cluster75 coverage-versus-contact plots are interesting as one of the clusters in both of these bins resulted in a structural template containing 16 proteins each from separate sequence families. In the case of the mainly-α cytokine superfamily, including this many structures resulted in a template with a large number of gaps and a poor agreement of consensus information. However, in this case, despite the alignment still containing a large number of gaps, a large number of conserved contacts were still identified.

One of the reasons for this difference in contact conservation is due to the flexibility of α-helices to shift and rotate within the core as mentioned previously, whereas the β-strands are anchored together by numerous hydrogen bonds, forming a stable β sheet. Another difference between these two superfamilies is the nature of contact patterns for helix-helix interactions and strand-strand interactions. Alignment shifts can seriously disrupt the regular dotted pattern of the mainly-α contact map as mentioned previously, however the blocks of contacts seen in mainly-β interactions remain relatively unaffected by such shifts. Although shifting these blocks of contacts relative to each other results in slightly fewer overlapping contacts around the periphery, the majority of the contacts remains overlapping. This not only provides structural templates with a better agreement of consensus information, but also provides a greater likelihood of matching homologous structures with higher

contact overlap scores.

It should also be noted that the most discriminatory coverage-versus-contact plots are seen in Cluster80 and Cluster85, which have a greater number of templates. Using the best matches from a series of smaller, more structurally coherent templates was seen to produce the most effective result.

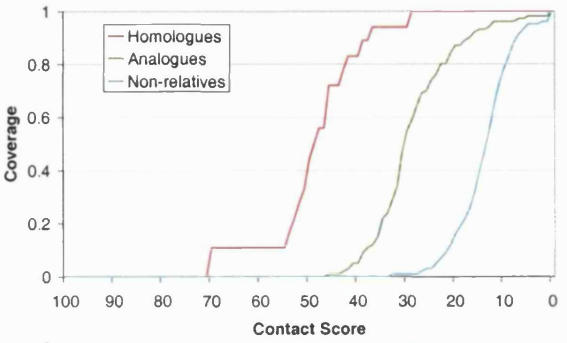
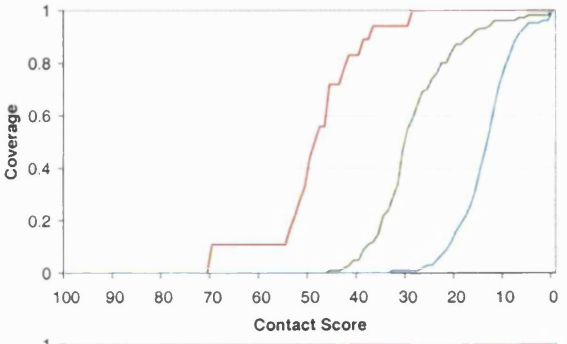
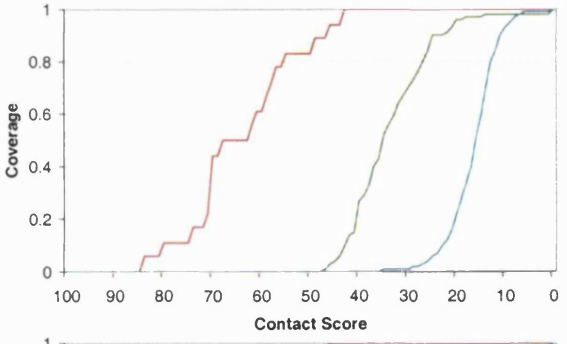
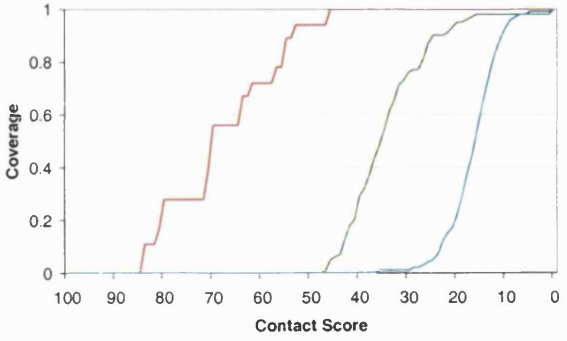
Cluster	Number of Templates (sequence families)	Coverage-Versus-Contact Plot
70	2 (2, 16)	
75	2 (2, 16)	
80	5 (2, 2, 3, 4, 6)	
85	6 (2, 2, 2, 2, 3, 3)	

Figure 3.16: Comparing the coverage-versus-contact plots for homologues, fold-relatives and non-relatives using the multiple templates from the superfamily 2.60.40.420, generated from four different cluster cutoffs (70, 75, 80, 85). The numbers in parentheses indicate the number of proteins in each sequence family.

3.3.2.3 $\alpha\beta$-Plait Superfamily (3.30.70.330)

The largest superfamily in the $\alpha\beta$-plait fold contained 14 non-identical representatives from 6 sequence families in CATH v1.7, the smallest number of structures of the four test superfamilies. The $\alpha\beta$-plait fold is a highly populated fold, containing 35 superfamilies and this provides the largest number of analogue structures of the four test superfamilies (46 sequence family representatives). The structures are relatively compact, with an average length of 90 residues and share a small, core anti-parallel β-sheet, recognisable in the contact map by a series of solid lines perpendicular to the diagonal (see figure 3.17).

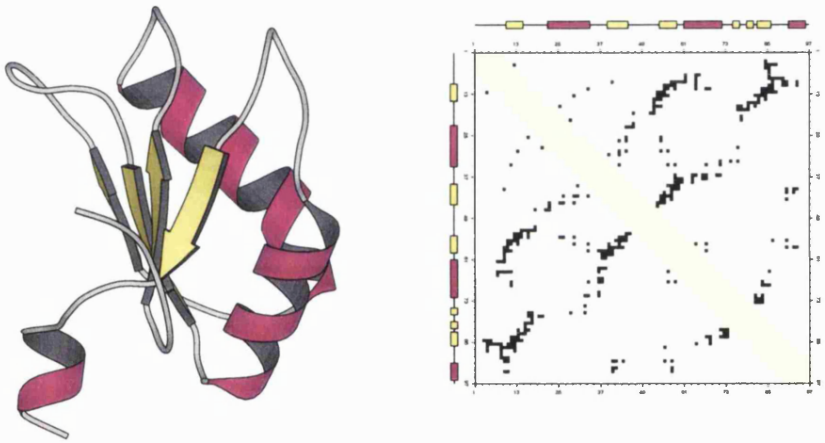


Figure 3.17: A MOLSCRIPT (Kraulis, 1991) description and contact map generated using COCOPLOT (I. Sillitoe, computer program) for a representative structure from the $\alpha\beta$-plait superfamily (1URN, chain A)

The sequence versus structure plot (figure 3.18) shows a high degree of structural similarity within the superfamily with only a small number of comparisons falling below a SSAP score of 80 and none falling below 75. This high level of structural coherence, helped by the relatively low number of structures in the superfamily, resulted in only one cluster being produced for all the cluster cutoff bins.

The coverage-versus-contact plots (see figure 3.19) show increased contact overlap scores for all three categories of query structures; homologues, analogues and non-relatives. This could be the result of several factors, most noticeably the prevalence of the anti-parallel β-sheet. Also, since these structures are relatively small, there are fewer opportunities for uniquely identifying structural features. Unsurprisingly, given their large number, the analogues showed good coverage at high

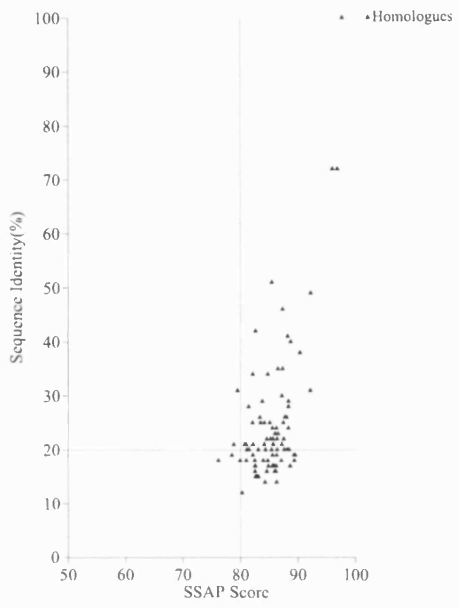


Figure 3.18: Sequence identity versus structural similarity plot for the $\alpha\beta$-plait superfamily (3.30.70.330).

contact overlap scores in all cluster cutoff bins. Despite the increased performance of the non-relatives, there was still clear discrepancy between the highest scoring non-relative and the lowest scoring homologue.

The Cluster70 and Cluster75 bins produced identical templates and therefore gave identical coverage-versus-contact plots (see figure 3.19). Cluster80 and Cluster85 produced templates that contained fewer structures and as a result the structures included in these templates gave higher contact overlap scores. However, narrowing the structural template in this way appeared to have little effect on recognising the homologues that were not included in the template as the coverage of these structures remained highly similar.

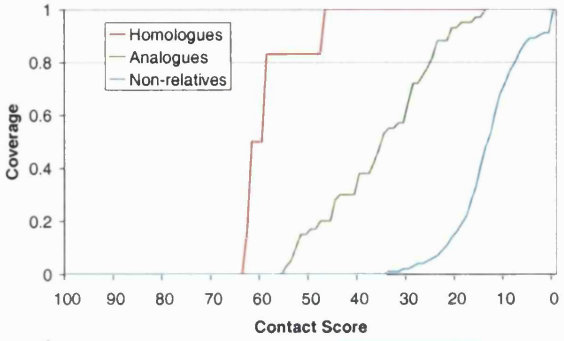
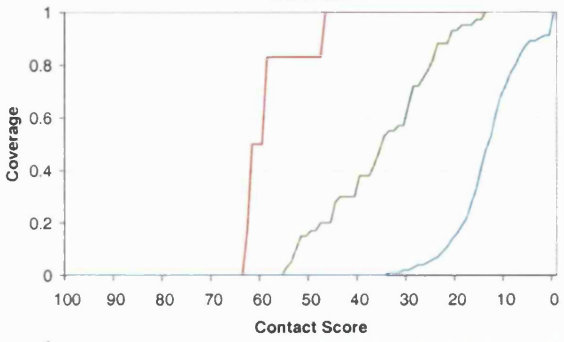
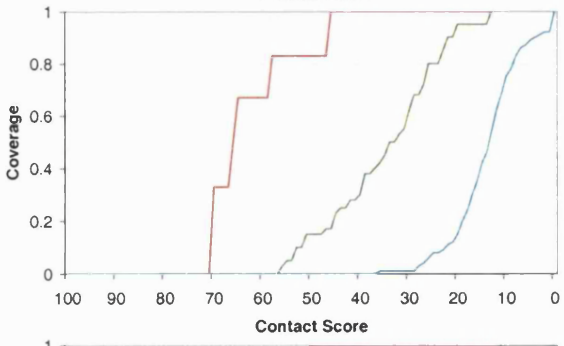
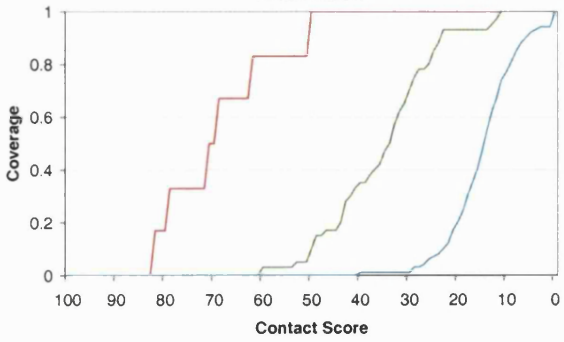
Cluster	Number of Templates (sequence families)	Coverage-Versus-Contact Plot
70	1 (6)	
75	1 (6)	
80	1 (4)	
85	1 (3)	

Figure 3.19: Comparing the coverage-versus-contact plots for homologues, fold-relatives and non-relatives using the multiple templates from the superfamily 3.30.70.330, generated from four different cluster cutoffs (70, 75, 80, 85). The numbers in parentheses indicate the number of proteins in each sequence family.

3.3.2.4 Rossmann Fold Superfamily (3.40.50.950)

This superfamily was the second largest in the Rossmann fold and contained 31 non-identical representatives from 19 sequence families in CATH v1.7. Structures have an average length of 293 residues, however the size varies widely within the superfamily from 197 to 452 residues. A representative structure and contact map is shown in figure 3.20, which shows the core $\alpha\beta\alpha$-sandwich architecture with typical embellishments consisting of helices in the periphery of the structure.

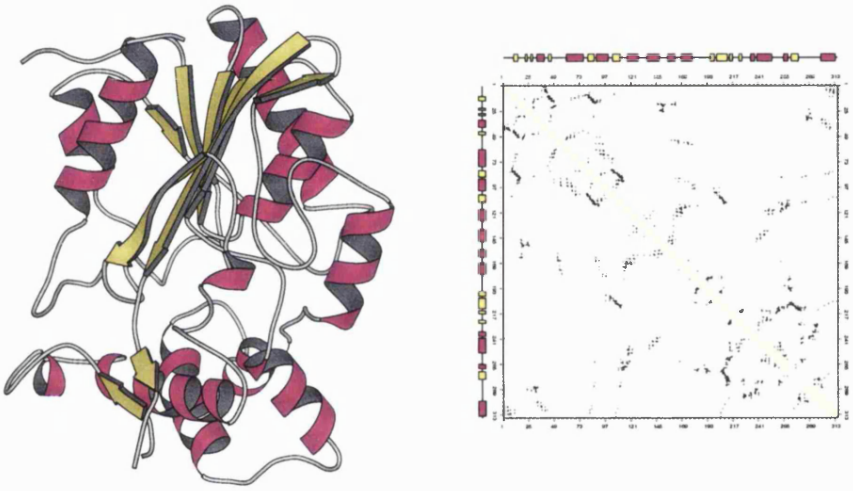


Figure 3.20: A MOLSCRIPT (Kraulis, 1991) description and contact map generated using COCOPLOT (I. Sillitoe, computer program) for a representative structure from the Rossmann fold superfamily (1CVL)

The high structural diversity is reflected in the sequence versus structure plot with the majority of pairwise comparisons having SSAP scores of less than 80 and some as low as 50 (see figure 3.21).

The coverage-versus-contact plots (see figure 3.22) each show similar results, however it can be seen that the Cluster80 again gives the highest discrepancy between homologues and non-homologues. Cluster85 only produced one template which contained two structures, yet this provides the second highest coverage of the different clusters. Since the proteins in Cluster85 are highly structurally conserved, the number of consensus contacts for these templates is also high, giving a greater chance of overlapping contacts when compared to homologous structures.

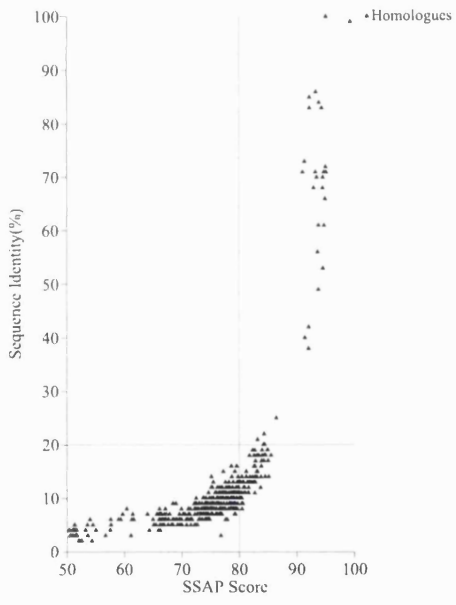


Figure 3.21: Sequence identity versus structural similarity plot for the Rossmann fold superfamily (3.40.50.950).

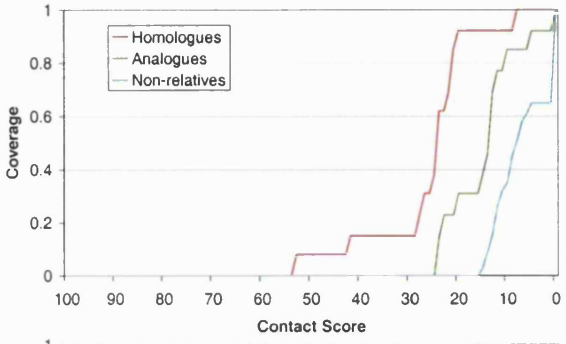
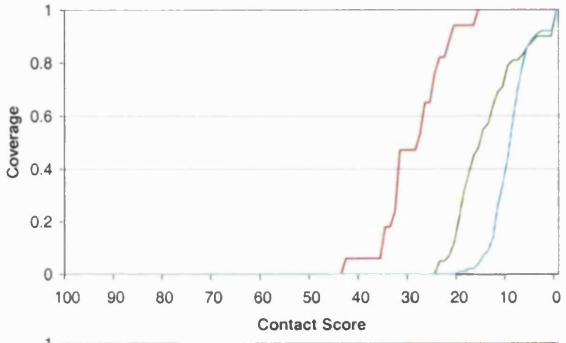
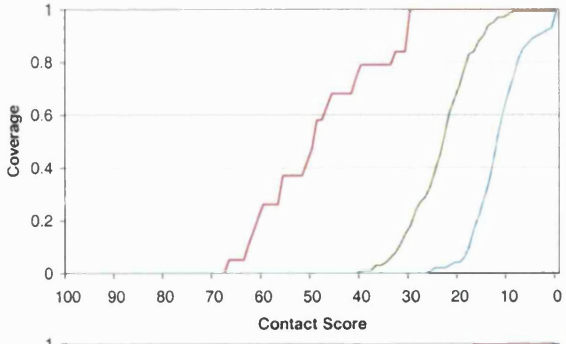
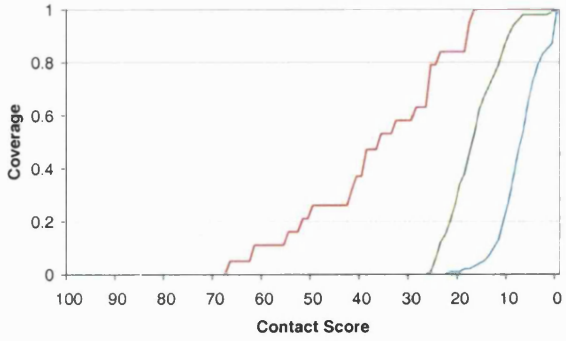
Cluster	Number of Templates (sequence families)	Coverage-Versus-Contact Plot
70	2 (3, 14)	
75	2 (2, 14)	
80	3 (2, 4, 6)	
85	1 (2)	

Figure 3.22: Comparing the coverage-versus-contact plots for homologues, fold-relatives and non-relatives using the multiple templates from the superfamily 3.40.50.950, generated from four different cluster cutoffs (70, 75, 80, 85). The numbers in parentheses indicate the number of proteins in each sequence family.

3.3.3 Examining Pre-Search Filters to Improve Sensitivity and Accelerate the Database Search.

By comparing easily obtained structural properties, such as length, secondary structure content or number of inter-residue contacts, it may be possible to judge whether a template-structure match is sufficiently similar to be worth the computational expense of a structure comparison. An effective pre-search filter would not only increase the speed of the database search by removing unnecessary comparisons, but could also improve the differentiation between homologues and non-relatives. Two such pre-filters were examined; a minimum size overlap cutoff and a minimum contact overlap cutoff.

3.3.3.1 Pre-Search Filter: Minimum Size Overlap

The minimum size cutoff was introduced to provide a fast and simple measure of template-structure compatibility (see equation 3.4). Since it is rare to find evolutionary relationships between proteins that are radically different in size, many sequence and structure comparison methods introduce this type of cutoff.

$$Cutoff_{size} = \frac{L_1}{L_2} * 100\% \tag{3.4}$$

Where
L_1= min ($Length_{structure}$, $Length_{template}$)
L_2= max ($Length_{structure}$, $Length_{template}$)

3.3.3.2 Pre-Search Filter: Minimum Contact Overlap

The second suitability measure to be examined was a minimum contact overlap (see equation 3.5). As the number of inter-residue contacts seen in a protein structure is dependent on a variety of structural properties such as size, secondary structure content and packing, it was postulated that the number of contacts could be used as a discriminating feature.

$$Cutoff_{contact} = \frac{C_1}{C_2} * 100\% \tag{3.5}$$

Where
C_1= min ($Contacts_{structure}$, $Contacts_{template}$)
C_2= max ($Contacts_{structure}$, $Contacts_{template}$)

3.3.3.3 Results of the Pre-Search Filters

A range of cutoff values were tested for both these filter types (20, 30, 40, 50, 60, 70 and 80%) on the results of each of the cluster cutoff bins for the four test superfamilies. The optimal cutoff value would be one that could remove the non-relatives from the coverage plot without affecting the homologues.

Figure 3.23 shows an example of the difference in coverage-versus-contact plots when these minimum size overlap thresholds have been applied to the results of the database scan for the $\alpha\beta$-plait superfamily 3.30.70.330 (Cutoff70). From this plot it can be observed that a size cutoff of 60% provides the optimal selectivity between homologues and non-relatives with minimal reduction in the coverage of homologues. The performance of this size cutoff was seen to be consistent in the majority of the plots for all cluster groups in all four test superfamilies and was therefore deemed suitable for inclusion in the final procedure.

An example of the effect of applying a contact overlap cutoff can be seen in figure 3.24. Although the minimum contact overlap filter was able to considerably reduce the number of analogues and non-relatives from the search list, the point at which the homologue coverage was adversely affected was quite variable. Due to this inconsistency across the four superfamilies, the minimum contact overlap was not included in the final procedure.

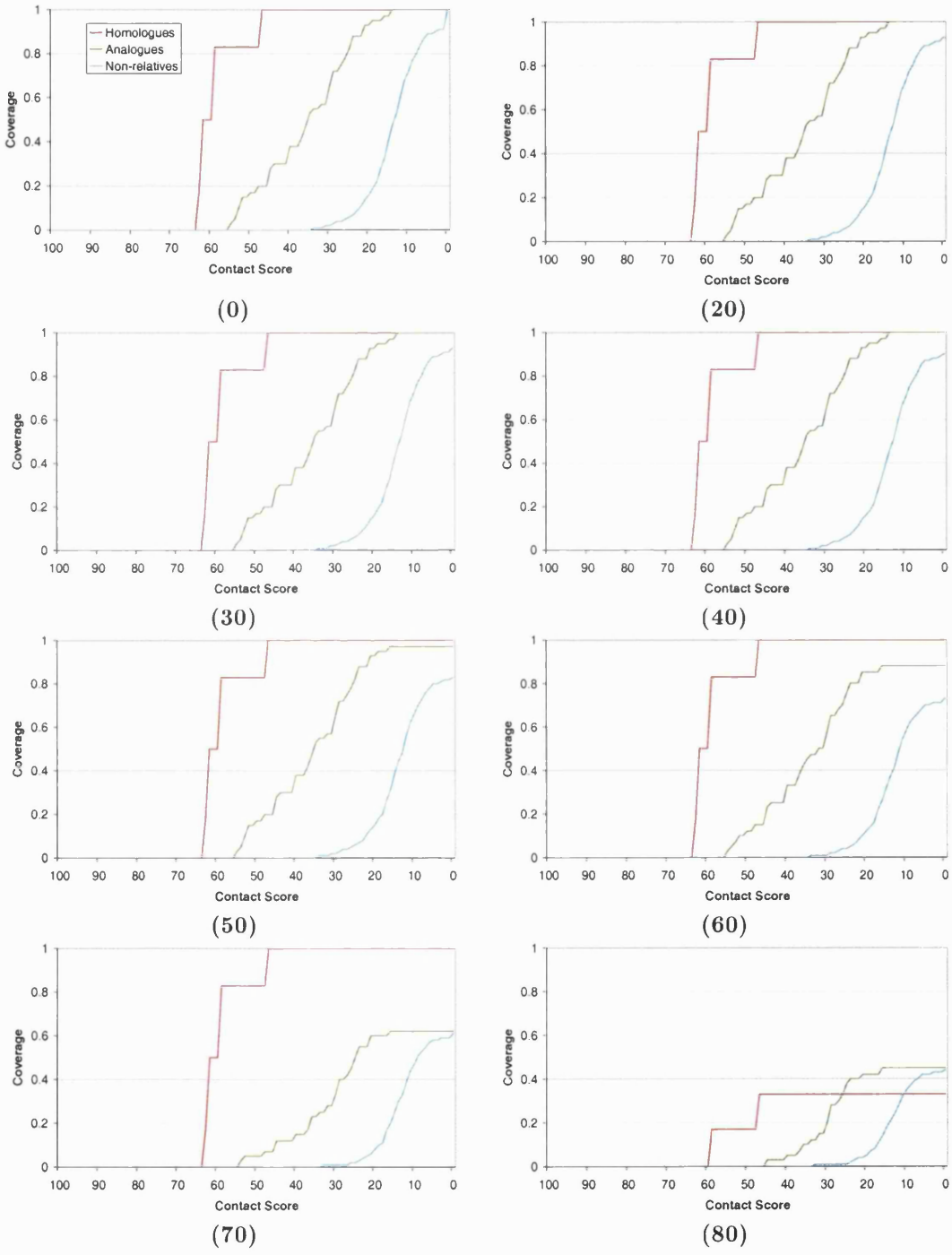


Figure 3.23: Introducing a minimum size overlap cutoff as a pre-search filter. These plots demonstrate the effect of applying an increasingly stringent size cutoff (values of 0, 20, 30, 40, 50, 60, 70, and 80 respectively) to the results from the 3.30.70.330 superfamily (Cluster70) database scan.

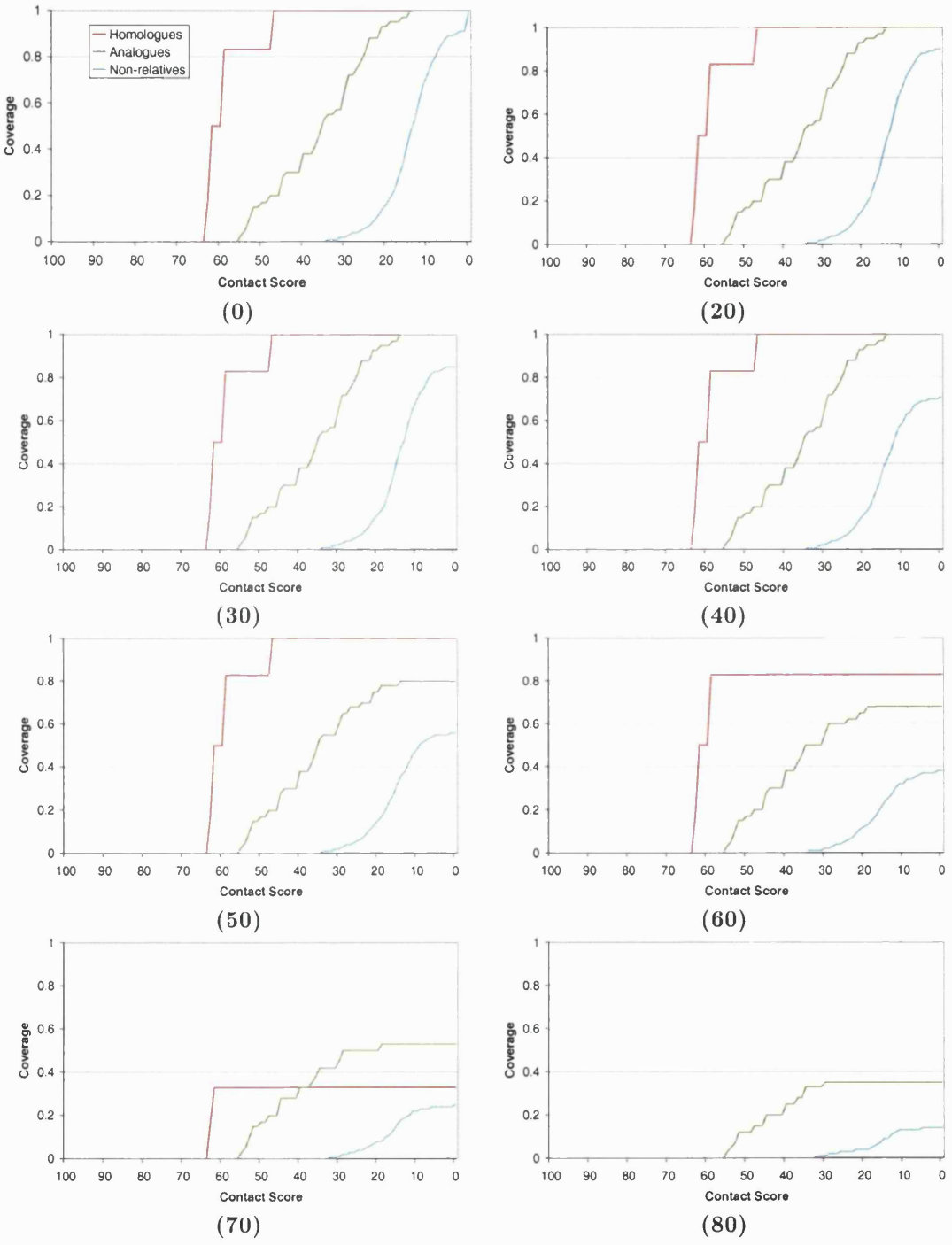
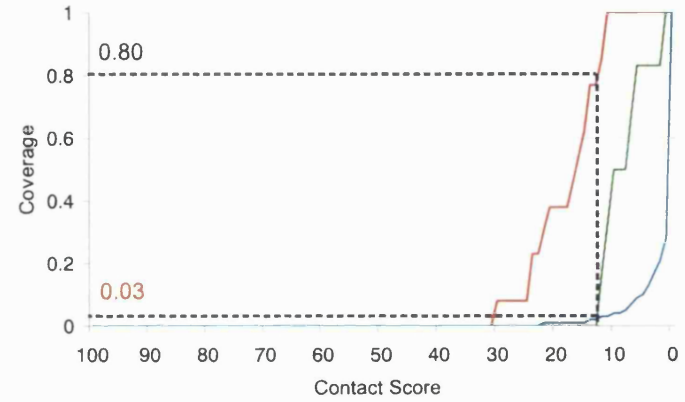


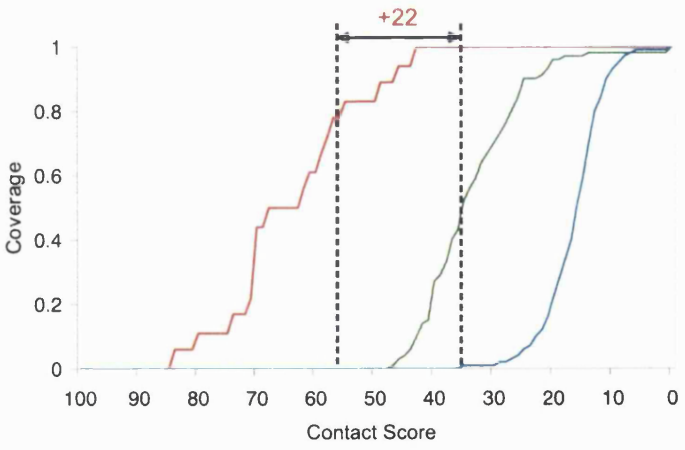
Figure 3.24: Introducing a minimum contact overlap cutoff as a pre-search filter. These plots demonstrate the effect of applying an increasingly stringent contact overlap cutoff (values of 0, 20, 30, 40, 50, 60, 70, and 80 respectively) to the results from the 3.30.70.330 superfamily (Cluster70) database scan.

