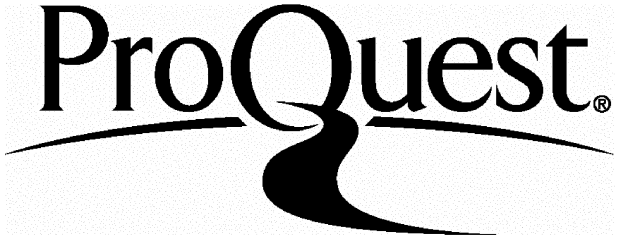
Protein Structure Prediction and Modelling

Robin Edward James Munro, M.Sc.

This thesis is submitted in partial fulfillment of the
requirements of the University of London
for the degree of Doctor of Philosophy

January 1999

Division of Mathematical Biology
National Institute for Medical Research
The Ridgeway, Mill Hill
London, United Kingdom



Abstract

The prediction of protein structures from their amino acid sequence alone is a very challenging problem. Using the variety of methods available, it is often possible to achieve good models or at least to gain some more information, to aid scientists in their research. This thesis uses many of the widely available methods for the prediction and modelling of protein structures and proposes some new ideas for aiding the process.

A new method for measuring the buriedness (or exposure) of residues is discussed which may lead to a potential way of assessing proteins' individual amino acid placement and whether they have a standard profile. This may become useful in assessing predicted models.

Threading analysis and modelling of structures for the Critical Assessment of Techniques for Protein Structure Prediction (CASP2) highlights inaccuracies in the current state of protein prediction, particularly with the alignment predictions of sequence on structure. An in depth analysis of the placement of gaps within a multiple sequence threading method is discussed, with ideas for the improvement of threading predictions by the construction of an improved gap penalty. A threading based homology model was constructed with an RMSD of 6.2Å, showing how combinations of methods can give usable results.

Using a distance geometry method, `DRAGON`, the *ab initio* prediction of a protein (NK Lysin) for the CASP2 assessment was achieved with an accuracy of 4.6Å. This highlighted several ideas in disulphide prediction and a novel method for predicting which cysteine residues might form disulphide bonds in proteins.

Using a combination of all the methods, with some like threading and homology modelling proving inadequate, an *ab initio* model of the N-terminal domain of a GPCR was built based on secondary structure and predictions of disulphide bonds.

Use of multiple sequences in comparing sequences to structures in threading should give enough information to enable the improvements required before threading can become a major way of building homology models. Furthermore, with the ability to predict disulphide bonds: restraints can be placed when building models, *ab initio* or otherwise.

Acknowledgments

I would like to thank my supervisor Willie Taylor for his help and support. Thank you to all the past and present members in the division for helpful discussions, in particular András Aszódi for his help with the finer points of DRAGON and Jaap Heringa for useful comments on the penultimate draft of this thesis.

I am grateful to Bob Bywater for inspiring the work on the GPCR N-terminus and for all his subsequent input.

Thanks to my supportive friends: Franca Fraternali, Beatrice Meanti and Patricia Novelli.

Thanks also to the CASP2 organisers for the Junior Fellowship and also to the GPC'97 organisers for their funding. The work for my Ph.D. was funded by the UK Medical Research Council.

I dedicate this thesis to the memory of my father: Dr. Hamish D. Munro (1942-1980) and to my mother for her invaluable support over the years.

Contents

List of Figures

List of Tables

Abbreviations & Programs

BLAST Basic Local Alignment Search Tool.
BLITZ Exhaustive similarity search.
CASP Critical Assessment of Techniques for Protein Structure Prediction.
CC C Compiler.
CHARMM Molecular Dynamics simulation package.
CMA Correlated Mutation Analysis.
CONES Conic Accessibility/Shieldedness algorithm.
DAC DSSP Accessibility Calculator.
DF Density Function.
DG Distance Geometry.
DP Dynamic Programming.
DRAGON Distance Regularisation Algorithm for Geometry Optimisation.
DSC Discrimination of Protein Secondary Structure Class.
FASTA Global alignment search tool.
GLP Glucagon Like Peptide.
GPCR G-Protein Coupled Receptor.
HEADGREP Protein header pattern matching program.
indels insertions and deletions.
K-S Kolmogorov-Smirnov Test.
MSF Multiple Sequence Format.
MST Multiple Sequence Threading.
MULTAL Multiple Alignment tool.
MD Molecular Dynamics.
NMR Nuclear Magnetic Resonance.
NNPREDICT Neural Network Protein secondary structure prediction.
OWL Non-redundant sequence database.
PADGREP Protein pattern matching program.
PAM Point Accepted Mutation.
PDB Protein Data Bank.
PHD PredictProtein at Heidelberg.
QUANTA Molecular visualisation tool.
RMSD Root Mean Squared Deviation.
SA Surface Accessibility.
SAP Structural Alignment of Proteins.
SCC Side Chain Centroid.
SS Secondary Structure (Sec. Str.).
SSPRED EMBL Secondary Structure Prediction.
UCLA University California, Los Angeles.
Un-wtd Un-Weighted.

NB: standard one and three letter codes are used for amino acids. Brookhaven Database codes are used for PDB entries, sometimes with an additional chain identifier.

Chapter 1

Introduction

The ultimate aim of protein structure prediction is to take a protein with unknown structure and, from its sequence alone, predict the tertiary, or 3D, structure. Despite the simplicity with which the basic problem can be stated, over the forty years that people have been considering it, no method has ever proved to be generally (some would say, even partly) successful. The intellectual challenge of the problem, despite its apparent intractability, has ensured that many have been (and still are) willing to look at it. Although no general method has resulted, all this effort has not been in vain as there are now many methods that, although they cannot predict a full tertiary structure, can provide insight into the sort of structure that a sequence might adopt. In the current situation, in which sequence data is being elucidated at an amazingly

rapid rate, any methods that can extract any structural features from sequence data alone is of great value.

The two major types of biological molecule, protein and nucleic acid (for simplicity, just DNA will be referred to), perform radically different functions; that of active-agents and data-archive respectively and this contrast is also manifest in their structure. DNA is regular, stable and inert, whereas proteins are asymmetric, plastic and active. This contrast was quite unexpected at the time the first structure was solved. The structure was that of myoglobin (a protein containing only α-helix structure) solved by John Kendrew and co-workers at the Medical Research Council Unit for Molecular Biology, the same institute where only three years earlier the Watson and Crick model of DNA was proposed. In his paper on the X-ray model of myoglobin, Kendrew said that "Perhaps the most remarkable features of the molecule are its complexity and lack of symmetry. The arrangement seems to be almost totally lacking in the kind of regularities which one instinctively anticipates and it is more complicated than has been predicted by any theory of protein structure" (Kendrew *et al.*, 1958). This complexity and flexibility of proteins is perhaps a fundamental necessity to permit the innumerable roles that they must fulfill in life; as too much inherent regularity would probably restrict the structures a protein might adopt. While necessary for life, complexity makes the job of structure prediction difficult.

Since myoglobin more than 7,000 structures have been solved by either crystallography or NMR. However, there has been a relative explosion in the number of protein sequences without structure. Computational structure prediction goes some of the

way to redress the balance of sequences over structure and develops faster methods for the solution of structure than those presently available in the "Wet Lab". Structures can play a large role in aiding scientists in the direction their work should take: allowing the ability to design drugs, to modify or interfere with proteins and the ways in which they work. These procedures are all facilitated by having a protein structure to work on. If the protein predictioner can provide a rough model of a structure, then information can be gained which might be vital for the fast development of a new drug. Only when scientists, from all fields of research, work together will the problem at hand be solved.

1.0.1 The current state of the field

Towards the end of 1998 there were just over 330,000 non-redundant sequenced proteins publicly available in various databases (all non-redundant GenBank CDS translations+PDB+SwissProt+SPupdate+PIR). SwissProt is a highly annotated database, but by no means contains a full set of sequences (Bairoch, 1990). Compared with around 7,300 protein structures (Diffraction+NMR, not models) in the PDB (Bernstein *et al.*, 1977). There are clearly many more sequences than there are structures. Many of the structures in the PDB are very similar; some 500 extra are theoretical and others are very low resolution or just backbone atoms. Various research groups have classified protein structures into a non-homologous database, in some cases up to 1000 proteins but more usually around 300 (Orengo, 1994a;

Holm and Sander, 1998). These are proteins that have little sequence and structural similarity and make up a good representation of the structurally solved folds. Using computational methods the aim of protein structure prediction is to redress the imbalance between the number of protein sequences and structure.

An unbiased survey of the power of the available prediction and modelling methods is provided in a prediction contest in which crystallographers (and others) tell in advance if they have solved, or are about to solve, a protein structure. Having provided only the sequence of their protein, it is then up to the predictors to determine the structure – by whatever means they can. The first gathering to assess such prediction results, called the Critical Assessment of Techniques for Protein Structure Prediction (CASP) was held at Asilomar, California, at the end of 1994 with a repeated and similar meeting held in 1996 as a culmination to the second experiment. The assessment covered the major areas of protein structure prediction: Comparative Modelling, Fold Recognition and *Ab initio* and also the prediction of associations between ligands and proteins (Docking). More details can be found in special issues of Proteins: Structure, Function and Genetics (CASP, 1995; CASP, 1997).

Chapters 3, 4 and 5 describe work which was submitted to the CASP2 meeting. The kind of methodology employed in making a structure prediction in this work is illustrated in Figure 1.1 and this introduction will give an overview of some of the ideas involved in predicting and modelling proteins.

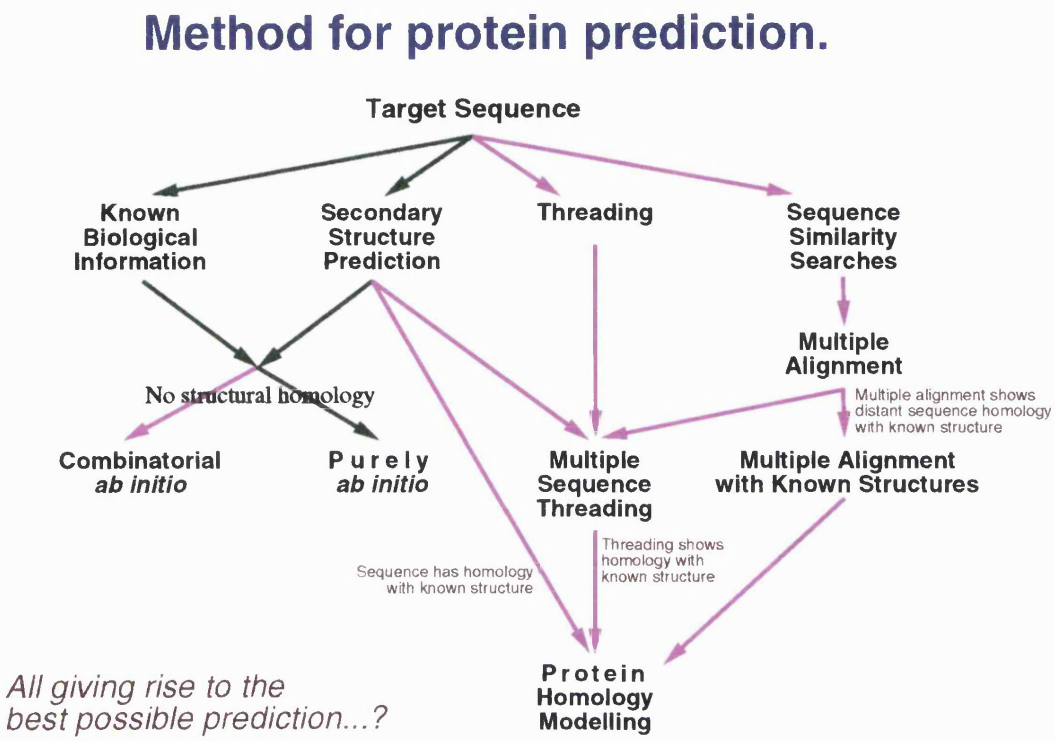


Figure 1.1: Flow diagram showing methods for CASP2 predictions.

1.1 Sequence similarity

The identification of any clear sequence similarity with a protein of known structure is the most certain way to infer the structure of a protein from just its sequence. This is a strong principle since, through evolution, the amino acid sequence can change (through conservative substitutions) much more than the structure itself, although exceptions do exist.Any sequence similarity therefore implies a similar structure and explains the importance of developing methods that can detect the most elusive of similarities, as even from these, some structural inference can be made.

Even without similarity to a protein of known structure, the alignment of other sequences can still be very helpful as they reveal the evolutionary constraints imposed

at each position of the sequence. As these constraints are often directly related to the local structure, then some idea of that structure can be inferred. Often the more sequences (across a wide phylogenetic range) that can be aligned, then (assuming a common structure) the greater is the information which can be gleaned about each position. With diverse alignments like these then there will be a better chance of predicting the structure.

Using methods to identify sequences with similarities to the sequence of interest can also give direct insight about function. The protein query found may be similar to a well characterised family of proteins about which the function, if not the structure is known. Function may also be inferred from alignment where conserved functional residues are identified.

1.1.1 Pairwise sequence alignment

For the structure of a protein to remain consistent with its functionality, a reasonable assumption which can be made is that the main core of a protein should remain well conserved and that its secondary structure elements are similarly arranged across a family of proteins. This arrangement of secondary structure elements is referred to as the protein architecture (i.e the spatial orientation of the elements). The order of connectivity of these elements is the fold or topology. The connections between secondary structure elements are referred to as loops and turns. These can be highly variable in length and are areas in which substitutions and deletions of sequence pref-

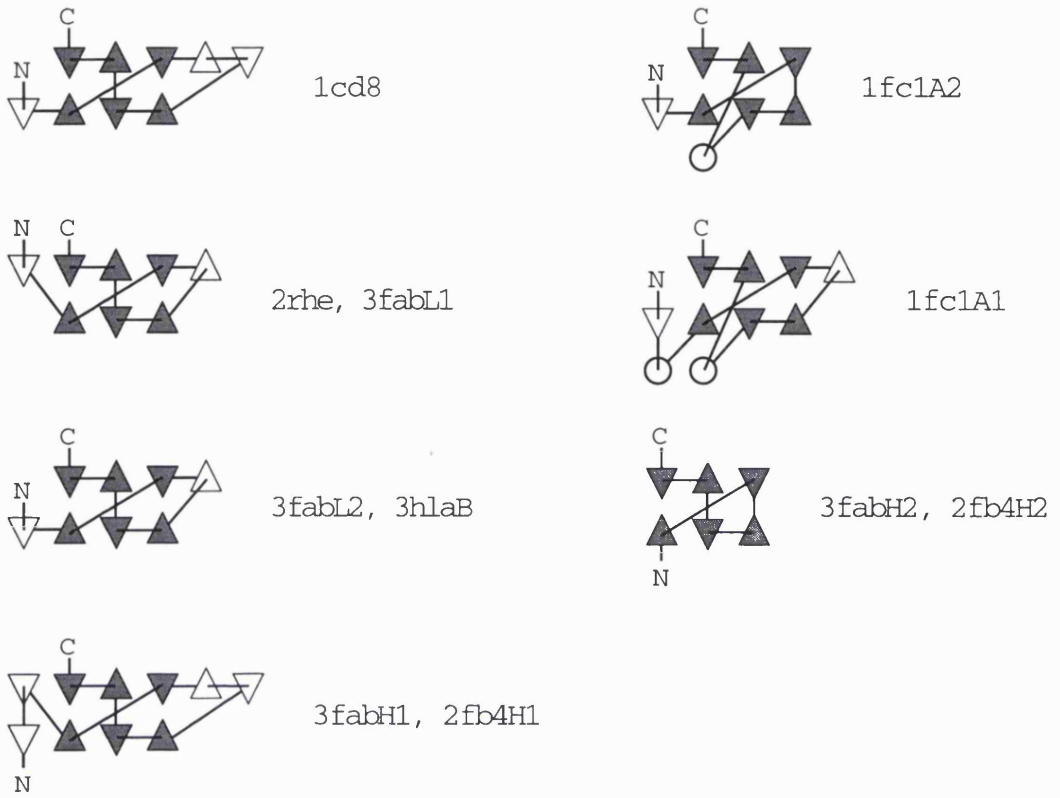


Figure 1.2: Topology Diagrams. Illustration of different domains from the immunoglobulin family of proteins. The folds represented as topology diagrams (TOPS representation (Flores *et al.*, 1994)). The conserved core of the domains, characteristic of the Ig family is shaded. Triangles represent a strand, with circles for helices; the lines show connectivity either above or below the secondary structure.

erentially occur. Whole secondary structure elements and even complete domains can be inserted within a family of folds, but the main "defining" core still exists. This can be best illustrated by Figure 1.2, which shows a series of topological representations of a family of proteins. Loops are often vital for protein function as they are typically comprised of the functional residues or may bind ligands and nucleotides. In such cases these loops are well conserved. One such example is the 'P' loop which is important in binding adenosine tri-phosphate (Saraste *et al.*, 1990).

Amino acid similarity When comparing two protein sequences (represented as strings of characters) some measure of amino acid similarity must be known. For nucleic acids, simply counting identical matches is sufficient, but with amino acids, the variety of chemical and physical properties that they exhibit requires a more graded matching scheme. Many have been devised over the years but those most commonly used are based on empirical counts of observed substitutions between related proteins.

When sequences are evolutionarily distant the problem of identifying the similarity (if indeed it exists) between the two sequences is a challenging one. The simplest model would be to construct an identity matrix where a high matching score would be achieved if, at a particular point in the comparison, the same amino acid type was found in each sequence. The more self-self matches achieved, the better the alignment. This idea is all very well for reasonably similar sequences (more than 50% similarity). However, when trying to identify the best pairwise alignment when the sequences have a lower similarity than this, then something different is called for. A more discriminating matrix system has been devised where the evolutionary likelihood of a particular amino-acid mutating into a different amino-acid has been estimated from aligned sequences.

The most widely known series of substitution matrices are the evolutionarily accepted point mutations (PAM) matrices (Dayhoff *et al.*, 1978) although a more recently developed series called the BLOSUM matrices are now also widely used (Henikoff and Henikoff, 1992). An updated PAM matrix called JTT (Jones *et al.*, 1992b) based on more recent sequence data is also sometimes used. Matrices have also been

generated over the entire protein sequence database (Gonnet *et al.*, 1992). In general all matrices contain scores for substitutions which are higher if size and hydrophobicity are conserved.

Gap penalty For more distantly related sequences it is necessary to introduce relative insertions and deletions into both sequences to attain a maximum matching of amino acids. However, the inclusion of such gaps cannot be be allowed to occur without some cost to the score, otherwise, to take an extreme example, a short protein aligned with very long sequence could insert gaps between every residue and eventually find a perfect match for every position.

To avoid such un-biological alignments, but still allow the possibility of insertions and deletions (indels), we introduce a gap-penalty. Sometimes this is a fixed penalty for a gap of any length (Needleman and Wunsch, 1970), but in most cases the penalty is made up of a gap creation penalty and a gap extension penalty.Usually a higher gap insertion and then once the gap is opened a lower, but still incremental, penalty is added for each further gap – this is the gap extension penalty. Hence the gap penalty can be written as: $g = a + bn$ (where g is the applied penalty, a and b are the opening and extending parameters and n the number of spaces in the gap) (Gotoh, 1982; Altschul and Erickson, 1986). Other methods have used a gap penalty for opening a gap and then penalise by the logarithm of the length of the gap (Miller and Myers, 1989).

When a gap penalty is too high the pairwise alignment lacks sensitivity. If too low the

alignment will be very dispersed making it hard to recognise a sensible evolutionary relationship. Often the alignment of distant sequences gains sensitivity when low gap penalties are used. In such cases large indels may be detected. There has been much work on the significant role a gap penalty plays in the comparison of sequences, but the only general rule is that the size of the penalty is dependent on the size of the numbers in the amino acid substitution scoring scheme. Further aspects and various alternative gap penalties are discussed in some reviews (Pascarella and Argos, 1992; Vingron and Waterman, 1994; Taylor, 1996).