3.3.4 Summary of Clustering Optimisation Results

The results from examining these four superfamilies indicate that the optimal threshold to generate structurally coherent clusters is using a SSAP score of 80. From manual inspection, the coverage-versus-contact plots at this threshold demonstrate the highest or equal highest discrimination between homologous matches and non-homologous matches. In two of the four superfamilies the improvement in discrimination for this threshold is clear (cytokine superfamily; 1.20.160.20, figure 3.11 and the thioesterase superfamily from the Rossmann fold; 3.40.50.950, figure 3.22).



(a) COVERAGE



(b) CONTACT SCORE

Figure 3.25: Quantifying the coverage-versus-contact plots. Homologous matches are shown in red, analogous matches are shown in green and non-related matches are shown in blue. The two measures of structural template discrimination between homologues and non-relatives: (a) coverage of non-relatives at 80% coverage of homologues and (b) difference in contact overlap score between 80% homologue coverage and the first non-relative.

These results can be quantified by examining the coverage of non-relatives, or false positives, at a given coverage of homologues, e.g. 80% (see figure 3.25a). This can be illustrated by comparing the coverage-versus-contact plots for Cluster70 and Cluster80 from the cytokine-like superfamily (CATH classification 1.20.160.30, figure 3.11). Using a threshold homologue coverage of 80%, Cluster70 has a non-relative coverage of 0.14 whereas Cluster80 has a non-relative coverage of 0.01. The reduced number of non-relatives, i.e. false positives, for the same number of homologues, i.e. true positives, indicates that Cluster80 has provided better discrimination. The results when using a homologue threshold coverage of 80% can be seen in table 3.3.

However, if the four clustering thresholds find all the homologues before finding any non-relatives then this measure of discrimination would not prove useful in selecting the optimal threshold. A more descriptive measure for these cases would be to assess the difference in contact scores between the 80% coverage of homologues, or true positives, and the first non-relative, or false positive (see figure 3.25b). The results when using this discrimination score can be also seen in table 3.3.

	(a) COVERAGE				(b) CONTACT SCORE			
	Cluster threshold				Cluster threshold			
Superfamily	*70*	75	80	85	70	75	*80*	85
1.20.160.30	0.14	0.15	**0.01**	0.03	-12	-12	**-1**	-10
2.60.40.420	**0.00**	**0.00**	**0.00**	**0.00**	+9	+9	**+20**	+19
3.30.70.330	**0.00**	**0.00**	**0.00**	**0.00**	**+25**	**+25**	+22	+22
3.40.50.950	**0.00**	**0.00**	**0.00**	0.02	+5	+4	**+6**	+2

Table 3.3: Summary of the optimisation results. The clustering thresholds are compared using two different measures for each of the four superfamilies: (a) the coverage of non-relatives at 80% coverage of homologues and (b) the difference in contact scores between 80% coverage of homologues and the first non-relative. The clustering threshold with the highest value for each differentiation measure is highlighted in bold for each superfamily.

3.3.5 Searching Novel Structures Against the Template Library

To assess the performance of the structural templates in recognition, a dataset of 303 remote structures, i.e. structures which had no detectable sequence similarity to structures in the templates, was identified (see section 3.2.5.2). A library of structural templates was also generated by clustering the structures in CATH v1.7 with a SSAP score of 80 as optimised in section 3.3.2.

Each of the 303 remote structures was then scanned against the library of structural templates. As described in section 3.2.5, a structural alignment was made for each template-structure comparison, then the contact overlap score was calculated based on that alignment. If a query structure had matched more than one structural template in a particular superfamily, then only the match with the highest contact overlap score was considered. A validated assignment of each structure in the remote dataset was given by the classification in CATH v2.0, so each match could be assigned as either homologous or non-homologous, i.e. a true positive or a false positive, simply by comparing the v2.0 classification code of the query structure with the v1.7 classification code of the matched structural template.

As mentioned in section 3.2.5.1, structural templates could not be generated for superfamilies that only contained either a single structure or highly similar structures. As a result, some of the 303 structures in the remote dataset were classified in superfamilies in CATH v2.0 which did not have a corresponding structural template in CATH v1.7, i.e. there was no true positive match from the database search. Of the 303 structures in the dataset, 228 were represented by at least one structural template in CATH v1.7. To account for this, a subset of the database scans was also taken that only included the results from the 228 structures with represented superfamilies. A coverage-versus-error plot was generated comparing the coverage from all 303 structures against the coverage for the 228 structures that could be assigned true positive hits (see figure 3.26).

It is important to remember that, although the majority of the structures in a superfamily may be represented by one or more structural templates, it is rarely true that all the structures in the superfamily are represented. When generating the templates, the clustering process inevitably provides some clusters that contain only a single structure and these single structure clusters are not converted into templates and are therefore not represented by the template library. This problem is made worse by the fact that a large proportion of selections of targets for experimental determination is based on proteins that have no detectable sequence similarity to a

previously solved structure. Therefore, these newly determined structures will tend to be very diverse relatives to structures already classified within the superfamily. This is partly due to the desirability of identifying novel folds and also due to the structural genomics initiatives that aim to provide a protein structure within homology modelling distance of every sequence ($\geq$35% sequence identity). However, this has provided a large number of 'islands' of structural space, otherwise known as singletons. To fully explore the evolution of protein structure, it will be vital that structural genomics initiatives continue to supplement the PDB with intermediate structures effectively providing bridges to connect these islands. Until the point arrives where the structure databases are sufficiently populated to allow these structural templates to model every diverse protein in every superfamily, the template library will not be expected to completely replace the pairwise SSAP comparisons. However, it typically takes around 90 minutes to search a 150 residue structure against the library of 407 structural templates compared to 12 hours to perform a pairwise search of 3,581 non-identical representatives. Therefore, all significant assignments made by searching the template library would be classified in CATH, thus leaving a much smaller set requiring pairwise SSAP comparisons. Also since these templates model the evolution of protein structure by combining distantly related proteins, they may also be expected to recognise more distant structural relationships than simple pairwise comparisons.

At an error rate of 0.1 the structures represented by the structural templates gave a coverage of 0.52 and at the same error rate the dataset containing all structures gave a coverage of 0.35. This demonstrates that the structural templates are able to correctly assign the homologous superfamily to more than 50% of the new structures where a correct assignment is possible. The maximum coverage of 0.62 for the represented structures was reached at an error rate of 0.20 and the maximum coverage of 0.49 for all structures was reached at an error rate of 0.35. The fact that the results for the dataset of represented structures did not reach full coverage implied that either the structural templates did not fully model the structural diversity or the testing protocol contained flaws, such as misclassifications in CATH. This issue is discussed in more detail in section 3.4.1.2. Obviously, since the assignments of homology for the dataset of 303 proteins were made using pairwise SSAP comparisons, performing SSAP comparisons against each structure from the correct homologous superfamily in CATH v1.7 would provide 100% coverage. Thus, augmenting the structural template library with the singleton structures involved in these islands of structural space could be expected to provide the most efficient structure searching tool.

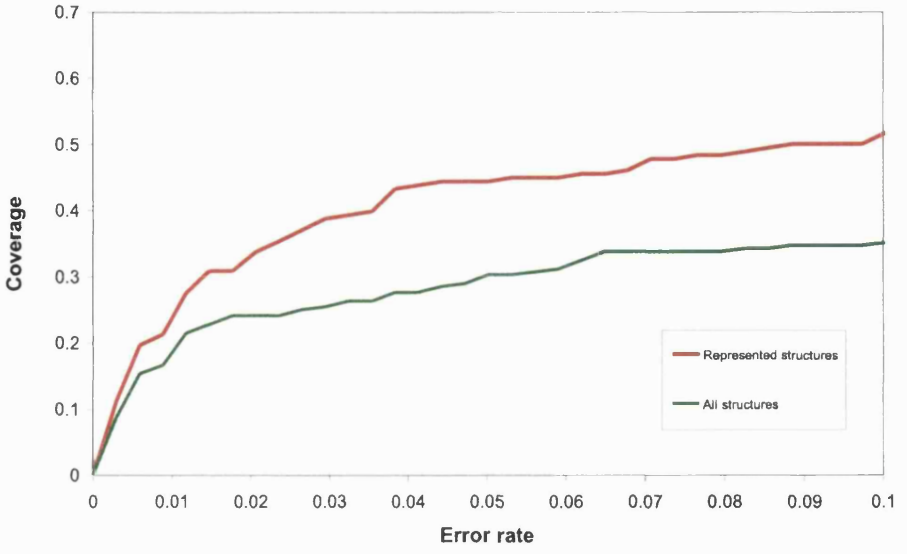


Figure 3.26: Performance of the structural templates in recognising structures within the same homologous superfamily. This coverage-versus-error plot compares the results from two sets of data. The results for 228 structures that are represented in the structural template library, i.e. have a possible true positive match, are shown in red. The results for all the structures in the dataset are shown in green.

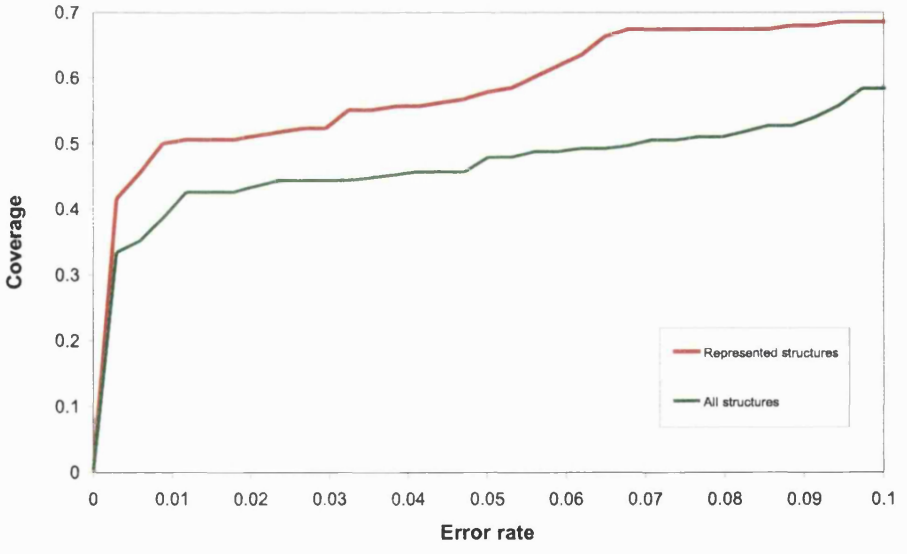


Figure 3.27: Performance of the structural templates in recognising structures within the same fold group in CATH. This coverage-versus-error plot compares the results from two sets of data. The results for 228 structures that are represented in the structural template library, i.e. have a possible true positive match, are shown in red. The results for all the structures in the dataset are shown in green.

The coverage in figure 3.26 was based on whether the structural templates could correctly assign the homologous superfamily for the dataset of remote structures. However, the templates were designed to recognise distant structural similarities and as such, it may also be useful to examine the ability of the structural templates to correctly recognise the general fold group, or topology in CATH, as well as the specific homologous superfamily. The coverage-versus-error plot based on the assignment of topology can be seen in figure 3.27. Again, the results for structures represented by the structural templates are compared against the results for all the structures. As expected, the performance for both sets of results increases when attempting to recognise the topology rather than the homologous superfamily. At an error rate of 0.1, the coverage with the represented structures increases from 0.52 to 0.69 and the coverage when using all the structures increases from 0.35 to 0.58.

3.4 Discussion

3.4.1 Overview

This chapter has presented a novel protocol that generates structurally coherent multiple structure alignments within homologous superfamilies in CATH. A multiple-linkage clustering algorithm was written and optimised that allowed multiple templates to represent structurally diverse superfamilies. The optimisation of this algorithm involved identifying a clustering threshold which would select a set of representative structures that sample a high degree of structural variability yet would not prove so diverse as to result in a poor structural alignment. A dataset of four structurally diverse homologous superfamilies was used in this optimisation procedure, with coverage-versus-contact plots employed to compare the performance of each clustering threshold. A library of structural templates was then generated using this optimised clustering protocol with the structures in CATH v1.7. The performance of this library was then assessed by attempting to recognise the homologous superfamily and topology of proteins with no detectable sequence relationship to structures in the template library.

The advantage of using structural templates to represent the structure database rather than the more traditional approach of using individual structures is two-fold. Since a structural template effectively represents a large number of individual structures, far fewer comparisons are necessary to cover an equivalent area of structural space. Also, the structural templates encompass evolutionary information that can be used to highlight important and therefore uniquely identifying structural features for a given superfamily and ignore highly variable regions. This should allow a more sensitive probe of distant structural relationships than using pairwise methods alone.

3.4.1.1 Errors in the Fold Recognition Performance of the Structural Templates

As mentioned in section 3.2.5.1, structural templates could only be generated for superfamilies containing sufficiently diverse structures to form clusters involving more than one protein. Although only 37% (340/903) of the superfamilies in the database met this criterion, the structural templates generated from these 340 superfamilies represented 55% of the sequence families in the database and 66% of the non-identical structures.

When the performance of the structural template library was assessed using the dataset of remote structures (see section 3.3.5), the homologous superfamily could

be correctly assigned in a maximum of 62% of the structures (with a representative template). Therefore 38% of these structures were not identified by the native structural template. However, the results of the coverage-versus-contact plots discussed in section 3.3.2 and illustrated in figure 3.11 demonstrate that in some cases the structural templates struggle to recognise even close structural homologues. In both these cases, the lack of recognition can be explained by one or more of the following reasons.

- **Quality of structural template**
  The structural template contain such structural diversity that no conserved contacts can be identified. This is addressed with the optimisation of the clustering threshold (see section 3.2.4).

- **Coverage of structural template**
  The structural templates do not fully represent the structural diversity in the homologous superfamily. When structures are clustered within the homologous superfamily inevitably some proteins will be left in single structure clusters and therefore may not be represented in the templates. The database composition of the structural templates is discussed in section 3.4.1.2.

- **Classification errors**
  Errors present in the CATH classification may result in database matches being classed as false positives when they are actually distant structural relatives. Since the structural templates attempt to use evolutionary information to identify more distant relationships than pairwise structural comparison, false positive matches scoring highly in the database scans may be the result of distant structural similarity. An example is illustrated in section 3.4.1.3.

3.4.1.2 Database Composition

Figure 3.28 illustrates the percentage of sequence families that are directly represented within the structural templates for homologous superfamilies in CATH v1.7. Every point in the graph corresponds to an homologous superfamily and the position on the x-axis is dictated by the number of sequence families within structural templates in that superfamily divided by the total number of sequence families. This distribution demonstrates that the majority of the homologous superfamilies either is fully represented or is not represented at all. However, a number of superfamilies fall in between these two extremes suggesting that the structural templates may not

fully represent the structures within those superfamilies, especially for superfamilies with as little as 25% of the sequence families represented.

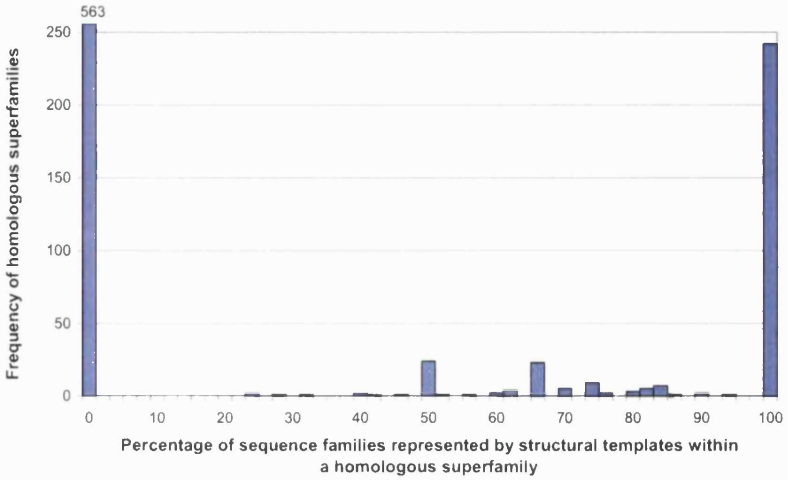


Figure 3.28: Representation of sequence families by the structural templates. The graph shows the distribution of homologous superfamilies based on percentage of sequence families represented by the structural templates.

Figure 3.29 applies this assessment of database composition to the results from the dataset of 303 remote homologues. The dataset of 303 structures was split into two categories; the 228 structures that were related to homologous superfamilies represented by the template library and the 75 structures related to homologous superfamilies not represented by the template library. The set of 228 represented structures, was split further into the 141 correct superfamily assignments, i.e. true positives (TRUE_POS), and the 87 incorrect superfamily assignments, i.e. false positives (FALSE_POS). Although some of the false positives belong to superfamilies that were completely represented by the template library, the relative distribution of false positive structures tends towards those belonging to the superfamilies with lower representation in the template library. This suggests that the coverage will increase further as the superfamilies become more populated.

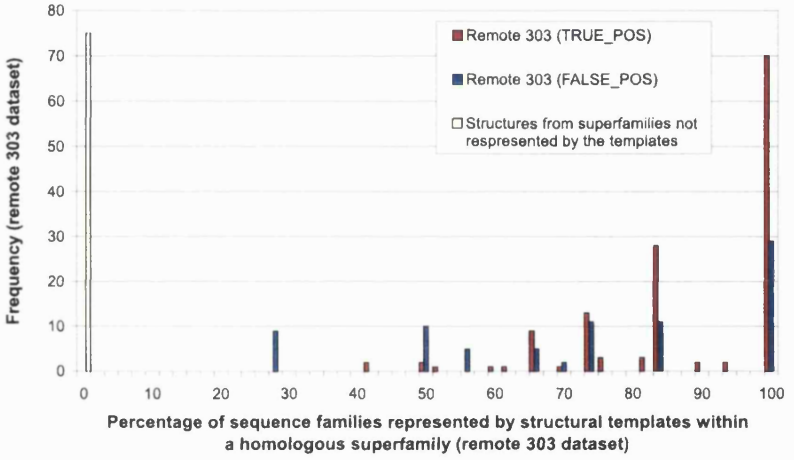


Figure 3.29: Recognition of the dataset of remote homologues in terms of representation in the template library. Each of the 303 structures from the remote homologue dataset was scored in terms of the percentage of sequence families represented by templates from the correct homologous superfamily. The red bars illustrate the structures correctly recognised from the scan of the template library, i.e. true positives (TRUE_POS), the blue bars illustrate the structures not recognised by the template library, i.e. false positives (FALSE_POS), The yellow bar represents the 75 structures from superfamilies that are not represented by any templates in the library.

3.4.1.3 Identification of Distant Structural Similarities

The structural templates were designed to provide a more sensitive probe of evolutionary relationships than pairwise structure comparison. However, the dataset used to assess the performance of the templates was validated with pairwise structure comparison (the structural similarity measure employed in the CATH classification protocol was calculated using the SSAP comparison algorithm). Therefore, if the templates were to identify evolutionary relationships that were too distant to be recognised with pairwise comparison, then the match would be considered an error. For this reason, matches deemed as false positives were necessarily treated with caution. A selection of high scoring false positive matches from these database scans were analysed for possible errors in classification due to distant structural relationships previously unrecognisable with pairwise structure comparison.

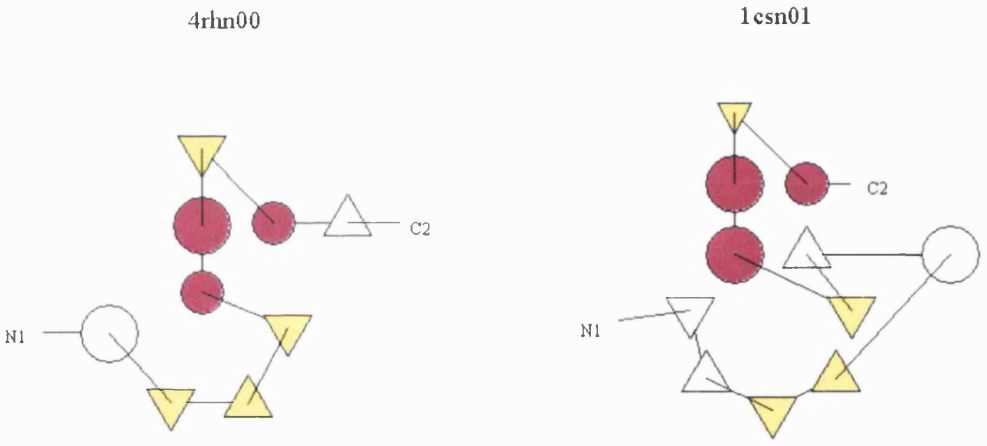


Figure 3.30: Distant structural similarities identified with the structural templates. Topology, or TOPS, diagrams (Westhead *et al.*, 1999) are displayed for the structural domains 1CSN, domain 1, and 4RHN.

An example of a putative structural relationship recognised by the template library can be illustrated by examining the highest scoring match between one of the structures in the remote dataset, the first domain of the kinase-like phosphotranferase protein (1CSN, domain 1; CATH classification 3.20.200.20), and a structural template from the histidine-triad nucleotide binding (HINT) superfamily (CATH classification 3.20.428.10). Figure 3.31 compares the structure 1CSN, domain 1, to one of the two structures from the high scoring structural template (4RHN). This figure also shows the comparison contact map based on the alignment between the single structure and the structural template with the overlapping contacts displayed

in red.

From the CATH classifications, it can be seen that these two structures have been assigned to different fold groups within the $\alpha\beta$-barrel architecture. However the structures 1CSN, domain 1 and 4RHN display a similar connectivity in the 5-layer up-down β-barrel in the structural core. The contact patterns between 1CSN, domain 1 and the structural template are highly similar as a result of this common β-sheet and this is reflected in the high degree of overlap between the two contact maps. The similarity between the aligned secondary structure of 1CSN, domain 1 and the consensus secondary structure for the structural template (3.30.428.10) can also be seen clearly from this comparison contact map. The topology, or TOPS, diagrams (Westhead *et al.*, 1999) for these two structures (figure 3.30) shows that the β-barrel in the structure 1CSN, domain 1, is more extensively curved than that of 4RHN, however the connections in the β-sheet are similar as is the $\beta\alpha\beta$ motif observed in the middle of the β-sheet.

3.4.1.4 Summary

In conclusion, the templates can correctly recognise the fold group for 70% of remote structures and the superfamily for 52% of remote structures provided the structure belongs to a superfamily represented in the template library. This represents a considerable saving of time for classifying new structures in CATH. Furthermore, many of the false positives may be remote similarities undetected by CATH. These have been subjected to manual evaluation by the CATH curator to check for missed relationships. Any new structures not recognised by the template library will be subjected to pairwise SSAP scans of all representative structures from CATH.

It has also been demonstrated that many of the missed homologous relationships were from diverse superfamilies that were not yet fully represented by the template library. This provides further evidence that the performance of this method will continue to increase as the structure databases become more populated, since the templates will be able to represent the superfamilies more completely.

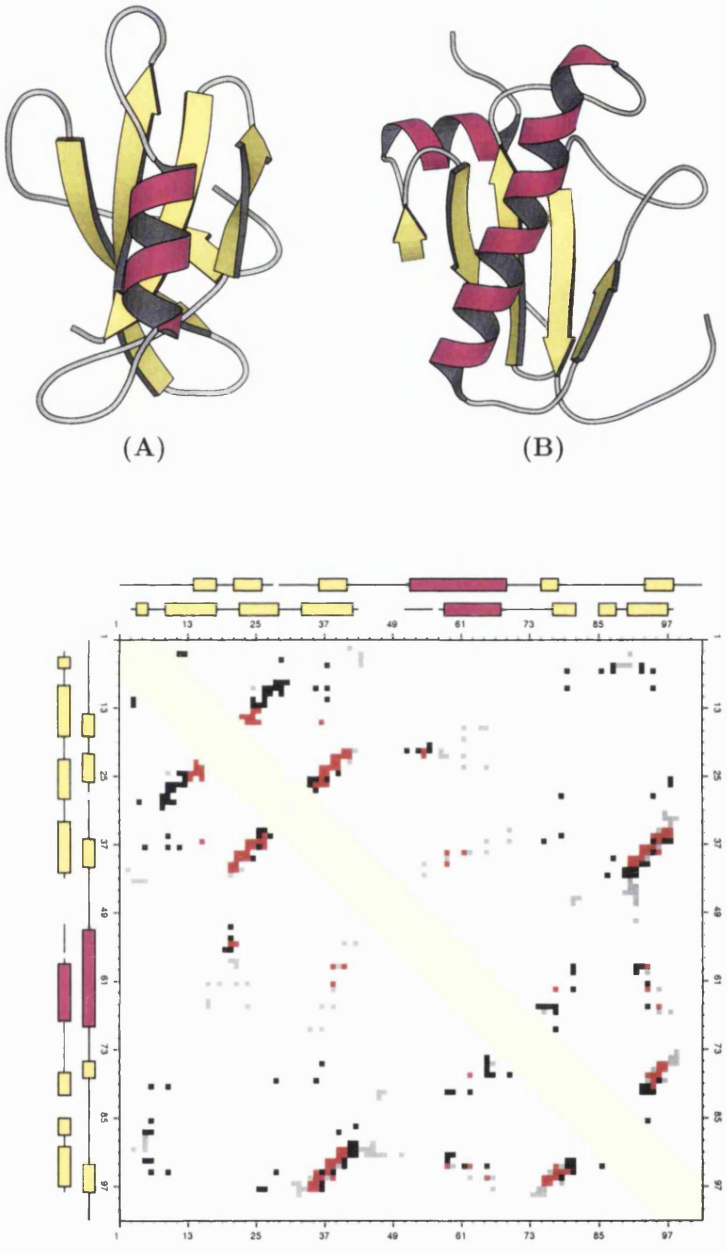


Figure 3.31: Identification of distant structural similarities in the CATH database. (A) shows one of the remote 303 query protein domains (CATH code 1csn01) classified in the protein kinase-like superfamily (CATH classification 3.30.200.20).(B) shows a representative for the highest scoring structural template which is from the histidine triad nucleotide-binding (HINT) superfamily (CATH classification 3.30.428.10). The template-structure contact comparison map can be seen in (C) with the consensus template contacts in grey, the query structure contacts in black and the overlapping contacts in red. The template and the query structure are classified in different fold groups in CATH therefore this highest scoring match is classified as a fold recognition error, despite structural similarities.

3.5 Appendix

3.5.1 Implementing the Structural Templates in the CATH Server

3.5.1.1 Background

In order to allow remote access to the CATH classification protocol, a web-based interface was generated and integrated into the CATH web site (http://www.biochem.ucl.ac.uk/bsm/cath). One of the main uses for a tool such as this is to allow a structural biologist to investigate structural homology of a protein that is either not yet classified in CATH, or not yet submitted to the Protein Data Bank (Berman *et al.*, 2000). Since the structural template library allows the structural database to be scanned far more quickly than an exhaustive pairwise comparison of all the single structures, this library was incorporated into the remote classification protocol.

3.5.1.2 Using the GRATH Algorithm as a Rapid Pre-Filter

The GRATH algorithm (Harrison *et al.*, 2002) uses a graph theoretical approach to perform a rapid pairwise comparison of protein structures (see section 1.2.7.3). A description of the protein structure is generated from the internal distances between secondary structures and graph theory is used to maximise the 'clique', or overlapping pattern of secondary structure distances, between the two structures. Although this method does not provide an accurate alignment, it presents a fast and discriminatory method of separating fold and non-fold matches from a large database of structures. When benchmarked against the CATH database, this algorithm gave a 95% accuracy in finding the correct fold within the top ten highest scoring comparisons from a database search.

3.5.1.3 Designing an Interface to the CATH Server

A web interface was written that allows users to upload their own PDB files to be structurally classified. The interface first manipulates these uploaded files to render them suitable for each of the algorithms in the classification protocol then executes and manages the output of these algorithms, checking for errors and displaying the results at each stage.

Considering the colourful variety of formats that have doubtless enriched the lives of all those fortunate enough to parse PDB files, it may not be surprising that

the front-end of the web interface is initiated with a client dialogue that validates and selects relevant sections from the uploaded PDB files. Since the algorithms in the protocol only function on single chain structures, and optimally on structures with a single domain, the user is also asked to specify which chain to include in the classification and given the option of providing manual domain boundaries.

The submitted structure is then scanned against the structure database using the fast GRATH algorithm, which often takes around one or two minutes. The results from this initial GRATH scan are used to identify the three most likely fold groups for the query structure. All the superfamilies within these three fold groups are then included in the more thorough search of the structural templates. These comparisons are then scored and ranked according to contact overlap and the results returned to the user. The web interface and protocol of the CATH server protocol (http://www.biochem.ucl.ac.uk/cath) is illustrated in figure 3.32.