1.1.2 Alignment algorithms

When aligning two sequences in order to gain an idea of the evolutionary relationship between them, a large number of gaps may or may not need to be inserted to get an optimal alignment. For one gap in a sequence of four letters there are five possible options. For two gaps there are 15 possible alignments to be considered. So generally there are many possible combinations and for large sequences and arbitrary gap sizes we have a combinatorial explosion. Fortunately, to cope with this problem, a very simple and effective computer algorithm has been developed called Dynamic Programming (DP) (Needleman and Wunsch, 1970; Sellers, 1974; Smith and Waterman, 1981b). DP is used (in one form or another) in many methods that align sequences (and even structures (Taylor and Orengo, 1989; Orengo and Taylor, 1996)). There are other methods which do not use DP, although the underlying

paradigm is the same (Notredame and Higgins, 1996). One method which is different uses graph theory (Reinert *et al.*, 1997).

Dynamic Programming With DP the placement of insertions and deletions (sometimes referred to as indels, but more commonly just gaps), as well as the similarity of different residues, can be taken into account. Thus, using DP, we calculate the highest scoring alignment between two sequences in a time proportional to the product of the lengths of the sequences.

To find the optimum pairwise alignment a matrix is constructed where one sequence is placed along each axis. For each element in the matrix a score is calculated based on the match between amino acids. DP will find the highest scoring path through the matrix (taking gaps into account) and, in theory, the best alignment for the given scoring scheme. When used in combination with a residue exchange matrix and gap penalties, DP affords a single optimal highest scoring alignment with its score as the summed exchange values over the matched position minus the penalty values for the inclusion of gaps.

The weakness in sequence alignment is not the DP algorithm but the uncertainty in what constitutes the best gap penalty and substitution matrix. Given this inherent uncertainty, there is a limit to the information which can be gained from a pairwise comparison; so, if possible, a multiple alignment should be considered. Pearson and Miller have a good review of DP algorithms in sequence comparison (Pearson and Miller, 1992).

1.1.3 Homologous sequence searching

Some fast methods for searching for sequences which are homologous to a sequence of interest have been developed over the last ten to 15 years. An early program was FASTA and FASTP (Lipman and Pearson, 1985; Pearson, 1990; Pearson and Miller, 1992) and has since been followed by BLAST (Altschul *et al.*, 1990). As DP is an intensive process, when used on thousands of sequences the results of a comparison are very slow. Both FASTA and BLAST contain shortcuts to reduce the computational time they take to run. With the advent of larger computers and parallelization of the algorithms it is now possible to run a full DP search over the sequence databank. This has been implemented as BLITZ (Smith and Waterman, 1981a) and, in theory, this would be the best method to search for similar sequences. However, there even exists BLAST servers running in parallel making sequence searches incredibly fast.

FASTA FASTA uses a fast technique roughly to locate regions of similarity, within which DP is then used to extract an alignment. The program uses hash coding of, typically, dipeptides in a query and scans the database counting the hits on each diagonal of the alignment. It then re-scores the areas of high score by allowing some amino acid substitution and shorter lengths of identity and then joins the best scoring regions from different diagonals. This is followed by a DP cycle to find the optimal alignment.

BLAST BLAST (which stands for Basic Local Alignment Search Tool) has a specific protein version (BLASTP) while other variants include BLASTN – for nucleotide searches and BLASTX – for nucleotide to peptide conversion before a peptide search and TBLASTN to search for a peptide in a nucleotide database which is converted to peptide. The program uses a very rapid search algorithm, not unlike FASTA, but more flexible. Developed by Altschul *et al.* (Altschul *et al.*, 1990), the basic algorithm is simple, robust and offers an increase in speed over FASTA. Methodological improvements in BLAST now take gaps into account and searches can be performed iteratively where the results of one search are used as a basis for the second search. These programs are called Gapped BLAST and PSI-BLAST, respectively (Altschul *et al.*, 1997). PSI-BLAST is now the accepted standard program to use for similarity searches.

BLITZ BLITZ uses the Smith and Waterman algorithm (Smith and Waterman, 1981a) and no pre-filtering of the data. A similarity matrix is used to compare the target sequence with those in the database, the algorithm searches for the highest local match and takes into account a gap penalty. The best results are ranked but only the highest match for any one sequence is given. Other searchers are also available which run on massively parallel machines (MasPar)(Collins and Coulson, 1990) or Biocellerators (BIC's).

Generally, all the methods give similar results when searching for proteins that are moderately related to the probe. It is worth submitting the sequence to several search methods so that the likelihood of missing a relevant match is reduced.

BLAST can be performed interactively on the WWW but the other methods generally need a search form which sends the results back by e-mail once the jobs have been completed on a remote computer.

1.1.4 Multiple alignment

To gain further information from protein sequences and the evolutionary information that they contain, a multiple alignment can provide a wealth of information. Several methods exist for going beyond a simple pairwise alignment; probably the most common method available for building multiple alignments is that of Tree or Hierarchical methods. The method assumes that a multiple alignment can be built from successive pairwise alignments. The first step in the method is a comparison of all the sequence pairs to align. So for N sequences there are $(N \text{ x } (N-1))/2$ pairwise comparisons. Cluster analysis of the comparisons is performed to give a *tree* or *hierarchy* of the sequences from which a multiple alignment is constructed by the pairwise comparisons indicated by the tree. The Neighbor-Joining evolutionary method (Saitou and Nei, 1987) is most often used for the cluster analysis as it gives a reasonable and fast comparison.

In theory, another method would be to extend a simple pairwise comparison, using DP, to a multiple comparison. So instead of constructing a 2D matrix and finding the optimal scoring path through it, the method would build a 3D or nD matrix (one dimension per sequence) and find the best route through the matrix. In practice this

is possible in 3D, but the time is proportional to N^5. Higher dimensional comparisons would take far too long to be practical, although with the use of shortcuts to localise the searches through the matrix it is possible to an extent. This method incorporating the shortcuts is used by the program MSA (Lipman *et al.*, 1989), which can handle alignments of up to eight sequences, each 200–300 amino acids in length.

The major packages for multiple alignment are CLUSTALW or CLUSTALX (Thompson *et al.*, 1994; Thompson *et al.*, 1997; Jeanmougin *et al.*, 1998) (formerly CLUSTALV) (Higgins *et al.*, 1992), PILEUP (Feng and Doolittle, 1987) (which is part of GCG) and CAMELEON (which incorporates `MULTAL`) (Taylor, 1988). It is more widely becoming the case that programs for multiple alignment (particularly CLUSTALW) are available over the `WWW`. Web sites such as SRS (Sequence Retrieval Service) are incorporating CLUSTALW into easy to use, form style, interfaces. Although these interfaces do not give the full range of options found on command line based versions, they are nevertheless a good starting point for constructing a multiple alignment. CLUSTAL is also available in an X interface format (CLUSTALX). For review see (Doolittle, 1990).

CLUSTAL and MULTAL both use a similar method for generating a multiple alignment. They use a hierarchical approach where all sequences are compared in a pairwise fashion using dynamic programming, a cluster analysis if then performed on the pairwise information and a hierarchy for alignment is generated. The multiple alignment is then built up by aligning the most similar pair of sequences and then the next most similar, according to the tree, until all sequences are aligned. MULTAL

also offers the possibility to recompute the guide tree at every step of the alignment and thus can incorporate a consensus sequence approach.

1.2 Secondary structure prediction

Using a multiple alignment, identifying well conserved areas and, in particular, areas of conserved hydrophobicity, can give a good indication of secondary structure. Secondary structure prediction methods have been much improved since the simple statistical methods of Chou and Fasman (Chou and Fasman, 1974; Chou and Fasman, 1978), particularly when they include the information gained by multiple sequence alignments. In the next section are some of the most widely used and available methods for predicting secondary structure.

In globular proteins in solution the more hydrophobic amino acids in a sequence will tend to lie towards the centre of a protein, away from the surrounding water. The hydrophobic side chains will pack into the interior of the molecule. For proteins of any reasonable size this creates a problem, because the backbone has polar atoms and is therefore hydrophilic and not easily buried in a hydrophobic protein core. To overcome this problem the amino (N-H) and carbonyl (C=0) groups in the backbone form hydrogen bonds and so their partial charges are neutralised. (The NH group is a H-bond donor and the C=O a H-bond acceptor). In the core of the protein, hydrogen bonds can form in two distinct ways: 1) a helix where the CO of residue n bonds to

the NH of residue $n + 4$; 2) a sheet, which is comprised of extended strands forming parallel or anti-parallel interconnections often between distant parts of the sequence. These are referred to as secondary structure.

Depending on how buried the secondary structure elements are in a protein, the pattern of conserved hydrophobic amino acids (indicating structural importance) can indicate what type of structure is present. For example a buried β-sheet will have a run of hydrophobic residues, whereas a partially exposed beta will contain alternating hydrophobic/hydrophilic residues. This is a slight over simplification, but it is possible successfully to predict secondary structure in this way.

1.2.1 Statistical methods

Analysis of proteins with known structures gives some idea of the propensities for different amino acids to occur in different secondary structures. For example Gly and Pro are often found at or in turns and at the ends of α-helices (Richardson and Richardson, 1988).

Chou-Fasman method Statistical methods for predicting secondary structure were developed by Chou and Fasman. They conducted a statistical study of protein structure to attempt to map the secondary structure from sequence (Chou and Fasman, 1974; Chou and Fasman, 1978). Not only did they examine whether a residue prefers to be in an alpha or beta state, but also classified the residues into six classes

depending on their likelihood to form or disrupt an alpha or beta structure.

GOR method The method, developed by Garnier, Osgothorpe and Robson in 1978, used four likelihood profiles to represent an alpha, beta, turn or coil (Garnier *et al.*, 1978; Gibrat *et al.*, 1987). Each likelihood profile is a function of a 17 amino acid window around the position of interest. To compute a probability for a position the 17 values and the corresponding positions of the surrounding residues are added up to give a score for each of the four states.

1.2.2 Neural networks

One of the earlier methods to use neural networks was based on a non-linear model (Qian and Sejnowski, 1988). Probably the most widely used method for predicting the secondary structure of proteins is the predict protein server based at EMBL, Heidelberg (Rost and Sander, 1993). The method, called `PHD` is based upon a series of trained neural networks. `PHD` is particularly useful as it constructs a multiple alignment and uses this additional information to predict the secondary structure. Additionally, several formats of multiple alignments can be submitted to `PHD` to give a secondary structure prediction of a specific alignment.

Another neural net based method is `NNPREDICT` (Kneller *et al.*, 1990).

1.2.3 Other secondary structure prediction programs

A fast and simple approach to secondary structure prediction from multiple alignments is available as SSPRED (Mehta *et al.*, 1995). The method uses aligned homologous sequences and structures to derive residue exchange statistics for each secondary structure type. The prediction for a given multiple alignment is calculated by correlating particular types of mutations known to prefer one of the secondary structure states, based on the derived statistics.

Another approach which uses the basic, but important, aspects of secondary structure prediction and then combines them using statistics is DSC (Discrimination of Secondary structure Class) (King and Sternberg, 1996). DSC combines GOR potentials, gaps in the multiple alignment, mean moment of conservation, mean moment of hydrophobicity and some other attributes to give a prediction. This approach has the advantage of being easy to understand and simple to run, when compared with the "black-box" workings of neural network methods.

1.3 Protein folds

More and more folds are being solved by the experimentalists, but in simple terms proteins can be classified into just a few classes and subclasses. SCOP and CATH classify protein folds at many levels, but the primary classification comes with the assignment of protein class (mainly alpha, mainly beta and alpha-beta).

There are more common motifs which occur with high frequency in the protein database. Many of these different fold types have now been classified. Shown in Table 1.1 is an example compiled by the CATH (Orengo *et al.*, 1997) database, just showing the major fold types in the mainly alpha, mainly beta, alpha beta and few secondary structures categories.

Mainly Alpha	Mainly Beta	Alpha Beta	Few Sec. Str.
Non-Bundle	Ribbon	Roll	Irregular
Bundle	Sheet	Barrel	
Few Sec. Str.	Roll	2 Layer s/w (ba)	
	Barrel	3 Layer s/w (aba)	
	Clam	3 Layer s/w (bba)	
	Sandwich	4 Layer s/w (abba)	
	Distorted s/w	Box	
	Trefoil	Horseshoe	
	Orthogonal Prism	Complex	
	Aligned Prism	Few Sec. Str.	
	4 propellor		
	6 propellor		
	7 propellor		
	8 propellor		
	2 solenoid		
	3 solenoid		
	Complex		

Table 1.1: Current CATH fold database categories. This table shows the current categories into which CATH subdivides all proteins. Sec. Str. = Secondary Structure. s/w = sandwich.

1.4 Structure comparison/classification

Most proteins have similarities with other proteins and many structural similarities are conserved better than their amino acid sequences. In general, indels in a se-

quence occur within loop regions. Therefore, fold families have related structure, but not necessarily highly related sequence. Similar folds without any significant sequence similarity are termed analogous, suggesting that the same fold has been arrived at from different evolutionary starting points; whereas if common evolutionary origin is implied by clear sequence similarity, the term homologous is used. It is usually the case that sequences with >30% similarity adopt the same fold. One of the major problems with sequence comparison at low levels of similarity is that remote homologues can also be picked up. It may be that by examining a family of sequences then the balance can be redressed and help with detection of similarity in this 'twilight-zone' (Taylor, 1995b).

Structure comparison between proteins can be carried out computationally. For different comparisons, the better methods involve the characterization of a local structural environment for each position in a protein (Taylor and Orengo, 1989; Sali and Blundell, 1990). In one of these methods a vector set of all inter-atomic distances for each point in the two structures was compared and from that the relative similarity of their respective positions. A similarity was derived for all pairs, and from this an optimum alignment of the structures obtained (Taylor and Orengo, 1989).

1.4.1 Structural classification databases

There are three major classes of proteins: all alpha, all beta, and alpha-beta (Levitt and Chothia, 1976). Further sub-groups and classifications can be made and the study of how proteins are related is kept up to date in publicly available databases.

There are about a few hundred classified folds we know of with an expected maximum estimate of 1400 from sequence analysis. Bearing in mind structural comparison a more conservative estimate of 1000 folds has been proposed (Chothia, 1992). There are several ways in which protein structures have been classified. Two widely available databases on the subject are SCOP and CATH.

SCOP is a highly comprehensive description of the structural and evolutionary relationships between all known protein structures (Murzin *et al.*, 1995; Hubbard *et al.*, 1997). The hierarchical arrangement is constructed by mainly visual inspection in conjunction with a variety of automated methods. The principal levels are Family, Superfamily and Fold. The Family category contains proteins with clear evolutionary relationships. The Superfamily have probable common origin; there may be low sequence identity, but structure and functional details suggest a common origin. The Fold level groups together proteins with the same major secondary structure elements in the same arrangement with identical topological connections.

CATH is a hierarchical domain classification of structures with a resolution better than 3.0Å and also NMR structures (Orengo *et al.*, 1997). The database is constructed wherever possible by automatic methods. The four levels in the hierarchy are: class

(C), architecture (A), topology (T) and homologous superfamily (H), a further level called sequence (S) families is sometimes included. "C" classifies proteins into mainly alpha, mainly beta and alpha-beta, which includes both α/β and $\alpha+\beta$. "A" describes the shape of the structure, or fold. "T" describes the connectivity and shape. "H" indicates groups thought to have a common ancestor, i.e. homologous. "S" structures are clustered on sequence identity.

SCOP and CATH can both be accessed via the `WWW`, respective URL's are:
`http://scop.mrc-lmb.cam.ac.uk/scop/` and
`http://www.biochem.ucl.ac.uk/bsm/cath/`.

1.4.2 Topological prediction

Predicting the way in which the secondary structure elements of a protein fold are connected can be difficult. For any one fold there are a number of ways to connect the elements together. One rule to observe, in general, is that of the chirality (or handedness) of the connections between secondary structure within proteins. In the majority of protein folds a right handed topology is maintained throughout. The handedness of proteins is thought to have been maintained throughout evolutionary time, since a hypothetical chance "decision" between left and right handed conformations at an early stage during evolution (Mason, 1984).

Chirality is the spatial arrangement of atoms, or super-secondary structures, such

that one structure is non-superposable on its mirror image. In proteins a helix is always right-handed, i.e. it has a clockwise rotation down its axis. Proteins also have axes, so looking down a protein or subunit along an axis and following the N-terminal to the C-terminal then if the direction is clockwise then the protein has right handedness. Figure 5.4 in Chapter 5 illustrates the handedness of four helix bundles.

1.5 Homology modelling, threading and *ab initio* modelling

When a sequence has high homology to a PDB structure then models can be generated using homology modelling programs with a very high accuracy. Often the models generated will be within 2Å RMSD of the X-ray structure when it is solved. The hardest problem is often modelling loops which do not occur in the template structure and where these may be functionally relevant then much of the modelling time can be concerned with these regions.

Where no sequence homology can be determined, then threading can be employed to find likely folds which may be suitable. If there is a high confidence in the predicted fold then a threading based homology model can be built, see Chapter 4. Although with a poor alignment, which is often a problem in threading methods, then relatively poor models are generated when compared to the crystal/NMR structure, perhaps models with 3 to 6 Å RMSD on just the backbone. Potentially,

though, if a correct fold is predicted then accurate models such as those produced by homology modelling can be produced. The hardest aspect of using threading based models is in being sure that a correct fold has been identified. Schoonman *et al.* have evaluated threading verses homology modelling and conclude that there is still a 2Å difference in accuracy between the two methods (J. *et al.*, 1998). All threading programs form a rank of predicted structures usually based on some normalised Z score, which is very much dependent on the quality of the potentials used in the program. Improvements in threading generally concentrate in improving the potentials used to score the threadings (Sippl, 1990; Nishikawa and Matsuo, 1993; Jones and Thornton, 1996).

Only when there is no structure with any sequence homology and no confident threading prediction do purely *ab initio* methods have to be employed, as in the case of Chapter 5. One of the worst cases would be to use a template model based on the incorrect fold, at the moment it is often difficult to determine when this may occur.

The next three sections discuss these different ways of modelling proteins.

1.6 Homology modelling

Sometimes it is possible to align sequences which already have a known structure. If this is the case then the alignment of the sequence with unknown structure and that of the known structure can be directly or indirectly aligned by the multiple sequence

alignment. This method is frequently also referred to as homology modelling. The first application of this idea was by Browne and co-workers (Browne *et al.*, 1969) and later refined by Greer *et al.* (Greer, 1981) and Blundell *et al.* (Blundell *et al.*, 1987; Blundell *et al.*, 1988).

Fragment based homology modelling is a technique where models of proteins can be constructed from separate fragments of other proteins. Areas where there are inserted residues, and no structure in the homologue, can be built using fragment matching (Jones and Thirup, 1986).

COMPOSER (Sutcliffe *et al.*, 1987a), a comparative modelling program, can derive an average framework from a series of homologous structures and then use that as a base for constructing a structure from homologous fragments. A following paper also by Sutcliffe *et al.* showed how rapidly to model side chain positions on such a model (Sutcliffe *et al.*, 1987b). Another method, MODELLER, optimally satisfies structural restraints derived from an alignment with one or more structures. These restraints are expressed as probability density functions (pdfs) for each feature, where a feature may be solvent accessibility, hydrogen bonding, secondary structure, etc. at residue positions and between residues (Sali and Blundell, 1993).

Comparative modelling has been reviewed many times (Greer, 1991; Sali, 1995). Where 40% sequence identity exits with a known structure then homology models with high accuracy can be generated, see the review by Sali (Sali, 1995).

Based on a multiple alignment with known homologous structures, distance restraints

can also be derived for a sequence with unknown structure and then solved using distance geometry (Havel and Snow, 1991; Havel, 1993; Srinivasan *et al.*, 1993; Brocklehurst and Perham, 1993; Sudarsanam *et al.*, 1994; Aszódi and Taylor, 1996).

A web-based homology modelling package is also available and can give some very good models based on a single sequence. Further refinement can also be made, such as adjusting the alignment and specifying which specific structures to use when model building (Peitsch, 1996). The program, called SWISS-MODEL is a widely available and can be accessed as a web based server for Homology modelling (Guex and Peitsch, 1997). Sanchez and Sali have recently modelled proteins on a very large scaler. They used an automated analysis on the yeast (*Saccharomyces cerevisiae*) genome and generated 1071 structures (Sanchez and Sali, 1998). Projects like this will enable easier analysis of the large amount of sequence data being generated by genome projects around the world.

Due to the large number of already solved folds it can be expected that more and more sequences will in fact have homologous structures. Hence, using these methods will play an ever more important role in model building.

1.7 Fold recognition

Fold recognition, or threading, is a process whereby a sequence with unknown structure is compared to a database of structures with different folds. In making the

comparison the structure which the unknown sequence finds a best fit for is taken to be the fold that the sequence will most likely form.

At the moment, where not every possible fold has been identified, one of the hardest areas of threading is to recognise when there is no fold similar for the sequence. In such a case one has to resort to *ab initio* prediction.

Fold recognition falls broadly into two categories, one method which uses pairwise energy/interaction potentials and the other which performs a 1D to 3D comparison.

1.7.1 Pairwise energy potentials

Pairwise potentials are any measure which can be used to classify a residue:residue interaction, or atom:atom interaction.

THREADER is a program which takes an empirical potential map of a protein and fits (or threads) the target sequence on to the structure of the known protein (Jones *et al.*, 1992a; Jones *et al.*, 1993). The targets are compared to a database of non homologous proteins, this is performed in 3-Dimensional space. The THREADER output for the target sequence can be ranked according to several scores and the structures which score significantly well may be correctly associated with the target sequence.

Another method derives knowledge-based force fields from known structures. The sequence is then compared to a fold database and the corresponding energies cal-

culated to give an indication of the predicted best-fit fold (Hendlich *et al.*, 1990; Sippl and Weitckus, 1992).

All the above methods use a single sequence. To use the information in multiple sequences, a Multiple Sequence Threading (MST) has been developed which compares multiple sequence information with a database of structures to determine the correct fold (Taylor, 1997; Taylor and Munro, 1997).

1.7.2 1D/3D comparison

Multiple sequence information is also used in the simpler 1D/3D fold recognition methods which perform a secondary structure prediction on the sequence of interest and then compares that secondary structure with all the secondary structures in sequences with known structures to find a possible match.

Methods which fall into this category include: TOPITS (Rost, 1995), MAP (Russell *et al.*, 1996) and H3P2 (Rice and Eisenberg, 1997).

Threadings for CASP2 are illustrated and discussed in Chapter 3.

1.8 *Ab initio* modelling

Ab initio modelling, or *de novo* folding, is not based on any template structures, but

rather on a secondary structure assignment and various sets of constraints. It does not mean that the methods try to mimic the biological folding process. Several approaches to the problem of *ab initio* modelling have been considered (Kim and Baldwin, 1982; Dill, 1985).

Many *ab initio* approaches involve simplifying the model of the protein to make it easier to handle. Once a rough model of the protein has been created a series of further steps can be applied to build the protein into a full atom representation.

Using high resolution crystallographic data, sets of fragments can be derived and then reassembled into new folds (Jones and Thirup, 1986; Jones and Thornton, 1993). This idea was carried forward when Jones modelled the NK lysin target for CASP2 (Jones, 1997). The protein has also been modelled outside the CASP2 assessment (Dandekar and Leippe, 1997). See Chapter 5 for my modelling prediction of this target.

The reason for using any *ab initio* method is that the models are not restricted to a known fold and can be used to model proteins with no known fold in the databank. Computing power and the complexity of the problem still limit the uses and success of *ab initio* in protein structure prediction. Probably the most common of these is the combinatorial approach.

1.8.1 Combinatorial method

One of the first steps in any prediction is to try and identify the secondary structure elements. Combinatorial methods try to explore all the combinations of arrangements of the secondary structure elements. These methods generally try to pack hydrophobic residues in the core of the protein (if it is a globular protein) (Cohen *et al.*, 1980; Sternberg *et al.*, 1982; Taylor, 1993). Some use a framework or lattice on which to base the secondary structure elements (Taylor, 1991). The most successful of these combinatorial approaches modelled α-helix proteins.

1.8.2 Lattice models

Other *ab initio* methods such as lattice models can be used to model proteins (Hinds and Levitt, 1992). These models are often preferred as they impose a reduced number of conformations a protein can take. Covell has folded simple protein chains into compact forms based on lattice folding and Monte Carlo methods (Covell, 1992) as have others (Skolnick and Kolinski, 1991; Kolinski and Skolnick, 1994; Reva *et al.*, 1997). Lattice models have been used to test potentials and force fields used in protein folding prediction (Reva *et al.*, 1998).

1.8.3 Distance geometry

Distance geometry has been in use for many years in the field of protein prediction and modelling (Mackay, 1974; Crippen and Havel, 1988; Kuntz *et al.*, 1989) but these techniques have rarely been applied to *ab initio* folding. Distance geometry is used more commonly applied to homology modelling (Havel and Snow, 1991; Havel, 1993; Srinivasan *et al.*, 1993; Sudarsanam *et al.*, 1994). NMR spectroscopists also use distance geometry for structure elucidation.

One such distance geometry method is incorporated into a program called `DRAGON` (Aszódi *et al.*, 1995a; Aszódi and Taylor, 1996). A simplified model chain is folded by projecting it into gradually decreasing dimensional spaces whilst subjecting it to a set of defined restraints, primarily secondary structure. In this way the geometry space is successfully explored to produce a protein backbone. The method generates many folds in a short time using an embedding algorithm incorporated into the program (Aszódi and Taylor, 1997).

1.8.4 `DRAGON`

`DRAGON` stands for Distance Regularisation Algorithm for Geometry OptimisatioN. The program has been developed in the lab by Drs. Aszódi and Taylor over a three year period.

`DRAGON` builds *ab initio* models from secondary structure predictions and multiple

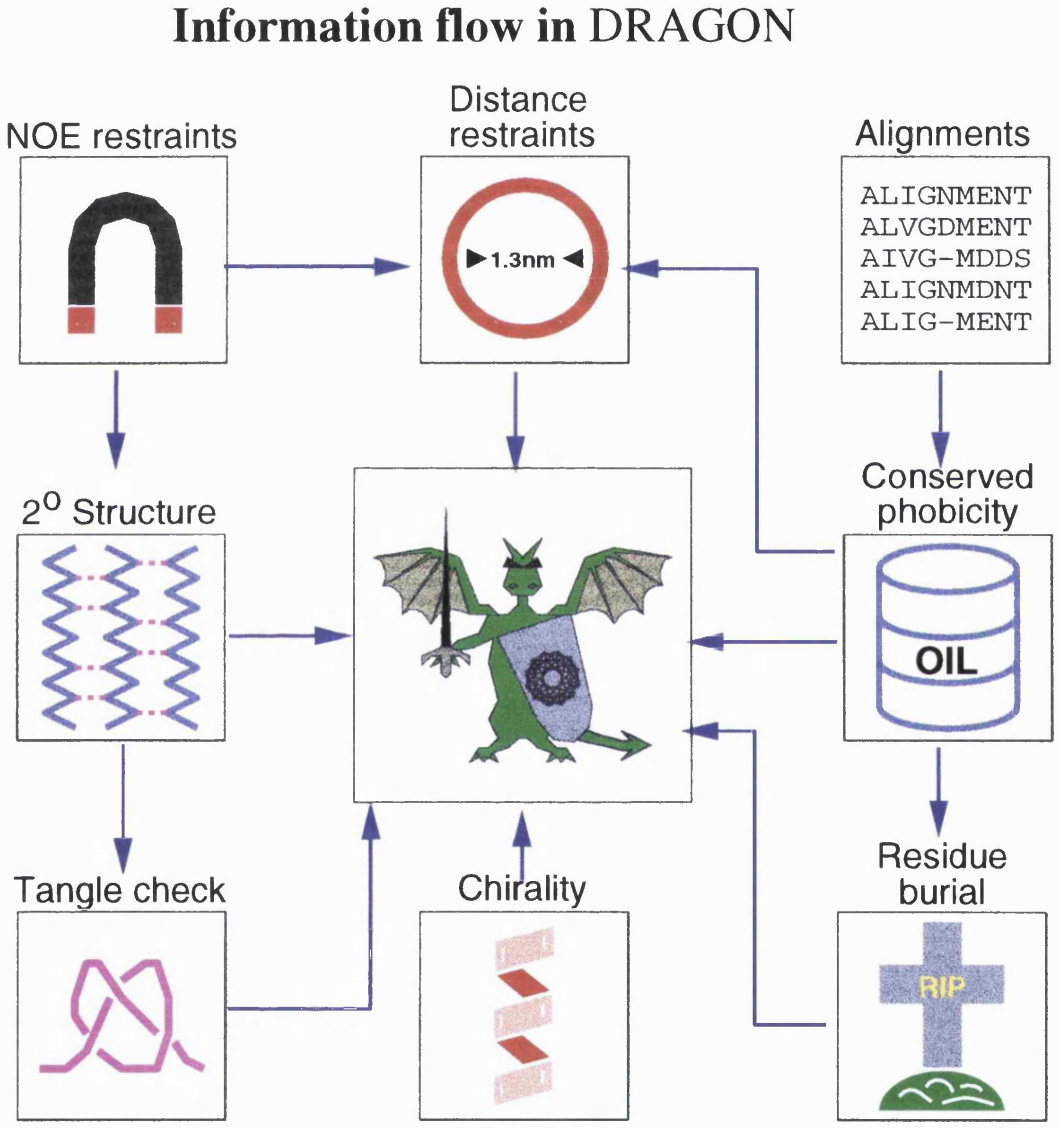


Figure 1.3: Flow diagram showing input to DRAGON. Figure courtesy of András Aszódi.

alignments using distance geometry, incorporating a hierarchical projection method. At the bare minimum all you really need to make a model is one sequence and a secondary structure prediction. In practice the more information the better. Figure 1.3 shows the different types of information which can be incorporated into DRAGON to produce a model.

Overview of the method

Hydrophobic residues are known to tend to cluster together, whereas hydrophilics prefer larger inter-residue distances. With this knowledge in mind a method to model proteins based on distance geometry was developed by Aszódi and Taylor (Aszódi *et al.*, 1995a; Aszódi and Taylor, 1996). A simplified model chain is folded by projecting it into gradually decreasing dimensional spaces whilst subjecting it to a set of defined restraints, primarily secondary structure. In this way it is possible to explore the geometry space to produce a range of protein backbones that satisfy the restraints.

The simple polypeptide chain has a fixed backbone distance of 3.8Å with a virtual bond angle of 1.82 rad. There are no restraints placed upon the bond torsion angle, based on Levitt (Levitt, 1976a).

Using distance geometry the geometric arrangement of any set of points can be defined by the distance matrix of the set of points. That is to say, all the inter point distances in distance space, although chirality is not characterised. From this a metric matrix

is calculated and then by finding its eigenvalues and eigenvectors it is diagonalised. A preset fraction of eigenvectors was kept (95%) and used in the projection. The projection causes the set of points to change shape due to the slight loss of information. A new distance matrix is constructed and the process cycles until a projection into three dimensions is achieved, by progressively losing eigenvectors until three remain. Accessibility adjustments are carried out in both distance space and Euclidean space. In the former, the *cones* method (as described in Chapter 2) operates on pairs of residues which are far apart, exposed and at least one of which is hydrophobic, the distance of which inter-atomic distance is then reduced. In Euclidean space exposed hydrophobics are moved towards the centroid and buried hydrophilics are moved outwards. By adjusting local conformations of hydrogen bonds to their ideal distance constraints, secondary structure elements can be defined. Bump violations between all the atoms are also considered, and are dependent on the side-chain relative C_β size. Other adjustments are made such as helical handedness, hydrogen bond handedness, hydrogen bond geometry and topology, residue density and hydrophobic interactions – these are described in detail in a paper by Aszódi and Taylor (Aszódi and Taylor, 1994).

In summary, at each iteration the eigenvalues and vectors of the metric matrix are calculated. If there are any negative eigenvalues then these are discarded, otherwise the dimensionality is decreased by removing the smallest eigenvalue. When three eigenvalues remain then three dimensions have been achieved. The program then does some further refinement in three dimensional space trying further to minimise

the violations within the distance restraints, the best model which satisfies most of the restraints imposed on the protein is then saved to a file. Each run of DRAGON produces one model and due to the speed of the algorithm many models can be generated in a short space of time.

Chapters 4, 5 and 7 show applications of DRAGON to several of the CASP2 targets and other proteins.

1.9 Classic modelling example

With the initial whirlwind of research into AIDS and HIV during the 1980's a great emphasis was put on getting high resolution crystal structures as fast as possible. An obvious protein to study is the HIV protease which cleaves several of the component proteins required in the construction of the HIV virus. If this protein could be disrupted in some way then the proliferation of the virus could be slowed or even halted. To gain some idea of what drugs would make good targets for interfering with the protease a structure was needed.

Prediction and modelling work carried out by Pearl and Taylor (Pearl and Taylor, 1987b) was able to produce a structure for HIV protease before any high resolution crystal structure was available and in this way aid in the development of a widely used AIDS drug.

As this chapter has suggested, the best way to go about this was to find similar sequences and construct an alignment. The multiple alignment was able to show that the HIV protease had an evolutionary relationship with a family of aspartic proteases – in particular an Asp-Thr-Gly motif conserved in the active sites of both families. One of the aspartic proteases had a known structure which enabled the construction of a model using an HIV protease sequence (homology modelling). The sequence was aligned on the backbone structure and several possible secondary structural elements were predicted and aligned with the homologue. This rough fitting method was used to find what they thought was the best alignment considering structure and was an early precursor to the threading method discussed earlier.

The final sequence was modelled on to the carbon backbone of the aspartic protease to give a model for the HIV protease.

Analysis of this model after a high resolution structure was published showed a good RMSD of 2.1Å for all atoms in the model. Particularly high accuracy was achieved at the important substrate binding site (1.2Å) (Weber, 1990).

1.10 Where does all this leave us?

All this sounds very promising, but is it really possible to predict a 3D atomic structure of a protein from its sequence alone? Homology modelling has shown itself to be successful, as outlined in the above example. As yet, little success has been achieved

from purely an *ab initio* standpoint. Methods for fold recognition may in time become more reliable with fewer false positives or false negatives. Important use can and should be made by incorporating information from experimental data into the modelling process.

Over the course of the next chapters I intend to describe several approaches which make use of many of the methods outlined in this introduction. These include an analysis of a method for measuring buriedness in proteins; some threading and threading based homology modelling; a purely *ab initio* prediction of a protein, which led on to ideas about predicting disulphides in proteins; investigations into trying to improve threading and also the construction of a model which may, or may not, give some insight into the role that Glucagon plays in type II diabetes.

Chapter 2

Conic residue accessibility

2.1 Introduction

Protein folding is driven by the hydrophobic effect. The idea was originally highlighted by Langmuir and then taken to the fore by Kauzmann. It is well described by Tanford (Tanford, 1997). Some amino acids have a propensity to like water, while others try to remain away from water, the universal solvent; although this is an extremely short range effect. Such hydrophobic residues will stay away from water and bury themselves in the core of the protein. For the protein predictioner this is a very useful aid in understanding the way in which protein structures may assemble.

The hydrophobic/hydrophilic property of amino acids has been coined hydropathy by Kyte and Doolittle (Kyte and Doolittle, 1982). Various hydropathy measures have been calculated (Chothia, 1984; Eisenberg and McLachlan, 1986; Kellis *et al.*, 1988; Lawrence and Bryant, 1991) and it is widely known which amino acids are more, or less, hydrophobic. The amino acids are sometimes grouped according to a variety of properties, and can be expressed in distance space (Taylor and Jones, 1993) or in a two dimensional Venn diagram approach (Taylor, 1986). All these classifications try to give more understanding into the properties of amino acids, which must in turn determine the way a protein forms.

2.1.1 Solvent accessibility

An amino acid's Solvent Accessibility (SA), or Accessible Surface Area (ASA) measure is widely used in the analysis and prediction of three dimensional protein structure. In other words the question "How much of any particular residue in a protein is in contact with the solvent, or water?" can be asked. The first and most widely used measure of solvent accessibility is that of Lee and Richards (Lee and Richards, 1971). Their method rolls a water molecule, represented by a 1.4Å sphere, over the protein surface and measures the amount of contact between that sphere and each amino acid in the protein. Obviously some of the buried amino acids will have no contact, while others may have a large surface exposed to the water. The `DSSP` (Dictionary of Secondary Structure of Proteins) method (Kabsch and Sander, 1983) uses geodesic

sphere integration to calculate the solvent accessible area.

SA has been used for many different purposes in the field of protein structure prediction. It is useful in determining protein folds (Bowie *et al.*, 1990) and in the elucidation of hydrophobic cores (Swindells, 1995) and may well be used in the future for domain determination. Some measures of exposure indicate only whether the residue is buried or exposed, but sometimes more complex levels of exposure are used (Rost and Sander, 1993).

Following earlier work by Finney (Finney, 1978), Gerstein and co-workers have analysed surface volumes using Voronoi polyhedra (Gerstein *et al.*, 1995). They found that the residues on the surface of proteins are not packed as tightly as those found in the core and also that the 1.4Å radius for a water molecule does not accurately represent true interactions with proteins and solvent.

In a similar study, Rose *et. al* calculated the normalised distribution functions of accessibility to solvent (Rose *et al.*, 1985). They analysed the surface area residues lose when becoming buried in a protein, along with a fractional accessibility as a measure of hydrophobicity. The study was extended to derive new scales for residues in proteins (Lesser and Rose, 1990). Teller shows that the overall SA of a monomeric protein varies as the 2/3 power of molecular weight (Teller, 1976). Whereas, Islam and Weaver found that accessible surface area is linearly related to molecular weight (Islam and Weaver, 1990).

As residue SA values are an important indication of a protein's fold, several methods

exist to predict them from sequence. For example, a simple three-state (buried, intermediate, exposed) method for exposure prediction from multiple protein sequence alignments would enable better *ab initio* predictions to be made (Pascarella *et al.*, 1998). Although simple to implement, this method of Pascarella performs less well than others (Holbrook *et al.*, 1990; Rost and Sander, 1994; Thompson and Goldstein, 1996). Wako and Blundell devised a method to predict site specific SA, which uses the known periodicity of hydrophobic patterns to predict secondary structure (Wako and Blundell, 1994a; Wako and Blundell, 1994b).

Fold recognition methodologies often use solvent accessibility as a way of matching folds to sequence (Bowie *et al.*, 1990; Rost, 1995). Hydrophobicity scores have also been used to try and improve the quality of sequence alignments, although this probably has no gain on a PAM matrix, particularly as the algorithmic accuracy seems to depend on sequence homology and even aligns sequences when they are not related (Kanaoka *et al.*, 1989).

2.1.2 *Cones* method

Currently most of the widely available methods for measuring the solvent exposure of proteins are quite slow. Here a new method for the rapid calculation of residue burial by Aszódi *et al.* (Aszódi and Taylor, 1994) is investigated and compared with the solvent accessibility measure as defined in `DSSP` (Kabsch and Sander, 1983), and with that of `NACCESS`, the program incorporating the Lee and Richards method (Lee

and Richards, 1971). The Aszódi *et al.* (*cones*) method is quoted and relies on a fast, simple and elegant algorithm.