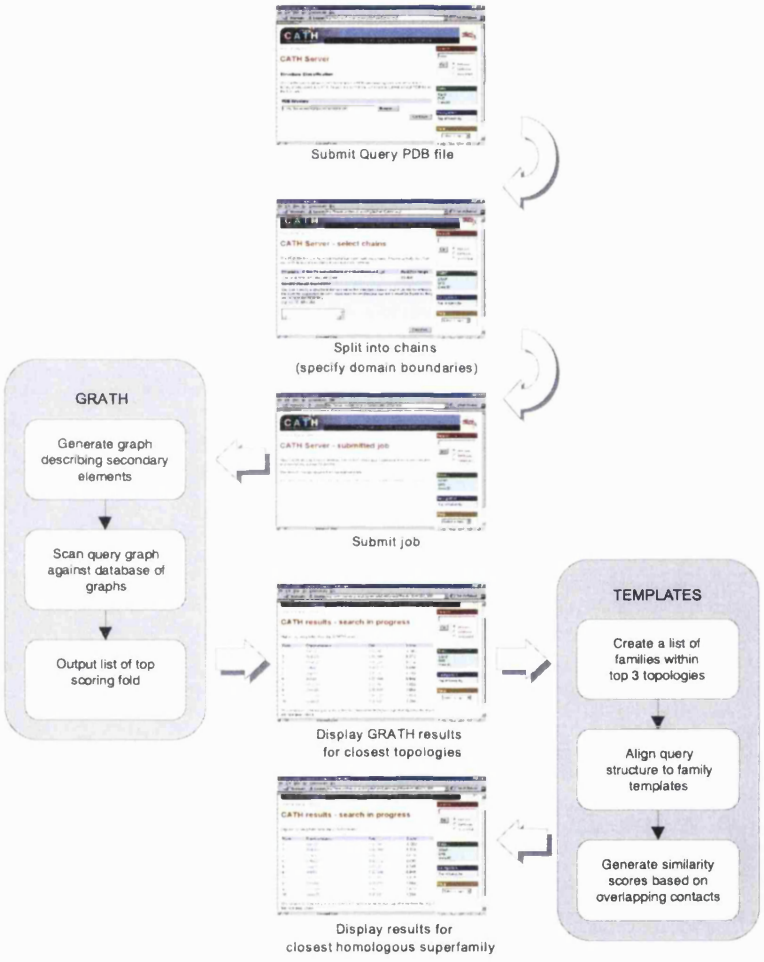


Figure 3.32: Flowchart describing the process of remote structure assignment using the CATH server. A protein structure is submitted to the server (PDB format) and split into chains. Structural domain boundaries can also be manually assigned at this point. The job is then submitted to the CATH server which runs an initial database search with the fast GRATH algorithm. The structural templates from the top three scoring folds are then scanned with the query structure. The matches are then ranked in decreasing order of contact overlap score and the alignments and structural superpositions made available for download.

Chapter 4

Structure Comparison Methods to Improve *ab initio* Protein Structure Prediction

4.1 Introduction

4.1.1 Background

Predicting the tertiary structure of a protein directly from its amino acid sequence, i.e. *ab initio*, has for a long time been regarded as the 'Holy Grail' in the field of Structural Biology. Since structure plays such an important role in understanding the biological function of a protein, a method that can predict structural features such as active sites, binding surfaces or the general folding arrangement solely from the readily available residue sequence would be most desirable.

Currently, the main source of structural information is that derived from experimental techniques such as X-ray crystallography and NMR spectroscopy. Generally these methods provide high quality three-dimensional (3D) protein structures showing ligand interactions and metal-ion binding sites in sufficiently high resolution for pharmaceutical application of drug design and target selection. However, both these experimental procedures have limitations when applied indiscriminately to uncharacterised protein sequences. For example, X-ray crystallography relies on obtaining pure crystals of the protein and some proteins can prove difficult to crystallise, especially membrane proteins and those containing regions of high flexibility. Also, large proteins (i.e. greater than 30KDa) are particularly hard to solve by NMR spectroscopy due to inherent limitations regarding the decreased sensitivity to Brownian motion of larger molecules. Many of the difficulties facing these experimental

methods can be overcome either from alterations in the structure, by making calculated changes to the protein sequence, or using more advanced analytical techniques. However, even without such technicalities proving problematic, protein structure determination by these experimental techniques is still a time consuming process. As a result, a procedure that allows some structural information to be identified from primary sequence data could be useful, at the very least, in the selection process for full experimental structure determination, or more optimistically, as a self-contained method for putative assignment of structure and function.

Since a variety of structural genomics projects aim to provide experimentally determined protein structures representing all the sequence families in the genomic sequence databases, it is distinctly possible that *ab initio* prediction may be sidelined as impractical and esoteric in the near future. However, a more fundamental reason for studying *ab initio* methods is that advancing the understanding of the chemical and physical properties is a worthy goal. Certainly, this detailed knowledge of protein structure would provide far greater scope for folding pathways in addition to protein design and engineering. Also, this level of understanding may prove vital when investigating and simulating more complex systems such as how proteins function in concert on a macromolecular scale.

Perhaps the importance that *ab initio* prediction methods still have to offer could be measured by the enormous investment by IBM in developing a 'petaflop' computer, i.e. capable of calculating 10^{15} floating-point operations per second, called 'Blue Gene' in order to tackle the protein folding problem (Butler, 1999).

4.1.2 Predicting Structural Features from Sequence

The *ab initio* prediction of protein structure still poses one of the most challenging problems in Structural Biology. This difficulty arises from two important factors: the enormous number of possibilities of conformational space (Dill, 1993) and the subtle interplay between the chemicophysical properties involved in protein structure stability, especially when regarding the co-operative effects of large networks of residues. As a result, current *ab initio* methods of protein folding are often limited by the huge computational effort involved in simulating the folding process even for small peptides. An additional problem is the difficulty in locating the energy minimum corresponding to the native conformation without converging on other non-native local energy minima.

There are four major types of structural predictions that can be derived from the amino acid sequence information alone. These are the prediction of class, secondary

structure, inter-residue contacts and tertiary structure. Each of these four areas of *ab initio* prediction will be discussed in the following sections.

4.1.2.1 Class Prediction

Many attempts have been made to predict general structural properties for proteins given the composition of the amino acid sequence. At the most basic level, composition is given as the fraction of each of the 20 amino acids in the sequences. This has also been extended to examine the composition of sequence fragments, for example using blocks of two or three residues, rather than the composition of individual residues.

This type of analysis has been used to predict the secondary structure content of a given protein sequence (i.e. percentage of helix, strand and coil) with a reasonable degree of accuracy, certainly of a comparable accuracy to the results from experimental methods such as circular dichroism (Rost & Sander, 1993; Eisenhaber *et al.*, 1996b,a).

4.1.2.2 Secondary Structure Prediction

When examining the secondary structure state (α-helix, β-strand or random coil) of residues in known structures, it can be noted that many amino acids display striking differences in propensity to adopt these different secondary structure states. For example, steric clashes between the pyrrolidine side chain of proline and the C_β atom of the preceding residue generally restricts this amino acid from being found within an α-helix (although it can appear at the first turn of the helix). These intrinsic propensities for secondary structure were analysed by Chou & Fasman (1974a) using a very limited dataset of protein structures (only 15 protein structures were available in 1974) and this was also used in a predictive method (Chou & Fasman, 1974b). A more successful method for using amino acid propensities to predict secondary structure was presented by Garnier, Osguthorpe and Robson in the GOR method (Garnier *et al.*, 1978). Instead of using propensities for single amino acids, this approach applied techniques taken from information theory to analyse a window of eight residues either side of the amino acid being predicted.

As the sequence database has grown many groups have attempted to use large alignments of related sequences to identify conserved patterns of amino acids typically seen in secondary structures. Examples of such patterns include repeats of hydrophobic and hydrophilic amino acids every three or four residues. Since there are 3.6 residues per turn of an α-helix, this recurring pattern often indicates a side

of an α-helix facing into the protein core and out to the solvent for hydrophobic and hydrophilic side-chains respectively. Also, positions in the sequence alignment with insertions and deletions usually coincide with random coil secondary secondary structure, often on the surface of the protein. It is only when many sequences are compared that the random evolutionary changes, i.e. noise, can be differentiated from conserved sequence patterns derived from conserved features in the protein structure.

The application of neural networks for the analysis of the sequence patterns in these multiple sequence families has so far proved the most successful method for automated prediction of secondary structure. The PHD method (Rost *et al.*, 1994) trained a neural network on profiles built from multiple sequence alignments. This was able to correctly assign the secondary structure states of around 70% of residues for previously unseen sequences. More recently, this accuracy score was increased to around 77% with the PSIPRED method (Jones, 1999b) by improving the quality of the sequence profiles that are used to train the neural networks.

4.1.2.3 Inter-Residue Contact Prediction

Since secondary structure can be predicted with reasonable accuracy, the next level of complexity in *ab initio* structure prediction is to predict how these secondary structure elements may pack together. To this end, considerable research effort has been spent on the prediction of interactions between residues within a protein from sequence alone. Knowledge of sufficient numbers of these points of contact could then be used to constrain the secondary structure elements and generate a reasonable model of the tertiary structure.

Prediction of these inter-residue contacts can be made by exploiting the phenomenon of correlated mutations (Gobel *et al.*, 1994; Taylor & Hatrick, 1994; Thomas *et al.*, 1996; Ortiz *et al.*, 1998; Fariselli *et al.*, 2001; Pollastri & Baldi, 2002). These correlated mutations arise due to the local steric and physicochemical environment changing following a given residue mutation. Mutations at positions close in spatial proximity acting to compensate for these changes are more likely to be accepted than the random changes observed in evolution. For this reason it is suggested that compensatory changes observed between two residues at a simultaneous point in the evolutionary ancestry may arise from the residue positions being close in the protein structure.

Again, many groups have attempted to recognise the sequence patterns resulting from correlated mutations by training neural networks on multiple sequence

alignments from known sequence families (Ortiz *et al.*, 1998; Fariselli *et al.*, 2001; Pollastri & Baldi, 2002). Having been trained, these neural networks are then used in a predictive capacity with previously unseen protein sequences. However, unlike secondary structure prediction, the prediction of inter-residue contacts by such methods has proved difficult and unreliable due to the enormous number of related sequences required to recognise such sequence patterns and the large number of false positives. One reason for the lack of substantial success with this approach is that compensatory mutations could occur across networks of residues rather than simply between two residues. This would make the sequence patterns for correlated mutations more likely to be specific for a given structural family rather than follow predictable rules across the structural space.

4.1.2.4 Tertiary Structure Prediction

As mentioned previously, the main goal of protein structure prediction is to obtain the tertiary fold directly from the amino acid sequence. Generally, most methods for predicting protein tertiary structure can be broken down into two parts.

- A procedure for generating a series of possible conformations of the protein chain.
- A potential energy function which can evaluate these conformations to correctly identify the native structure.

A general difficulty with *ab initio* prediction is the enormous number of conformations that a protein chain can possibly adopt. Many groups have chosen to simplify this problem by restricting the residues in the chain to discrete points on a 3D lattice (Hinds & Levitt, 1994; Kolinski & Skolnick, 1994; Park & Levitt, 1995) or by restricting the protein chain to a small number of allowed torsion angles (Dandekar & Argos, 1994; Srinivasan & Rose, 1995).

True *ab initio* methods then evaluate these predicted structures based solely on the fundamental physicochemical properties of amino acid residues, e.g. size and charge. However, a more pragmatic approach is to introduce knowledge-based techniques, i.e. methods that incorporate information from databases of known structures. The advantage of true *ab initio* methods is that, when successful, the results would be independent of any bias present in protein structure databases. Methods which rely on knowledge-based approaches alone, e.g. threading (see section 4.1.2.5), will have the inherent limitation that they can only provide accurate models for sequences adopting previously observed folds.

4.1.2.5 Fold Recognition

As mentioned in section 4.1.2.4, predicting an initial conformation for the protein chain can present a difficult problem due to the large number of conformational possibilities. To avoid this, a method was proposed that 'threaded' the query sequence into conformations adopted by experimentally solved protein structures or templates (Jones *et al.*, 1992). Each of these threaded structures were then assigned a global energy by comparing the distances between amino acids on this template structure with the distances seen in known structures. To allow for insertions and deletions, a double dynamic algorithm (see chapter 1) was employed to find the optimal alignment between the query sequence and template structure, i.e. the alignment that provided the lowest global energy using the same knowledge-based potential.

Therefore, threading methods avoid the computational expense of the first step of many *ab initio* procedures, i.e. generating putative conformations of the protein chain, by using known structures as templates. As a result threading offers a fast method for recognising sequences that adopt known structural folds.

Profile-based sequence comparisons also provide a means of fold recognition using sequence information alone. A sequence profile provides a highly detailed description of the observed residue changes for each position of a large multiple sequence alignment (discussed in more detail in chapter 5). The variability of residue substitutions observed at each position in the sequence alignment reflects the flexibility of these positions in 3D space. As a result, these sequence profiles implicitly incorporate a great deal of structural information that is specific to the family of proteins they describe. The most powerful profile-based sequence methods, such as SAM (Karplus *et al.*, 1998), can reach levels of recognition comparable to structure-based threading methods (Orengo *et al.*, 1999).

4.1.3 Aims

In a typical *ab initio* prediction method, a general packing arrangement of secondary structures is predicted, then this approximate protein structure undergoes a series of refinement stages. Often a large number of these models are generated using small variations in the parameters, then each is assessed for native-like structural features, such as solvent accessibility, good secondary structure packing and favourable inter-residue interactions. This step is used to determine whether each model is a likely candidate or should be discarded from the refinement process. This refinement process can prove extremely time consuming and computationally expensive since the protein chain can adopt so many conformational possibilities for each of these structures (see figure 4.1).

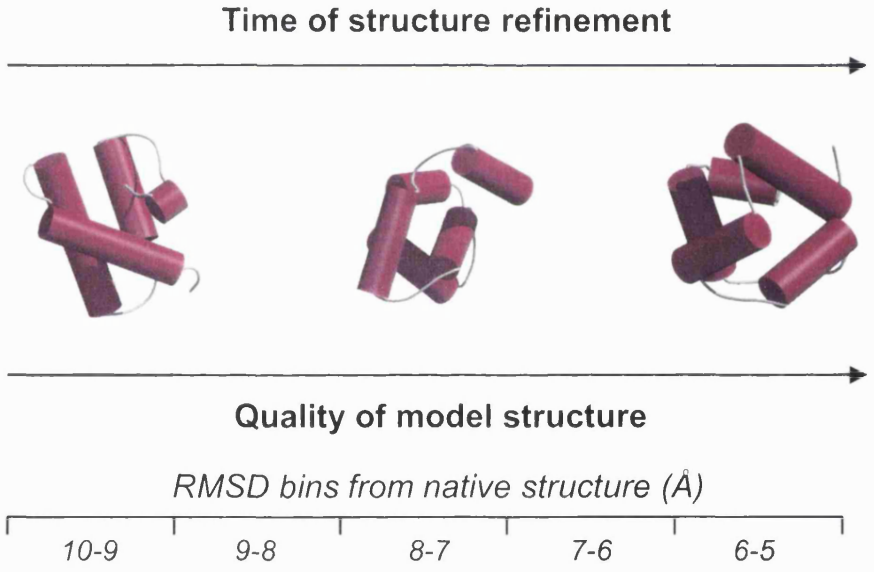


Figure 4.1: Overview of the structure refinement procedure of protein models predicted by *ab initio* methods. The iterative process of the structure refinement can take a large amount of processing time and resources. The aim of this method is to identify the native fold at an early stage of refinement in order to both accelerate the procedure and identify structural features from related structures that could improve the model.

The flowchart in figure 4.2 describes a general overview of the steps involved in the prediction and refinement of protein tertiary structure from amino acid sequence. From a given protein sequence, established *ab initio* methods would be employed to generate a number of low resolution structural models, i.e. predictions at the start of the structure refinement process. The method proposed in this chapter then attempts to recognise the most likely fold of this target structure by comparing

these approximate models to a database of known structures. A consensus of the results from the database searches of all these models is then taken in order to assign the most likely fold. After the native fold has been identified by this method, it is then proposed that further structural refinement could be driven by constraining the models with highly conserved structural features identified from the related superfamilies within the native fold. This last step will not be covered in this chapter, however the identification of conserved structural features such as inter-residue contact is discussed in detail in chapter 2.

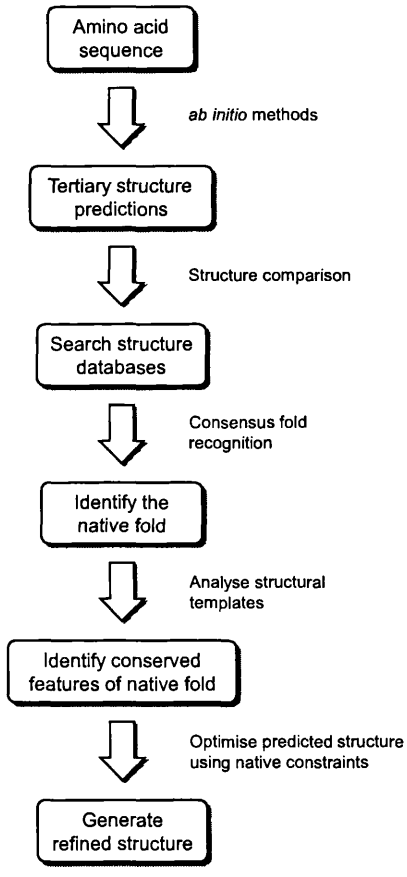


Figure 4.2: Flowchart describing a procedure to generate high quality *ab initio* predicted structures.

In summary, a method is presented which aims to recognise the native fold of a set of *ab initio* predicted model structures during an early stage of structure refinement, thus reducing the time and increasing the accuracy of further refinement. This protocol presents an alternative method of fold recognition that complements the more established threading methods currently in use. Also the accelerated refinement time, together with advancing *ab initio* methods, could enable an application

of *ab initio* approaches to far larger datasets of protein sequences such as genomic data.

The method presented has been developed to assist fold recognition for proteins with structural relatives using *ab initio* approaches. Threading methods have already been shown to perform well for some such targets. However, for very distant homologues or more diverse analogues, the potentials used in threading may not model the sequences sufficiently to distinguish the correct fold. *Ab initio* approaches using more flexible approaches, rather than the static templates used in threading, may perform better. Therefore, the fold recognition performance of this method was compared to the performance of traditional threading results.

The work discussed in this chapter was conducted in collaboration with Xavier de la Cruz at University College, London and was published in *Proteins: Structure, Function and Genetics* (de la Cruz *et al.*, 2002).

4.2 Methods

4.2.1 Definition of Terms

Methods for assessing the consensus fold recognition protocol presented in this chapter can be separated into two parts. The first describes the procedure for generating the different datasets of protein models that the fold recognition procedures will be applied to. This is discussed in more detail in section 4.2.2. The second part describes the two different structure comparison procedures used for the fold recognition (see section 4.2.3).

Figure 4.3: Example of a fold, or topology, relationship in CATH. Both PDB structures, 1b54 and 1aj0, share highly similar folding arrangements and are classified in the same TIM-barrel topology (3.20.20 in CATH). However there is insufficient evolutionary evidence to guarantee a common ancestor so are classified into two different homologous superfamilies in CATH

Throughout this chapter the topology, or fold, of a protein is defined by the first three numbers in the CATH classification database. Structures classified in the same topology in CATH share the same general spatial arrangement and connectivity of secondary structures. This can be illustrated by comparing the PDB structures 1b54 and 1aj0 (see figure 4.3), which are both contained in the triose phosphate isomerase (TIM) barrel fold in CATH, classification code 3.20.20. The first structure, 1b54, is a hypothetical protein found in Baker's yeast, *Saccharomyces cerevisiae*, which binds pyridoxal-5'-phosphate (vitamin B6 complex). It is classified in the alanine racemase superfamily in CATH with the classification code 3.20.20.10. The second structure, 1aj0, is a dihydropteroate (DHP) synthetase enzyme from *E. Coli*, and is a member of the DHP synthetase superfamily in CATH (classification code 3.20.20.20). Both these proteins are enzymes and both have highly similar folding

arrangements, however since there is currently insufficient evolutionary evidence to guarantee that they diverge from a common ancestor, they are classified into the same TIM-barrel topology, but different superfamilies, in CATH.

Since fold recognition aims to identify structural, rather than specifically evolutionary relationships, matching a relative in the correct topology is considered correct recognition. For clarity, it should be mentioned that the protein models that are used to search the structure database are referred to as query structures throughout this chapter.

Also, when assessing the fold recognition protocol, care was taken to ensure that any structural relationships in the structure database that could have been identified simply by sequence similarity were removed, unless explicitly stated otherwise. This was accomplished by removing any structures from the database search that had $\geq$35% sequence identity to the query structure.

4.2.2 Generating the Datasets

4.2.2.1 Summary of Datasets

The ability to recognise the native fold from non-native structures was tested by examining protein models derived from three sources, covering the most frequently used techniques in different *ab initio* methods.

- Low resolution versions of native structures provided by Xavier de la Cruz (de la Cruz *et al.*, 1997).
- *ab initio* predictions kindly provided by the David Baker group (Simons *et al.*, 1997).
- *ab initio* predictions by various methods from the CASP3 protein structure prediction competition.

4.2.2.2 Low Resolution Versions of Native Structures

The aim of this dataset was to provide a set of protein structures that would approximate models generated by *ab initio* methods. Due to the enormous size of conformational space that a polypeptide chain can possibly adopt, many *ab initio* methods attempt to limit this search by restricting the chain to certain states, such as restricting torsion angles to a given set of values or restricting the position of residues to the nearest points in a 3D lattice. Work by de la Cruz *et al.* (1997) suggested a protocol to build a range of low resolution protein structures from the

native experimental structures. Using this protocol, this dataset of approximate protein structures was generated and provided by Xavier de la Cruz.

Selecting the Dataset

The dataset of proteins that would be reconstructed as low resolution models were chosen based on two criteria. First, the experimental structures were required to be solved at high resolution (less than 2Å) to ensure that the results would not be affected by the quality of the native structures. Also, to reflect the type of proteins traditionally selected for *ab initio* methods, smaller structures were given preference over large structures, i.e. fewer than 250 residues. Table 4.1 provides a structural description for each of these proteins.

PDB code	Class	Architecture	Residues	Resolution (Å)
1bvc	α	non-bundle	153	1.5
1csu	α	non-bundle	108	1.8
1hcrA	α	non-bundle	52	1.8
2wrpR	α	non-bundle	104	1.7
4icb	α	non-bundle	75	1.6
1rbs	$\alpha\ \beta$	2-layer sandwich	155	1.8
1shaA	$\alpha\ \beta$	2-layer sandwich	103	1.5
2bopA	$\alpha\ \beta$	2-layer sandwich	83	1.7
121p	$\alpha\ \beta$	3-layer ($\alpha\ \beta\ \alpha$) sandwich	166	1.5
1aba	$\alpha\ \beta$	3-layer ($\alpha\ \beta\ \alpha$) sandwich	87	1.5
1hfc	$\alpha\ \beta$	3-layer ($\alpha\ \beta\ \alpha$) sandwich	157	1.6
1bgh	β	barrel	85	1.8
1cbs	β	barrel	137	1.8
1hviA	β	barrel	99	1.8
2tgi	β	ribbon	112	1.8
1aaj	β	sandwich	105	1.8
1flrL	β	sandwich	219	1.9
1fna	β	sandwich	91	1.8
1ppfI	β	single-sheet	56	1.8

Table 4.1: Description of the 19 structures in the dataset for low resolution models. The class and architecture descriptions are taken from the October, 1998 release of CATH.

The database composition for these 19 proteins is described in table 4.2. As an example, the query structure 1bvc has 34 proteins within the same sequence family (>35% sequence identity), 11 sequence families within the same homologous superfamily (based on more distant evolutionary relationships) and 20 homologous superfamilies within the same topology (based on structural but not evolutionary relationships).

	Database composition		
PDB code	**C.A.T.H.S**	**C.A.T.H**	**C.A.T**
1bvc	34	11	20
1csu	15	16	0
1hcrA	14	20	0
2wrpR	1	38	0
4icb	6	0	0
1rbs	6	2	0
1shaA	5	0	4
2bopA	5	44	1
121p	72	70	3
1aba	16	0	0
1hfc	6	0	3
1bgh	5	22	0
1cbs	10	0	11
1hviA	7	9	1
2tgi	6	0	0
1aaj	26	279	5
1flrL	56	89	6
1fna	11	288	3
1ppfI	12	0	2

Table 4.2: Database composition for the 19 structures in the dataset for low resolution models. The figures given are the number of relatives for a given query structure, for given levels in CATH. CATHS is the number of representatives in the same sequence family (clustered at >35% sequence identity), CATH is the number of relatives in the same homologous superfamily and CAT is the number of relatives in the same fold or topology.

Generating the Approximate Models

For each of these 19 proteins, eight simplified representations of the 3D structure were constructed using increasing degrees of complexity (S1 to S8, with S8 being the most complex). These simplified models were generated by restricting the geometry between Cα backbone atoms (defined by the θ_1, τ and θ_2 angles shown in figure 4.4) to a set of discrete states.

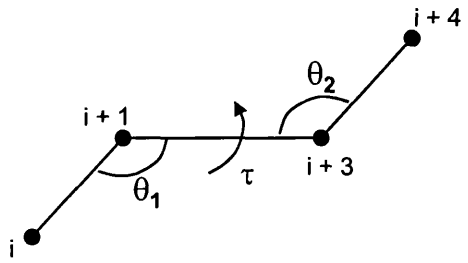


Figure 4.4: Definition of θ_1, τ and θ_2 angles involved in connecting four consecutive Cα atoms (i, i+1, i+2, i+3).

The allowed values for the angles in a given set were taken from analysis of the distribution of values observed in known structures (de la Cruz *et al.*, 1997). A simplified model of the distribution of the θ_1 and θ_2 angles can be seen in figure 4.5. This distribution is separated into four regions labelled A, B, C and D which are a result of the secondary structure preference of the residue fragments. Generally, region A encompasses all the angles for residues involved in α-helices, conversely region B contains angles for residues in β-strands. Region C represents angles from residues involved in the transition from α-helix to β-strand and region D represents angles from residues involved in the transition from β-strand to α-helix.

The most simple set of discrete states, S1, restricts every θ_1 and θ_2 angle to the central values of one of these four regions. However, each of these four regions has a unique distribution for the angle τ. Thus, the angle τ in the most simple set of discrete states, S1, is taken as the most highly populated value of τ for each of the four regions A, B, C and D. The other models, S2–S8, are generated by allowing different sets of allowed values of τ. Models S1, S2 and S3 use a similar number of discrete states, i.e. values of τ, however the regions occupied by these additional allowed values of τ (A-D) are varied. Models S4–S8 generally increase the number of allowed values of τ, which results in a greater degree of flexibility for the chain and allows a more accurate model of the native structure.

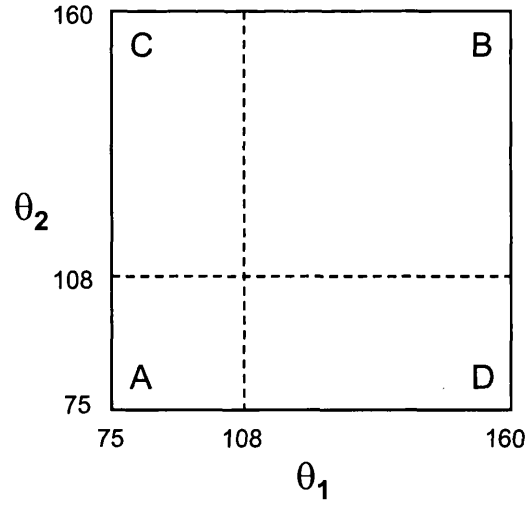


Figure 4.5: Simplified model of the distribution of θ_1 and θ_2 angles observed in known structures. The θ_1-τ-θ_2 points have been projected onto the θ_1-θ_2 plane for ease of visualisation. This plot is approximately symmetric about the $\theta_1 = \theta_2$ axis and consists of four main concentrations of points labelled A-D, separated by the lines $\theta_1 = \theta_2 = 108$.

The effect on the resolution for each of these approximate models is demonstrated in figure 4.6. This figure shows a plot of the RMSD with respect to the experimental structure for each of the different resolution models for the 19 proteins in the dataset. This plot is best illustrated by describing the different resolution models for a given structure, e.g. for 1aaj. As mentioned previously, this protein was reconstructed into eight models (S1–S8), each generated by restricting the allowed (θ_1, τ, θ_2) angles to a specific set of discrete states. Although the values of θ_1 and θ_2 are always restricted to one of four values, the discrete sets generally grow in complexity by allowing different values of τ for each of these four points, thereby allowing the original structure to be modelled more accurately.

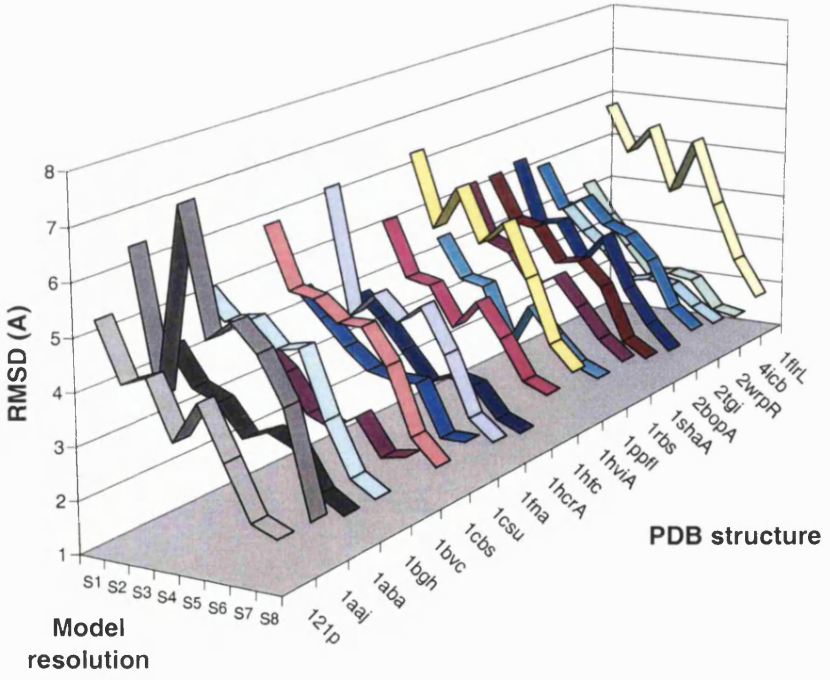


Figure 4.6: Comparison of RMSD from native for the dataset of 19 proteins. The RMSDs are given using eight different representations of the C_α chain (S1-S8, increasing resolution).

The resolution of these models was measured by comparing the model structure to the experimental structure and calculating the root mean square deviation, or RMSD. The RMSD values for the S1 to S8 models of the all-β protein, 1aaj, are 6.4, 3.7, 7.3, 4.9, 5.3, 4.8, 3.9 and 1.9 respectively. In this case the high RMSD for the S3 model is particularly interesting, especially when compared to the low RMSD for the S2 model. This is due to the fact that 1aaj is comprised of a high proportion of

β-strands and the extra values of τ for the S2 model are found in the θ_1, θ_2 position responsible for β-strands (region B in figure 4.5). This subsequently allows β-strands to be modelled with higher accuracy thereby giving a better overall structure and a lower RMSD with respect to the experimental structure. Conversely, the S3 discrete states allow more flexibility in modelling α-helices and as a result does nothing to improve the quality of the all-β structure.

4.2.2.3 Structures Predicted by Simons *et al.* (1997)

The second dataset of model structures was taken from the *ab initio* method presented by Simons *et al.* (1997) which proved the most successful *ab initio* prediction method in CASP3 (see section 4.2.2.4) (Orengo *et al.*, 1999). They simulate a model of protein folding which suggests that local amino-acid sequence restricts the conformational possibilities of local structure and non-local interactions preferentially stabilise the native conformation.

The modelling process begins with a sequence search of the closest 25 relatives in the PDB for each nine-residue fragment in the query protein. The conformation of each fragment is assigned from the nearest relative and these fragments are then spliced together using a simulated annealing procedure. The resulting structure is then evaluated with a knowledge-based scoring function to assess whether the model displays the structural features present in native protein structures (for example compactness, torsion angles, solvent accessibility). For each query protein, 100 models were generated using a variety of simulated annealing conditions.

The 100 predicted models of four structures (1ctf, 2cro, 2gb1 and 4icb) were classified into 1Å RMSD bins ranging from 10Å to 5Å with a random selection of 10 models placed in each bin.

4.2.2.4 Predicted Models from CASP3

The critical assessment of methods for protein structure prediction (CASP) is a community-wide event which provides an opportunity to assess the current structure prediction methods on a number of target sequences. These sequences are taken from structures that are either in the process of being solved experimentally or have already been solved but have not been published. This 'blind test' allows an unbiased comparison of the performance of the current state-of-the-art structure prediction methods. The results from these biannual experiments are published in a special supplement of the journal *Proteins: Structure, Function and Genetics* (Moult *et al.*, 1995, 1997, 1999; Zemla *et al.*, 2001) and the data for target structures and submitted

predictions are made available online (http://predictioncenter.llnl.gov/casp3/).

The query structures for this dataset were taken from the third CASP experiment (CASP3). In this assessment the targets were separated into three categories based on decreasing levels of similarity to known structures.

- **Comparative modelling**
  Targets have detectable sequence similarities to known structures.
- **Fold recognition**
  Targets have little or no detectable sequence similarity but do have structural similarity to known structures, thus can be recognised by threading based methods for example.
- ***Ab initio* prediction**
  Targets adopting a novel fold which therefore have neither sequence or structural similarity to known structures.

Since the method proposed in this chapter was to be assessed for fold recognition, it was necessary to use targets that could be assigned 'correct' answers for fold assignment, i.e. structures that belonged to fold groups already classified in CATH. Seven of the targets classified in the fold recognition category had predictions submitted by *ab initio* groups and many of these groups submitted more than one prediction for each target structure (table 4.3 provides a summary of these CASP3 targets). This final dataset of query structures comprised of predictions submitted by *ab initio* methods for these seven targets.

CASP3 target	Topology	Description
43	3.30.70	2 layer sandwich between a 4 stranded β-sheet and 2 helices
46	2.60.40	mainly-β, 2 layer sandwich
59	2.30.30	mainly-β roll
61	1.10.?	non-compact, α-orthogonal structure
63	2.40.10/2.40.50	Small β-roll
75	1.10.472	4 helix bundle architecture (helix packing is less regular than classical bundle structures)
77	3.30.?	2 layer, α β sandwich with one β α β motif and one split β α β motif

Table 4.3: Topology and description of the CASP3 Targets. Targets with a known architecture but unknown topology in CATH are marked 'x.x.?'.

Five *ab initio* groups submitted predictions for these models and many submitted the maximum of five structural models for each CASP3 target (see table 4.4). A brief description of the methods used by each of these groups for these CASP predictions is given in table 4.5

	CASP3 target structure						
	43	**46**	**59**	**61**	**63**	**75**	**77**
Baker	NR	RT	NR	NR	NR	NR	NR
<RMSD>	16.3	15.8	10.7	10.1	15.8	14.0	12.6
Models	5	5	5	5	5	5	5
Osguthorpe	-	NR	NR	NR	-	NR	NR
<RMSD>	-	17.2	14.6	13.6	-	15.2	13.6
Models	-	2	3	2	-	2	2
Samudrala	RT	NR	NR	RA	NR	RA	-
<RMSD>	16.0	15.5	12.7	10.2	15.8	11.9	-
Models	5	5	5	5	5	5	-
Scheraga	-	-	-	RA	-	-	-
<RMSD>	-	-	-	8.6	-	-	-
Models	-	-	-	4	-	-	-
SkolOrtKol	-	-	RA	-	-	-	NR
<RMSD>	-	-	11.5	-	14.2	14.1	8.6
Models	-	-	3	-	4	4	5

Table 4.4: CASP3 *ab initio* predictions for seven of the fold recognition targets. The RMSD values are the average RMSD over all the models. The fold recognition for each target is classified as follows: NR, no recognition; RT, protein topology recognition; RA, protein architecture was recognised; -, no prediction was provided for the target.

Group	Participants	Method Description
Baker	Baker, Bystroff, Ruczinski, Bonneau, Simons	Structure prediction using simulated annealing of structural fragments (see section 4.2.2.3).
Osguthorpe	Osguthorpe	Simplified flexible geometry model for protein folding using reduced representation model and simplified force field (Osguthorpe, 1999).
Samudrala	Samudrala, Xia, Huang, Levitt	Combined method for low-resolution *ab initio* prediction. The procedure starts with a simple lattice model then conformations are built with increasing detail using empirical energy functions with increasing complexity. Low energy conformations are examined using contact energy function and all-atom models are generated using predicted secondary structure. The final predictions are generated using a consensus distance geometry procedure (Samudrala *et al.*, 1999).
Scheraga	Lee, Liwo, Ripoll, Pillardy, Scheraga	Hierarchical and classical physics-based *ab initio* approach using conformational space annealing and electrostatically driven Monte Carlo methods (Lee *et al.*, 1999).
SkolOrtKol	Skolnick, Ortiz, Kolinski	*Ab initio* folding using restraints derived from multiple sequence alignments (Ortiz *et al.*, 1999).

Table 4.5: Brief descriptions of the *ab initio* prediction methods in CASP3 that generated the models used in this chapter.

4.2.3 Consensus Fold Recognition Protocol

As mentioned in section 4.1.2.4, many *ab initio* methods work in a two-stage process. The first stage is to generate a large number of structures for the query sequence and the second is to assess which of these structures is the most likely. Obviously, the aim of this method is to provide a single, accurate prediction of the native structure for the query sequence. However, in reality this procedure often produces a series of predicted structures that have similar protein-like qualities (e.g. measured by compact, globular core, secondary structure content, solvent accessibility, favourable residue pair contacts). A simple procedure of just selecting the first structure in the list and ignoring the other predictions could easily result in the correct fold being missed. With this in mind, a consensus approach was taken to the fold recognition of these models.

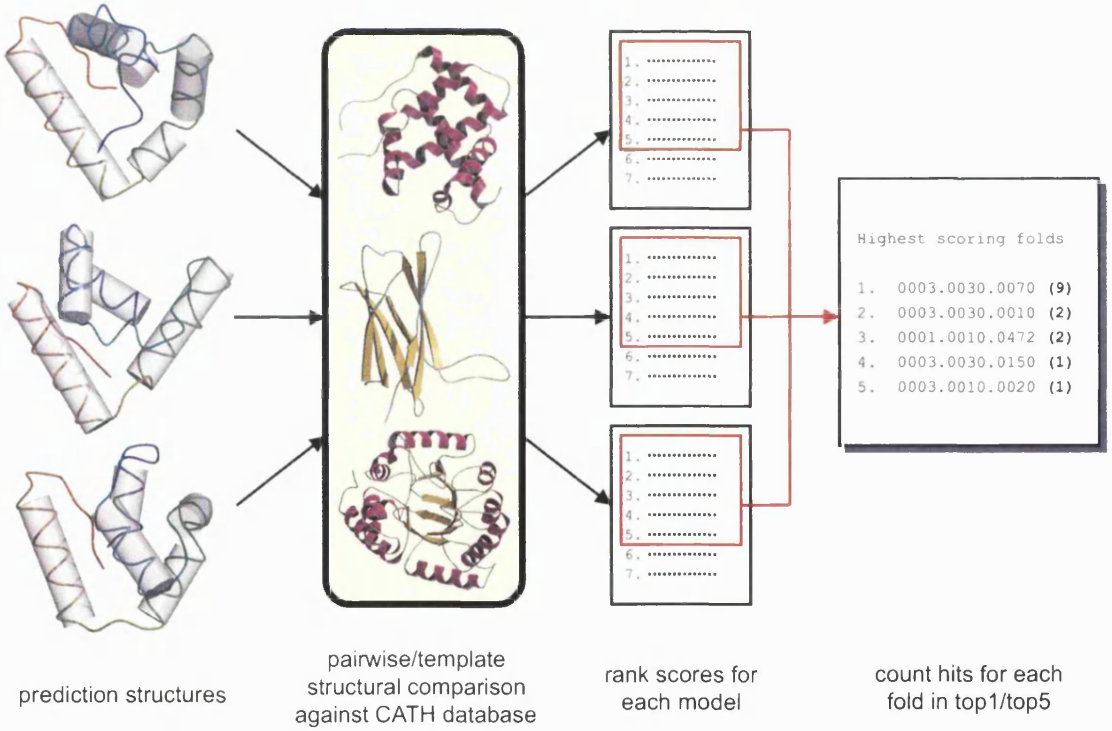


Figure 4.7: A flowchart demonstrating the consensus fold recognition protocol. A set of predicted query structures are searched against the structure database then the folds of the highest scoring database matches are accumulated to give an overall fold assignment.

This procedure begins by performing a structure-based search against a structural library for each model in a set of predictions and the results of these searches are all ranked by decreasing structural similarity. The topologies observed in the high scoring matches for all these searches are then accumulated and the most frequently noted topology assigned as the most likely fold (see figure 4.7).

Two different approaches were taken to provide the structural similarity scores. The first was to search a library of representative structures in a pairwise manner using the SSAP structure comparison algorithm (Orengo & Taylor, 1990). The second approach was to use the CONALIGN protocol based on structural templates generated from multiple structure alignments (described in more detail in section 4.2.3.2 and chapter 2).

4.2.3.1 Pairwise Comparison

As mentioned above, the double dynamic programming algorithm SSAP (Taylor & Orengo, 1989) was used for the pairwise structure comparisons (see section 1.2.7.3). The first step of this algorithm is to use dynamic programming to compare the intramolecular C_β vectors, i.e. structural environments, of all potentially similar residues between the two structures. The result of each pairwise residue comparison is a path describing the best alignment of the structures based on the structural environments as viewed by each residue. If the residue environments are sufficiently similar (i.e. the residue-level alignment path scores above a given threshold), then the scores for the alignment path are accumulated in a summary matrix. A second pass of dynamic programming is applied to provide the optimal path through this summary matrix and therefore the optimal alignment between the two structures. This final pass also provides a normalised similarity score between 0–100 with identical structures returning a score of 100.

The structure database used for these pairwise comparisons was taken from the CATH classification (October 1998 release). A non-redundant list of structural domains was taken for the database (N-level) giving a searchable library of 2,819 structures. A typical pairwise database search of this size took around 11 hours on a MIPS10000 processor for a query structure of around 200 residues.

4.2.3.2 Template Comparison

The CORA algorithm (Orengo, 1999) was used to generate a structural template which describes highly conserved structural features within a series of related proteins (see chapter 3 for more details). These structural templates can be used as iden-

tifying fingerprints to represent homologous superfamilies in the CATH database. Instead of performing pairwise comparison against a non-redundant library containing around 3,000 structures, the templates allow the number of comparisons to be reduced to the number of superfamilies in CATH that contain more than one structure (362 for this version of CATH). This method does present potential problems when considering superfamilies that contain either a single structure or a few very similar structures, as discussed in chapter 3, however this will prove less problematic as the structure databases continue to populate.

The search procedure starts by aligning the query structure to each of the templates using the double dynamic algorithm CORALIGN (Orengo, 1999). A structural similarity score was then generated for this alignment by comparing the contacts seen in the query structure with highly conserved contacts seen in the structural template (see chapter 2 for more details).

Figure 4.8 illustrates the use of comparison contact maps to assess structural similarity based on the alignment between a query structure and a multiple structure template. The comparison contact map in figure 4.8a shows the native structure, 2gb1, aligned against the multiple structure template from the native superfamily (immunoglobulin-binding domain, CATH code 3.30.70.330). In this figure, contacts in the query structure are coloured black and consensus contacts in the structural template are coloured grey with overlapping contacts shown in red. The overall score for the comparison is given by the number of overlapping contacts divided by the larger number of contacts between the homologous family template and model.

A similar comparison is also shown between one of the predicted structures for 2gb1 and the structural template for the native superfamily (see figure 4.8b). Despite this predicted structure displaying fewer overlapping contacts (seen in red) than the experimental structure, the native structural template still proved the highest scoring template.

For cases where the native family had a representative in the template database, care was taken to avoid generating this template with any structure that shared more than 25% sequence identity with the query protein. The only exception was for the structural template of the repressor-like DNA-binding domain, 2cro. This contained a structure with 52% sequence identity to 2cro, necessarily included to make the minimum number of structures for the structural template. Instead of removing this example from the dataset, it was used as an error checking exercise. If the proposed method was not able to correctly recognise predicted structures using a structural template containing a related structure then it would not be likely to recognise more distant structures.

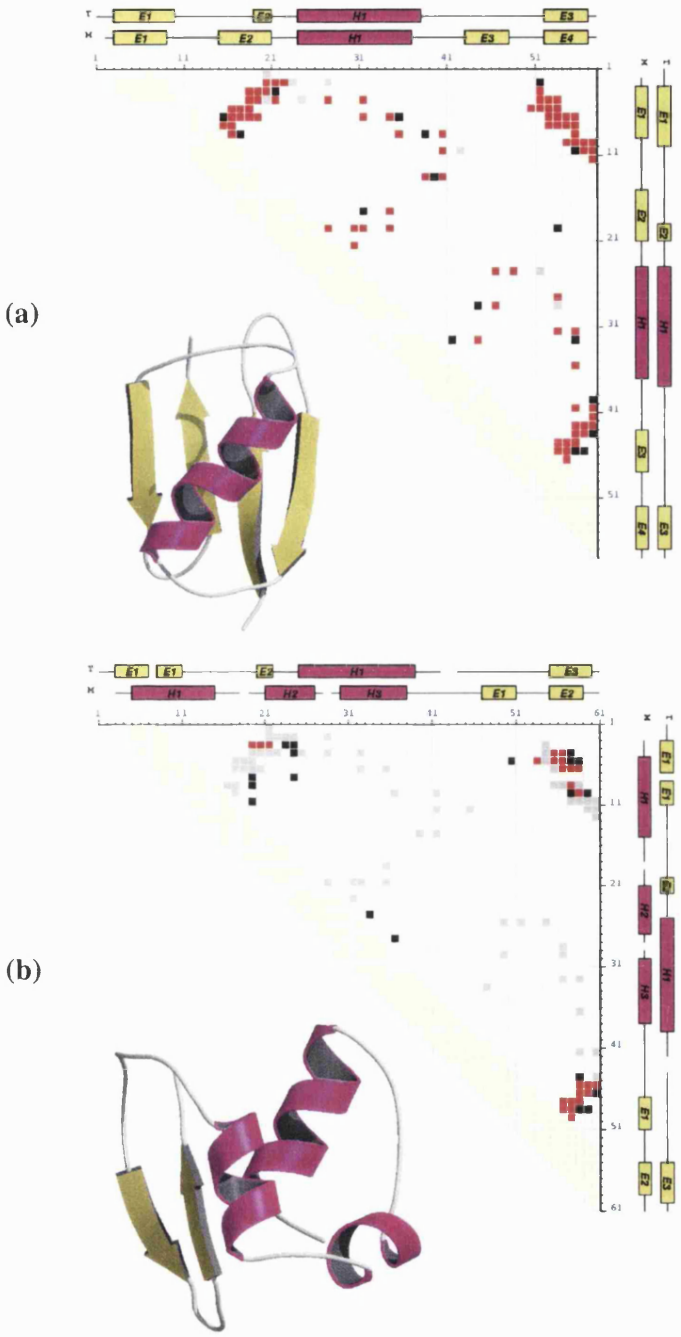


Figure 4.8: Comparison contact maps for native structure and a predicted model. These show similarities between a single structure (coloured black) and consensus contacts in the homologous superfamily template (coloured grey) with overlapping contacts shown in red. The overall score for the comparison is given by the number of overlapping contacts divided by the larger number of contacts between the homologous family template and model. A schematic representation of the alignment between homologous family template (T) and model (M) is shown along both axes. (a) Comparison contact map for the native structure of 2gb1 aligned to its homologous family template in CATH. (b) Comparison contact map for a predicted model of 2gb1 (6-7 Å RMSD bin) aligned to the correct homologous family template, which was the highest scoring structure template.

4.3 Results

4.3.1 Overview of Results

The objective of this work was to assess whether low-resolution predictions of a target protein could be used to identify a structural relative that could provide the correct identification of the target topology. To test this approach, two scenarios were investigated.

1. The query structures are low-resolution versions of the experimental structure (as described in section 4.2.2.2).

2. The query structures are taken from two separate sets of *ab initio* predictions.

    - Predicted structures are based on four native proteins taken from the Simons *et al.* (1997) method (see section 4.2.2.3) spanning a resolution range of 5–10Å RMSD.
    - Predicted structures submitted to the CASP3 experiment using a variety of *ab initio* methods. These predictions ranged from low to very low resolution structures (see section 4.2.2.4).

The first dataset was examined using SSAP to compare the models against the structural library of 2,819 non-identical protein structures. The second dataset proved a more thorough test of the fold recognition capabilities of the structural library. Therefore the library of structural templates was used in addition to the pairwise SSAP scores for these distant models.

4.3.2 Fold Recognition Using Low Resolution Versions of Native Structures

Database searches were carried out using SSAP in a pairwise manner for the sets of eight low resolution models for each of the 19 proteins. Each comparison from the structure database was either classed as related or non-related depending on whether the match belonged to the same topology in CATH (i.e. shared the first three classification levels) to the query structure. Figure 4.9 shows an example for the distributions using the native structure (121p) compared to the eight low resolution models.

It can be seen that the distributions between native structure and low resolution models are different. Searching the database with low resolution models gives lower

SSAP scores for both relatives and non-relatives than with the native structure. For example, the mean SSAP scores for the native structure is 67.5 and 48.2 for relatives and non-relatives respectively whereas the mean SSAP scores of relatives and non-relatives for the reduced model S4 is 43.1 and 39.8. From this it can also been seen that there is also less differentiation between the distributions of scores for related and non-related matches when using the reduced model structures rather than the native structures. However, it is still possible to assign the correct fold for most of the low resolution models as some relatives still rank either in the first or in the top five positions.

The percentage of correct database matches in the top ranking position and the top five ranking positions was calculated across the different model resolutions (S1–S8) for all 19 proteins (see figure 4.10). The results for the fold recognition performance for all relatives, including close sequence relatives, can be seen as the white bar. The results for all these relatives have then been broken down into subsets, showing the fold recognition performance based on the relationship between the query structure and database match. These subsets are; all structural relatives (homologues and analogues but not close sequence relatives), homologues only and analogues only.

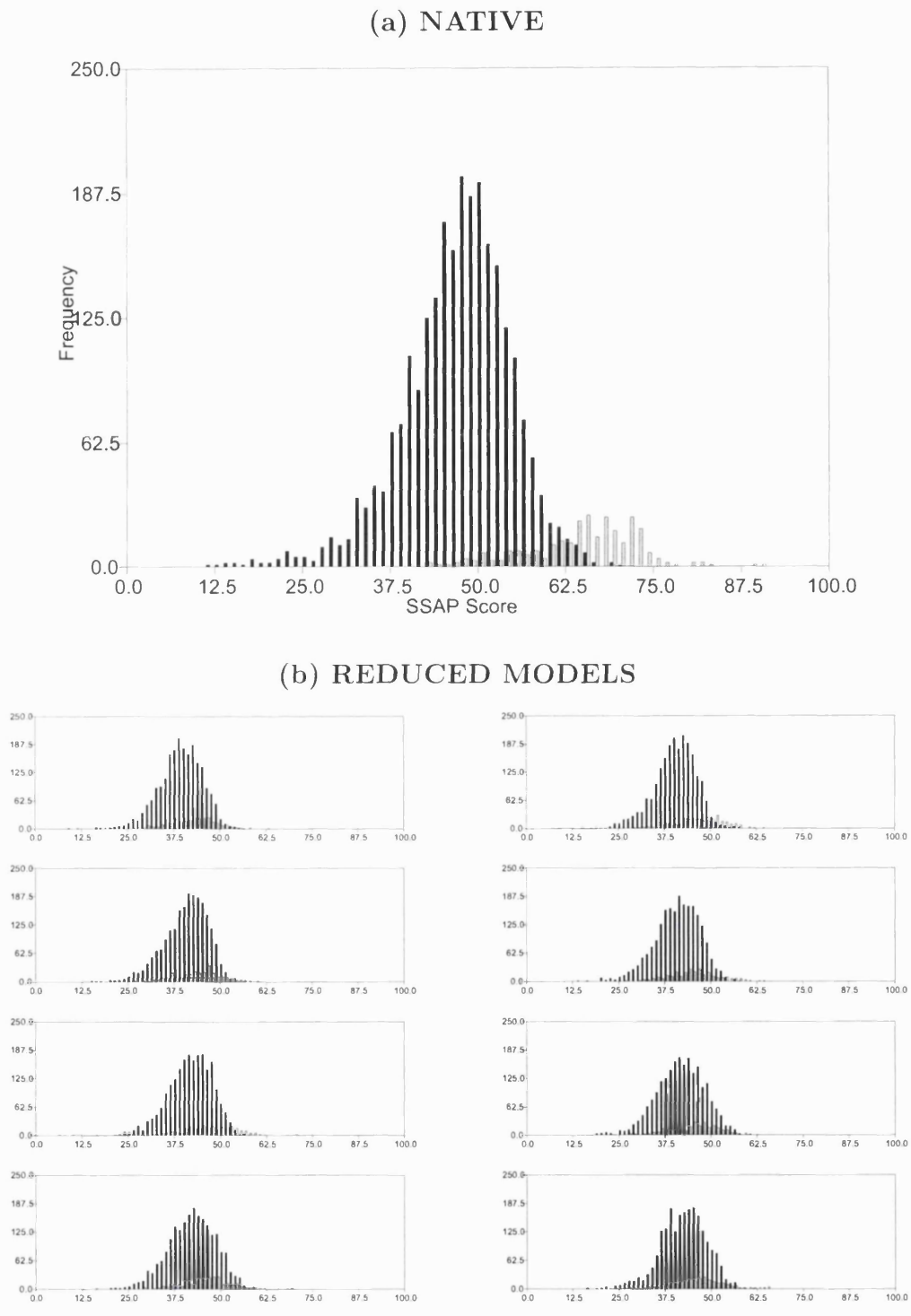
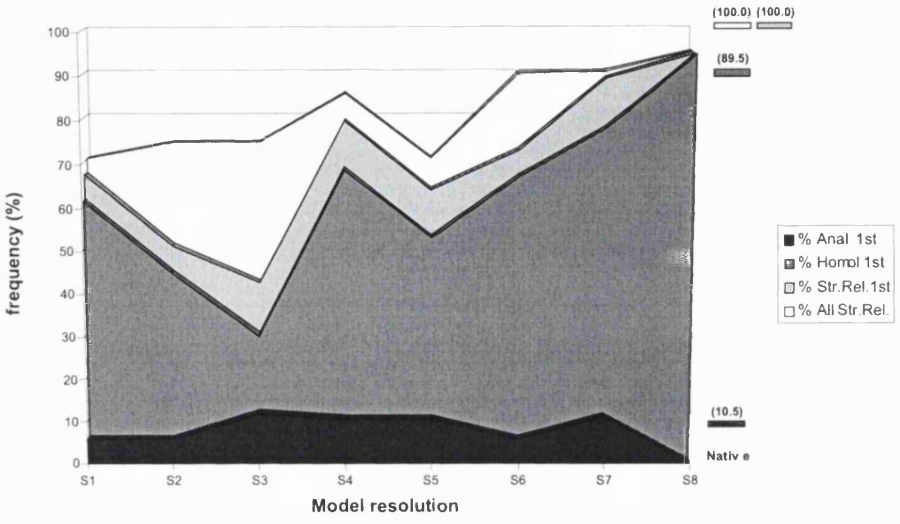


Figure 4.9: Distributions of pairwise structural comparison (SSAP) scores for the native (a) and reduced models (b) (S1-S8, increasing resolution from left to right and top to bottom) of 121p against the CATH database of 2,819 non-identical structures (October 1998 release). Structures related at the toplogy level are shown in light grey, non-related structures are shown in black.

(a) TOP RANKING POSITION



(b) TOP FIVE RANKING POSITIONS

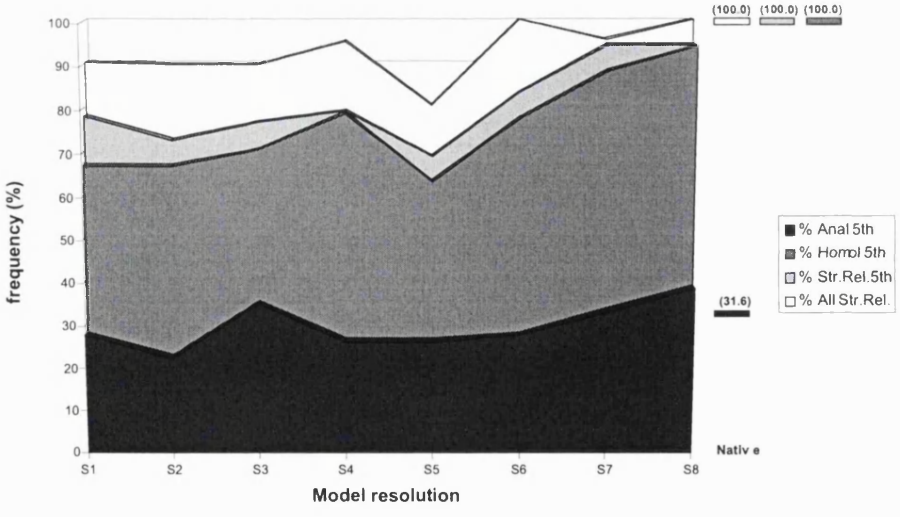


Figure 4.10: Recognition rates for pairwise structural comparisons of reduced models (S1-S8, increasing resolution) based on structures ranked in first place (a) and in the top five places (b) for all 19 models. The frequency of analogues, i.e. those structures sharing the same fold as the query structure but not the same homologous superfamily, are shown in dark grey, homologous structures are shown in medium grey, structural relatives (homologues and analogues) in light grey and all structural relatives (including closer sequence relatives, homologues and analogues) in white. For comparison the recognition rates when using the native structure are also highlighted to the right of the graph.

When using the chain representations S1, S2 and S3 (4–5Å RMSD from native), the success rate is approximately 40-60% for finding a correct topology match in the top ranking position (figure 4.10a). As the structure representations become more accurate, using the S4, S5 and S6 models (3–4Å RMSD), the recognition rate increases to 65–85%. With the highest quality models, S7 and S8 (1.5–2.5Å RMSD), the recognition again increases to 90–95%. If the recognition rate is calculated on finding the correct topology within the top five ranking positions (figure 4.10b) then this score increases for almost all of the chain representations. This is demonstrated by the recognition rate of around 70% for even the very lowest resolution models.

When the results for the database searches were analysed it could be seen that some analogues of the native structure matched with a higher similarity score than any homologues (summarised in table 4.6). This was a surprising result since homologues, almost by definition, should be more structurally similar than analogues. It is likely that this was either an accidental result, i.e. the low resolution chain representation by chance happened to be rebuilt in a conformation closer to analogues, or due to the fact that some topologies have greater structural similarity between homologous superfamilies than others.

Target[a]	**Model**[b]	**RMSD**[c]	**Match**[d]
1fna	S1	6.15 (5.1 ± 0.8)	1faiL
2bopA	S2	3.96 (4.1 ± 0.6)	1psdA
1hcrA	S3	2.80 (4.2 ± 1.1)	1mbg
2bopA	S3	4.11 (4.2 ± 1.1)	1sphA
121p	S4	3.32 (3.4 ± 0.8)	1ordA
2bopA	S4	3.30 (3.4 ± 0.8)	1mla
1rbs	S5	3.54 (3.7 ± 0.9)	1hpm
2bopA	S5	3.89 (3.7 ± 0.9)	1dcoA
2bopA	S6	2.77 (3.1 ± 0.6)	1mla
121p	S7	2.08 (2.2 ± 0.5)	1rcf
2wrpR	S7	2.01 (2.2 ± 0.5)	1fipA

Table 4.6: Cases where an analog of the query protein ranked in the top position.
[a]PDB code of the target protein
[b]residue model used to build the query structure
[c]RMSD between the experimental and query structures of the target protein. The query structure is an approximate version of the experimental structure, rebuilt using the residue model mentioned in the second column (see Methods). The average and standard deviation values for the 19 proteins in the test are shown in parentheses.
[d]analogue recovered after the database query.

As expected, the results seen in figure 4.10 demonstrate that including the close sequence relatives of the query structure improves the fold recognition performance, especially for the low resolution models. This suggests that database composition, i.e. number of structural relatives in the database, would have an important impact on the final performance of this method. To investigate the effect of database composition, the set of 19 proteins was separated into 10 proteins that had more than 20 relatives and 9 proteins that had less than 20 relatives in the database. The performance between these two subsets was again compared by plotting the frequency of a structural relative occurring in the top position of the database search for each resolution bin (see figure 4.11). From this graph it can be seen that this fold recognition method produced higher performances for query proteins that contained more than 20 relatives in the database. Indeed, the correct fold was identified for all query proteins with more than 20 relatives even at the lowest resolution bin (S1).

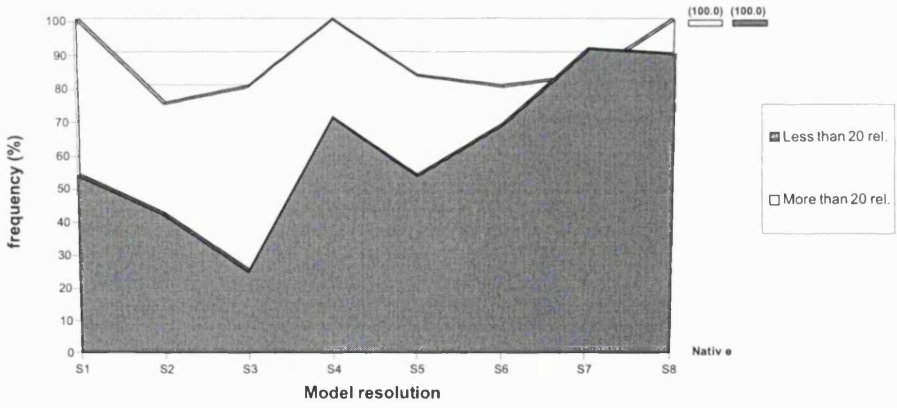


Figure 4.11: Effect of database composition on fold recognition rates. Recognition rates for the 9 proteins with more than 20 relatives in the database are shown in white and recognition rates for the 10 proteins with less than 20 relatives are shown in grey.