The method first described by Aszódi and Taylor (Aszódi and Taylor, 1994) and subsequently modified (Aszódi *et al.*, 1995b) has been developed to aid distance geometry based basic fold construction. It is compared here with other measures of accessibility to assess the possibility of using this method on a more widespread basis.

By working out the area in contact with a solvent molecule for each residue an accessibility measure can be defined and it is this that both `DSSP` and `NACCESS` use. The *cones* method, on the other hand, uses a measure of angles to define how deeply buried a protein is. It has the advantage over the former two methods in that it can give an idea of how deeply buried residues are within the core of a protein. `DSSP` can give an accessibility of zero which only tells you that the residue is not in contact with any solvent. Whereas the *cones* method gives information about the positions of completely buried residues within a protein and thus distinguishes between residues that would obtain a `DSSP` SA of zero.

The extra information on the deepness of completely buried core residues as conferred by the *cones* method is important for understanding processes associated with conformational changes, such as docking. In these cases the knowledge about the potential of residues to become exposed is crucial. For protein folding and protein core formation, such information about the buriedness of residues in the protein core is essential.

2.2 Methods

All methods were compared using the non-homologous set first compiled by Orengo (Orengo, 1994b) and comprises of 201 nonhomologous proteins[1].

2.2.1 The 'DSSP' SA calculation

Both the methods of calculating SA in NACCESS (Lee and Richards, 1971; Hubbard and Thornton, 1993) and in the DSSP work out the surface area of the path made by the centre of a rolling sphere of given radius. I take a solvent radius of 1.4Å, which is the accepted approximate size of a water molecule, which is in fact the size of an oxygen atom, ignoring the two hydrogen atoms (Chothia, 1976).

[1]Note – under certain circumstances not all the proteins have identical residue numbers in both the DSSP database and Brookhaven PDB database. This is because some PDB files have terminal residues with incomplete side chains, these are ignored by the DSSP accessibility program and result in one amino acid being ignored at the terminus. The 201 proteins are listed using their PDB codes and chain identifiers in upper case where applicable.
135l, 1aak, 1abk0, 1aco, 1add0, 1ads, 1ak3A, 1ala, 1alkA, 1aozA, 1apa, 1arb, 1atnA, 1ayh, 1bbpA, 1bbt2, 1bgh0, 1bllE, 1bmv1, 1bmv2, 1brnL, 1btc, 1cauA, 1cde, 1cdg, 1cewI, 1cmbA, 1cobA, 1colA, 1cpcA, 1cseE, 1cseI, 1ctf, 1cy3, 1d66A, 1dhr, 1dmb, 1dsbA, 1eca, 1ede, 1end0, 1etrL, 1ezm, 1fbaA, 1fc2D, 1fkb, 1fnr, 1fus, 1fxd, 1gal, 1gd1O, 1gdhA, 1gky, 1gluA, 1gly, 1gof, 1gpb, 1gpr, 1hleA, 1hmy0, 1hoe, 1hsbA, 1hstA, 1hyp0, 1ifc, 1ipd, 1isuA, 1lct, 1lfi, 1lis0, 1lmb3, 1ltsA, 1ltsC, 1mat0, 1minB, 1mypA, 1mypC, 1nar, 1nipA, 1noa, 1nscA, 1ofv, 1omf, 1ovb, 1pfkA, 1pgd, 1pgx, 1pha, 1phh, 1pii, 1pkp0, 1plc, 1poa, 1ppn, 1prcC, 1prcH, 1prcL, 1ptf, 1pyaA, 1pyaB, 1pyp, 1raiA, 1raiB, 1rcb, 1rfbA, 1rhd, 1ribA, 1rro, 1rveA, 1sbp, 1shaA, 1shg0, 1sim, 1sryA, 1stp, 1tbpA, 1ten, 1thg, 1tml0, 1tnfA, 1tta, 1ula, 1utg, 1vsgA, 1wsyB, 1xis, 1ycc, 1ysaC, 1zaaC, 256bA, 2aaiB, 2bbkH, 2bbkL, 2bopA, 2bpa1, 2bpa2, 2cas, 2cba, 2cdv, 2cmd, 2cpl0, 2ctc, 2ctvA, 2cyp, 2er7E, 2gstA, 2hpdA, 2lbp, 2ltnB, 2mev4, 2mnr, 2msbA, 2nckL, 2ohxA, 2ovo, 2pia, 2pmgA, 2polA, 2rhe, 2rn2, 2sicI, 2sn3, 2sns, 2stv, 2tgi, 2tmdA, 2tmvP, 2tscA, 2yhx, 3blm0, 3cla, 3dfr, 3ebx, 3ecaA, 3gapA, 3grs, 3il8, 3mdsA, 3monA, 3monB, 3pgk, 3pgm, 3rubS, 3sc2A, 3sc2B, 4enl, 4fgf, 4gcr, 4mt20, 4sbvA, 4sgbI, 5fd10, 5p21, 5pti, 5timA, 6insE, 7aatA, 7catA, 7rsa, 8rxn, 9wgaA.

2.2.2 The *cones* method

A cone is placed on the k-th residue in a protein, centred on the atom at the cone's apex, with the cone's axis passing through the centroid of the protein (or set of points). The cone is expanded to encompass all the other atoms within the protein and may invert in the process. This is illustrated in Figure 2.1. The angle defined by the cone (α_k) is a direct indication of the buriedness of the atom k.

In some cases large scale effects such as clefts in proteins, may incorrectly place the cone. This scenario could come about if an atom was exposed on the edge of a 'buried' cleft. To overcome this problem only a subset of the atoms in the protein are examined, a radius of 8Å around the atom of interest. The *cones* measure gives a scale between -1 and +1, a range of fully exposed to fully buried respectively.

A further simplification to the *cones* method was investigated. Instead of calculating the side chain centroid, the C_β was placed along the bisector of the α-carbon virtual bond angle at a fixed distance from the α-carbon. Alternatively an average side chain length could also be imposed.

This side chain length (C_β distance) was added by calculating the position of the centroid for three consecutive C_α atoms. The vector between the centroid and the central α-carbon is calculated, and then scaled to translate the centroid beyond the middle C_α, to the desired length of side chain (Levitt, 1976b).

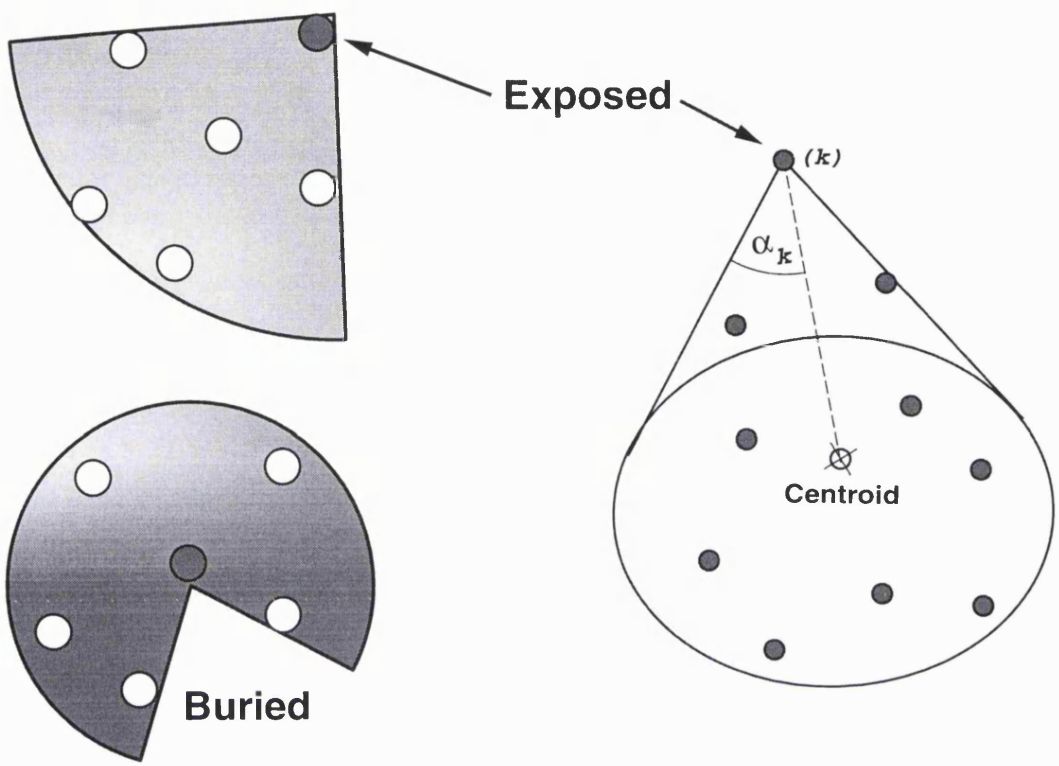


Figure 2.1: Illustration of the *cones* method. Figure courtesy of András Aszódi.

2.2.3 Density function

For each amino acid using both DSSP SA and *cones* methods a density function (DF) was calculated. With the *cones* method the shieldedness is scaled from -1 to +1. For the DSSP SA, I normalised the scores to the range 0–1. Using a specially designed program to calculate the CDF for the data, the results were converted to density functions using a graphics package (xmgr).

Normalisation of DSSP scores Due to the unrestricted maximum value of exposure which can be obtained using the DSSP measure, it was preferable to convert the areas to a percentage measure using a cut-off. To do this a theoretical maximum SA for each amino acid was calculated by constructing a fully extended backbone of

glycines into which were placed the amino acids. A chain such as G-G-**A**-G-G-**C**-G-G etc. was used and the SA calculated, using the DSSP method. This was compared to a Gly rich helix, where each residue was separated by ten Gly's so as to place them on alternating sides of a helix. The values from the helix were lower than the measurements taken from the poly-glycine chain, so the chain maximum accessibilities were taken as maximal. A proline was not included in the chain but was calculated separately, because of the inherent bend in the amino acid which might have affected the results of the other amino acids. They compared reasonably with a similar such calculation made by Miller *et al.* (Miller *et al.*, 1987). (See Table 2.1).

Amino Acid		Max.Acc.	Percent overflow
Ala	(A)	113	0.95
Cys	(C)	143	0.12
Asp	(D)	171	0.25
Glu	(E)	201	0.43
Phe	(F)	218	0.10
Gly	(G)	85	1.31
His	(H)	196	0.28
Ile	(I)	179	0.04
Lys	(K)	216	0.80
Leu	(L)	175	0.28
Met	(M)	200	0.94
Asn	(N)	168	0.31
Pro	(P)	145	0.60
Gln	(Q)	205	0.23
Arg	(R)	255	0.27
Ser	(S)	133	0.99
Thr	(T)	153	0.35
Val	(V)	157	0.12
Trp	(W)	262	0.00
Tyr	(Y)	243	0.06

Table 2.1: Maximum accessibilities ($\AA^2$). These were calculated for each amino acid in a poly-glycine chain, using the DSSP measure. Also shown are the percentage of amino acids with a SA which exceeded the calculated maximum SA.

Any amino acids which had an accessibility equal or greater than the expected maximum accessibility were assigned a relative accessibility of 1. None were significantly higher than the theoretical maximum values calculated from the poly-glycine chain.

2.2.4 Programming

All programs were run on Silicon Graphics workstations, written in C and compiled using the standard C compiler. The normalised DF's were converted to simple density functions by using the forward difference in `ACE/gr (xmgr) v3.01`. This program was also used for the calculations of the regression and correlation statistics.

The `DSSP` program was originally written in Pascal. I converted it to C and modified it to calculate only the SA. From this point it was an easy matter to implement new features, such as the ability to calculate SA for a number of different solvent sizes (e.g. 0.0 to 3.0 in 0.1Å steps).

The access program is written in FORTRAN, scripts were used to run the program for each non-homologous protein. Using a C program I was then able to extract the relevant details from the Relative Surface Accessibility files, created by the access program.

2.3 Results

2.3.1 Comparison of NACCESS and DSSP SA

Although calculated in slightly different ways, both programs give effectively the same result. The Lee and Richards NACCESS program (in Fortran) gives SA to two decimal points, whereas the Kabsch and Sander DSSP give to the nearest integer. Plotting the two SA calculations give a very highly correlated set of points (rho=0.996), with little spread around the calculated regression line, see Figure 2.2.

As both measures are very similar I have concentrated on using the DSSP SA alone for the comparative analysis. However, Figure 2.3 does show all three measures compared.

2.3.2 Comparison between *cones* and DSSP accessibilities

For all proteins in the non-homologous set the accessibilities were plotted for each residue against the relevant *cones* calculation. The *cones* method used a C_α chain with side chain centroids. A comparison of the two methods is shown in Figure 2.4. The graph has a coefficient of correlation of -0.864. A regression line can be fitted to the points but it is apparent from the graph that the fit could be improved if there was a slight curve in the line, although exponential fits deviate more latterly.

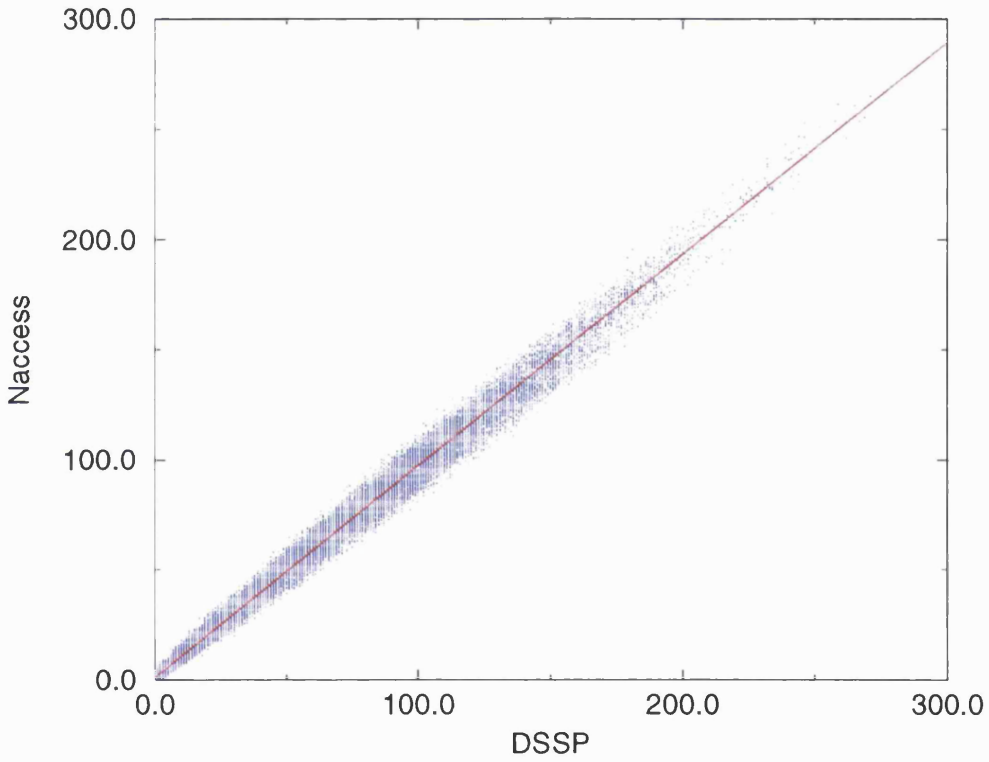


Figure 2.2: Regression analysis of DSSP plotted against NACCESS. Each dot is one residue in the non homologous data set. Correlation coefficient of DSSP vs. NACCESS is 0.996.

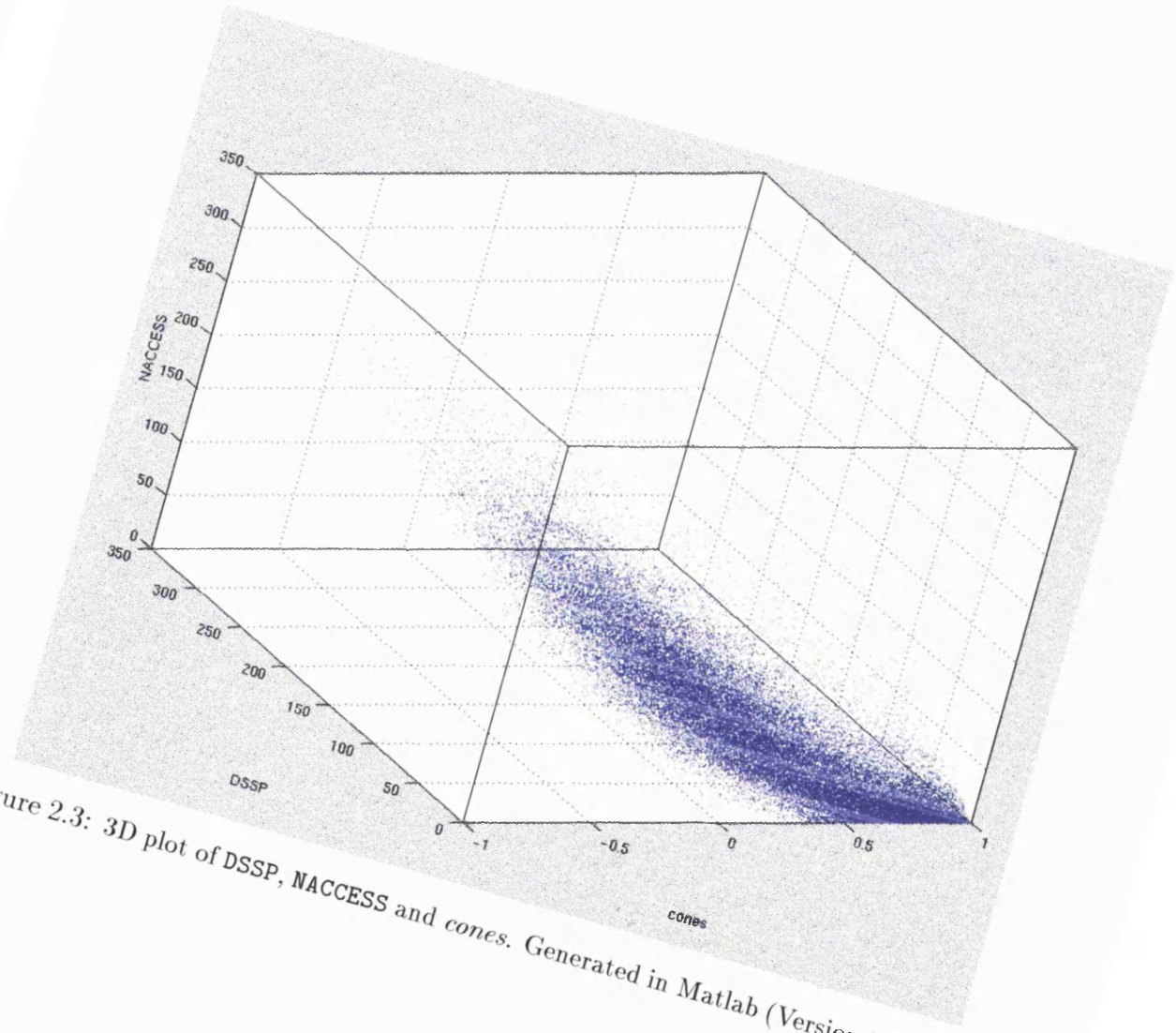


Figure 2.3: 3D plot of DSSP, NACCESS and *cones*. Generated in Matlab (Version 5.2.1.1421)

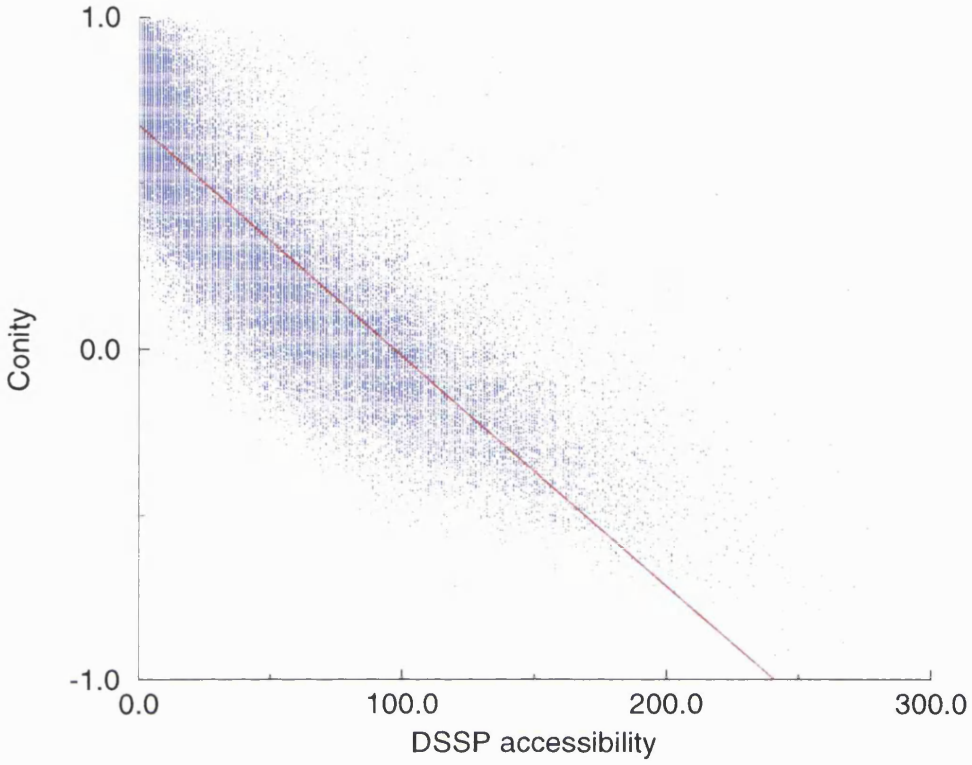


Figure 2.4: Comparison of DSSP accessibility and *cones*. Correlation coefficient of DSSP vs. *cones* is -0.864

Correlation per residue type is discussed later with the use of density functions.

A further simplification in the side chain representation was tested using the *cones* method. Instead of the side chain being the centroid, it was set to a fixed distance from the backbone (in Figure 4.3 $d_{\alpha\beta}$ is set to 2 Å). This, unsurprisingly, gives slightly less good correlation with the DSSP measure. A variety of differing C_β side chain distances (1, 2, 2.5, 3, 4 and 5Å) were compared to DSSP SA (for small data sets this is not accurate). It can be seen that a side chain length of 2Å is the best approximation to the measure which takes into account the actual protein information, having a

correlation coefficient of -0.771.

2.3.3 Density function plots

Density functions were calculated for each of the amino acids using the *cones* method, as shown in Figure 2.5. They were also calculated using *cones* with the simplified side chain of 2Å. These are shown as blue lines in Figure 2.5.

Figure 2.6, shows all the SCC density functions of the conic accessibilities, overlaid and coloured. This is effectively figure 2.5 but plotted on one graph, so that the similarities between the different amino acids are more clearly visible. In particular this shows how the similarities appear to cluster into three or four different classes, as might have been expected.

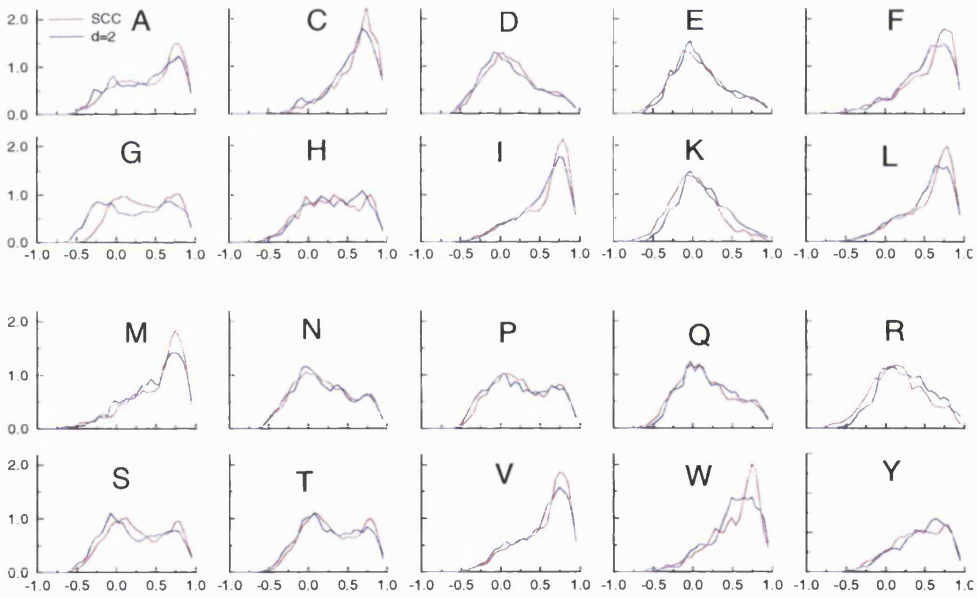


Figure 2.5: C_β side chain conic accessibilities. Density functions of the conic accessibilities plotted for each amino acid. Shown are the measurements for side chain centroid (SCC) and a fake side chain of length 2Å (d=2).

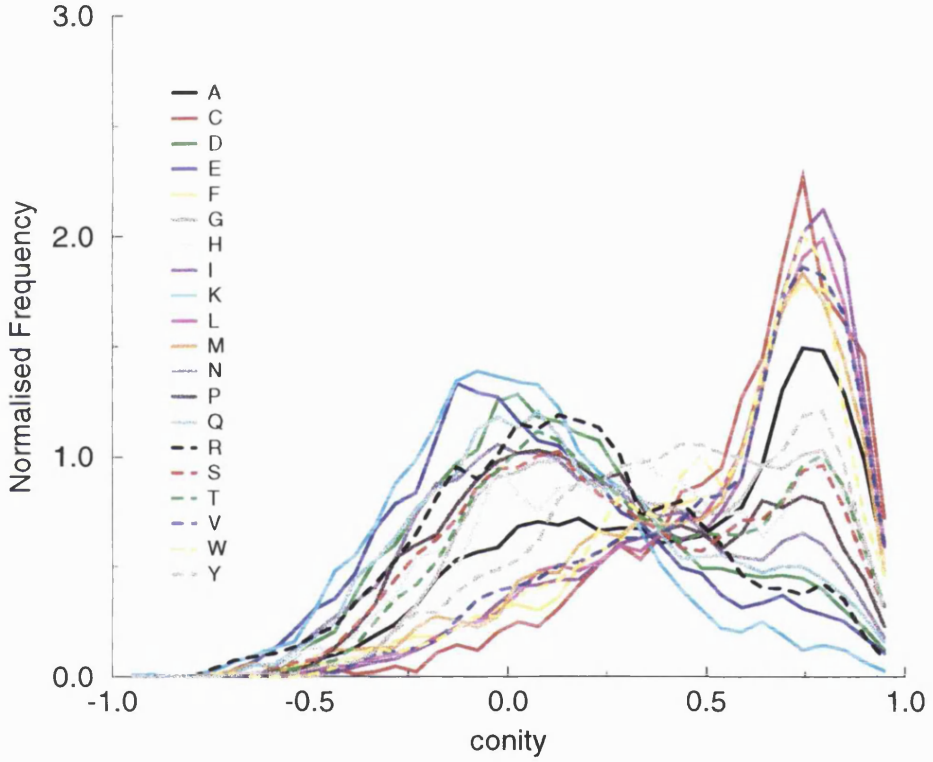


Figure 2.6: All C_β side chain conic accessibilities. All SCC density functions of the conic accessibilities overlaid to show the obvious differences between the different amino acid types.

From the density functions it is possible arbitrarily to group the amino acids, based on the location of the peaks in the density functions. For example: C, F, I, L, M, V, W and Y, all peak towards the right in Figure 2.6 and belong in the same class (buried) as too might N, P, S which have bimodal distributions and T. D, E, K, Q and R which peak to the left and may be classified as exposed leaving A, G and H. From what is known about the properties of amino acids, these divisions would make sense from a physiochemical point of view.

Kolmogorov-Smirnov (K-S) Test The Kolmogorov-Smirnov test can be used to see if two distributions are different. It was therefore possible to compare the distributions created using the *cones* DF and the simplified *cones* DF. The K-S test was implemented and carried out for each amino acid. It showed that there was no significant difference between the distributions for any amino-acid. The data is shown in Table 2.2.

Amino Acid	KS stat.	p
ALA	0.150	0.724
CYS	0.250	0.139
ASP	0.175	0.531
GLU	0.150	0.724
PHE	0.125	0.893
GLY	0.125	0.893
HIS	0.100	0.983
ILE	0.175	0.531
LYS	0.150	0.724
LEU	0.150	0.724
MET	0.125	0.893
ASN	0.200	0.361
PRO	0.175	0.531
GLN	0.150	0.724
ARG	0.100	0.983
SER	0.200	0.361
THR	0.150	0.724
VAL	0.175	0.531
TRP	0.125	0.893
TYR	0.150	0.724

Table 2.2: K-S test results. Small Values of p show that DF of data set1 is significantly different from data set2 - the statistic p is never less than 0.01 so there is no significant difference between the distributions calculated by the differing *cones* methods.

2.3.4 DSSP density function

During the calculation of the DF's any observed residues with a greater than maximum SA were noted. Only in the case of glycine did the accessibilities exceed the theoretical maximum by one percent (see Table 2.1). Figure 2.7 shows the scaled DSSP surface accessibility density functions.

2.3.5 Individual protein analysis

Figure 2.8 shows how the DSSP SA differs according to the size of the water molecule. From the graph it is clear that the small solvent size is less discriminating, whereas the largest size is not all that sensitive for deeply buried residues. The "best" size appears to be that of the 1.4Å sphere.

Figure 2.9 and Figure 2.10 show examples of DSSP SA and *cones* for individual proteins. Figure 2.9 is a direct comparison for each residue in a small protein, whereas Figure 2.10 is a similar comparison but for a much larger protein. For clarity the plot has been smoothed over a window of five residues.

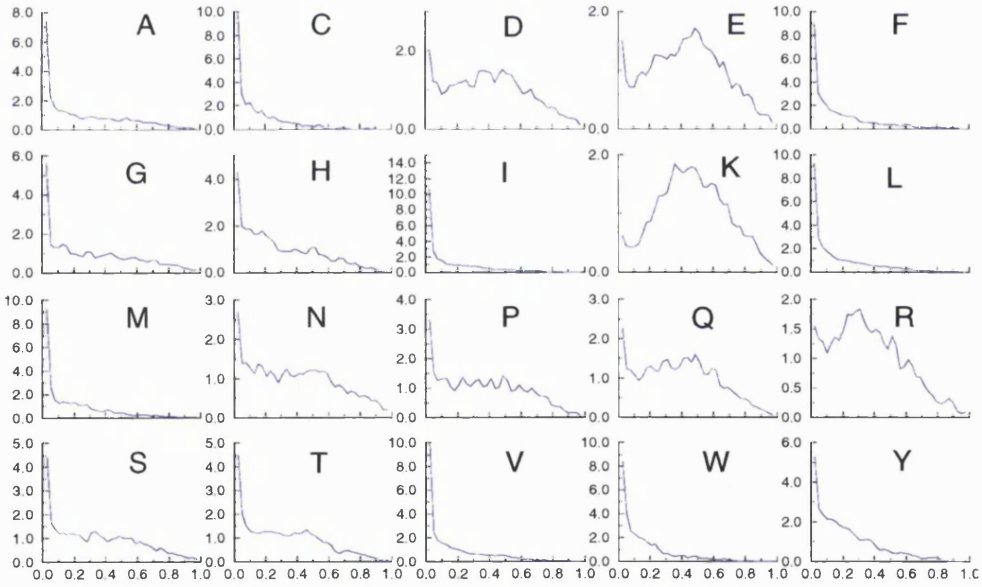


Figure 2.7: Scaled DSSP surface accessibilities. Density functions of the DSSP surface accessibilities plotted for each amino acid. Compared with the figure for the *cones* method these plots appear less discriminating.

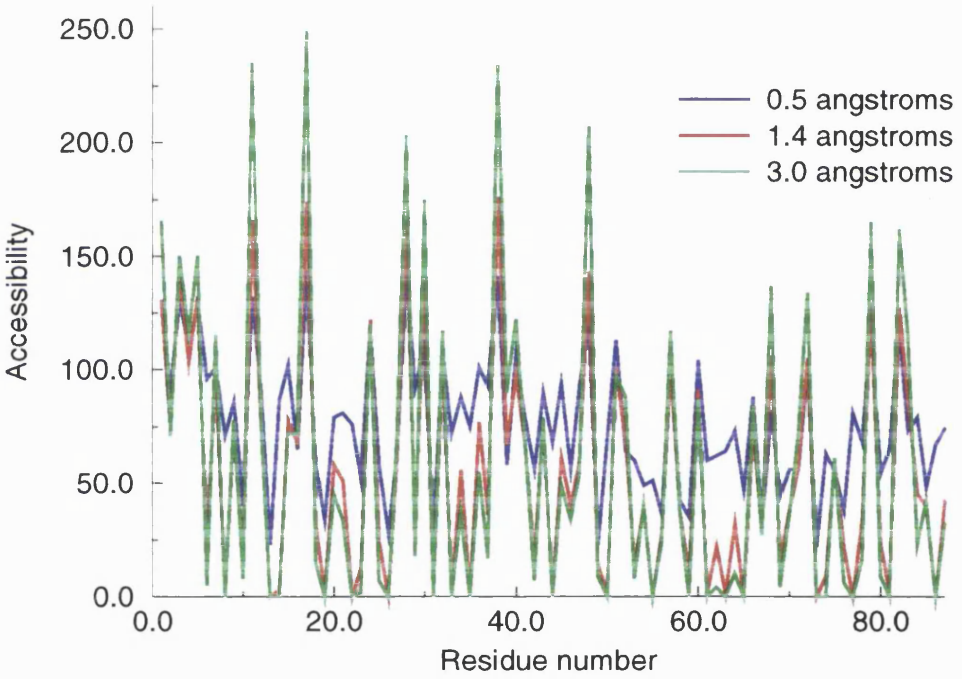


Figure 2.8: Comparison of different sized water spheres. Comparison between 0.5, 1.4 and 3.0 angstroms for the rolling sphere are shown here for a phosphotransferase protein (1ptf). It is clear that the plots are related, but some sizes of sphere are less revealing about the accessibility.

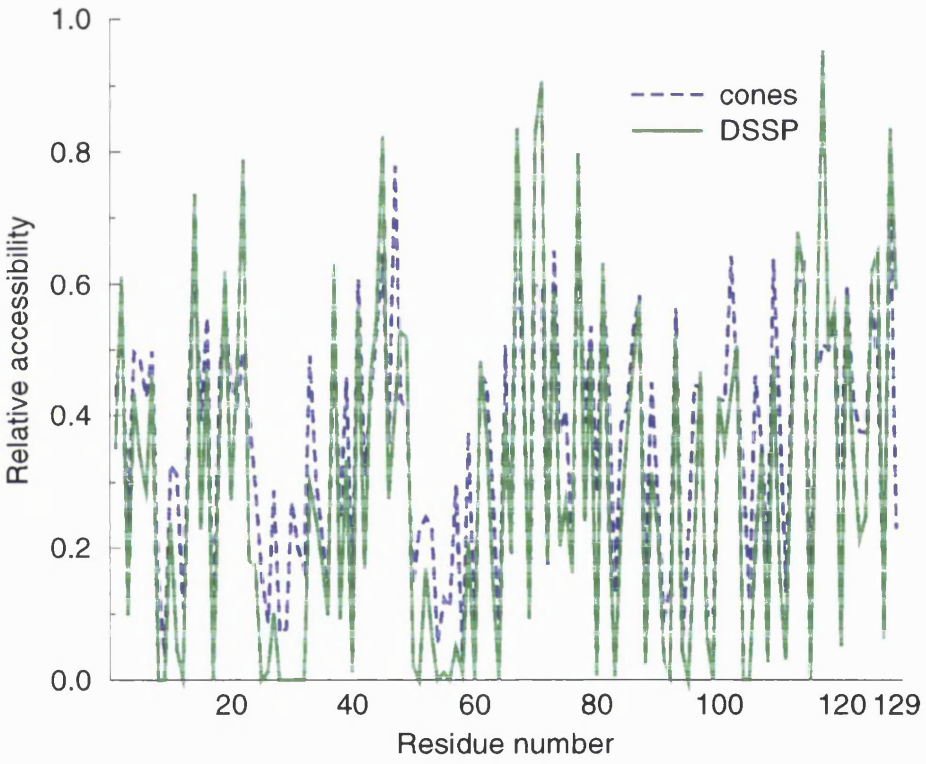


Figure 2.9: Plot of DSSP-SA vs. *cones* for protein 135L. Trace for comparison between SA and *cones*. No smoothing, protein is only 129 residues long. The two traces are comparable. Where the DSSP values are close to zero the *cones* measure shows greater variation, showing the true differences in the atom positions within the protein, such as residues 27-31.

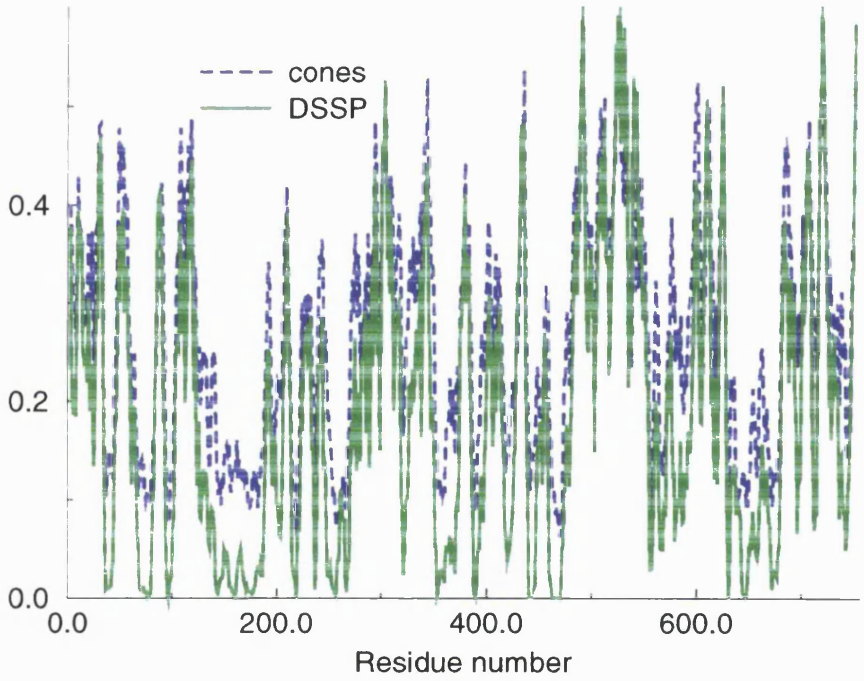


Figure 2.10: Plot of DSSP-SA vs. *cones* for protein 1ACO. Smoothed trace for comparison between SA and *cones*. Sliding average window of 5 residues. Due to the smoothing it is clearer to see the similarities between the two plots. Similarly to Figure 2.9, the *cones* method is better when DSSP is close to zero, this is clearer to see on the un-smoothed plot when only part of the protein length is plotted.