4.3.3 Fold Recognition Using Models from *Ab initio* Structure Prediction

4.3.3.1 Overview of fold recognition results from *ab initio* models

This section discusses the results of the consensus fold recognition approach when applied to real *ab initio* structure predictions based on four native proteins; 1ctf, 2cro, 2gb1 and 4icb (Simons *et al.*, 1997) (see section 4.2 for more details). The predictions for each protein were classified into 1Å RMSD resolution bins in the 5–10Å range with 10 predicted structures selected for each bin. These predicted structures were then searched against the structure database using two different structure comparison methods.

- Conventional pairwise structure comparison using the SSAP algorithm
- Comparison of structural templates derived from multiple structure alignments

A running total of the topologies occurring in the first position and top five positions of the database search was kept for each resolution bin. The performance of each method was then assessed by stating the position of the correct topology after ranking each topology in this running total, i.e. position 1 means the native topology was correctly identified. The results for both structure comparison approaches have have been summarised in table 4.7.

4.3.3.2 Pairwise Comparisons

The structures predicted for the all-α protein, 4icb, gave the correct fold as the most frequently occurring topology for every resolution bin (see table 4.7a). The structures for the other all-α protein, 2cro, also gave good results as the correct fold was recognised in four out of the five resolution bins when assessing the top ranking topologies. When using the top five ranking topologies, the native fold gave consistently high results but only two of the resolution bins were able to provide correct identification.

The α-β proteins, 1ctf and 2gb1, gave much weaker results using just the top ranking topology as the native fold was not the highest ranking topology in any resolution bin. However, using the topologies seen in the top five positions helped the performance by recognising the correct fold in highest resolution bin for 1ctf. Unfortunately, no predictions were available for the 2cro protein within the 5–6Å resolution.

A. Pairwise comparison

	Top ranking candidates					Top five ranking candidates				
RMSD	5–6	6–7	7–8	8–9	9–10	5–6	6–7	7–8	8–9	9–10
1ctf	2	-	-	3	-	1	5	-	-	5
2gb1	*	-	-	-	-	*	3	-	-	-
2cro	1	1	1	-	1	1	2	1	2	3
4icb	1	1	1	1	1	1	1	1	1	1

B. Template comparison

	Top ranking candidates					Top five ranking candidates				
RMSD	5–6	6–7	7–8	8–9	9–10	5–6	6–7	7–8	8–9	9–10
1ctf	1	1	1	1	1	1	1	1	1	1
2gb1	*	3	2	2	1	*	1	4	5	2
2cro	1	1	3	-	3	1	1	-	-	2
4icb	1	1	2	1	-	1	1	2	-	-

Table 4.7: Results when querying the structure databases with *ab initio* predictions. This table assesses the performance of the method to identify the native topology of a target protein with decreasing quality of the structure predictions used to query the database. The numbers shown in the table designate the position of the correct fold based on the consensus fold recognition protocol for each RMSD resolution bin
*No prediction models were available within this resolution range.

4.3.3.3 Structural Template Comparisons

The structural templates capture the conserved structural features of a group of related proteins. Therefore they describe the structural variability of each position in the multiple structure alignment rather than treating variable and conserved features equally, which is inevitable in the pairwise method. Thus, it was hoped that the template comparison would help push the fold recognition to even lower resolutions. From the results seen in table 4.7b, this structure comparison method appeared to provide a more sensitive probe of all the predicted structures.

When considering the top five ranking folds, the native fold was correctly identified in all four proteins across the 5–7Å range. With only one exception, this result was also seen when using just the top ranking topology. Unlike the pairwise comparison method, the correct fold was recognised for 1ctf at all resolutions.

4.3.4 Fold Recognition Using *Ab initio* Structure Predictions From CASP3

The performance of the consensus fold recognition protocol was also assessed using different *ab initio* protocols based on predictions submitted to the CASP3 experiment (see methods section 4.2). The submitted predictions were generally of a lower quality than the models in the other two datasets used in this chapter, having an RMSD from experimental structure in the range 8.6–16.3Å. Since each group submitted more than one prediction for the targets, the consensus approach was based on the fold recognition for a given target using the set of models submitted from a given group. The results are summarised in table 4.8.

	CASP3 Target						
Fold recognition method	43	46	59	61	63	75	77
Consensus	RT	RT	NR	RA	NR	RA	NR
Threading	NR	RT	NR	NF	RT	RA	NR

Table 4.8: Comparison of the consensus fold recognition protocol to established threading methods using CASP3 targets. The consensus method results are taken from the method discussed in this chapter and the results from threading submissions to CASP3. For both methods, the results shown are based on the most successful group submissions for each target. The results are described as: NR, No recognition; RT, Protein topology recognition; RA, Protein architecture was recognised; NF, New fold predicted.

As expected, the recognition rates for these targets were not good due to the low resolution of the predicted structures. The correct fold was assigned in only two of the cases (targets 43 and 46) and the correct architecture was identified in two of the remaining targets (targets 61 and 75). However, these correct fold and architecture identifications were taken from a range of different methods rather than one *ab initio* method consistently producing accurate predictions. These results confirm the previous findings that the fold recognition performance of this method decreases considerably as the resolution moves beyond 8–9Å RMSD.

These results were also compared with the results from the threading predictions, with the most successful threading results and most successful consensus fold recognition results taken for each target. The overall performance for these two approaches are similar with the best threading methods also managing to identify the correct topology for only two of the seven targets (46 and 63), with only one of the remaining targets having the correct architecture assignment (target 75). Target 61 could not be assigned a clear fold so the prediction of the novel fold given by the Jones group should also be seen as a successful result.

4.4 Discussion

The work presented in this chapter describes a novel application of structure comparison methods for fold recognition of low resolution protein structures. This is intended to be the first step in an optimisation procedure that could produce better quality predictions of protein structures directly from amino acid sequence. After the correct fold has been identified, it should then be possible to refine the predictions by constraining the structures with the highly conserved structural features observed in the native fold.

This fold recognition protocol took groups of low resolution protein structures and searched them against the structural database. A consensus approach was then employed to identify the most commonly occurring fold from the results of the structural comparisons. The method was tested using low resolution approximations of native protein structures and low resolution structural models predicted from *ab initio* methods, all covering a wide range of RMSD from the experimental structure.

Results indicate that structure comparison methods can be used to correctly identify the native fold of low resolution protein structures for RMSD values of $\leq$7Å from experimental structure. The fold recognition at even lower resolutions, i.e. higher RMSD, is certainly possible but far less consistent. However, the success of some of these very low resolution cases suggests that this approach can be applied to improve protein structures derived from high throughput experimental techniques such as cryo-electron microscopy.

Using RMSD may not always provide the most accurate measure of structural similarity, especially when dealing with more distant relationships. However, when the optimal structural alignment can be guaranteed, as seen in this case with a one-to-one alignment between predicted and experimental structures, RMSD can be viewed as a useful measure to compare the similarity of related structures. Having said this, RMSD will not describe differences between two structures for any structural feature other than the 3D atomic co-ordinates. For example, a large number of the *ab initio* predictions displayed little, if any, agreement of secondary structure assignment when compared to the experimental structure. These discrepancies could have resulted in incorrect fold assignment for manual assessment and some automatic protocols of fold recognition. However, a strength of the proposed fold recognition method was that it was unaffected by secondary structure assignment. This was illustrated in the correct fold recognition for all resolution bins of the mixed-$\alpha\beta$protein 1ctf even though none of the *ab initio* predicted structures contained β-strands. Similarly, only two predictions of 2gb1 in the 6–7Å RMSD range

predicted at best a two stranded β-sheet compared to the four stranded β-sheet of the experimental structure (see figure 4.8). Despite these apparent differences in structure, the native fold was still correctly identified for this resolution bin. The latter example may be of particular importance since the prediction of β-sheets has proved one of the more challenging aspects in the field of *ab initio* prediction field (Dandekar & Argos, 1996; Eyrich *et al.*, 1999; Simons *et al.*, 1999).

It is perhaps inevitable that parallels will be drawn between this method and the more traditional fold recognition methods such as threading. After all, threading has been optimised to the point where it has been been applied to entire genomes (Jones, 1999a). However for all its strengths, it is estimated that threading can only recognise the correct fold for a given sequence 40–60% of the time (Jones *et al.*, 1999). The results seen in table 4.8 suggest that since the proposed protocol and threading methods were successful for different CASP3 targets, the two approaches could be used to complement each other for increased fold recognition.

Since the fold recognition procedure involves searching a database of known structures, it is unsurprising that the performance has a dependency on the degree of representation of the target protein for both the pairwise and structural template comparisons. This was illustrated in figure 4.11 by the increase in performance of the fold recognition for query structures with more than 20 structural relatives in the database (relatives with high sequence identity were not included in this search). In addition to this dependence on the composition within the structure database, this method can only hope to recognise the fold of *ab initio* predictions if they belong to folds that already exist in the database. However, this method may even prove useful for sequences adopting novel folds as the results seen in table 4.8 suggest that the method may be used for assignment of architecture. Knowledge of such architectural assignments could help to guide the refinement stage by imposing constraints derived from analysis of secondary structure packing.

The proposed method, along with threading methods, will benefit enormously from the structural genomics projects effectively 'filling in the gaps' of the protein folding universe. Until the structure databases are sufficiently populated to place all known protein sequence within homology modelling distance of a related structure, the proposed method provides a novel and useful approach to fold recognition that complements established threading methods.

Chapter 5

Derivation of Structure-based Sequence Models to Detect Remote Evolutionary Relationships

5.1 Introduction

5.1.1 Background

The rate at which new protein sequences are being uncovered from genomics initiatives far outweighs the rate at which any structural information can be gathered. In cases where only the amino acid sequence is known, the first step towards understanding the biological role of a novel protein usually begins by examining the sequence databases in an attempt to identify relationships to known proteins. Furthermore, identifying an evolutionary relationship between the sequence of the novel protein and the sequence of a well characterised structure in the database often allows structural and possibly functional information to be inferred.

During the process of evolution, protein sequence can diverge beyond all recognition yet structure is often highly conserved due to structural and functional constraints. As a result, methods comparing proteins based on structural rather than sequence features are often able to identify more distant relationships. Also since there is so much more information when considering 3D structure rather than sequence, structural alignments between evolutionary distant proteins often prove more accurate than sequence alignments. However, these detailed structural comparisons are

usually far more computationally expensive than sequence comparisons and the protein sequence is far easier to obtain than the protein structure. Therefore, it is still highly desirable to investigate techniques that push the limits of remote homologue detection using sequence comparison methods.

5.1.2 Pairwise Sequence Alignment

5.1.2.1 Coping with Insertions and Deletions

Evolutionary relationships can be identified most simply by aligning sequences in a pairwise manner then scoring the resulting alignments. Aligning the sequences of highly similar proteins is trivial, however more distantly related proteins can have insertions and deletions (indels) in addition to single mutations of the amino acid sequence. A more sophisticated algorithm is required to be able to cope with indels and provide a reliable alignment in these cases.

5.1.2.2 Rigorous Alignment Algorithms

An example of an algorithm that accounts for indels is the dynamic programming algorithm. This provides an optimal global alignment between two sets of data and is discussed at length in section 1.2.4.2. This algorithm was first applied to sequence comparison by Needleman and Wunsch (Needleman & Wunsch, 1970). A further modification of this algorithm was introduced by Smith and Waterman (Smith & Waterman, 1981) which focused on providing local, rather than global, alignments. This implementation of the dynamic programming algorithm explores all possible alignment paths then identifies and aligns similar fragments between the two sequences rather than attempting to provide a single global alignment.

5.1.2.3 FASTA

The FASTA algorithm (Pearson & Lipman, 1988) employs a simple and fast approach to sequence alignment by initially searching the two sequences for small segments having n identical residues (known as n-tuple fragments). A hash table (a general computer programming technique) is used as an efficient means of storing and searching all the n-tuple fragments from a large database of sequences. This provides a rapid method of screening a large number of sequences for likely matches. Once the putative matches have been identified, the chains of aligned segments are then entered into a two-dimensional matrix as before and a more thorough dynamic

programming algorithm is used to string the segments together for a global alignment.

5.1.2.4 BLAST

The BLAST (Basic Local Alignment Search Tool, Altschul *et al.* (1990)) presents an alternative approach for the detection of distant but biologically sensitive relationships. This algorithm begins by separating the protein sequence into tripeptide fragments, e.g. ACE. The resulting list of tripeptide fragments is then expanded to include a series of closely related fragments of similar length, e.g. ACE is expanded to ACE, GCE, GME, AME. These fragments are identified by scoring all permutations of triplets with the BLOSUM substitution matrix (see section 1.2.3.2) and incorporating only the tripeptides scoring over a given threshold value.

The query sequence is then searched against the sequence database to identify fragments matching identically to the expanded list of tripeptide fragments. Having identified a database sequence with a matching tripeptide fragment, this matching fragment is treated as a 'seed' that is extended in both directions along the sequence in order to identify the highest scoring segment pairs (HSP). The segment pair with the highest score is called the maximum segment pair (MSP) and represents the highest scoring sequence match in the database. The overall scores assigned to the resulting sequence matches are based on the probability that an equivalent matching fragment could have emerged by chance.

In order to identify more distant evolutionary relationships, various scoring methods have been tested, for example using chemicophysical properties as an added comparison criterion, rather than just specific residue identity. This is based on the assumption that it is the chemicophysical role of the amino acid in the structure that is conserved rather than the specific amino acid identity, so matching properties rather than identities of amino acids provides a more sensitive probe for more distant evolutionary relationships.

5.1.3 Profile-based Sequence Comparison

5.1.3.1 Background

One approach to improve the performance of these sequence comparison methods is to identify features that are conserved during the process of evolution by examining multiple sequence alignments of related protein sequences. The advantage of using this approach is that the variation of observed amino acids can then be modelled

for each position in the alignment in a sequence 'profile'. A profile can assign significance to each alignment position based on the degree of conservation at that position, whereas a simple pairwise sequence comparison gives all positions in the alignment equal weighting. Emphasising the importance of highly conserved regions and reducing the importance of poorly conserved regions during the search procedure allows more accurate alignments and provides more discriminating scoring schemes (Barton & Sternberg, 1987; Taylor, 1987; Rice & Eisenberg, 1997; Park *et al.*, 1998; Kelley *et al.*, 2000).

A profile can be formally defined as a consensus primary structure model consisting of position-specific information (Eddy, 1996). Several methods have been developed to generate sequence profiles and use them to identify distantly related sequences (Taylor, 1986b; Gribskov *et al.*, 1987; Barton & Sternberg, 1990). These sequence profiles effectively reflect the likelihood of finding a given amino acid or a gap at a specific position in the alignment. In the method proposed by (Gribskov *et al.*, 1987) these profiles are generated by summing Dayhoff exchange matrix values (Dayhoff, 1978) for every position in the sequence alignment. To model insertions and deletions, the penalty for introducing a gap in the alignment is reduced for the positions in the sequence model containing large numbers of gaps.

5.1.3.2 Hidden Markov Models

Sequence profiles can be implemented using a statistical modelling technique known as a hidden Markov model (HMM) which allow sequences to be aligned against the model in a probabilistic manner. To use an analogy, HMMs can be considered as sequence generating factories capable of producing many different sequences with different probabilities. Internally, the HMM works by representing each column in the multiple sequence alignment by three states; match, delete and insert (see figure 5.1). The match state models the distribution of residues allowed at a specific column of the sequence alignment, the delete state models having no residue at this column and the insert state models an insertion of one or more residues after this column. These states are connected by state-transition probabilities and a sequence of states is generated by moving from the start to the end point according to these probabilities. At each state, a residue is emitted according to the emission probability distribution and this creates an observable sequence of residues. The sequence of these internal states is hidden, hence the name hidden Markov models, therefore the most likely state sequence must be inferred from an alignment between the HMM and the query sequence (Eddy, 1996).

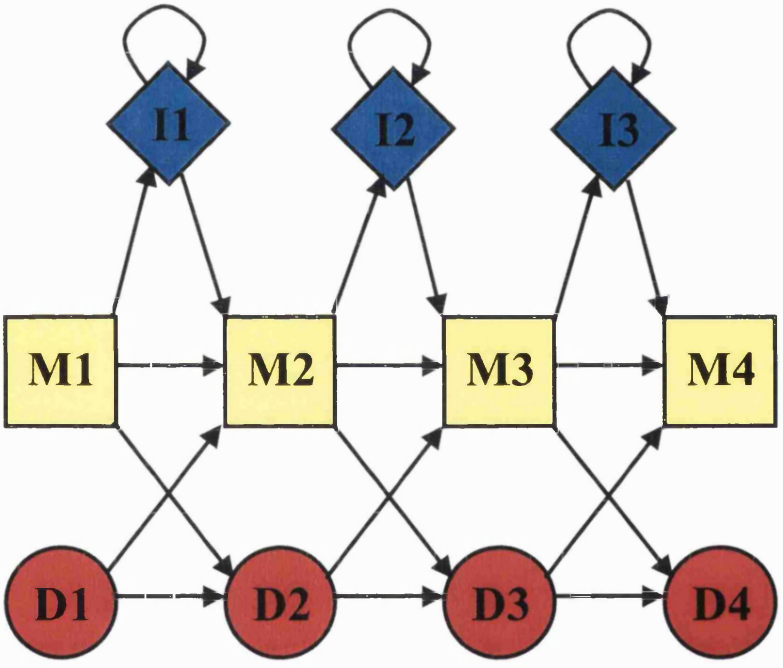


Figure 5.1: Overview of the profile hidden Markov model (HMM). This is characterised by its match (M), delete (D) and insert (I) states and the allowed transitions (arrows) between them

5.1.3.3 SAM-T99

The SAM-T99 method, based on the earlier SAM-T98 protocol (Karplus *et al.*, 1998), builds a HMM from either a single seed sequence or a reliable seed alignment using a large sequence database such as the non-redundant translated GenBank sequence database (NRDB) (Benson *et al.*, 2000). After the initial scan of the sequence database, a model is generated from the alignment of these related sequences and this model is then used for a further database scan (see figure 5.2). Every added sequence provides the model with more detail on the acceptable sequence variability at each position within the sequence family. As a result, therefore allows greater sensitivity for identifying more distantly related sequences.

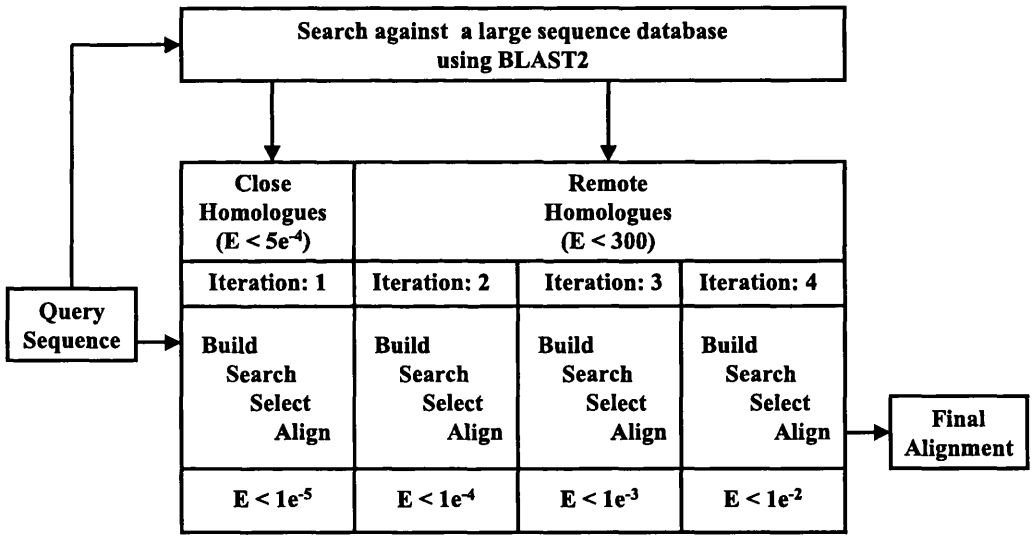


Figure 5.2: Overview of the SAM-T99 protocol for detecting remote homologues.

5.1.3.4 PSI-BLAST

PSI-BLAST (Altschul *et al.*, 1997) uses an iterative approach that begins with a simple pairwise BLAST search of a sequence database. This identifies a set of close relatives from which a multiple sequence alignment is generated. Instead of searching with a single sequence, the database is now searched with a profile derived from the multiple sequence alignment, thus identifying more distant hits. Again, the sequence information from these distant relatives is then incorporated into the growing alignment and the process repeated until either no more sequences are found or a specified number of iterations has been reached.

5.1.4 Intermediate Sequence Searching

The percentage of relatives identified by any of the sequence search methods can be increased by scanning against protein family libraries, or intermediate sequence libraries (ISLs) rather than libraries containing single sequences. Many structure and sequence databases cluster sequences into families according to sequence, structural and/or functional similarity, e.g. Pfam (Bateman *et al.*, 2000), PRINTS (Attwood *et al.*, 1998), CATH (Pearl *et al.*, 2001b) and SCOP (Lo Conte *et al.*, 2000). Each of these families can then be represented in an ISL by providing a single representative sequence for each family. This ISL can then be searched with a query sequence to identify putative homologous relationships to these representative sequences. Since the homology within the protein families is well defined, identifying a homologous relationship to a representative sequence implies a homologous relationship between the query sequence and all other sequences within the sequence family.

This concept of intermediate sequence searching (ISS) was first introduced by Park *et al.* (1997). The FASTA sequence comparison algorithm (Pearson & Lipman, 1988) was used to provide sets of intermediate sequences, i.e. protein families, for a dataset of representative PDB sequences (clustered at 40% sequence identity). Following an all-against-all FASTA sequence comparison of the dataset of PDB sequences, the number of true homologous matches (validated by sequence and structural similarities) recognised by the ISL and the single sequence library were compared. At a low rate of error (1%), the ISL library was seen to recognise 70% more homologous relationships than the pairwise method (pairwise recognised 15% and ISL recognised 26% of the evolutionary relationships).

5.1.5 CATH Protein Family Database: CATH-PFDB

5.1.5.1 Incorporating Genomic Sequences into the CATH Database

Within the CATH structure database, proteins are clustered into sequence families if they share at least 35% sequence identity (CATH-S35). This conservative threshold ensures that each sequence in the cluster is closely related, however the application of such a strict cutoff also results in many homologous relationships being missed at the sequence level. This is compensated for in CATH by using structural and functional information to group more distantly related proteins into the same homologous superfamily.

A more recent development within the CATH database was the application of PSI-BLAST to provide structural annotation for genomic sequences (Pearl *et al.*,

2001b). This involved using PSI-BLAST with conservative thresholds to identify homologous sequences from the translated GenBank-NRDB for all the structures in CATH. Once identified using PSI-BLAST, these genomic sequences are clustered into CATH homologous superfamilies using the pairwise Needleman-Wunsch sequence comparison algorithm (Needleman & Wunsch, 1970).

5.1.5.2 Using the CATH-PFDB as an Intermediate Sequence Library

The CATH-PFDB provides validated sequence families for all structures in the CATH database. Since all the sequences within a protein family are known to be related by evolution, identifying a relationship between a query protein sequence and a protein in a sequence family of the CATH-PFDB infers a relationship between this query protein and all the other proteins in the same family. As these sequence families all include at least one protein with known structure, finding a match to any one of the genomic sequences in a sequence family automatically provides a structural assignment for the query protein. This method of inferring homology through an intermediate, i.e. using a rule of 'two degrees of separation', allows more distant relationships to be identified than using simple pairwise sequence comparison methods.

5.1.6 Performance of the Sequence Comparison Algorithms

The relative performance of pairwise sequence comparison and profile-based sequence comparison was first assessed by Park *et al.* (1998). The authors first selected a representative dataset of sequences from the known structures in the SCOP database (Murzin *et al.*, 1995) clustered at 40% or less sequence identity. The SCOP structure classification was used to provide a structurally validated assignment of remote homologous relationships for sequences classified into the same superfamily. An all-against-all sequence comparison was made with the proteins in the dataset using a variety of pairwise and profile-based sequence comparison methods. For a given rate of errors, the percentage of distantly related sequences that each method was able to find was then compared using a coverage-versus-error plot (see section 5.2.4.5).

This work found that profile-based methods that could use additional genomic sequences (e.g. from the translated GenBank NRDB) were able to recognise around twice the number of relationships detected by pairwise methods that were searching the PDB alone. When considering more remote relationships, where the pairs of compared sequences had less than 30% sequence identity, the performance of profile-

based methods was three times the performance of pairwise methods. Of the profile-based methods, the SAM-T98 approach was able to identify the highest percentage of homologous relationships.

5.1.7 Structure-Based Sequence Alignments

5.1.7.1 Extending the Profile-Based Methods

The results by Park *et al.* (1998) together with results from additional research within the CATH group (Pearl *et al.*, 2001a; Buchan *et al.*, 2002) suggest that sequence profiles that use additional genomic sequences provide the most effective sequence comparison tool for the detection of remote evolutionary relationships. However, the performance of these profile-based approaches is dependent on the quality of the multiple sequence alignment used to generate the profile. Including more distantly related sequences provides more information for these profiles since the highly conserved sequence features are more likely to be the result of structural or functional constraints rather than simply an artifact of sampling proteins that are close in evolutionary time. Also, ensuring the accuracy of an alignment becomes increasingly difficult as the sequences become more distant. As a result, compromise is often sought when building sequence profiles between only including similar sequences, in order to guarantee a high quality alignment, and allowing more remote sequences which provides a more descriptive and therefore more sensitive profile.

An alternative approach that avoids this compromise is by providing a more accurate alignment between distant sequences by structural comparison. In this way, the remote sequences are included yet the alignment quality is still retained which should result in an accurate and highly descriptive sequence profile.

5.1.7.2 3D-PSSM

The 3D-PSSM method (Kelley *et al.*, 2000) provides an example of a method that uses structural information to improve the alignment of distant sequences (see figure 5.3). This method takes a structural sequence, called the master sequence (A0), and uses PSI-BLAST (see section 5.1.3.4) to identify sequence relatives for all structural sequences in a given SCOP superfamily (A0, B0, ...). This produces a series of sequence alignments whose parents are related by structure but not necessarily sequence similarity (A0, A1, A2 and B0, B1, B2). This sequence alignment is then converted to a position specific score matrix (1D-PSSM) which encodes the sequence variation at each position of the alignment.

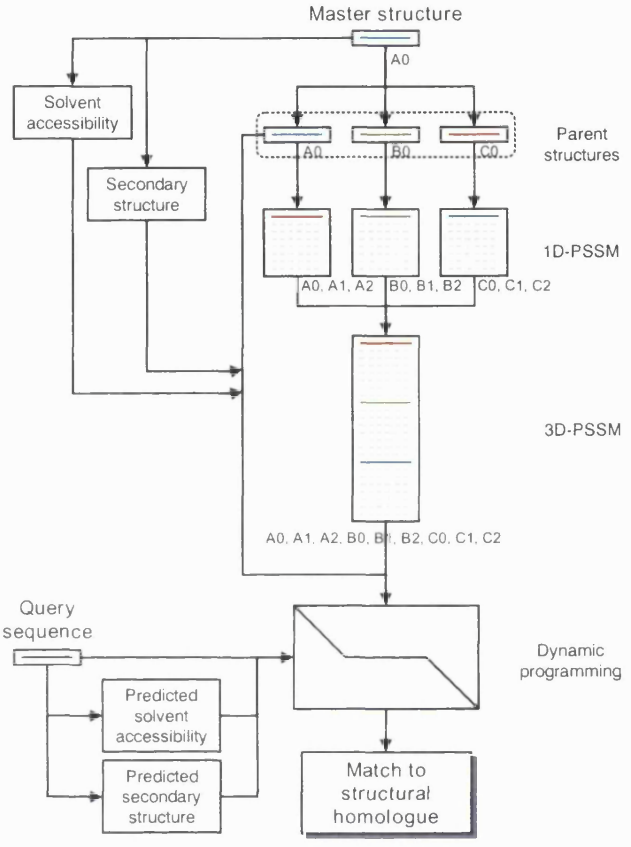


Figure 5.3: Overview of the 3D-PSSM protocol. Four types of information are generated for each master protein (A0) in the library; solvent accessibility, secondary structure state, 1D-PSSM and 3D-PSSM. The solvent propensities and predicted secondary structure are predicted for the query sequence. An all-against-all comparison of residue features is made and the scores entered into a dynamic programming matrix. From this, an alignment and global similarity score are generated.

The SSAP algorithm (Taylor & Orengo, 1989) is then used to perform an all-against-all structure comparison of the parent structures (A0, B0, C0, ...). From a superposition based on the structural alignment, any structures with more than 6Å RMSD with respect to the master structure (A0) are removed. Then, starting with the most similar structural match (e.g. A0 and B0), the sequence alignments for each parent are combined by introducing gaps throughout the individual sequence alignments where gaps occur in the structural alignment of the parents. This procedure is repeated for all structures in the superfamily to create a large multiple sequence alignment based around the master structure (A0, A1, A2, B0, B1, B2). A PSSM is then generated for this structure-based sequence alignment (3D-PSSM).

A query protein is searched against a structure in the PSSM library by comparing the sequence of the query with the 1D-PSSM and the 3D-PSSM generated from the library structure. The sequences are compared by first populating a score matrix (see section 1.2.4.2) with values based on the sequence similarities between each residue of the query sequence and each position in the PSSM. Then dynamic programming is used to find the optimal alignment through this score matrix and therefore the optimal alignment between the query sequence and library structure.

This work also examines the effect of including structural information that can be predicted from the query sequence in the comparison procedure. Secondary structure state and solvent accessibility are predicted using automated protocols and assigned to each residue in the query sequence. The comparison of these predicted values with the observed values in the master structure is also included in the score matrix. This provides additional information for the query and master sequences to be aligned.

The 3D-PSSM protocol was benchmarked using a set of 136 sequences whose homology could not be detected by PSI-BLAST. Incorporating structural information in the comparison protocol was found to increase the recognition by 14% (19 out of 136 extra correctly recognised homologies at an equivalent error rate of 0.05). These results clearly show that this protocol has a higher performance than using PSI-BLAST alone to select sequence relatives and build the initial sequence profiles.