2.3.6 Speed issues

On a Silicon Graphics Workstation (SGI R10000 175 MHZ IP32 Processor and 128 Mbytes RAM) DSSP SA calculation for the first 10 proteins from the non-homologous set took 9.6 seconds (execution time). The corresponding *cones* calculation for the set took 5.0 seconds and thus was almost twice as fast compared to DSSP, 1998 version. Table 2.3 shows comparative speeds for different programs and versions of the same programs on the first ten proteins in the homologous data set.

Program	Description	Time taken (s)
CONES	1998 version	5.0
DSSP	1998 version	9.6
DSSP	1988 version	58.3
DSSP	p2c version	121.6
DAC	SA algorithm from above	82.6
NACCESS	1993 (S.J. Hubbard) (Hubbard and Thornton, 1993)	60.2

Table 2.3: The speed of different programs. All were run on an SGI R10000 175 MHZ IP32 Processor. All times quoted are the execution times on an unloaded system.

2.4 Discussion

In this study a non homologous set of 201 proteins was used. These were all small globular proteins and were assumed to be a representative sample of proteins which is useful for a general analysis such as this.

The character of amino acids in general could be easily distinguished by eye when looking at the density functions for each amino acid. For example, a highly hydrophobic amino-acid such as valine gives a profile where the majority of the occurrences of a valine residue occurs in a buried position within proteins and so the distribution is skewed to 1 (buried). Conversely, polar amino acids tend to be more skewed the other way.

It could be argued that less information is retained due to the simple representation of the protein chain and actual volume of a part of a side chain is not considered, only its relative location in the protein. This is important for exposed residues, so DSSP may well show better resolution in the case of these amino acids.

Using DSSP it was possible to scale the SA over a range of $0 \rightarrow 1$ (buried $\rightarrow$ exposed) by finding the maximum possible SA for a given amino acid and scaling accordingly. The resultant DF's showed similar results to the *cones* approach, and are detailed in Figure 2.7. Comparing the two methods it appears that the distributions are similar, although the DSSP plots tend to be less discriminating and have a large number of residues with zero SA. The conic measure has some advantages over DSSP and the

like. It is able to distinguish something about every residue in a protein. For example a residue which is deep in the core will have zero SA, while with the *cones* measure it is possible to measure how deeply located it is. This is apparent in the density functions which are more normally distributed using the *cones* method compared to DSSP. By examining Figure 2.5 and Figure 2.7 comparisons can be drawn between the two methods. Qualitatively there is less information in the DSSP measure than the *cones* method. Any of the hydrophobic residues which are completely buried cannot be measured by DSSP. Take alanine and phenylalanine for example, in the DSSP measure the two look similar to each other. The *cones* method, on the other hand shows a distinct difference in the distributions. When the side chain of each residue is of a fixed length ($d = 2$Å) the distributions are similar, amino acids such as tryptophan have a more squat distribution because they effectively lose half their C_α:SCC distance. From the aspect of modelling it is good to be able to measure a property of an amino acid under any circumstance and not just if it is on the surface of a protein.

The *cones* method on the other hand reduces the amount of information required in the calculation by taking the side chain centroid and using it and the C_α position to calculate buriedness. This reduction in information makes for a much faster analysis, but information is lost about the exposure of residues. Another advantage of the *cones* side chain centroid method is that it can be applied to simple 'non-real' proteins, i.e. structures created by homology modelling programs and the like. It is possible to have a measure of buriedness without having to build specific side chains and their

associated rotamers. Using an average C_β as a side chain centroid, the buriedness can be calculated, whereas SA methods cannot be applied.

The DSSP SA and *cones* DF distributions could not be compared empirically as the abscissa are different, so the K-S test was not employed. Comparing the plots of DSSP and *cones* on individual proteins has shown them to be very similar. From the analysis characteristic density functions for each amino acid are calculated. These show the differences between amino acid types and they can be characterised them into separate groups.

I used a Kolmogorov-Smirnov test to compare the distributions created using a fixed length C_β in place of a side chain centroid (which is residue dependent). It was found that the best approximation to DSSP SA was with the side chain centroid. There is bound to be an inherent loss of information when the side chain length is fixed.

The allocation of side chains to the C_α chain took into account the position of glycine's, which have no side chain. This would not occur if the backbone was to be used for any threading applications, as an unbiased chain for any amino acid would be required. Similarly, amino acids with long side chains, such as arginine had marginally different distributions from the fake C_β side chains.

The conicity is incorporated into DRAGON which has been shown to give quite good models (Aszódi *et al.*, 1997a). Chapters 4, 5 and 8 outline some modelling which has been carried out using this program. In future rankings of models could be made using the amino acid profiles generated by the *cones* algorithm. Rankings could be

produced according to the conicity of the models' amino acids and how standard or non-standard their compositions might be.

2.4.1 *Cones* or SA?

Over the years since DSSP was first implemented the algorithm itself has been speeded up significantly, as is clearly shown in Table 2.3. The two measures are now very similar in speed, but the *cones* still has advantages over the other method. Particularly because of the fact that it is a buriedness measure and not just a surface accessibility measure. Comparing the Speed of DSSP with only the accessibility algorithm, implemented as DAC then roughly 68% of the time is spent on the SA calculation. Assuming that this is still the case in the latest version of DSSP, then the SA calculation would take about 6.5 seconds, leaving the *cones* method still slightly faster at 5.0 seconds. Advantages of SA are such that the measure can be used to assess solvation energy, where *cones* cannot. Also the resolution of partly exposed residues must be greater for DSSP as it uses more information.

Further work should examine the standard deviation of *cones* from DSSP for each residue. This would show where *cones* variation is greatest compared with DSSP. This could also be done the other way round showing the deviation of DSSP versus *cones*. This would show which measure has better resolution for exposed and buried residues. The methods are sufficiently different that they should be used in conjunction with each other when possible.

Chapter 3

CASP fold recognition

3.1 Introduction

When protein sequences have no known homology to a PDB structure then we have to find alternative ways of building models. One solution is to identify a fold which the sequence of interest may adopt. This is called Fold Recognition, or sometimes threading. Using this fold as a starting configuration, it might then be possible to build a 3D model for the query protein sequence. This idea has been put into practice in Chapter 4.

There are many different methods available and a number of workers are continuing to

look at improving our ability to recognise folds from sequence. There will come a time in the not too distant future when most folds which proteins adopt have been solved. This would make *ab initio* prediction obsolete if the fold could be correctly identified. At the moment a hard problem is knowing when there is no fold for the sequence, as all threading programs recognise a best fold. Usually there is an associated statistic such as a Z score to aid in the decision of whether or not to accept the answer.

One of the first computer programs designed for fold recognition was called `THREADER` (Jones *et al.*, 1992a) and is described briefly in Chapter 1. Rather than using conventional distance based potentials, Zimmer *et al.* have devised a method based on Voronoi contacts (Zimmer *et al.*, 1998). Compared with Sippl's potentials (Sippl, 1990), they report that the Voronoi contact potentials to improve recognition by over at least 9%. HMM's (hidden markov models) have also been used for fold recognition with moderate success in CASP2 (Francesco *et al.*, 1997; Karplus *et al.*, 1997). See also articles on the assessment of the CASP2 predictions by the assessors (Marchler-Bauer and Bryant, 1997; Marchler-Bauer *et al.*, 1997; Levitt, 1997).

Lattice based potentials have been derived (Reva *et al.*, 1997) and compared by the 'hide and seek' threading test (Hendlich *et al.*, 1990). These lattice adjusted potentials worked well on lattices up to 3.8Å spacing, compared to non lattice based potentials. Other methods also use lattices (Chiu and Goldstein, 1998).

Rooman and Gilis derived different potentials to test their predictive power in fold recognition (Rooman and Gilis, 1998), but concluded that a universally successful

single database derived potential does not exist.

In a recent publication by Russell and Saqi *et al.* (Russell *et al.*, 1998), their method (FOLDFIT) uses a 1D/3D approach. To find the best fold they use matrices to match accessibility and secondary structures between the sequence and the structure database. Compared directly with a potentials based threading method (THREADER) (Jones *et al.*, 1992a) their method does better. However it does not do as well against the latest available version (THREADER2). A similar 1D/3D approach is described by Rice and Eisenberg (Rice and Eisenberg, 1997).

A hydrophobic contact potential has been generated by Huang *et. al* (Huang *et al.*, 1996) which can discriminate between native and grossly miss-folded proteins and they postulate to discriminate between native and near native structures in the future.

Many other methods and potentials have been developed to predict the correct fold (Rost, 1995; Fischer *et al.*, 1996b; Huber and Torda, 1998). For a review, see (Jones and Thornton, 1996). An improvement in sequence to structure alignments, as well as non-physical force fields based on parameter optimisation, will greatly improve fold recognition (Torda, 1997).

3.1.1 CASP2

The first Meeting on the Critical Assessment of Techniques for Protein Structure Prediction was held at Asilomar, California at the end of 1994. A similar assessment was repeated and a meeting was held at the end of 1996 (CASP, 1995; CASP, 1997). The exercise covered the major areas of protein structure prediction (Comparative Modelling, Fold Recognition, *Ab initio* and also the prediction of associations between ligands and proteins (Docking)). CASP2 involved the submission of sequences which have newly solved structures that have not yet been released. Researchers were then invited to predict the structure of the proteins and submit their findings. At the end of 1996 the predictions were assessed by comparison with the previously solved structures.

Over the next three chapters some of the work carried out for CASP2 will be presented and analysed. This chapter will concentrate on the fold recognition predictions made.

Threading was carried out on the fold recognition targets using the method of Jones (Jones *et al.*, 1992a; Jones *et al.*, 1993), implemented as `THREADER`. This was run in conjunction with Multiple Sequence Threading (`MST`) developed by Taylor (Taylor, 1997).

A variety of methods were used to build up a picture about the protein of interest. By examining the sequence, its predicted secondary structure, the relatedness to other proteins, the results of various threading methods and the known function of the protein a putative structure for the protein could be deduced. Some use was made of

the THREADER program, but a more specific run of MST was used to thread multiple alignments on to the structures of particular interest.

Here, I describe the methods that I used for the CASP2 protein structure prediction assessment, concentrating on the threading side of the experiment.

3.2 Methods

3.2.1 Similarity searches

Use of programs to search databases for similar sequences:

`BLAST` searching on an NCBI non-redundant database, OWL database.

`BLITZ` searching on the Swiss-Prot database.

`FASTA` searches at the EBI.

The programs `BLAST`, `BLITZ` and `FASTA` are designed to perform sequence searches in large databases of sequences. Their different algorithms are briefly described in the introduction.

It is useful to use all these methods as they may not give exactly the same results. There are now better methods for sequence searching which are based on an iterated search process, where the results are built up over a series of profile based searches. Programs such as `PSI-BLAST` (Altschul *et al.*, 1997), `QUEST` (Taylor, 1998a) and `PROBE` (Neuwald *et al.*, 1997) are based on this idea and can often identify distantly related sequence hits.

3.2.2 Multiple alignments

The multiple sequence alignment program MULTAL (Taylor, 1988) was used to align sequences which were related to the target sequence. BLAST, BLITZ and FASTA were used to find similar sequences. MULTAL was used to align these sequences, often with a gap opening penalty of between 15 and 20 with a PAM120 matrix (no gap extension penalty was used). The resulting alignment was then examined to find patterns of conserved residues which had aligned across the sequences. The motifs found were used to search an OWL database file for sequences containing that motif. With a larger alignment there may be a better idea of a target sequences derivation. Often these larger alignments were pruned by removing sequences which were very similar. These proteins were threaded along with the target sequence.

Elimination of sequences which were highly similar (over 90% identity) was carried out to reduce extraneous information and make the alignment easier to view. This should remove any possibility of sequence weighting.

Sequences of known structures could sometimes be aligned with relatively distant homologues of the target sequence. Many of these structures were first identified by threading methods.

BLAST was performed interactively on the WWW but the other queries were run a search form which sends the results back via e-mail once the jobs have been completed on a remote computer.

3.2.3 Secondary structure prediction

Use of various secondary structure prediction packages:

PHD Protein Prediction (Rost and Sander, 1993). SSPRED EMBL Secondary Structure Prediction (Mehta *et al.*, 1995). SOPM Self Optimized Prediction Method (Geourjon and Deleage, 1995). NNPREDICT Protein Secondary Structure Prediction (Kneller *et al.*, 1990).

By using these programs it is possible to gain an idea of the secondary structure of the protein sequence of interest. The consistency over the various methods can give insight into which regions might be most reliably predicted. These consensus alignments show those secondary structures which have been predicted by all or most of the programs. PHD is particularly useful as it constructs a multiple alignment before using it to predict the secondary structure. MSF format multiple alignments can be submitted to PHD and DSC to give a SS prediction unique to that alignment. This is much better than using a single sequence in the prediction. A good alignment can increase the chances of a correct prediction.

3.2.4 Fold recognition

THREADER is a program which takes a potential map of a protein and threads the target sequence on to the structure of the known protein. The targets were threaded on to a database of 941 proteins. The THREADER output for the target sequence and

any other related sequences which were also threaded were ranked according to Z score for the core-shuffled pairwise energies and the structures which scored well were examined. Frequently a structure would be picked up which had a high score for all the threaded, related sequences.

The PDB filenames were selected from an index of all the header/compound information in the PDB and the protein names printed for easy identification of the `THREADER` results.

The possible structures indicated by `THREADER` were visually compared, taking into account the knowledge of the secondary structure prediction methods as well as a "by eye" examination of the alignment.

Occasionally, if the situation warranted it, some of the PDB structures were compared with each other to see if any areas in the proteins were similar and a likely place for the threading to have picked up. Using a modification of `SSAP` a structure comparison program (Orengo and Taylor, 1996), now referred to as `SAP`, the proteins can be superimposed and areas of similarity highlighted.

Generally, the PDB files were examined using `ProtDraw`, an in-house PDB viewer with high speed hardware manipulation (Aszódi). Using this it is possible to view different properties in a protein or view properties such as occupancy which can relate to whatever the user decides to colour code. In the case of the `MST` program the occupancy shows the areas of deletion and insertion to which the threaded sequence can fit, as well as likely loop areas.

A knowledge of the function of the protein was often helpful, indicating it to fall into various families, of which there may have been known structure. Also by identifying any areas of known importance in the sequence of the protein the likely structure could also be predicted. For example the location of disulphide bonds can give an easy idea of whether or not a protein structure is feasible with a given sequence.

A program was written to convert the `MULTAL` alignment output into an MSF file which could then be parsed by the `PHD` program. This allowed a `MULTAL` alignment to be edited slightly by hand before a secondary structure prediction was re-assessed. The SS prediction was also modified by hand, in many cases where alpha, beta surface and beta buried motifs could be recognised.

The aligned sequences were studied for conserved residues and in particular conserved hydrophobics, V, I, L, F, M, A, W, H, C, T, G. These are hydrophobic (in approximate decreasing order of magnitude). Cysteines where applicable were scrutinised for disulphide formation. Acidity may also be conserved - such as E and D. Large residues such as K and R can be partly hydrophobic due to their large size and may be part of a hydrophobic pattern consistent of a secondary structure.

Areas of secondary structure are often delimited by breaks in the sequence alignment where insertions and/or deletions have occurred in the loops joining the SSs. These areas may also be glycine rich and contain prolines. Gaps however do not always occur between SSs, but G and P are still indicators of loops. Proline is also sometimes found in N-cap helices. Pattern of conserved hydrophobicity can give a good idea of

secondary structure.

Further work recently published by Taylor (Taylor, 1998b) could be used to introduce sophisticated and fast domain recognition system into the threading methodology.

MST output essentially consists of the alignment and then the threading of that alignment made for each structure of interest. An example of the MST output is shown in Figure 3.1, more details on the method are in Chapter 6, where improvements in the MST algorithm have been investigated, with particular regard to the placement of insertions and deletions in the sequence and structure being compared.

```
structure                sequence

     :
  11 E     *    0.0      H      F       1
  12 Y     *    0.0      H      A  V    2
  13 R          35.1      H AAAMR  A    3
  14 N     *    36.0      H EEESTKDL    4
  15 T E   *    38.8  *   H IIVIHVLN    5
  16 Q E   *    38.6      H EEEEAKQK    6
  17 I E   *    36.9  *   H VVAVIPPK    7
  18 Y E   *    36.2  *   E GGGGGGGG    8
  19 T E   *    38.1      E RRVSQDMA    9
  20 I E   *    35.7  **  E VIIKIVII   10
                  :
  21 N     *    27.9      E TAKQPEET   12
  22 D E   *    34.7  *   E GGGGGGGG   13
  23 K E   *    40.1      E KKKKKTAK   14
  24 I E   **   40.4  **  E VVVIVVVV   15
  25 L E   *    35.3  *   E TTTTTQTT   16
  26 S E        38.5      E RRRGKRNA   17
  27 Y E   *    35.7  **  E IILILLVV   18
  28 T E        30.8  *   E VVATVVTD   19
  29 E E   *    28.1      E DDDNPSNA   20
     :
  36 E E   *    10.3  **  E FFFFFFFK   21
  37 M E   *    36.3  *   E GGGGGGGG   22
  38 V E   **   41.4  **  E AAAAAAAA   23
  39 I E   *    39.3  **  E FFFFFFFT   24
  40 I E   **   38.0  **  E VVVVVVVV   25
  41 T E   *    33.6      E AAAEREDE   26
  42 F     **   32.1  **    IIILVIIL   27
  43 K      *   24.4        GGVPELGA   28
  44 S          23.9        GGGGEPVD   29
  45 G          21.5        GGNGGGHG   30
  46 E E   *    24.8        KKKSIVQV   31
  47 T E   *    26.2        EEETEEDE   32
  48 F E   **   31.2  *     GGGGGGGG   33
  49 Q E   *    26.6  **  E LLLLLLLY   34
  50 V E   *    20.2  **  E VVVVVVVL   35
  51 E E   *    16.8      E HHHHHHHR   36
  52 V E   *    15.3  **  E IIIIIIIA   37
     :
  59 D          33.8      E SSSSSSSS   38
  60 S H        41.3      H QQQEEQSE   39
  61 Q H   *    42.5  **  H IIIVLILA   40
  62 K H   *    38.2  *   H AAAAASSS   41
  63 K H        38.0      H DDEDENNR   42
  64 A H        40.7      H KKENRKKD   43
  65 I H   *    36.7      H RRRYHHFR   44
  66 E H   *    39.1  **  H VVVVVIVV   45
  67 R H        41.5      H EEEKEGEE   46
  68 M H   *    39.3      H KKKDVTDD   47
  69 K H   *    38.8  *   E VVVIPPPA   48
  70 D H        45.1      E TASNDHHT   49
  71 T H   *    50.3      E DDDDQETL   50
  72 L H   **   51.9  *   E YYYHVVVV   51
  73 R H   *    44.0  **  E LLLLVLVL   52
  74 I H        40.8      E QQQKAEKS   53
  75 T H   *    36.0  **  E MVVVVEAV   54
  76 Y H   *    26.6  *   E GGGGGGGG   55
     :
  78 T          19.1      E QQQDDQDD   56
  79 E      *   22.3      E EEEQDTIE   57
  80 T E    *   38.9  **  E VTVVAVVV   58
  81 K E        46.3      E PSNEMKKE   59
  82 I E   **   50.4  **  E VVVVVVVA   60
  83 D E   *    54.3      E KKKKKKKK   61
  84 K E   *    58.0  **  E VVVVVVVF   62
  85 L E   **   56.7  **  E LLVIILLT   63
  86 C E   *    47.8      E EEENDDEG   64
  87 V E   **   47.9  **  H VIIVIVVV   65
  88 W E   *    35.7      H DDDEDNDD   66
  89 N     *    33.4      H RRRKLELR   67
  90 N     *     0.0        QQQDENQK   68
  91 K     *     0.0        GGGGRERN   69
  92 T          37.9      H     REKR   70
  93 P     *    36.7        RRRKRRRA   71
  94 N     *    47.2  **  E IVIIIIII   72
  95 S E   *    48.9      E RRRGSSAS   73
  96 I E   **   51.1  **  E LLLLLLLL   74
  97 A E   *    42.7        SSTSSSTS   75
  98 A E   *    42.8  **    IIMILMMV   76
  99 I E   **    0.0        KKKKKRRR   77
 100 S E   *    38.9        EEDKAELA   78
 101 M E   *    38.4        AALADLDK   79
 102 K E   *    35.1        TTAKQEED   80
 103 N E        39.0        EAPDREQE   81

pcthit       inf   Drawing structure
in pack.out
```

Figure 3.1: An example of the output generated my the MST program. Structure is compared with sequence, shown are the observed and predicted secondary structures (H = α helix, E = β strand) and the solvent exposure where ** is very buried and * is partly buried. The gaps indicated by a : are deletions in sequence or structure. The numbers between the sequence and structure sides indicate the score attained by the match of the positions.