5.1.8 Aims

The research in this chapter aims to provide a sensitive tool that can recognise remote homologous relationships between a query sequence and a known structure. Identifying a homologous relationship to a characterised structure can allow structural and even functional features to be inferred to the query sequence. This type of procedure was inspired by the improved performance observed from 3D-PSMM and is highly applicable for the classification of new structures in the CATH database and as a potential method for genome annotation.

The structural alignments generated in chapter 3 provide a framework that will allow distant sequence alignments to be combined in a structurally validated manner. Incorporating these distant sequences into the same sequence profile provides a greater level of description of the observed evolutionary changes. Thus these structure-based sequence profiles are expected to recognise more remote homologous relationships than profiles based on sequence alone.

This work aims to build on the concepts seen in the 3D-PSSM protocol by attempting to make improvements in a number of key areas. Firstly, the structural alignments used in the 3D-PSSM protocol were built from chaining together a series of pairwise structure alignments. One possible problem when chaining together pairwise alignments is that the global alignment tends to include a high proportion of gaps. This is because gaps are only considered within the pairwise alignment used to align the new structure rather than using information from all the structures in the growing alignment. The method discussed in this chapter uses the CORA multiple structure comparison algorithm (Orengo, 1999) (see section 3.1.2.2) rather than a chain of pairwise alignments to provide the multiple structure alignment. As a result this approach should provide a more accurate alignment that contains fewer gaps.

Also, the 3D-PSSM method selects the proteins to include in the structure alignment based on RMSD ($<$ 6Å RMSD to the master protein). Although RMSD is a useful measure when examining highly similar structures, it can prove inconsistent when examining more distantly related structures. The structures selected for inclusion in the proposed method are the result of an optimised clustering protocol based on a normalised global structural similarity (SSAP) score (this clustering protocol is described in section 3.2.4). Since this protocol used multiple linkage clustering, the clusters generated were guaranteed to contain significant structural similarity to all other proteins in the cluster. Producing such structurally coherent clusters provides a consistently high quality of multiple structure alignments across the range of protein families in the structural database. Generating the structural alignment

in such a thorough manner plays a crucial role in this protocol since the quality of the structural alignment is vital to the quality of the final sequence profile.

Another difference between the 3D-PSSM method and the method presented is the protocol used to generate the initial sequence alignments. The 3D-PSSM method uses PSI-BLAST whereas the method presented here uses SAM-T99, which has been shown to provide the highest performance for detection of remote homologues (see section 5.1.6).

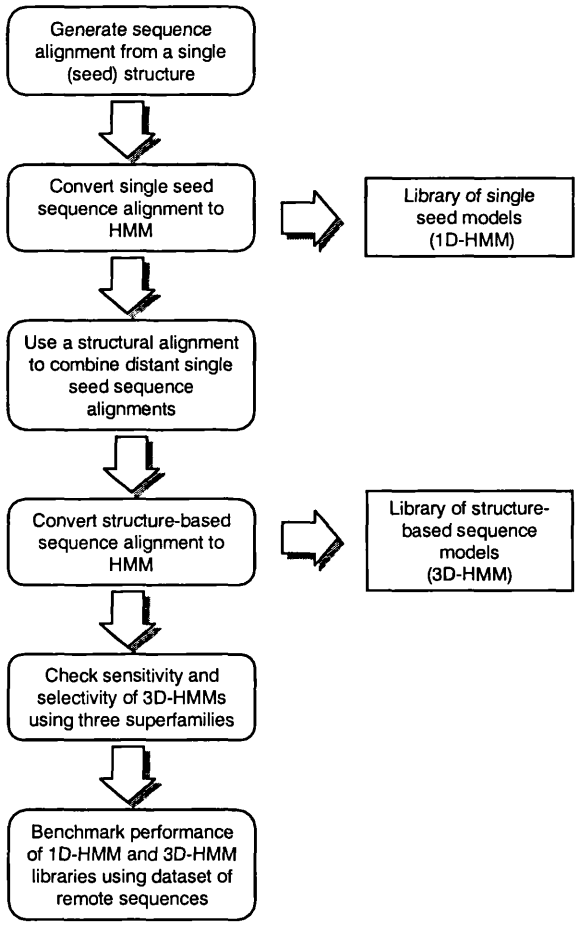


Figure 5.4: Overview of the work described in this chapter. Two sets of model libraries are generated, one consisting only of sequence information (1D-HMM) and one that uses structural information to combine distant sequence alignments (3D-HMM). The procedure for generating a sequence augmented model of structural alignment has been termed as the SAMOSA protocol. The 3D-HMM library is tested on a small dataset of three homologous superfamilies then both libraries are fully benchmarked using a set of structurally validated remote sequences.

The protocol proposed in this chapter generates these structure-based sequence alignments initially by using sequence comparison methods to generate a series of

accurate sequence alignments from closely related sequences (see figure 5.4). These sequence alignments are converted into hidden Markov models (HMM) to provide the 1D-HMM library. The more distant relationships between these sequence alignments are then modelled by using multiple structure comparison methods to allow the sequence alignments can be combined in a reliable manner. The resulting sequence alignment is again converted into a HMM to produce the 3D-HMM library. The procedure for generating these structure-based sequence models has been termed sequence augmented models of structural alignments, or SAMOSA, protocol.

The additional performance of the 3D-HMM generated by this protocol when added to the 1D-HMM library was assessed by comparing the performance of the 1D-HMM library alone. The performance of these two model libraries was assessed by attempting to recognise distant homologous relationships validated using the CATH database. Results from these models were visualised in the form of coverage-versus-error plots and were benchmarked against the SSEARCH sequence comparison method (Pearson, 1991).

5.2 Methods

5.2.1 Overview of Methods

The first part of this section describes the procedure for adding sequence information to a structural alignment, the SAMOSA protocol. This protocol was used to generate a library of 1D-HHM and 3D-HMM sequence models (see figure 5.5).

The second half of this section covers the process of benchmarking the 1D-HMM and 3D-HMM libraries. To provide a thorough benchmark it was necessary to generate a dataset of sequences that had no detectable sequence similarity to sequences used in the model libraries. The procedure for selecting these remote sequences is discussed in addition to the coverage-versus-error used to analyse the results.

5.2.2 The SAMOSA Protocol

5.2.2.1 Overview of the SAMOSA Protocol

A summary of the SAMOSA protocol is shown in figure 5.5. The first step was to generate sequence alignments using SAM-T99 with a single structural sequence as a seed (see section 5.2.2.2). These sequence alignments were then combined by using a multiple structural alignment of the seed structures (CORAXplode protocol, see section 5.2.2.3). The single seed alignments and the structure-based alignments were converted into HMMs and incorporated into a searchable library of sequence models (see section 5.5).

5.2.2.2 Generating the 1D-HMM Library

The SAM-T99 software was used to generate the 1D-HMM sequence models in a two stage process. First, a single structural sequence was used as a seed to search the genomic sequence database (translated GenBank-NRDB, released March 2000). The `target99` script in the SAM-T99 software identifies a set of related genomic sequences and generates a multiple sequence alignment. These sequence alignments are saved for future use (see section 5.2.2.3) and converted to HMMs to produce the 1D-HMM library.

The full 1D-HMM library used in this chapter contained 1D-HMMs generated from all 3,581 non-identical representatives in CATH v1.7. Since the non-identical representatives are clustered at less than 95% sequence identity, this library was given the full title of 1D-HMM-S95. In order to investigate the effect of reducing the sequence redundancy of this library, a subset of this 1D-HMM-S95 library was

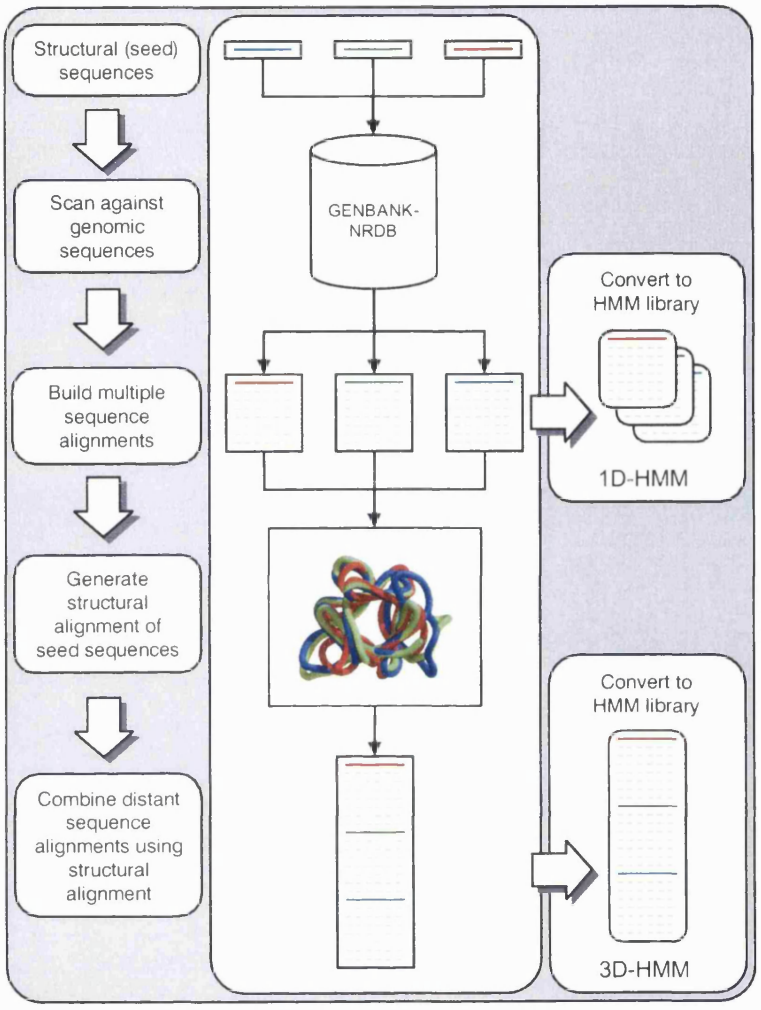


Figure 5.5: Flowchart summarising the SAMOSA protocol. Single structures are initially used as seeds to search a genomic sequence database. The resulting sequence alignments are then converted into HMMs (1D-HMM) using the SAM software. These distantly related sequence alignments are then combined by referring to a multiple structural alignment of the seed structures. Again the resulting structure-based sequence alignments are converted to HMMs (3D-HMM).

taken only using models seeded from sequence families representatives in CATH. Since sequence families are clustered at less than 35% sequence identity, this library was given the full title 1D-HMM-S35 and contained 1,798 models.

5.2.2.3 Generating the 3D-HMM Library

A protocol for generating a sequence model based on a multiple structural alignment, built using the CORA algorithm (Orengo, 1999), was encoded in a program called

CORAXplode (I. Sillitoe, computer program, 2002). In summary, the program takes a set of similar structures from a homologous superfamily (selected using clustering criteria optimised in section 3.2.4). A multiple structure alignment of these seed proteins is then generated using the CORA algorithm. As discussed in section 5.2.2.2, sequence alignments of these structures are generated by using SAM-T99 to search the translated GenBank-NRDB. These initial sequence alignments were then condensed by ignoring any alignment positions with a gap in the seed structure. This step avoided the complication when combining the sequence alignments of attempting to align genomic sequences that could not be referenced back to positions in the structural alignment. The truncated sequence alignments were then combined by inserting gaps throughout the sequence alignment where gaps occurred in the structural alignment. A flowchart of this program is shown in figure 5.6.

The resulting sequence model constructed by CORAXplode was then converted to a HMM (3D-HMM). The 3D-HMM contains the same sequence information as the 1D-HMMs generated from the seed structures in the multiple structure alignment. However, by combining these sequence data into a single CORAXplode model, consensus sequence patterns for even more distant relationships are accurately represented. Aligning such distant sequences without structural validation could provide inaccurate alignments and therefore lower the sensitivity of the models. Since the sequences were combined in a manner consistent with established evolutionary relationships, the added sequence information would be expected to provide a far more detailed model of the possible variations in sequence. As a result, the increased detail of the 3D-HMM would be expected to prove more effective at recognising very remote sequences than using the individual 1D-HMMs.

5.2.3 Measuring Performance

5.2.3.1 Searching Sequences Against the HMM Libraries

The `hmmscore` program (included in the SAM-T99 package) was used to search a large database of sequences against a library of HMMs. This program identifies sequences from the database that are similar to a given model and provides a significance score, or E-value, and an alignment for each of these matches. The E-value is the probability that a match of the same sequence identity and length could occur by chance. This score therefore provides an indication of the quality and the confidence associated with this match, with more significant matches returning lower E-values.

Matches from scanning the HMM library with a query sequence were classed as

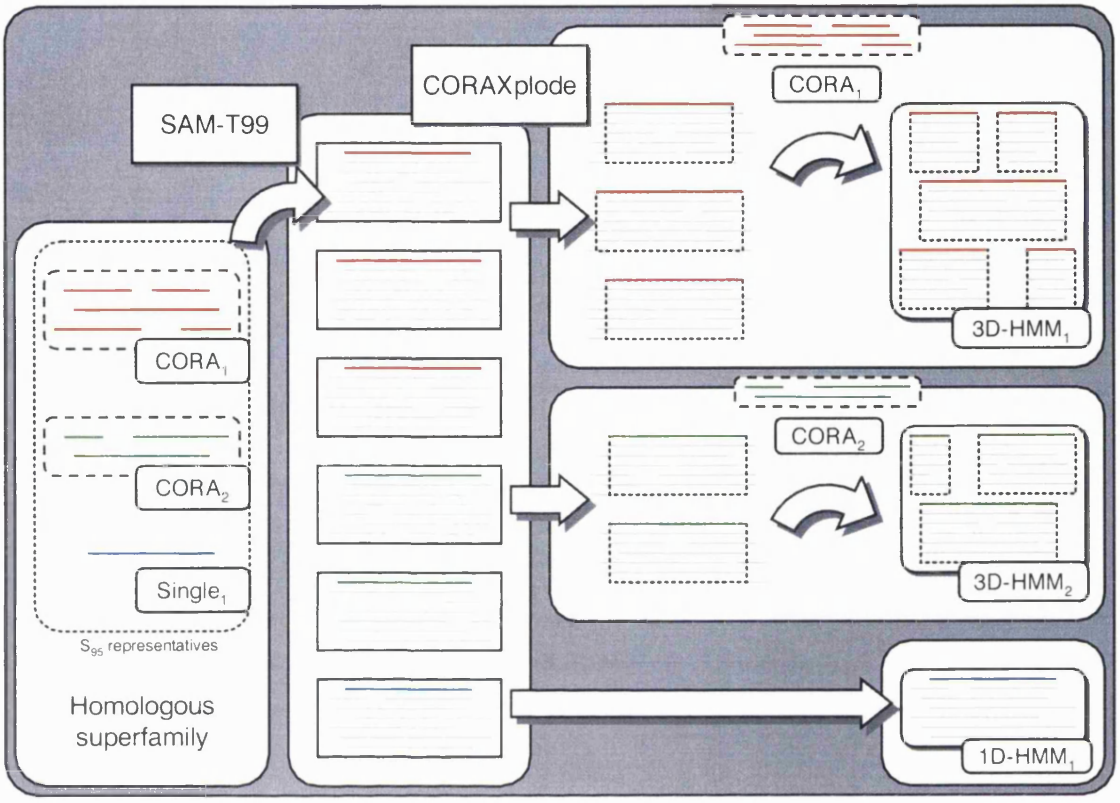


Figure 5.6: Flowchart describing the CORAXplode program. This procedure first takes the structural sequences from a given structural subclusters within a homologous superfamily. A multiple structural alignment of the non-identical representative (S95) structures within the clusters is generated using CORA. A set of sequence alignments is also generated using each of these S95 structural sequences as a seed using SAM-T99. The sequences in these alignments are then truncated with respect to the seed structure so alignment positions corresponding to gaps in the seed structure are removed. These blocks of sequences are then combined using the structural alignment to form a structure-based sequence alignment.

true positives (TP) if the homologous superfamily classification in the CATH-PFDB matched the classification of the query model. These matches were classed as false positives (FP) if the classifications did not match. Homologous sequences that did not match the models were classified as false negatives (FN) and all non-homologous sequences that were not identified were seen as true negatives (TN). The number of possible true positives (i.e. TP + FN) was the total number of sequences in the given superfamily as dictated by the CATH-PFDB. If a query sequence matches more than one HMM from a given superfamily in the HMM library then only the best scoring match (lowest E-value) is used. This 'one-to-many' (Muller *et al.*, 1999)

approach of assigning a homologous relationship avoids artificially exaggerating the number of recognised homologies for each query sequence.

5.2.3.2 Coverage-Versus-E-value Plots

To investigate the degree of sequence representation of the 3D-HMM library and the number of errors identified when searching CATH-PFDB v1.7, a plot was used to compare the percentage of true positive matches over expected matches, i.e. coverage, and number of false positive matches, i.e. errors, at various E-value thresholds (see equation 5.1). The coverage of true positive matches was plotted using the scale on the left and number of false positive matches were plotted using the scale on the right of the plot (see figure 5.7). The quality of different HMM libraries could then be checked by comparing the coverage and error results on the same graph, as discussed in section 5.2.4.3. For this quality assessment exercise, the coverage-versus-E-value plot was favoured over the more commonly seen coverage-versus-error plot (see section 5.2.4.5) since full coverage was often reached before any errors were introduced for all libraries.

$$\text{Coverage (score)} = \frac{\text{TP (score)}}{\text{TP (score)} + \text{FN (score)}} * 100\% \tag{5.1}$$

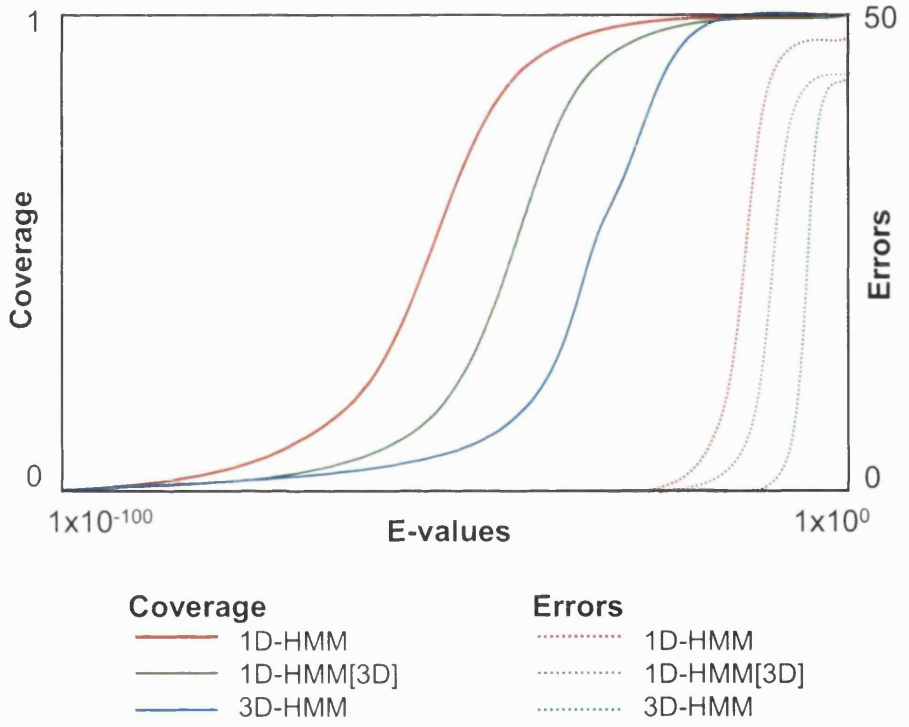


Figure 5.7: Coverage-versus-E-value plot. The plot displays the results for the three different HMM libraries; 1D-HMM in red (HMM seeded by a single sequence), 1D-HMM[3D] in green (subset of the 1D-HMM library based on sequences represented in the structural template library) and 3D-HMM in blue (SAMOSAs built from the combination of distant sequence alignments). Both the coverage (left axis, solid lines) and number of errors (right axis, dotted lines) at different E-value thresholds are displayed.

5.2.4 Selecting Datasets to Test the HMM Library

5.2.4.1 Generating the Intermediate Sequence Library

The CATH-PFDB v1.7 reference sequence library contained 160,316 genomic sequences based on 903 homologous superfamilies with known structures (see section 5.1.5.2). The redundancy in this sequence database was removed by taking one representative sequence for clusters at 60% sequence identity. This smaller sequence database contained 28,578 proteins (PFDB-S60) and was used as the reference sequence library for the ISL library. The rigorous Smith-Waterman algorithm, SSEARCH, could then be used to search this extensive sequence database. The highest scoring match could then be traced back to the known structure for that superfamily.

5.2.4.2 Selecting the Benchmark Sequences

In order to examine the HMM libraries, it was necessary to generate a set of query sequences that shared a homologous relationship with at least one HMM in the library. The sequences could then be used to query the HMM library and the resulting hits assessed for evidence of homology. The CATH-PFDB provides such a resource as it contains a library of protein sequences that are related to structural domains in CATH (see section 5.1.5). This sequence database was generated by using PSI-BLAST (with the standard BLOSUM62 matrix) to scan the sequences of all the structural domains in CATH against the translated GenBank-NRDB. Conservative thresholds for PSI-BLAST were used (E-values less than $5x10^{-4}$) to ensure that homology could be confidently assigned (Pearl *et al.*, 2001b). The CATH-PFDB v1.7 (built from the structural domains in CATH v1.7) contained over 200,000 domain sequences, clustered into 2,327 sequence families and 863 homologous superfamilies.

Since each sequence in the CATH-PFDB was assigned to a homologous superfamily, this provided a validation of matches resulting from scanning the HMM library. A match was considered a true evolutionary relationship if the superfamily classification of the sequence matched the superfamily classification of the seed structure in the case of the 1D-HMM or the structural alignment in the case of the 3D-HMM.

5.2.4.3 Quality Assessment of the 3D-HMM Library

In order to investigate the quality of the models in the 3D-HMM library, i.e. the ability to recognise all the homologous sequences before non-homologous sequences,

a dataset containing three example superfamilies was used (selected based on their high sequence and structural diversity). A 1D-HMM library was generated for each superfamily using the non-identical representative structures as seeds for the sequence alignment. A 3D-HMM library was also generated for each superfamily using the SAMOSA protocol described in section 5.2.2.3. To assess the quality of the 3D-HMM library, three libraries were generated for the three superfamilies being examined.

- **1D-HMM library**
  Containing 1D-HMMs built from all structural sequences in the superfamily.
- **1D-HMM[3D] library**
  This is a subset of the 1D-HMM library which only contains models seeded by structural sequences present in the 3D-HMM library.
- **3D-HMM library**
  Containing the combined structural sequences for all clusters in the superfamily.

All sequences from CATH-PFDB v1.7 were then searched against these libraries and the results ranked by decreasing E-value. Since the sequences sampled in both the 1D-HMM and 3D-HMM library were constructed from the same NRDB release as the CATH, it was expected that both sets of models would identify a large percentage, if not all of the sequences for the corresponding homologous superfamily. As a result, this exercise could not be expected to provide a comparable measure of performance for the 3D-HMMs. Rather, it was used as a quality assessment exercise to ensure that the 3D-HMMs, despite combining distant sequence families, could still recognise homologous sequences without the introduction of errors. These results were analysed using a coverage-versus-E-value plot (see section 5.2.3.2).

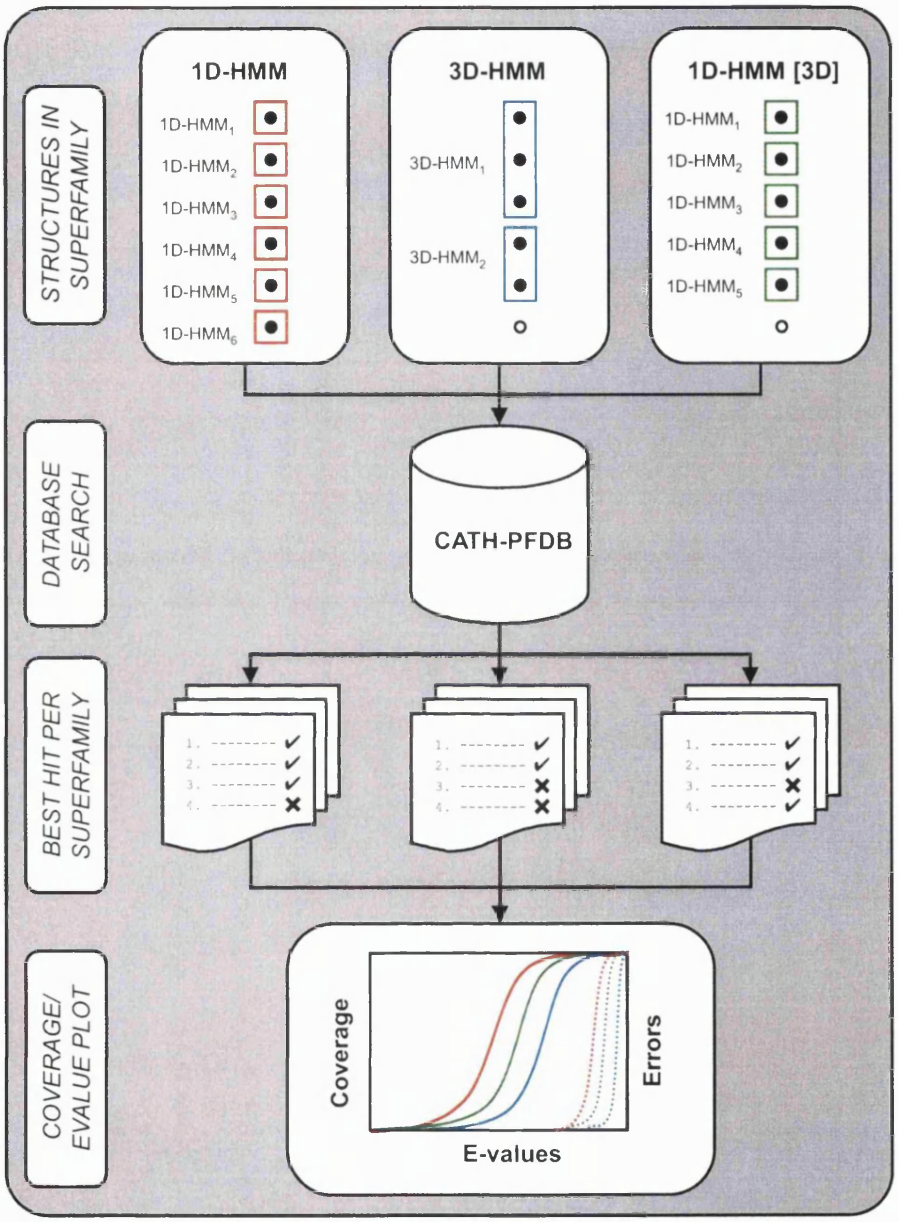


Figure 5.8: Flowchart describing the process of checking the 3D-HMMs. Three sets of models were generated for a single superfamily; 1D-HMMs, 3D-HMMs and the 1D-HMMs just for the single structures involved in the 3D-HMMs, called 1D-HMM[3D]. These sets of models were scored against the CATH-PFDB sequence database with the best score (i.e. lowest E-value) per superfamily taken for each sequence match. The CATH-PFDB classification provided a validated assignment of true positive or false positive for each hit which related to coverage and errors respectively in the final plot.

5.2.4.4 Performance of the 3D-HMM Library

In order to provide an accurate benchmark for the performance of the 3D-HMM library it was necessary to use a dataset of sequences that had no detectable sequence similarity to sequences used to generate the library. The process for selecting these remote sequences (summarised in figure 5.9) started by comparing two versions of the CATH classification database and identifying the structures that were present in the more recent release (v2.0) but not found in the older release (v1.7). Filtering these sequences to remove redundancy provided 1,284 sequences that did not share any more than 35% sequence identity with any other member of the dataset. In order to assign homologous and non-homologous relationships, only those sequences that belonged to a superfamily existing in the older version were kept in the dataset, reducing the number of sequences to 816.

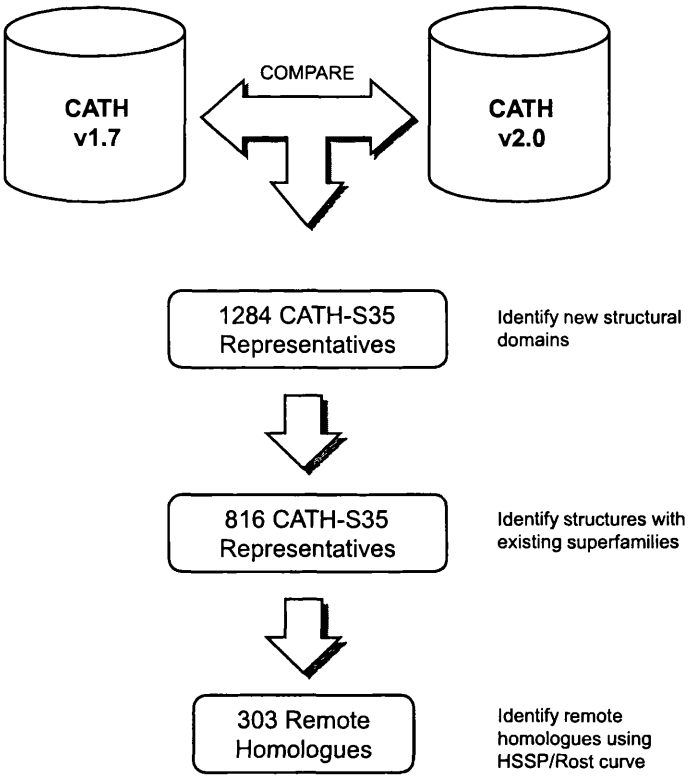


Figure 5.9: Generating the dataset of 303 remote sequences. The procedure starts by comparing two versions of the CATH database and identifying the new structural domains. The sequence redundancy within these new structures was removed by selecting representatives sharing no more then 35% sequence identity (CATH-S35). This dataset was further reduced by only selecting structures that were involved in superfamilies that existed in the older version of CATH. In a final step the HSSP/Rost curve was used to separate the remote homologues from the close homologues.

Since the sequence of a new structure could already be represented in the extended sequence library that the models were generated from, these 816 sequences were then scanned against the CATH-PFDB (based on CATH version 1.7) using the pairwise sequence comparison method SSEARCH (Smith & Waterman, 1981; Pearson, 1991). To provide a reliable measure of remote homology, the results of these matches were plotted on a graph of sequence identity against the number of aligned residues. As the true homologous superfamily was known for each of these 816 sequences, each match to the CATH-PFDB could also be assigned as either homologous or non-homologous. A HSSP/Rost curve (Sander & Schneider, 1991; Rost, 1999) was then empirically determined from this graph to provide a threshold above which no non-homologous matches, could be found (see figure 5.10). This HSSP/Rost curve was used rather than the more simplistic sequence identity threshold to account for the fact that a match with a sequence identity of 35% over 20 aligned residues is far less significant than a match with the same sequence identity over 200 residues.