3.3 Results

Here I illustrate the results of the top threadings, showing figures detailing the threading and where insertions and deletions are predicted. Also shown are the "correct alignments" with a SAP structural superposition showing the overlaid structure with the threading identified protein. Bear in mind these are the predicted folds and not the best folds which could be identified in hindsight. These are also shown separated to give a clearer idea of the similarities. The threading is colour-coded to show areas of insertion, in white, deleted structure, in blue, and hydropathy; with hydrophobic in red and hydrophilic in green. Multiple alignments for the targets are shown in the appendix at the end of this chapter. A summary of the proteins is given in Table 3.1

Id	Decision		Len.	Difficulty	Method	PDB code	Description
T0004	Right	(Ch4)	84	MED	NMR	1SRO	S1-motif of polyribonucleotide nucleotidyltransferase, E.coli
T0005	Wrong	(Ch3)	268	NEW	Xray	1FIB	C-terminal domain of gamma-fibrinogen, H. sapiens
T0006	n/a	(Ch3)	269	-	-	-	outer membrane phospholipase A, E.coli
T0011	Wrong	(Ch3)	220	NEW	Xray	1AH8	hsp-90, c-terminal domain, S. cerevisiae
T0014	Right	(Ch3)	252	MED	Xray	-	3-dehydroquinase, S. typhi
T0020	Wrong	(Ch3)	320	HARD	Xray	1AK1	Ferrochelatase, B. subtilis T. reesei
T0023	n/a	(Ch3)	284	-	-	-	KDO8P Synthase, E. coli
T0030	Right	(Ch3)	66	NEW	NMR	1FGP	domain 1 of protein g3, filamentous phage fd
T0031	Right	(Ch3)	242	EASY	Xray	1EXF	Exfoliative toxin A, S. aureus
T0037	Wrong	(Ch3)	109	NEW	Xray	1AA2	calponin-homology domain of beta-spectrin, H. sapiens
T0038	Right	(Ch3)	152	MED	NMR	1ULO	Cellulose-Binding domain, Endoglucanase C, C. fimi
T0042	Right	(Ch5)	78	NEW	NMR	1NKL	NK-lysin, S. scrofa

Table 3.1: Summary of CASP2 targets discussed. The second column indicates whether the fold was identified or a structure prediction was made correctly. In the case of new folds the decision not to submit a prediction was classified as correct. NEW – are new folds so impossible to predict by fold recognition; The other folds are classed as easy, medium or hard (Marchler-Bauer and Bryant, 1997). In the case of target 6, the structure has not been released. References are as follows: T0004 (Bycroft *et al.*, 1997), T0005 (Yee *et al.*, 1997), T0014 (Shrive *et al.*,), T0020 (Al-Karadaghi *et al.*, 1997), T0030 (Holliger and Riechmann, 1997), T0031 (Vath *et al.*, 1997), T0037 (Carugo *et al.*, 1997), T0038 (Johnson *et al.*, 1996) and T0042 (Liepinsh *et al.*, 1997).

3.3.1 Target T0004

Potential folds were easily recognised for this target, and due to the confidence of the results models were built based on three folds, as described in the next chapter, consequently T0004 was not submitted as a fold recognition target (but as an *ab initio* homology model). Chapter 4 describes in detail the identification of the fold of this protein and the consequent threading based homology modelling.

3.3.2 Target T0005

Figure 3.2 shows the threading and superposition between target 5 and 8fab Threading results for the target T0005, compared against a Database containing 319 Structures (extended UCLA benchmark set) See Jaroszewski *et al.* for comparison of this benchmark with various methods of fold recognition (Jaroszewski *et al.*, 1998). Only structures with more than 20% beta and more beta than alpha were considered. The model structure was taken from several high scoring alternatives on the basis that it could (be made to) form the known disulphide connections: J—J, U—U and O—N-terminus (BETA-chain only), as identified on sequence FIBB_BOVIN in the multiple alignment.

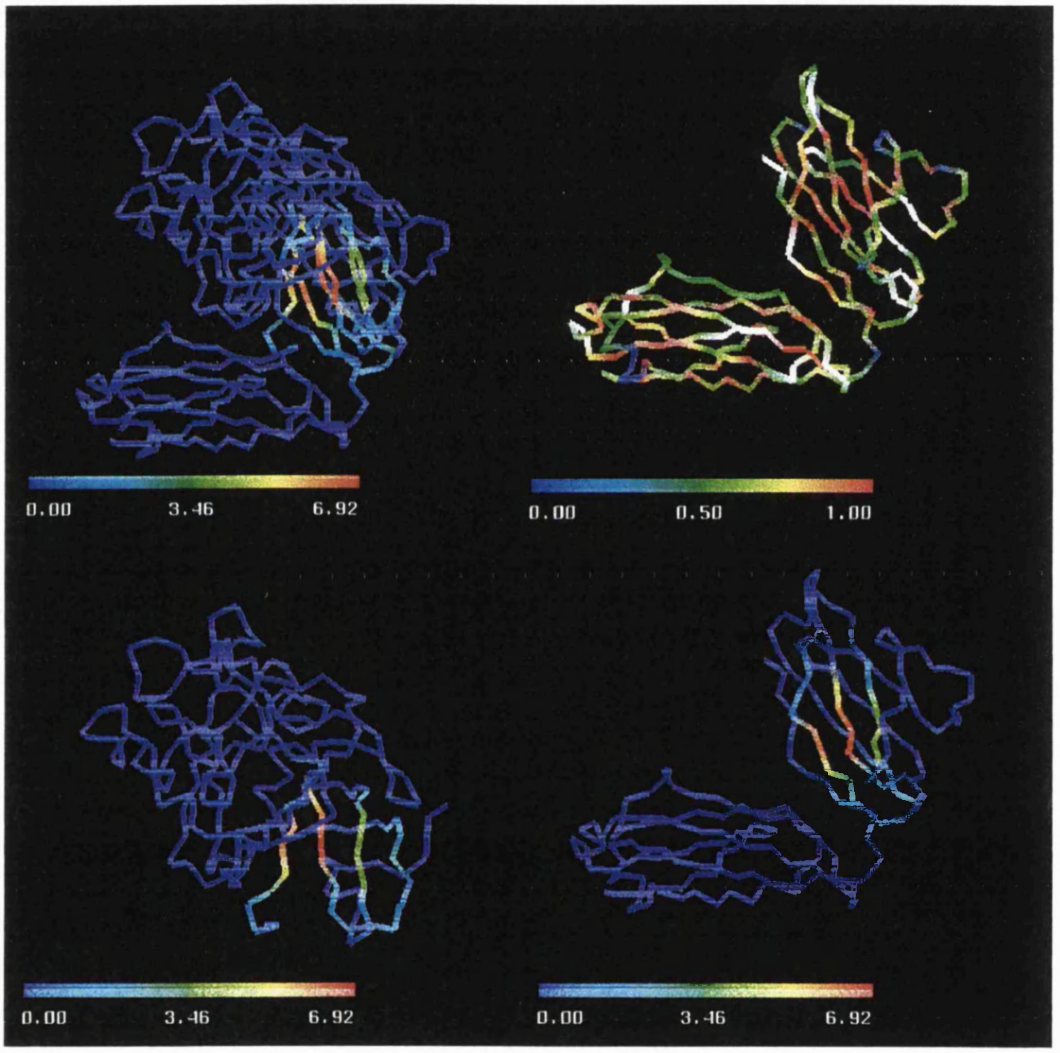


Figure 3.2: Threading of target 5. Showing the threading (top-right) and superposition (top-left) between target 5 and 8fab. The two figures at the bottom show the proteins split from the superposition, but still coloured according to the strength of the match. The orientation in all pictures is consistent and the predicted fold is on the right again underneath its threading prediction.

3.3.3 Target T0006

This target was examined and due to insufficient sequence information and a general lack of confidence in any of the prediction methods this was not submitted to CASP2. In fact the structure has not yet been released so no comparison could have been made.

3.3.4 Target T0011

Figure 3.3 shows the threading and superposition between target 11 and 2dnj. This figure allows for easy visual comparison of how well the threading results compare to a SAP structural comparison.

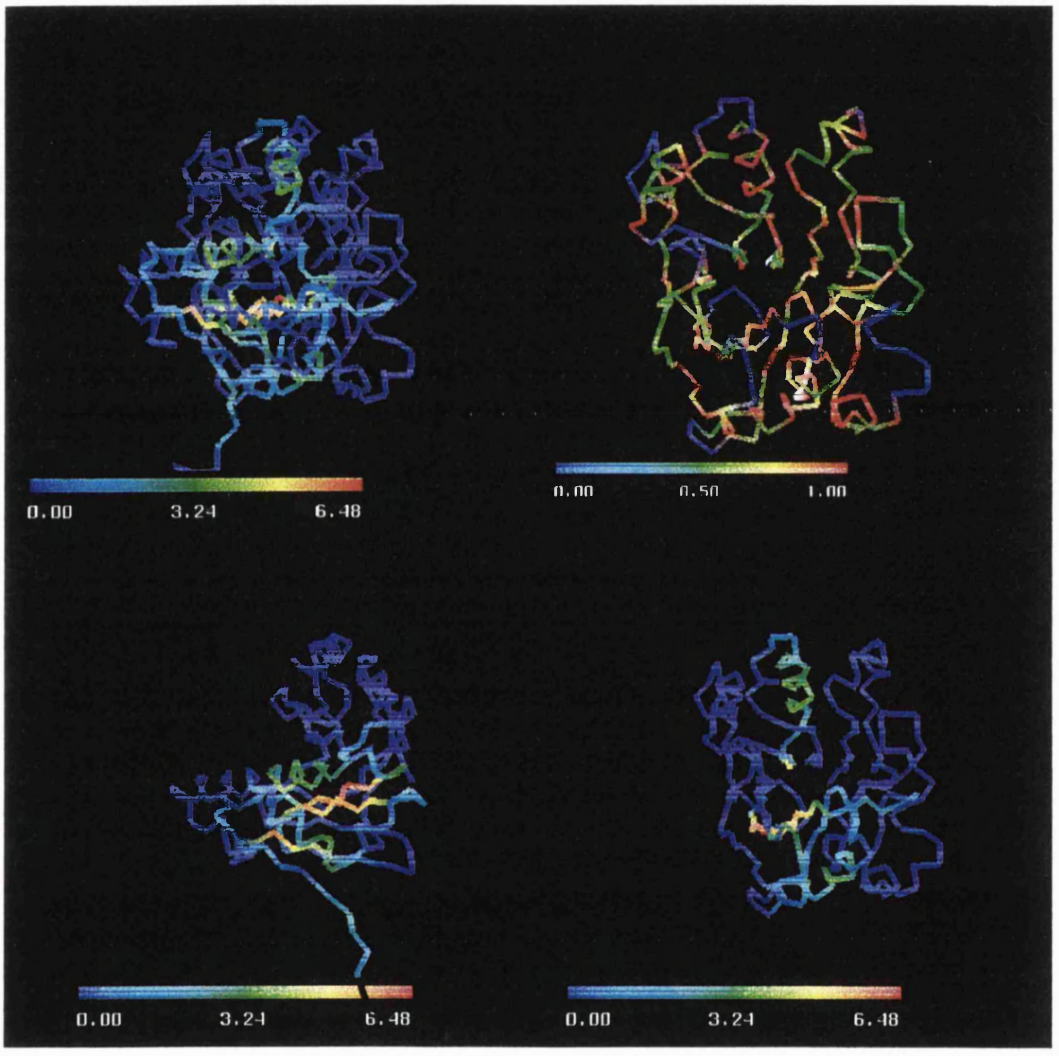


Figure 3.3: Superposition and threading of target 11 with 2dnj. The top left figure is a SAP superposition, highlighted by the weight matrix, so the "hotter" the picture the better the alignment. This figure has then been split to show the individual structures. The correct structure is shown bottom left. On the bottom right is shown the fold which was recognised by the fold recognition process. The top right figure is the threading made of the target sequence on to the template, coloured with blue for deleted structure, white for inserted sequence, and red is hydrophobic while green is not.

3.3.5 Target T0014

Figure 3.4 shows the threading and superposition between the CASP coordinates and 1tre, chain A. This figure has been split to show the individual structures. The correct structure is shown on the bottom left. On the right is shown the fold which was recognised by the fold recognition process. The top right figure is the threading made of the target sequence on to the template, coloured as mentioned earlier. Bottom right is the same structure coloured according to the SAP superposition.

The RMSD of the structure and 1tre is 5.5Å over all the matched atoms (217).

A prediction of another fold 1ipd was also made and it to shows structural homology with target 14, as assessed through the structure superposition method DALI which was run by the CASP2 assessors. However, this is not correct as 1ipd shows the Rossmann type fold, whereas the target is a TIM barrel. This shows how even structure comparison results should not be accepted without looking closely at the results.

Figure 3.5 shows an alignment of the target sequence to the predicted fold by the MST and by SAP structural alignment. It can be seen that the alignment predicted by the threading is not identical most of the time. Areas of perfect alignment are boxed.

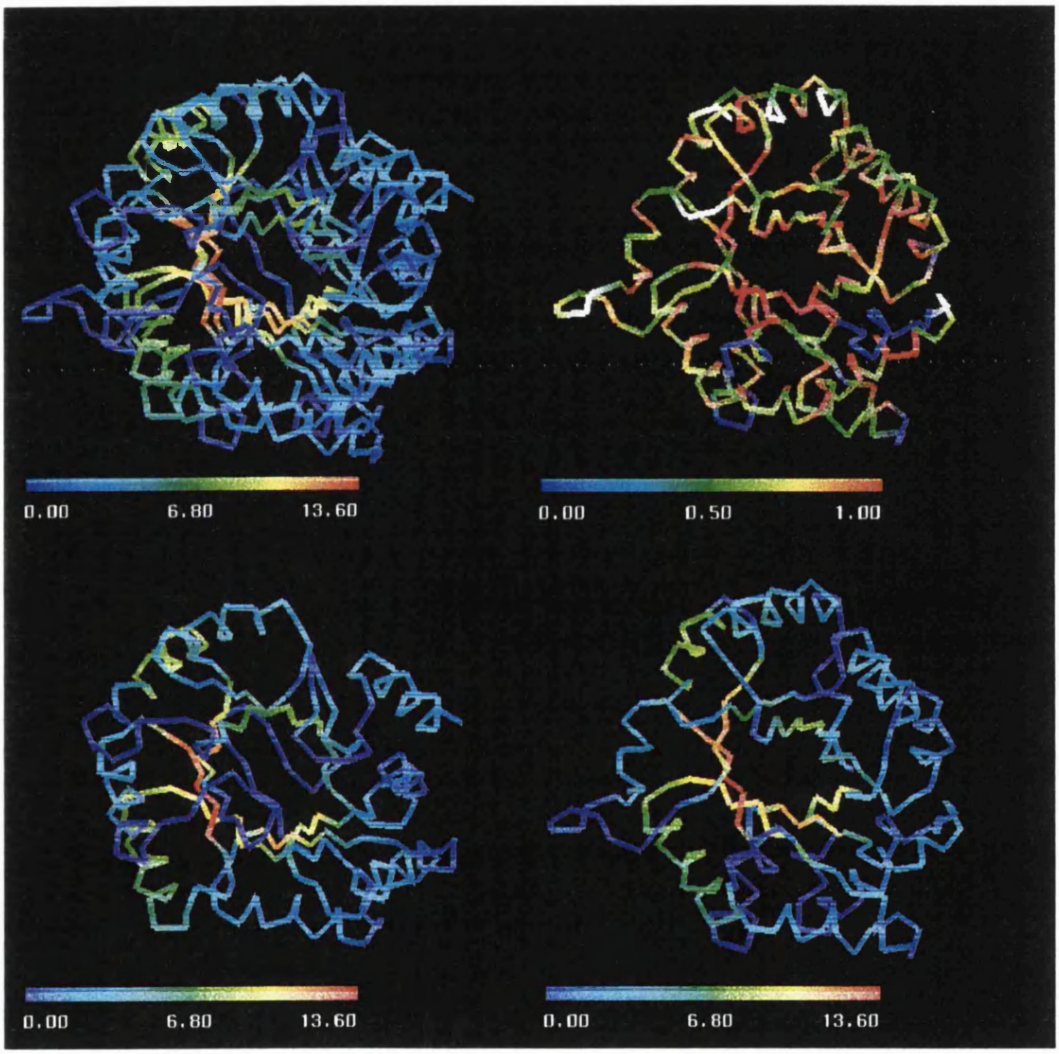


Figure 3.4: Superposition and threading of target 14 on 1tre, chain A. The superposition (top-left) and threading (top-right) are shown, along with a decomposition of the superposition shown below. The predicted fold is shown under its threading prediction highlighted for the structural superposition with the target structure solution. An alignment comparison between the SAP superposition shown here and the threading result is shown in Figure 3.5.

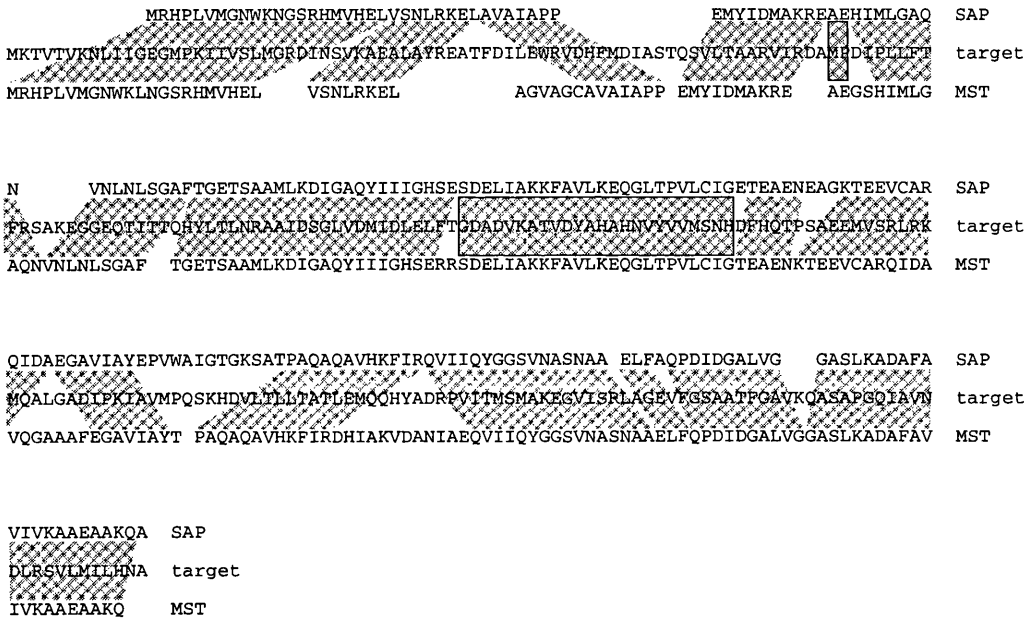


Figure 3.5: Alignment comparison of target 14 and 1tre. The target sequence (in this case T0014) is shown in the centre of the three sequences. Also shown is the mapping of 1tre by structural comparison (SAP) and by predicted threading (MST). Taking the SAP:1tre alignment as the best possible it is a simple case to compare how good the threading alignment is by showing how the sequence is shifted, or not. This is highlighted by the grey shading. A perfect threading alignment is boxed. The greater the shift in the alignment, the larger the slant of the grey shading. The less shift there is in the alignment, the better.