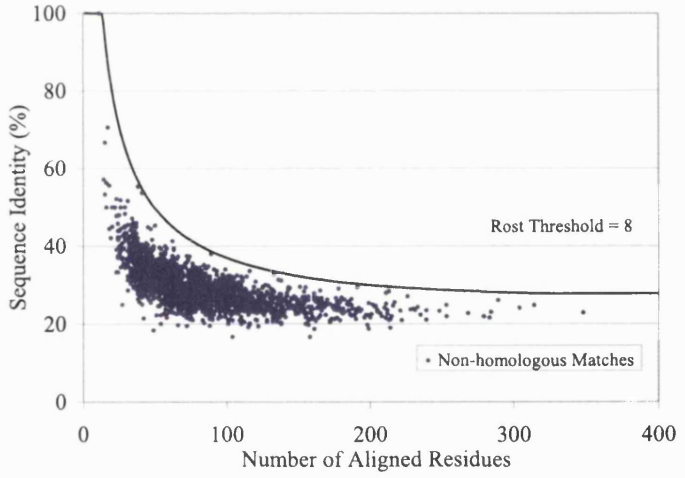


Figure 5.10: Plot of sequence identity against number of aligned residues for the non-homologous pairs observed from scanning the 816 non-redundant sequences against the CATH-PFDB sequence library. The Rost equation was used to provide a boundary above which no non-homologous matches can be found and is represented as the black line in the graph.

When considering homologous matches this threshold also provides a boundary above which the relationships can be considered as close, or easily recognisable. Conversely, the matches that appear below this boundary can be considered remote, or more difficult to recognise, since they cannot be reliably distinguished from non-homologous matches using pairwise sequence comparison. Separating the sequences in this way left a smaller dataset of 303 remote sequences (see figure 5.11).

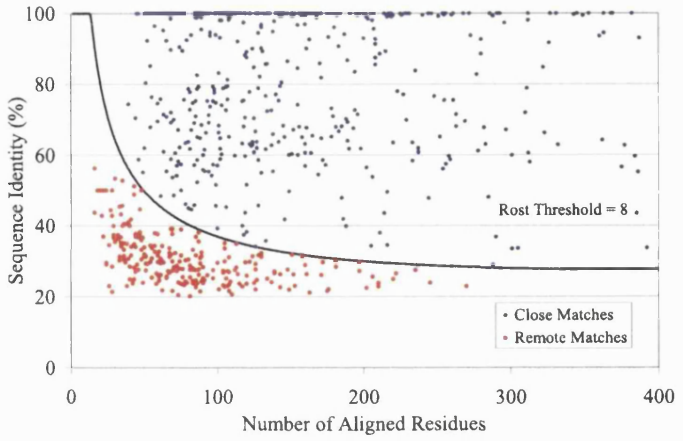


Figure 5.11: Plot of sequence identity against number of aligned residues for the homologous pairs observed from scanning the 816 non-redundant sequences against the CATH-PFDB sequence library. The Rost equation was used to provide a boundary above which no non-homologous matches can be found and is represented as the black line in the graph. The 303 proteins that match a CATH-PFDB sequence below this line are considered remote matches.

5.2.4.5 Coverage-Versus-Error Plot

A useful measure of performance for search algorithms or HMM libraries is to compare the number of remote homologies that a particular method or library can identify (i.e. coverage) within an acceptable level of error (i.e. error rate). In a similar manner to the coverage-versus-E-value plot described in section 5.2.3.2, the values of coverage and error rate are defined by first classing a real homologous match as a true positive (TP) and a non-homologous match as a false positive (FP). True homologous matches that have not been recognised by the HMM library can be termed as false negatives (FN) and non-homologous proteins that have not been matched by the HMM-library can be termed true negatives (TN). Coverage can then be seen as the percentage of true matches (TP) over the total number of homologues (TP + FN) (see equation 5.2). Error rate can be seen as the percentage of false matches (FP) over the total number of non-homologues (FP + TN) (see equation 5.2).

$$\text{Coverage (score)} = \frac{\text{TP(score)}}{\text{TP (score)} + \text{FN (score)}} \tag{5.2}$$

$$\text{Error Rate (score)} = \frac{\text{FP (score)}}{\text{FP (score)} + \text{TN (score)}} \tag{5.3}$$

Plotting the results of the sequence search algorithms on a graph of coverage versus error rate (coverage-versus-error) allows the performance of different search procedures to be compared as long as both the query sequences and the reference

sequence database are kept constant. Thus, if one algorithm or HMM library provides higher coverage than a counterpart at an equivalent error rate then it can be said to have a higher performance.

5.3 Results

5.3.1 Overview of Results

The results have been organised into two sections. In section 5.3.2, the general premise of combining individual sequence alignments with reference to a multiple structural alignment is examined. This involved checking that models based on a combination of very distant sequence alignments did not result in severely disrupted alignments with a subsequent reduction of coverage for sequences from the same superfamily. Since the sequences being used to query these models were in many cases either present or similar to sequences present in the models themselves, these results should be viewed as a quality assessment exercise rather than a true performance evaluation.

Section 5.3.3 describes a true benchmark for the SAMOSAs by attempting to correctly identify a set of validated remote homologous sequences. The results of the performance test for the SAMOSAs were compared to the single structure models and a benchmark using the pairwise sequence alignment method SSEARCH.

5.3.2 Quality Assessment of the 3D-HMM Library

The process of clustering the structures within a superfamily to provide the groups of structures used as seeds for the 3D-HMMs inevitably left some clusters containing only a single structure. As the 3D-HMMs were based on representative structures found in the multiple structural alignment, it was likely that some sequence families within the superfamily were not represented in any of the 3D-HMMs. To account for this, three separate HMM libraries were generated for each of the three superfamilies.

- **1D-HMM**
  Built from all structural sequences in the superfamily.
- **1D-HMM[3D]**
  Built from all structural sequences found in the 3D-HMM library.
- **3D-HMM**
  Built for all multiple structure clusters in the superfamily.

The performance of each of these libraries was assessed for the three superfamilies by attempting to recognise all the homologous relationships defined by the CATH-PFDB. The sequence homologies in the CATH-PFDB had been previously identified using the PSI-BLAST algorithm with each structural sequence in a superfamily as a

seed. The 1D-HMM library was generated in an analogous protocol with SAM-T99 (see section 5.2.2.2) and as such would be expected to find all these homologous relationships.

5.3.2.1 Cytokine Four-Helix Bundle Superfamily

A summary of the number of models generated to represent this superfamily is shown in table 5.1. The superfamily contained 15 non-identical representatives which were used as seeds for the 1D-HMMs. Clustering of this superfamily generated 3 3D-HMMs which included 9 of the 10 sequence families. The 1D-HMMs seeded from the 9 structural sequences in the 3D-HMMs were used as the 1D-HMM[3D] library.

3D-HMM	1D-HMM[3D]	1D-HMM (sequence families)	Homologous Sequences
3	9	15 (10)	589

Table 5.1: Summary of the model data for the cytokine four-helix bundle superfamily. The models in the 1D-HMM[3D] library are built from the single structures within the 3D-HMMs and the final column is the 1D-HMMs for all the non-identical representatives in the superfamily with the number of sequence family representatives given in parentheses.

The coverage-versus-E-value plot for this superfamily, seen in figure 5.12, shows that the 1D-HMMs find all the sequences in the superfamily before an E-value of 10^{-30}. As mentioned in section 5.2.4.3, this result is expected since the dataset is not jack-knifed, i.e. sequences included in the dataset may be present in the models. However, of more interest was the comparison between the results for the 3D-HMM library and the 1D-HMM[3D] library. It can be seen that the coverage plot for the 1D-HMM[3D]s follows that of the 1D-HMM at very low E-values then peaks out at a coverage of around 0.9 (526/589 possible homologous sequences). This is consistent with the fact that the 1D-HMM[3D] library represents 90% of the sequence families in the superfamily (see table 5.1). Similarly, the coverage for the 3D-HMM library peaks at exactly the same level although it does so at a higher E-value of around 10^{-15} rather than 10^{-30} for the equivalent 1D-HMM[3D] library. The shift of coverage to higher E-values is significant and is due to the high divergence in the 3D-HMM library. Since the 3D-HMMs are designed for detection of remote homologues, it follows that the 3D-HMMs will score less well against close homologues than a combination of the more specific SSMs.

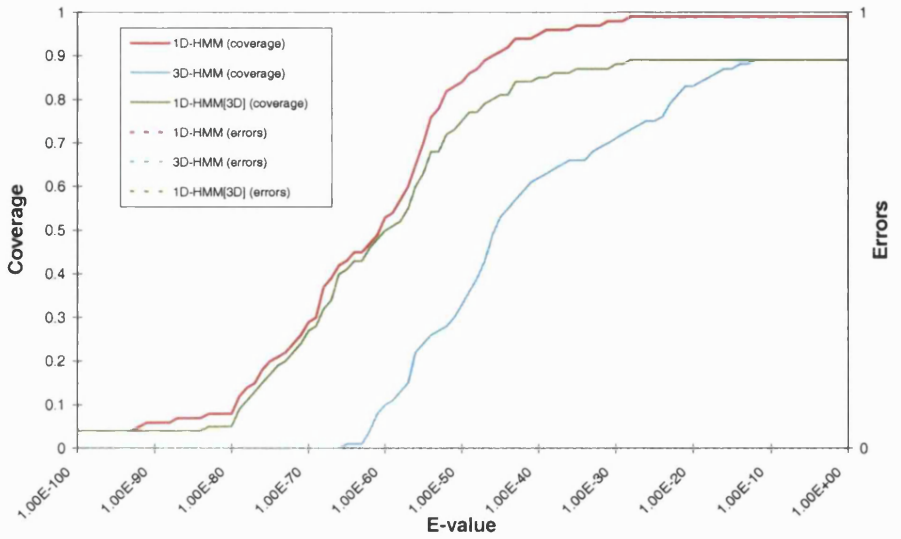


Figure 5.12: Coverage-versus-E-value plot for the cytokine four-helix bundle superfamily. The three sets of were taken from the 1D-HMM library based on all the non-identical structural sequences in the superfamily (red), the 3D-HMM library (blue) and the 1D-HMM[3D] library based on the structural sequences in the 3D-HMM library (green). Coverage is calculated as the fraction of true positives over possible number of homologous matches for a given E-value and is represented by the solid lines and the left side axis. The number of errors, or non-homologous matches, for the same E-value threshold is represented by the broken lines and the right side axis (in this graph there are no errors with E-values < 1.

None of the model sets produced matches that were deemed as errors, i.e. belonging to a different superfamily in the CATH-PFDB, for an E-value of less than 1. This also gives an indication that the quality of the 3D-HMMs has not deteriorated to the point of recognising large numbers of false positives.

5.3.2.2 Cupredoxin Superfamily

The summary of the libraries of models used to represent this superfamily can be seen in table 5.2. This is a large superfamily containing 42 non-identical structural domains and 17 sequence families. The 3D-HMM library contains five models which together include sequences from 16 of the 17 sequence families.

The results for the coverage versus E-value plot for this superfamily can be seen in figure 5.13. For this superfamily, all model sets provide full coverage despite the fact that the 3D-HMM, and therefore 1D-HMM[3D], libraries only cover 16 of the 17 sequence families. Again, the SAMOSA coverage is shifted to higher E-values, although this is less evident than seen in the results for the previous superfamily

3D-HMM	1D-HMM[3D]	1D-HMM (sequence families)	Homologous Sequences
5	16	42 (17)	299

Table 5.2: Summary of the model data for the cupredoxin superfamily. See table legend 5.1 for more details.

(figure 5.12).

Each model set has matches deemed as errors at E-values less than 1 and are represented by the dashed lines in the graph. Although the number of errors recognised by the 3D-HMM library is very similar to that of the 1D-HMM and 1D-HMM[3D] libraries, the errors generated using the 3D-HMM library are shifted to higher E-values (approximately 10^{-5} rather than 10^{-15}). Since these errors are identified with lower confidence they are less likely to be mistaken for true positive matches. Therefore, for this superfamily, the 3D-HMM library provides an equivalent coverage to the 1D-HMM library with a larger distinction between true positive matches and false positive matches.

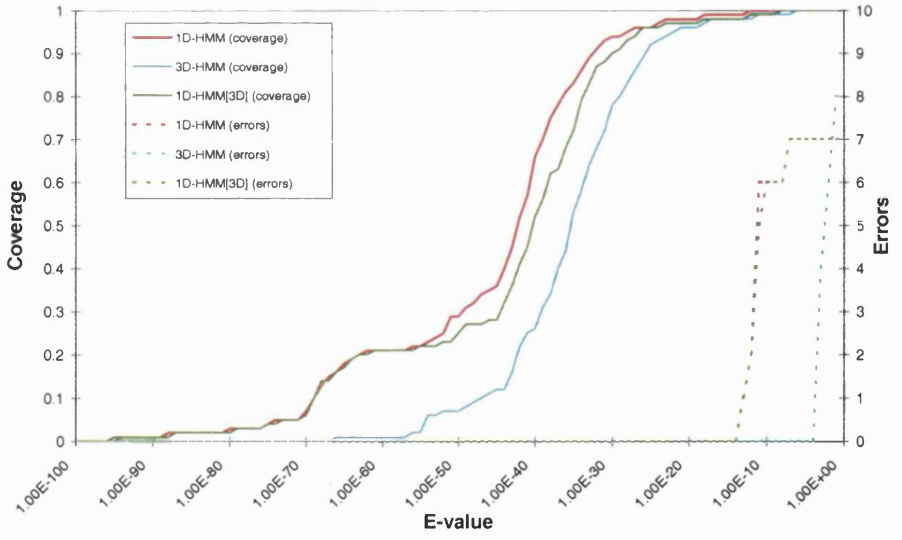


Figure 5.13: Coverage-versus-E-value plot for the cupredoxin superfamily. Comparing the effect of introducing structural information into the sequence alignments.

5.3.2.3 $\alpha\beta$-Hydrolase Superfamily

This superfamily has 24 non-identical structural domains covering 13 sequence families (see table 5.3). Only 6 of these 13 sequence family representatives were represented in the 2 models in the 3D-HMM library. Despite this lack of representation, both the 3D-HMM and 1D-HMM[3D] library provide over 90% coverage before any errors are introduced.

3D-HMM	1D-HMM[3D]	1D-HMM (sequence families)	Homologous Sequences
2	6	24 (13)	1744

Table 5.3: Summary of the model data for the $\alpha\beta$-hydrolase superfamily. See table legend 5.1 for more details.

Interestingly, even the 1D-HMM library does not reach full coverage as only 1693 of the 1744 homologous sequences in the CATH-PFDB are found (see figure 5.14). All of the 51 homologous sequences not identified were clustered into CATH-PFDB-S35 families (clustered at 35% sequence identity) that did not contain a structural representative. Thus all these missed homologies were distant sequence relationships identified by PSI-BLAST. This highlights differences between the PSI-BLAST algorithm, used to identify homologous sequences in the CATH-PFDB, and SAM-T99 used to generate these HMM libraries (discussed further in section 5.4).

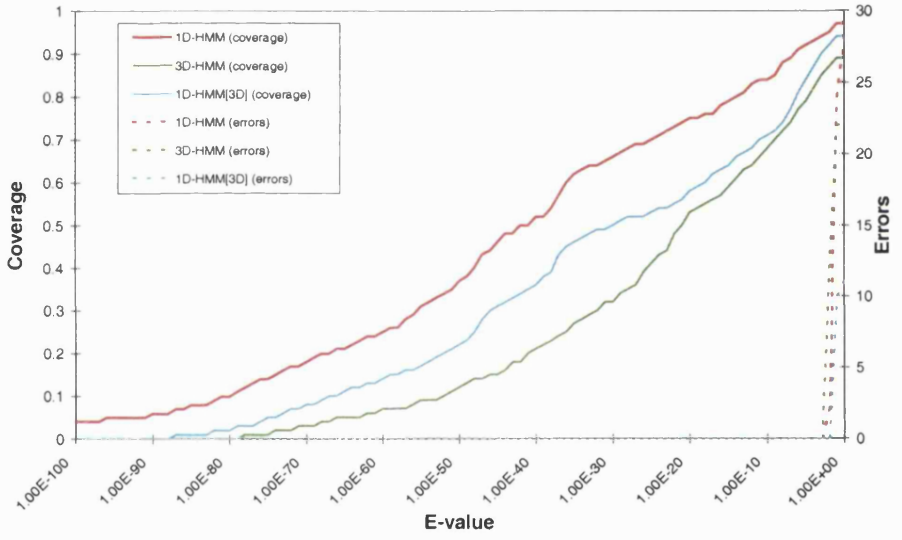


Figure 5.14: Coverage-versus-E-value plot for the $\alpha\beta$-hydrolase superfamily. Comparing the effect of introducing structural information into the sequence alignments.

5.3.3 Benchmarking the 3D-HMM library

5.3.3.1 Overview of the Benchmarking Procedure

The performance of the SAMOSA protocol was measured by evaluating the ability of the 3D-HMM library to recognise a dataset of remote homologous sequences. These remote sequences were selected based on the criteria that a homology could be assigned through structure comparison but would be difficult to assign using sequence methods. Using this prior knowledge of homology based on structure comparison, it was possible to search the library of sequence models with the remote query sequences and classify the resulting matches as true positives (same homologous superfamily) or false positives (different homologous superfamily). Thus, this set of sequences represents a blind dataset of remote sequences that can allow the performance of different sequence alignment procedures to be compared.

5.3.3.2 Comparison of Pairwise and Profile Search Methods

In this analysis, the profile-based method of SAM-T99 was benchmarked against an ISL (intermediate sequence library) method using the pairwise SSEARCH algorithm (discussed in section 5.1.5). In order to compare the profile-based approach with the pairwise approach directly with coverage-versus-error graphs, it was essential that the number of expected true positive matches (TP + FN) and the number of possible false matches (FP + TN) were the same for both methods. By selecting only the best match per superfamily for each query sequence, there is always 1 expected true positive and 902 possible false positives (903 homologous superfamilies minus the correct answer) for both the pairwise or profile-based method.

As mentioned in section 5.2.2.3, the SAMOSA protocol could only generate 3D-HMMs for superfamilies that had sufficient structural diversity to provide a multiple structure alignment. Since this structural diversity was present in only 406 of the 903 superfamilies, the number of superfamilies represented by the 3D-HMM library was less than half that of the 1D-HMM library and ISL library. This meant that the number of possible false positive matches for a given query sequence was 405 $(406-1)$ for the 3D-HMM library rather than 902 $(903-1)$ when using the full set of superfamilies seen in the 1D-HMM and ISL libraries. As a result it was not possible to compare the performance of the 3D-HMM library with that of the 1D-HMM and ISL libraries directly.

To avoid this problem, the results for the 3D-HMM library were combined with the results for the 1D-HMM library. Again, only the lowest E-value score per superfamily was taken for each query sequence, using either the 3D-HMM library or

the 1D-HMM library. This ensured that all homologous superfamilies were still represented in the library and provided a constant number of possible errors. The results of this combined library were then compared to the results from the 1D-HMM library. Any improvement in the performance would then be attributed to an increased number of homologous relationships detected by the 3D-HMM library and any degradation in the performance due to an increased number of errors.

Also to investigate the effect of sequence redundancy in the full 1D-HMM library, that is the 1D-HMM library containing sequences seeded from the non-identical representatives (1D-HMM-S95), a further subset of the results was taken just using the models seeded from sequence family representatives clustered at 35% sequence identity (1D-HMM-S35). Since the 1D-HMM-S35 library contained 50% (1,798/3,581) fewer models than the 1D-HMM-S95 library, the faster database scans of the smaller library would be more desirable if the performance between the two libraries were comparable.

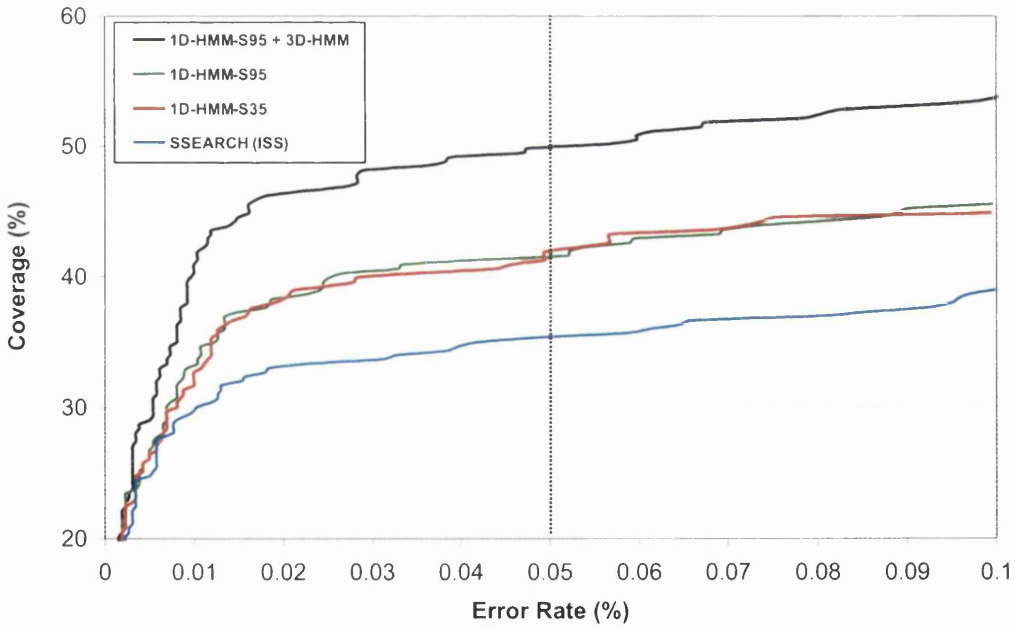


Figure 5.15: Results of the SAMOSA benchmark. Coverage-versus-error plot comparing the performance of the combined 1D-HMM and 3D-HMM library with the 1D-HMM library (using two thresholds of sequence redundancy 1D-HMM-S35 and 1D-HMM-95) and ISL library.

The error-versus-coverage plot for the results of this benchmark can be seen in figure 5.15. The performance of the pairwise SSEARCH method (ISL) can be seen in blue, the 1D-HMM-S35 library is shown in red, the 1D-HMM-S95 library in green

and the combined 1D-HMM-S95 and 3D-HMM library is shown in black. Again, the best performance is the method with the highest coverage at a fixed error rate.

Error Rate	Coverage (%)			
	SSEARCH	**1D-HMM-S95**	**1D-HMM-S35**	**1D-HMM-S95 + 3D-HMM**
0.05	35.5	41.1	42.1	50.0
0.10	38.8	45.5	44.8	53.8

Table 5.4: Comparison of the results from the SAMOSA benchmark. The percentage of true positives, or coverage, is given for each sequence library at error rate thresholds of 0.05 and 0.10, or confidence levels of 95% and 90% respectively.

Table 5.4 quantifies these results by providing the coverage of the four libraries at an error rate of 0.05% and 0.10%, or a confidence level of 95% and 90% respectively. At an error rate of 0.05% the SSEARCH method provided a coverage of 35.5% which is around 6% lower than both the S95 and S35 1D-HMM libraries (41.1% and 42.1% respectively) and 15% lower than the combined 1D-HMM-S95 and 3D-HMM library (50.0%). From these values it can also be seen that adding the 3D-HMM models to the 1D-HMM-S95 library provided a 9% increase in coverage. This demonstrates a significant increase in performance over the traditional SAM-T99 method. Another interesting result was that the smaller, less redundant 1D-HMM-S35 library provided almost identical coverage to the 1D-HMM-S95 library, suggesting little advantage in using the larger library.

These findings are reiterated by the results when using the error rate of 0.10%. The SSEARCH method obtains a coverage of 38.8% and again this is around 6% less coverage than the S95 and S35 1D-HMM libraries which reach 45.5% and 44.8% coverage respectively. The coverage for the combined 1D-HMM-S95 and 3D-HMM library (53.8%) was 15% higher than the SSEARCH method and around 9% higher the two 1D-HMM libraries.

5.4 Discussion

A sequence profile can be described as a summary of the consensus features of a multiple sequence alignment. Hence the more varied the sequences contained within the alignment, the more information the model will contain. Including remote sequences in the alignment allows more remote relationships to be found since conserved patterns are more likely to occur by design rather than by chance between distant homologues. However, this presents an obvious problem of attempting to align very dissimilar sequences based on their sequence similarity, which inevitably results in a poor alignment. Thus, these distant sequences can only be accurately aligned by using a measure of protein homology that is more conserved over evolutionary time than sequence similarity, i.e. structural similarity. The work presented in this chapter demonstrates that the validated structure alignments discussed in chapter 3 could be used as a framework to align distant sequence alignments. Also this work indicates that hidden Markov models derived from these structure-based sequence alignments can be used to recognise more distant homologous relationships and extend the performance of the current state-of-the-art method SAM-T99 (see section 5.1.6).

The success of this method is most likely to be the result of a strict and rigorously tested protocol that provided high quality structural alignments even for highly diverse superfamilies. To summarise, the quality of these structural alignments was ensured by paying particular attention to the following features of the structural alignment.

- **Structurally coherent clusters of representative proteins**
  Proteins in a superfamily were carefully grouped into structurally coherent clusters with a multiple-linkage algorithm to ensure that the multiple structural alignments would not drift.

- **Model structurally diverse superfamilies**
  More than one structural alignment was used to represent structurally diverse superfamilies. This allowed the quality of the structural alignment to be retained without sacrificing the representative coverage of the superfamily.

- **Rigorous multiple structure alignment algorithm**
  A rigorous, residue-based algorithm was used to generate the structural alignments. This involved adding proteins to a growing consensus template rather than simply chaining together pairwise structural alignments.

From the results of the quality assessment exercise of the 3D-HMM library (section 5.3.2), it could be seen that these SAMOSA sequence models were able to recognise a similar, if not identical, number of PSI-BLAST relatives when compared to the results using an equivalent set of individual sequence models in the 1D-HMM[3D] library. So, despite aligning distantly related sequences, the 3D-HMM library could provide similar coverage to an equivalent 1D-HMM library with no extra errors being introduced. In fact the results from the 3D-HMM library based on the cupredoxin superfamily display fewer errors than both 1D-HMM libraries (4 rather than 7). Also, the errors for this superfamily appear at a higher E-value in the 3D-HMM library ($1 * 10^{-03}$ rather than $1 * 10^{-13}$). Thus the 3D-HMM actually select against non-homologous relationships in this case, providing more differentiation between homologous matches and non-homologous errors.

However when processing these results, some consideration should be given to the criteria defining a homologous match and an 'error' for these database scans. As discussed in section 5.2.3.1, the definition of a true homologous match is taken from the CATH-PFDB classification where relationships between structural sequences in CATH and genomic sequences are identified using PSI-BLAST with conservative thresholds. It follows that an error is defined as a match between the HMM and a sequence classified as non-homologous in the CATH-PFDB. Therefore, this 'error' could either be due to the fact that the matching sequence is genuinely non-homologous or is just a case of the PSI-BLAST algorithm failing to identify this homologous relationship in the CATH-PFDB. The results when checking the HMM libraries from the α β-hydrolase superfamily (table 5.3) showed that the SAM-T99 models failed to find 51 of the 1744 sequence relationships identified by PSI-BLAST. PSI-BLAST allows up to 20 iterations when adding and incorporating new sequences into the growing sequence model. This highlights differences between PSI-BLAST and SAM-T99. When considering these 51 missed homologies, or false negatives, 47 were clustered within just 2 S35 families. It is possible that PSI-BLAST may have identified two marginal sequence homologies during an early iteration then as a result of these sequences being incorporated into the model, many more sequences from related families could be identified as homologous. Obviously this is the very feature that makes PSI-BLAST so powerful in detecting remote homologies, however if a non-homologous sequence is wrongly identified and incorporated into the model then the error becomes magnified as further non-homologous sequences are assimilated into the model.

A conservative threshold of a maximum E-value of 0.0005 was used when identifying homologous sequences in the CATH-PFDB for this very reason. However when

considering matches for such distant relationships, sequence comparison methods can only provide an opinion of homology. This homology can only be truly verified using more sensitive probes of evolutionary relationship, such as structural similarity, or based on expert knowledge and manual inspection. It is for this reason, and due to the lack of 'jack-knifing' (i.e. sequences used to test the models were related to sequences in the models themselves) that these results could only be used as an exercise in quality assessment rather than a true measure of performance.

An unbiased and accurate benchmark of the SAMOSA protocol was achieved by selecting a series of structural sequences which were unrelated to any individual sequences in the libraries and whose homology was validated by structural comparison methods. An interesting finding from these results was that there was little difference in performance between the 1D-HMM library built from the 3,581 S95 representatives structural sequences and the library built from 1,798 S35 representatives. This suggests that since there is no advantage of using the larger library the smaller, and therefore faster, 1D-HMM-S35 library would be more appropriate to use for the rapid identification of homology in the classification procedure.

The results from the quality assessment procedure demonstrated that the 3D-HMM library generally provided the same coverage as an equivalent 1D-HMM library (seeded from the same sequences as the 3D-HMM library). However the 3D-HMM library did not provide as much coverage as the full 1D-HMM library based on all sequences in the superfamily rather than just those seen in the 3D-HMM library. Thus by combining the full 1D-HMM library with the 3D-HMM library it was hoped that the coverage of the 1D-HMM library would complement the sensitivity of the 3D-HMM library. From the results in figure 5.15, this certainly was observed to be the case as the combined library provided a 10% increase of coverage at an equivalent error rate. This represents a significant advance in performance over the current methods of sequence comparison and could therefore provide an important tool in assigning structural domains to novel genomic sequences.

Chapter 6

Discussion

In order to keep pace with the exponential increase in the number of protein structures deposited into the PDB each year, structure classification databases will need to improve the speed and sensitivity of automated protocols to detect evolutionary relationships. Classifying novel structures into families that are known to be related by evolution provides insights into the features of the protein structure and sequence that are important for the structural stability and function. For example, aligning the protein sequences within a given family allows the conservation of sequence identity to be examined for each residue position. If a residue identity at a given position in the alignment is found to be highly conserved, especially when examining distantly related proteins, it is likely that this position plays a crucial role in the structure or function of the protein family. These important consensus sequence features can be described in a sequence profile which then acts as a unique and identifying 'fingerprint' for each protein family, allowing even more distant relationships to be identified.

Since protein structure is more conserved than protein sequence, aligning proteins based on similarities in structure, rather than sequence, allows even more distant evolutionary relationships to be explored. Generating a multiple structure alignment of distantly proteins allows the highly conserved structural features of a protein family to be identified. For example, highly conserved inter-residue contacts can be identified by examining equivalent residue positions of the proteins in the structural template. In an analogous manner to sequence profiles, taking into account the conservation of structural features reduces the 'noise' (i.e. ignoring comparisons to highly variable features of the template) when searching for further structural similarities. This effectively provides a more sensitive probe of evolutionary relationships.

The work in this thesis has aimed to identify highly conserved inter-residue contacts from superfamilies in the CATH database. This work was originally inspired by the work of Russell & Barton (1994) and others who suggested that inter-residue contacts could be used to differentiate between related and unrelated structures. This information has been applied to the analysis, comparison and alignment of protein structures (see figure 6.1). The work in chapter 2 has described a computational tool written to identify and manipulate these sets of contact data (COCOPLOT). This chapter has also introduced a novel structure comparison algorithm (CONALIGN) that can align protein structures based on inter-residue contacts. This algorithm was able to recognise the correct fold of structures using as little as 10% of the contacts observed in the native structure. This method has applications for fold recognition where only limited structural data is known and will become increasingly important as the accuracy of predicting contact data directly from sequence continues to improve.

Chapter 3 discussed the clustering protocol used to generate high quality multiple structural alignments for all superfamilies in CATH (release 1.7). These multiple structure alignments were all converted to structural templates, thus providing a searchable library of 'average' structural features which encompassed the variability observed within each structural alignment. A protocol was then established to enable a query structure to be scanned against the template library. The structural similarity of each structure–template alignment was assessed with the COCOPLOT software, using a score based on matching contacts in the query structure to highly conserved contacts observed the structural template. Using a structurally validated dataset of distantly related structures, this protocol was able to recognise the correct fold group for 70% of the structures and the correct homologous superfamily for 52% of the structures, provided the structure belonged to a superfamily represented in the template library.

The library of structural templates contained far fewer entries (407 templates representing 340 homologous superfamilies) than the library of structures used for the pairwise comparisons (3,581 non-identical representatives). Therefore, a typical scan of the template library could be achieved far more rapidly than a pairwise scan (approximately 1.5 hours for the template library compared to approximately 12 hours for the pairwise library). This method will be integrated into the CATH classification protocol in order to quickly assign a large percentage of novel structures. Also, the performance of this method will greatly improve as the structure database becomes more populated, i.e. from the large scale structural genomics projects. As more examples of structures are included in the superfamilies, more

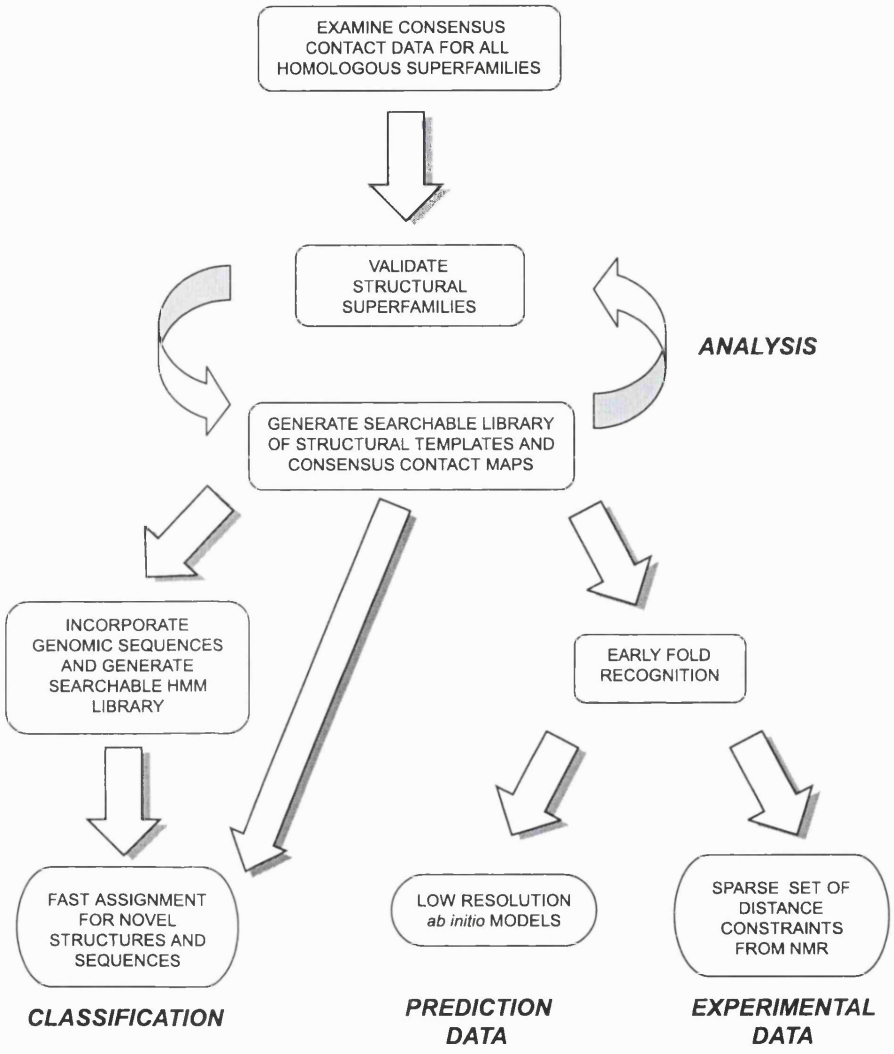


Figure 6.1: Summary of work presented in this thesis.

of these superfamilies will be represented by the structural templates. Also, the existing structural templates will become more descriptive and will provide better coverage of each superfamily, including all the distant homologues.

Chapter 4 applies this protocol of scanning the library of structural templates to identify distant structural similarities from matching consensus contacts, to recognise the fold of protein models predicted directly from amino acid sequence, i.e. *ab initio*. Identifying the likely folding arrangement from a low resolution model predicted at an early stage in the prediction should greatly increase the speed of the subsequent refinement procedure. The conserved structural features from the matching fold could also be used to constrain the protein model in order to improve

the resolution of the model itself. This protocol was able to recognise the correct fold from predicted models displaying little obvious structural similarity (in half of the structures tested, the correct fold was recognised using models between 9–10Å RMSD from native structure). When applied to predicted models from CASP3 submissions, this protocol was seen to provide a performance comparable to threading approaches. Also, since the successes using this method and the threading methods were different (i.e. they correctly recognised the fold of different CASP3 target sequences), combining elements of these different fold recognition approaches may present a method with even greater performance.

Finally, chapter 5 used the high quality structural alignments generated in chapter 3 to improve the performance of sequence profiles when applied to the detection of distant evolutionary relationships. The high quality multiple structure alignments allowed a series of highly diverse SAM-T99 sequence alignments to be combined in the SAMOSA protocol. The resulting sequence alignment provided descriptions of remote sequence similarities that would be too distant to model from sequence information alone. Thus, adding a library of hidden Markov models generated from these diverse sequence alignments was expected to increase the coverage of the existing sequence libraries in CATH. Using a structurally validated set of remote sequences, the performance of recognising remote evolutionary relationships was seen to increase by 10% when including these SAMOSA models in the existing sequence libraries.

Future Work

It will be of great importance to develop protocols that will allow both the library of structural templates and the library of SAMOSA sequence models to be frequently updated in order to ensure optimal performance in recognising evolutionary relationships. Since both libraries were generated from release 1.7 of the CATH database, such update protocols will be applied to generate libraries for the latest version of the CATH database (release 2.4).

Another area of future work will be to generate a more statistically robust score when scanning a query structure against the template library. In the work described in this thesis, the measure of structural similarity (contact overlap score) was calculated as the number of overlapping contacts as a percentage of the maximum number of contacts between the two structures being compared. This score was shown to work well when applied to the identification of structural homology (see chapters 2, 3 and 4). However, this score was sensitive to the number of conserved contacts observed in the template (see section 3.3.2). A more reliable scoring scheme

would be to provide significance scores, i.e. Z-scores (see section 1.2.5.2) for each structure–template comparison. This could be achieved by scanning each structural template against a series of unrelated structures generating a distribution of 'random' scores specific to the template. When scanning a query structure against a structural template, the significance score would then be calculated by comparing each contact overlap score with the distribution of random scores for that template. Employing these significance scores would be expected to increase the recognition rates when using the structural template library even further. Recent expansion of the Linux farm used for classifying structures in CATH would make this technically feasible.

The computational tools described in this thesis would also allow a comprehensive analysis of the role of inter-residue contacts in the evolution of protein structure. By comparing the contacts observed in pairs of structures and also the conservation of contacts in multiple structure alignments, the structural analysis by Russell & Barton (1994) could be updated and extended using a much larger dataset. One interesting analysis would be to examine the relationship between the conservation of sequence at positions in the multiple structure alignment which are involved in highly conserved contacts. This effectively inverses the problem of predicting contacts directly from sequence and should provide an estimate of the maximum amount of information that could be gained from *ab initio* contact prediction.

List of Abbreviations

Abbreviation	Details
BLAST	Basic Local Alignment Search Tool
BLOSUM	Blocks Substitution Matrices
CASP	Critical Assessment of Methods of Protein Structure Prediction
CATH	Class, Architecture, Topology, Homologous Superfamily
CATH-PFDB	CATH Protein Family Database
CORA	Conserved Residue Attributes
DDP	Double Dynamic Programming
DHP	Dihydropteroate
DNA	Deoxyribonucleic Acid
DP	Dynamic Programming
EBI	European Bioinformatics Institute
EVD	Extreme Value Distribution
FSSP	Fold classification based on Structure-Structure alignment of Proteins
HMM	Hidden Markov Models
HOMSTRAD	Homologous Structure Alignment Database
HSP	High Scoring Segment Pairs
HTML	Hypertext Markup Language
ISL	Intermediate Sequence Library
ISS	Intermediate Sequence Search
LCS	Longest Continuous Segment
MAR	Minimum Alignment Ratio
MCR	Minimum Contact Ratio
MDM	Mutation Data Matrix
MMDB	Molecular Modelling Database
NAD	Nicotinamide Adenine Dinucleotide
NCBI	National Center for Biotechnology Information
NMR	Nuclear Magnetic Resonance
NRDB	Non-redundant Database
PAM	Point Accepted Mutation
PDB	Protein Data Bank
PSSM	Position Specific Score Matrices
PSI-BLAST	Position Specific Iterated-BLAST
RMSD	Root Mean Squared Deviations
RNA	Ribonucleic Acid
SAM	Sequence Alignment and Modelling
SCOP	Structural Classification Of Proteins
SSAP	Sequential Structural Alignment Program
SSE	Secondary Structure Element
STAMP	Structural Alignment of Multiple Proteins
TIM	Triosephosphate Isomerase
VAST	Vector Alignment Search Tool
WWW	World Wide Web

Bibliography

Altschul, S., Gish, W., Miller, W., Myers, E. & Lipman, D. (1990). Basic local alignment search tool. *J Mol Biol*, **215**, 403–10.

Altschul, S., Madden, T., Schaffer, A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*, **25**, 3389–402.

Attwood, T., Beck, M., Flower, D., Scordis, P. & Selley, J. (1998). The PRINTS protein fingerprint database in its fifth year. *Nucleic Acids Res*, **26**, 304–8.

Barton, G. & Sternberg, M. (1987). A strategy for the rapid multiple alignment of protein sequences. Confidence levels from tertiary structure comparisons. *J Mol Biol*, **198**, 327–37.

Barton, G. & Sternberg, M. (1990). Flexible protein sequence patterns. A sensitive method to detect weak structural similarities. *J Mol Biol*, **212**, 389–402.

Bateman, A., Birney, E., Durbin, R., Eddy, S., Howe, K. & Sonnhammer, E. (2000). The Pfam protein families database. *Nucleic Acids Res*, **28**, 263–6.

Benson, D., Boguski, M., Lipman, D. & Ostell, J. (1996). GenBank. *Nucleic Acids Res*, **24**, 1–5.

Benson, D., Karsch-Mizrachi, I., Lipman, D., Ostell, J., Rapp, B. & Wheeler, D. (2000). GenBank. *Nucleic Acids Res*, **28**, 15–8.

Berman, H., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T., Weissig, H., Shindyalov, I. & Bourne, P. (2000). The Protein Data Bank. *Nucleic Acids Res*, **28**, 235–42.

Bray, J. (2001). Predicting the structure and function of genomic sequences using the CATH database. *Thesis*.

Brenner, S., Chothia, C. & Hubbard, T. (1998). Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. *Proc Natl Acad Sci U S A*, **95**, 6073–8.

Buchan, D., Shepherd, A., Lee, D., Pearl, F., Rison, S., Thornton, J. & Orengo, C. (2002). Gene3D: structural assignment for whole genes and genomes using the CATH domain structure database. *Genome Res*, **12**, 503–14.

Butler, D. (1999). IBM promises scientists 500-fold leap in supercomputing power...and a chance to tackle protein structure. *Nature*, **402**.

Chothia, C. (1992). Proteins. One thousand families for the molecular biologist. *Nature*, **357**, 543–4.

Chothia, C. & Lesk, A. (1985). Helix movements and the reconstruction of the haem pocket during the evolution of the cytochrome c family. *J Mol Biol*, **182**, 151–8.

Chothia, C. & Lesk, A. (1986). The relation between the divergence of sequence and structure in proteins. *EMBO J*, **5**, 823–6.

Chou, P. & Fasman, G. (1974a). Conformational parameters for amino acids in helical, beta-sheet, and random coil regions calculated from proteins. *Biochemistry*, **13**, 211–22.

Chou, P. & Fasman, G. (1974b). Prediction of protein conformation. *Biochemistry*, **13**, 222–45.

Dandekar, T. & Argos, P. (1994). Folding the main chain of small proteins with the genetic algorithm. *J Mol Biol*, **236**, 844–61.

Dandekar, T. & Argos, P. (1996). Identifying the tertiary fold of small proteins with different topologies from sequence and secondary structure using the genetic algorithm and extended criteria specific for strand regions. *J Mol Biol*, **256**, 645–60.

Dayhoff, M. (1978). Matrices for detecting distant relationships. *Atlas Protein Seq. Struct.*, **5**, 353–358.

de la Cruz, X., Mahoney, M. & Lee, B. (1997). Discrete representations of the protein C alpha chain. *Fold Des*, **2**, 223–34.

de la Cruz, X., Sillitoe, I. & Orengo, C. (2002). Use of structure comparison methods for the refinement of protein structure predictions. I. Identifying the structural family of a protein from low-resolution models. *Proteins*, **46**, 72–84.

Dembo, A., Karlin, S. & Zeitouni, O. (1994). Limit distribution of maximal non-aligned two-sequence segmental score. *Ann. Prob.*, **22**, 2022.

Dengler, U., Siddiqui, A. & Barton, G. (2001). Protein structural domains: analysis of the 3Dee domains database. *Proteins*, **42**, 332–44.

Dill (1993). Folding Proteins. Finding a needle in a haystack. *Curr Opin Struct Biol*, **3**, 99–103.

Doolittle, R. (1986). *Of URFs and ORFs: A primer on how to analyse derived amino acid sequences*. University Science Books.

Eddy, S. (1996). Hidden Markov models. *Curr Opin Struct Biol*, **6**, 361–5.

Eisenhaber, F., Frommel, C. & Argos, P. (1996a). Prediction of secondary structural content of proteins from their amino acid composition alone. II. The paradox with secondary structural class. *Proteins*, **25**, 169–79.

Eisenhaber, F., Imperiale, F., Argos, P. & Frommel, C. (1996b). Prediction of secondary structural content of proteins from their amino acid composition alone. I. New analytic vector decomposition methods. *Proteins*, **25**, 157–68.

Eyrich, V., Standley, D. & Friesner, R. (1999). Prediction of protein tertiary structure to low resolution: performance for a large and structurally diverse test set. *J Mol Biol*, **288**, 725–42.

Fariselli, P. & Casadio, R. (1999). A neural network based predictor of residue contacts in proteins. *Protein Eng*, **12**, 15–21.

Fariselli, P., Olmea, O., Valencia, A. & Casadio, R. (2001). Prediction of contact maps with neural networks and correlated mutations. *Protein Eng*, **14**, 835–43.

Finkelstein, A. & Ptitsyn, O. (1987). Why do globular proteins fit the limited set of folding patterns? *Prog Biophys Mol Biol*, **50**, 171–90.

Flores, T., Orengo, C., Moss, D. & Thornton, J. (1993). Comparison of conformational characteristics in structurally similar protein pairs. *Protein Sci*, **2**, 1811–26.

Garnier, J., Osguthorpe, D. & Robson, B. (1978). Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *J Mol Biol*, **120**, 97–120.

Gobel, U., Sander, C., Schneider, R. & Valencia, A. (1994). Correlated mutations and residue contacts in proteins. *Proteins*, **18**, 309–17.

Gotoh, O. (1996). Significant improvement in accuracy of multiple protein sequence alignments by iterative refinement as assessed by reference to structural alignments. *J Mol Biol*, **264**, 823–38.

Gribskov, M., McLachlan, A. & Eisenberg, D. (1987). Profile analysis: detection of distantly related proteins. *Proc Natl Acad Sci U S A*, **84**, 4355–8.

Grindley, H., Artymiuk, P., Rice, D. & Willett, P. (1993). Identification of tertiary structure resemblance in proteins using a maximal common subgraph isomorphism algorithm. *J Mol Biol*, **229**, 707–21.

Harrison, A., Pearl, F., Mott, R., Thornton, J. & Orengo, C. (2002). Quantifying the similarities within fold space. *J Mol Biol*, **323**, 909–26.

Henikoff, S. & Henikoff, J. (1991). Automated assembly of protein blocks for database searching. *Nucleic Acids Res*, **19**, 6565–72.

Henikoff, S. & Henikoff, J. (1992). Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A*, **89**, 10915–9.

Henikoff, S. & Henikoff, J. (1993). Performance evaluation of amino acid substitution matrices. *Proteins*, **17**, 49–61.

Hinds, D. & Levitt, M. (1994). Exploring conformational space with a simple lattice model for protein structure. *J Mol Biol*, **243**, 668–82.

Holm, L. & Sander, C. (1993). Protein structure comparison by alignment of distance matrices. *J Mol Biol*, **233**, 123–38.

Holm, L. & Sander, C. (1998). Touring protein fold space with Dali/FSSP. *Nucleic Acids Res*, **26**, 316–9.

Holm, L., Ouzounis, C., Sander, C., Tuparev, G. & Vriend, G. (1992). A database of protein structure families with common folding motifs. *Protein Sci*, **1**, 1691–8.

Jones, D. (1997). Progress in protein structure prediction. *Curr Opin Struct Biol*, **7**, 377–87.

Jones, D. (1999a). GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. *J Mol Biol*, **287**, 797–815.

Jones, D. (1999b). Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol*, **292**, 195–202.

Jones, D., Taylor, W. & Thornton, J. (1992). A new approach to protein fold recognition. *Nature*, **358**, 86–9.

Jones, D., Tress, M., Bryson, K. & Hadley, C. (1999). Successful recognition of protein folds using threading methods biased by sequence similarity and predicted secondary structure. *Proteins*, **37**, 104–111.

Jones, S., Stewart, M., Michie, A., Swindells, M., Orengo, C. & Thornton, J. (1998). Domain assignment for protein structures using a consensus approach: characterization and analysis. *Protein Sci*, **7**, 233–42.

Karplus, K., Barrett, C. & Hughey, R. (1998). Hidden Markov models for detecting remote protein homologies. *Bioinformatics*, **14**, 846–56.

Kelley, L., MacCallum, R. & Sternberg, M. (2000). Enhanced genome annotation using structural profiles in the program 3D-PSSM. *J Mol Biol*, **299**, 499–520.

Kendrew, J., Bodo, G., Dintzis, H., Parrish, R., Wyckoff, H. & Phillips, D. (1958). A three-dimensional model of the myoglobin molecule obtained by X-ray analysis. *Nature*, **181**, 662–666.

Kolinski, A. & Skolnick, J. (1994). Monte Carlo simulations of protein folding. I. Lattice model and interaction scheme. *Proteins*, **18**, 338–52.

Kraulis, P. (1991). MolScript: a program to produce both detailed and schematic plots of protein structures. *Journal of Applied Crystallography*, **24**, 946–950.

Lee, J., Liwo, A., Ripoll, D., Pillardy, J. & Scheraga, H. (1999). Calculation of protein conformation by global optimization of a potential energy function. *Proteins*, **37**, 204–208.

Lesk, A., Lo Conte, L. & Hubbard, T. (2001). Assessment of novel fold targets in CASP4: predictions of three-dimensional structures, secondary structures, and interresidue contacts. *Proteins*, **Suppl 5**, 98–118.

Lo Conte, L., Ailey, B., Hubbard, T., Brenner, S., Murzin, A. & Chothia, C. (2000). SCOP: a structural classification of proteins database. *Nucleic Acids Res*, **28**, 257–9.

Lund, O., Frimand, K., Gorodkin, J., Bohr, H., Bohr, J., Hansen, J. & Brunak, S. (1997). Protein distance constraints predicted by neural networks and probability density functions. *Protein Eng*, **10**, 1241–8.

Madej, T., Gibrat, J. & Bryant, S. (1995). Threading a database of protein cores. *Proteins*, **23**, 356–69.

Maizel JV, J. & Lenk, R. (1981). Enhanced graphic matrix analysis of nucleic acid and protein sequences. *Proc Natl Acad Sci U S A*, **78**, 7665–9.

Mirny, L. & Shakhnovich, E. (1999). Universally conserved positions in protein folds: reading evolutionary signals about stability, folding kinetics and function. *J Mol Biol*, **291**, 177–96.

Moult, J., Pedersen, J., Judson, R. & Fidelis, K. (1995). A large-scale experiment to assess protein structure prediction methods. *Proteins*, **23**, ii–v.

Moult, J., Hubbard, T., Bryant, S., Fidelis, K. & Pedersen, J. (1997). Critical assessment of methods of protein structure prediction (CASP): round II. *Proteins*, **Suppl 1**, 2–6.

Moult, J., Hubbard, T., Fidelis, K. & Pedersen, J. (1999). Critical assessment of methods of protein structure prediction (CASP): round III. *Proteins*, **Suppl 3**, 2–6.

Muller, A., MacCallum, R. & Sternberg, M. (1999). Benchmarking PSI-BLAST in genome annotation. *J Mol Biol*, **293**, 1257–71.

Murzin, A., Brenner, S., Hubbard, T. & Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol*, **247**, 536–40.

Needleman, S. & Wunsch, C. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol*, **48**, 443–53.

Orengo, C. (1999). CORA–topological fingerprints for protein structural families. *Protein Sci*, **8**, 699–715.

Orengo, C. & Taylor, W. (1990). A rapid method of protein structure alignment. *J Theor Biol*, **147**, 517–51.

Orengo, C. & Taylor, W. (1996). SSAP: sequential structure alignment program for protein structure comparison. *Methods Enzymol*, **266**, 617–35.

Orengo, C., Flores, T., Taylor, W. & Thornton, J. (1993). Identification and classification of protein fold families. *Protein Eng*, **6**, 485–500.

Orengo, C., Michie, A., Jones, S., Jones, D., Swindells, M. & Thornton, J. (1997). CATH–a hierarchic classification of protein domain structures. *Structure*, **5**, 1093–108.

Orengo, C., Bray, J., Hubbard, T., LoConte, L. & Sillitoe, I. (1999). Analysis and assessment of ab initio three-dimensional prediction, secondary structure, and contacts prediction. *Proteins*, **Suppl 3**, 149–70.

Ortiz, A., Kolinski, A. & Skolnick, J. (1998). Fold assembly of small proteins using monte carlo simulations driven by restraints derived from multiple sequence alignments. *J Mol Biol*, **277**, 419–48.

Ortiz, A., Kolinski, A., Rotkiewicz, P., Ilkowski, B. & Skolnick, J. (1999). Ab initio folding of proteins using restraints derived from evolutionary information. *Proteins*, **37**, 177–185.

Osguthorpe, D. (1999). Improved ab Initio predictions with a simplified, flexible geometry model. *Proteins*, **37**, 186–193.

Park, B. & Levitt, M. (1995). The complexity and accuracy of discrete state models of protein structure. *J Mol Biol*, **249**, 493–507.

Park, J., Teichmann, S., Hubbard, T. & Chothia, C. (1997). Intermediate sequences increase the detection of homology between sequences. *J Mol Biol*, **273**, 349–54.

Park, J., Karplus, K., Barrett, C., Hughey, R., Haussler, D., Hubbard, T. & Chothia, C. (1998). Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. *J Mol Biol*, **284**, 1201–10.

Pearl, F., Lee, D., Bray, J., Buchan, A., D Shepherd & Orengo, C. (2001a). The CATH Extended Protein Family Database: providing structural annotations for genome sequences. *Protein Science*, accepted.

Pearl, F., Martin, N., Bray, J., Buchan, D., Harrison, A., Lee, D., Reeves, G., Shepherd, A., Sillitoe, I., Todd, A., Thornton, J. & Orengo, C. (2001b). A rapid classification protocol for the CATH Domain Database to support structural genomics. *Nucleic Acids Res*, **29**, 223–7.

Pearson, W. (1991). Searching protein sequence libraries: comparison of the sensitivity and selectivity of the Smith-Waterman and FASTA algorithms. *Genomics*, **11**, 635–50.

Pearson, W. & Lipman, D. (1988). Improved tools for biological sequence comparison. *Proc Natl Acad Sci U S A*, **85**, 2444–8.

Pennisi, E. (1998). Taking a structured approach to understanding proteins. *Science*, **279**, 978–9.

Phillips, D. (1970). Development of crystallographic enzymology. *Biochem Soc Symp*, **31**, 11–28.

Pollastri, G. & Baldi, P. (2002). Prediction of contact maps by GIOHMMs and recurrent neural networks using lateral propagation from all four cardinal corners. *Bioinformatics*, **18 Suppl 1**, S62–S70.

Rice, D. & Eisenberg, D. (1997). A 3D-1D substitution matrix for protein fold recognition that includes predicted secondary structure of the sequence. *J Mol Biol*, **267**, 1026–38.

Richards, F. & Kundrot, C. (1988). Identification of structural motifs from protein coordinate data: secondary structure and first-level supersecondary structure. *Proteins*, **3**, 71–84.

Rost, B. (1999). Twilight zone of protein sequence alignments. *Protein Eng*, **12**, 85–94.

Rost, B. & Sander, C. (1993). Prediction of protein secondary structure at better than 70% accuracy. *J Mol Biol*, **232**, 584–99.

Rost, B., Sander, C. & Schneider, R. (1994). PHD–an automatic mail server for protein secondary structure prediction. *Comput Appl Biosci*, **10**, 53–60.

Russell, R. & Barton, G. (1992). Multiple protein sequence alignment from tertiary structure comparison: assignment of global and residue confidence levels. *Proteins*, **14**, 309–23.

Russell, R. & Barton, G. (1994). Structural features can be unconserved in proteins with similar folds. An analysis of side-chain to side-chain contacts secondary structure and accessibility. *J Mol Biol*, **244**, 332–50.

Sali, A. & Blundell, T. (1990). Definition of general topological equivalence in protein structures. A procedure involving comparison of properties and relationships through simulated annealing and dynamic programming. *J Mol Biol*, **212**, 403–28.

Samudrala, R., Xia, Y., Huang, E. & Levitt, M. (1999). Ab initio protein structure prediction using a combined hierarchical approach. *Proteins*, **37**, 194–198.

Sander, C. & Schneider, R. (1991). Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins*, **9**, 56–68.

Simons, K., Kooperberg, C., Huang, E. & Baker, D. (1997). Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *J Mol Biol*, **268**, 209–25.

Simons, K., Ruczinski, I., Kooperberg, C., Fox, B., Bystroff, C. & Baker, D. (1999). Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins. *Proteins*, **34**, 82–95.

Sippl, M. & Weitckus, S. (1992). Detection of native-like models for amino acid sequences of unknown three-dimensional structure in a data base of known protein conformations. *Proteins*, **13**, 258–71.

Skolnick, J., Kolinski, A. & Ortiz, A. (1997). MONSSTER: a method for folding globular proteins with a small number of distance restraints. *J Mol Biol*, **265**, 217–41.

Smith, T. & Waterman, M. (1981). Identification of common molecular subsequences. *J Mol Biol*, **147**, 195–197.

Smith-Brown, M., Kominos, D. & Levy, R. (1993). Global folding of proteins using a limited number of distance constraints. *Protein Eng*, **6**, 605–14.

Srinivasan, R. & Rose, G. (1995). LINUS: a hierarchic procedure to predict the fold of a protein. *Proteins*, **22**, 81–99.

Standley, D., Eyrich, V., Felts, A., Friesner, R. & McDermott, A. (1999). A branch and bound algorithm for protein structure refinement from sparse NMR data sets. *J Mol Biol*, **285**, 1691–710.

Stryer, L. (1995). *Biochemistry*. W. H. Freeman and Company, New York, 4th edn.

Taylor, W. (1986a). The classification of amino acid conservation. *J Theor Biol*, **119**, 205–18.

Taylor, W. (1986b). Identification of protein sequence homology by consensus template alignment. *J Mol Biol*, **188**, 233–58.

Taylor, W. (1987). Multiple sequence alignment by a pairwise algorithm. *Comput Appl Biosci*, **3**, 81–7.

Taylor, W. & Hatrick, K. (1994). Compensating changes in protein multiple sequence alignments. *Protein Eng*, **7**, 341–8.

Taylor, W. & Orengo, C. (1989). Protein structure alignment. *J Mol Biol*, **208**, 1–22.

Thomas, D., Casari, G. & Sander, C. (1996). The prediction of protein contacts from multiple sequence alignments. *Protein Eng*, **9**, 941–8.

Thompson, J., Higgins, D. & Gibson, T. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*, **22**, 4673–80.

Todd, A., Orengo, C. & Thornton, J. (2001). Evolution of function in protein superfamilies, from a structural perspective. *J Mol Biol*, **307**, 1113–43.

Westhead, D., Slidel, T., Flores, T. & Thornton, J. (1999). Protein structural topology: Automated analysis and diagrammatic representation. *Protein Sci*, **8**, 897–904.

Zemla, A., Venclovas, Moult, J. & Fidelis, K. (2001). Processing and evaluation of predictions in CASP4. *Proteins*, **Suppl 5**, 13–21.