3.3.6 Target T0020

Figure 3.6 and Figure 3.7 shows the threading and superposition between target 20 and the first and second fold predictions of 1dmb and 2lbp. The latter is a correctly recognised fold although only part of the alignment is correct, as can be clearly seen in figure 3.7. The first choice, corresponding to a slightly higher scoring fold, is incorrect. This is shown in figure 3.6 `MST` scores of 16325 and 15965, respectively).

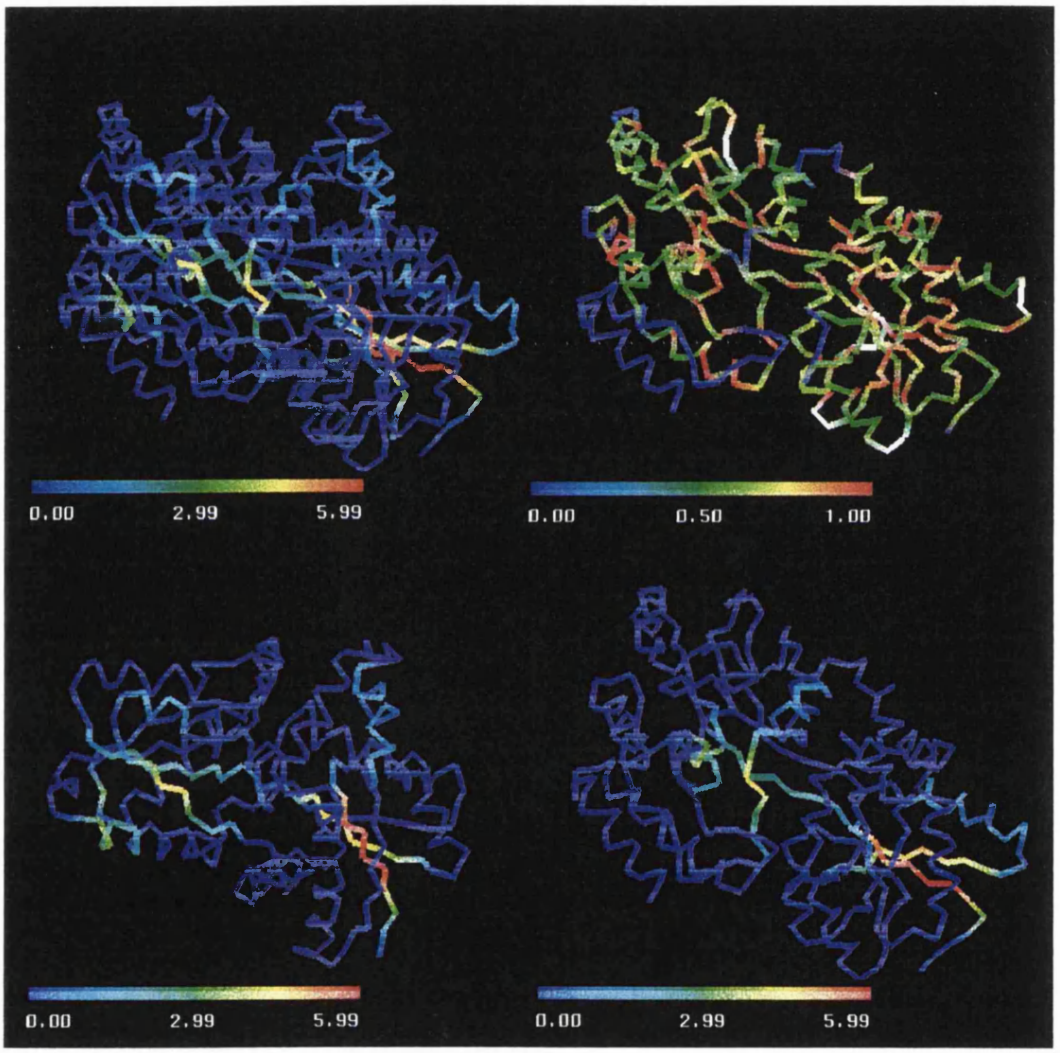


Figure 3.6: Superposition and threading of target 20 with 1dmb. The proteins are arranged in the same fashion as the previous figures. Easy comparison can be made between the predicted threading (top-right) and the superposition match on the same protein (bottom-right).

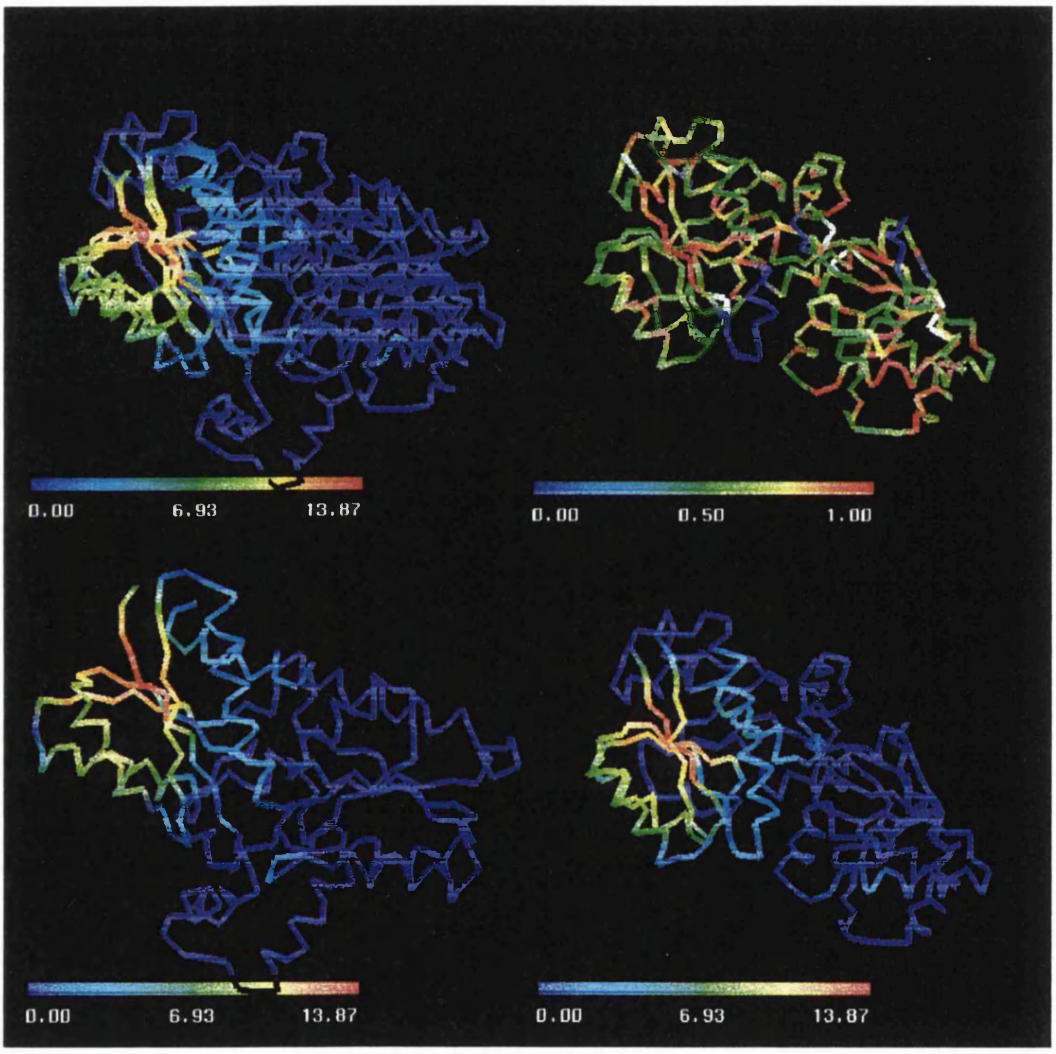


Figure 3.7: Superposition and threading of target 20 with 1lpb. This shows a better CASP prediction of this target compared to the previous figure.

3.3.7 Target T0023

The predicted fold for this target was 1ald. This target has yet not had a structure released by the experimentalists so no further analysis has been carried out.

3.3.8 Target T0030

After examining this protein it was decided that the fold could not be predicted. After release of the structure, it was found to be a new fold. I could take this as a correct answer as fold recognition could not have been correct in this case. As mentioned earlier, this is an inherent threading problem.

3.3.9 Target T0031

Figure 3.8 shows the superpositions between target 31 and 1gctA, 1hcgA, 1hf1, 1lmwB, 1smfE and 2cha. These protein folds were identified by sequence analysis, rather than threading. The sequence alignment can be seen in the appendix. Only one of the folds was submitted to CASP2, see figure 3.9. Figure 3.8 shows a multiple superposition of – top left: Xray structure of T0031; bottom left: superposition of 1gctA and correct structure; top right: highlights the 2 proteins in different colours; and bottom right: shows a multiple superposition of some more of the folds identified by the multiple alignment.

A threading output format (TALIGN) has been included for the first structure in the alignment (1GCT, A chain) see Figure 3.9, but not for any of the others. They would be fairly similar, as can be seen from the `MULTAL` alignment.

Target T0031 Figure 3.10 shows a plot similar to that of T0014 (Figure 3.5), illustrating areas where the `SAP` structural alignment of 1gctA with T0031 and the sequence alignment (in appendix) match up to the correct structure. Much of the alignment again shows inaccuracies, highlighting the major problem with this method.

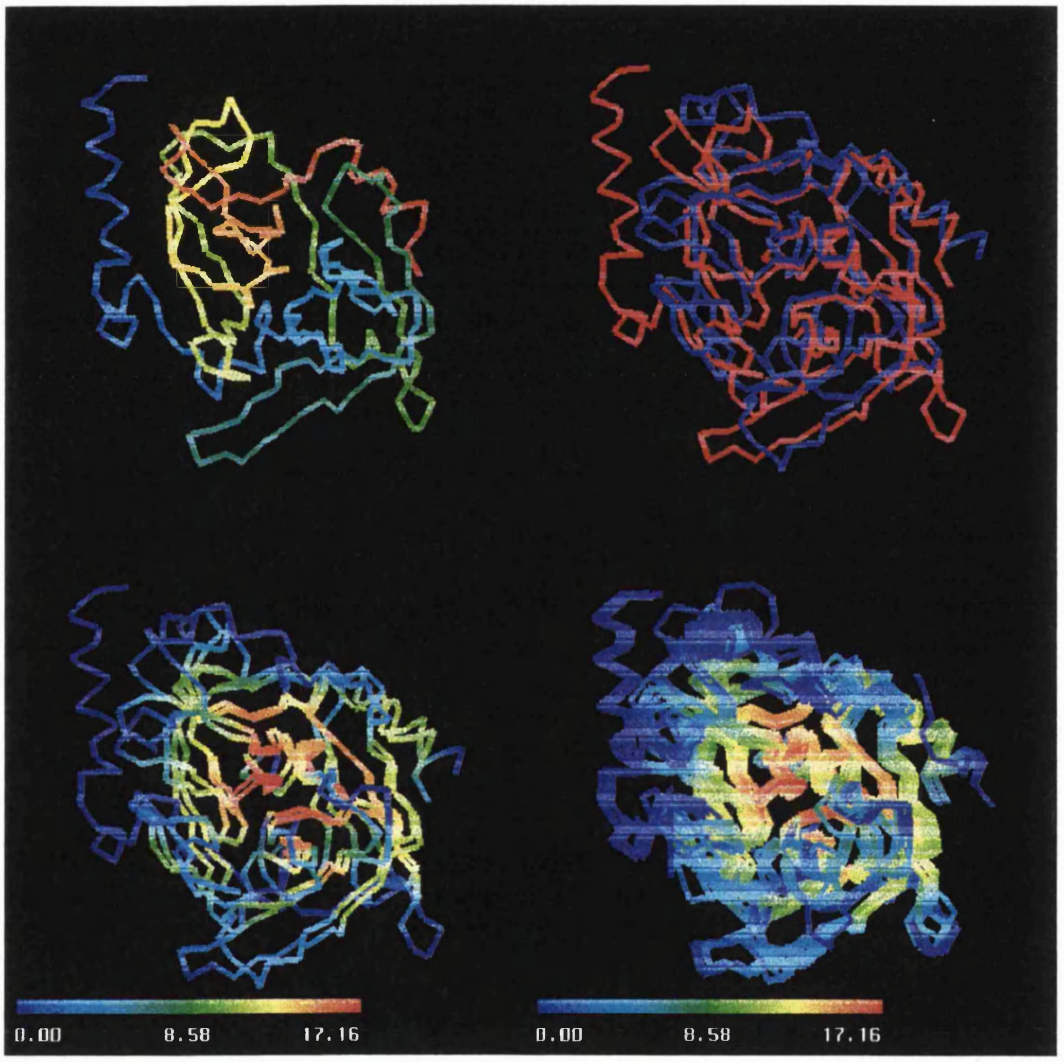


Figure 3.8: Superposition comparison of predicted structures of target 31. Shown is a superposition of the structure identified by multiple alignment with closest sequence homology (1gctA) and the solution (top-right, bottom-left). The "answer" is shown top-left and a multiple superposition of six identified folds. As can be seen 1gctA was not the identified fold with closest backbone homology.

```
================================================================================
TSCORE T0031 0  1.0     1GCT A 0
================================================================================
TALIGN T0031 0   33   39        1GCT A 0         16   22        1.0     1
TALIGN T0031 0   41   52        1GCT A 0         23   34        1.0     1
TALIGN T0031 0   53   74        1GCT A 0         38   59        1.0     1
TALIGN T0031 0   79   80        1GCT A 0         60   61        1.0     1
TALIGN T0031 0   81   93        1GCT A 0         63   75        1.0     1
TALIGN T0031 0   94   98        1GCT A 0         77   81        1.0     1
TALIGN T0031 0   99  115        1GCT A 0         83   99        1.0     1
TALIGN T0031 0  118  129        1GCT A 0        100  111        1.0     1
TALIGN T0031 0  130  141        1GCT A 0        113  124        1.0     1
TALIGN T0031 0  142  155        1GCT A 0        131  144        1.0     1
TALIGN T0031 0  156  174        1GCT A 0        155  173        1.0     1
TALIGN T0031 0  175  183        1GCT A 0        176  184        1.0     1
TALIGN T0031 0  185  185        1GCT A 0        188  188        1.0     1
TALIGN T0031 0  189  203        1GCT A 0        189  203        1.0     1
TALIGN T0031 0  205  213        1GCT A 0        208  216        1.0     1
TALIGN T0031 0  214  216        1GCT A 0        218  220        1.0     1
TALIGN T0031 0  217  240        1GCT A 0        222  245        1.0     1
REMARK
================================================================================
```

Figure 3.9: Threading alignment (TALIGN) prediction of target 31 and 1GCT, chain A. The format shows the corresponding predicted alignment between unknown sequence and fold prediction.

Figure 3.10: Alignment comparison of target 31 and 1GCT. See legend to Figure 3.5 for details. The only difference with this plot is that the predicted alignment was made with a multiple sequence alignment method (`MULTAL`) instead of the threading method.

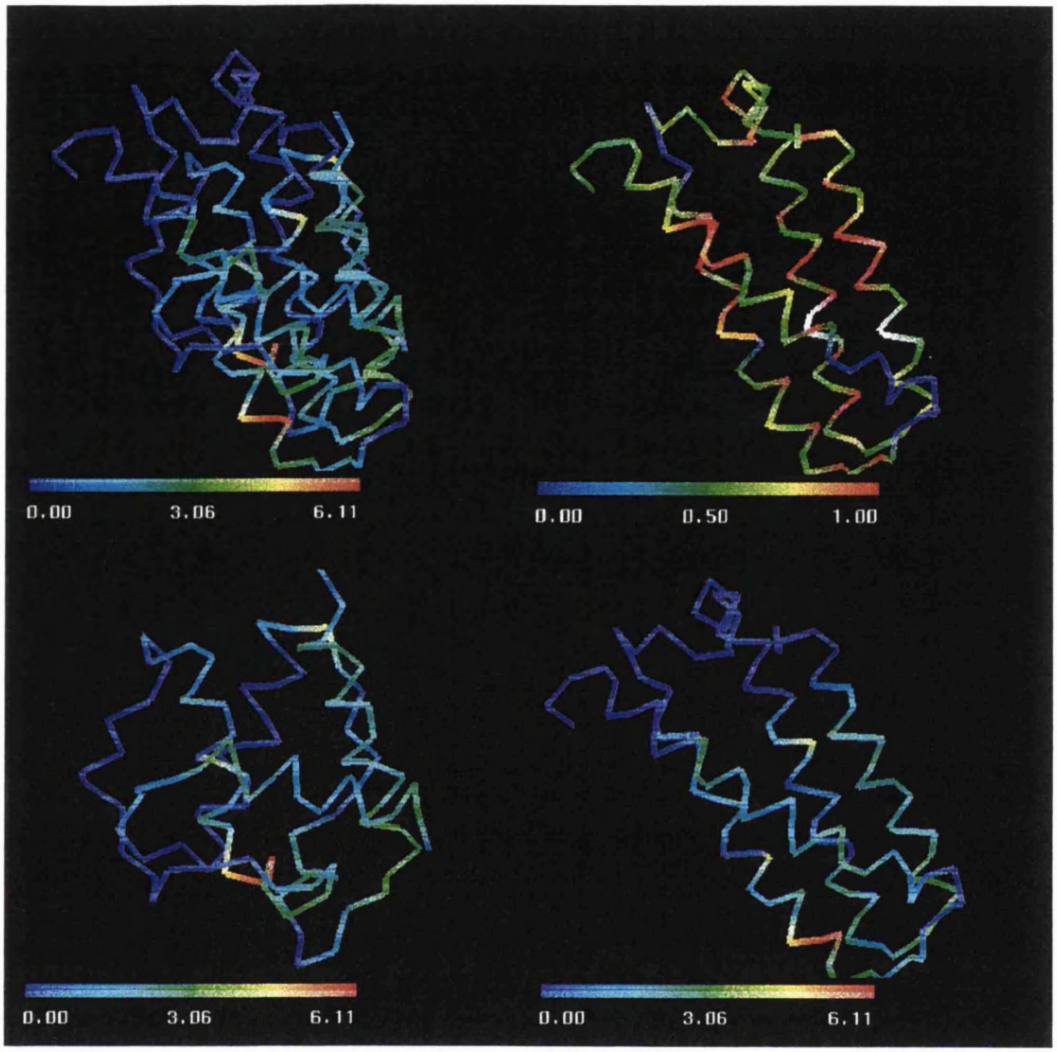


Figure 3.11: Superposition and threading of target 37 with 1cgo. This is a new fold and could not be identified by fold recognition. From the diagram it can be seen that some of the secondary structure elements do correspond, but this is not too surprising.

3.3.10 Target T0037

Figure 3.11 shows the threading and superposition between target 37 and 1cgo. Neither VAST nor DALI showed any similarity in the CASP assessment. So the recognised fold was incorrect. Structurally speaking, from the SAP analysis shown in figure 3.11 there are similarities between the proteins, in that they both contain helices, although the RMSD over the best 47 matched residues is 9.5Å.

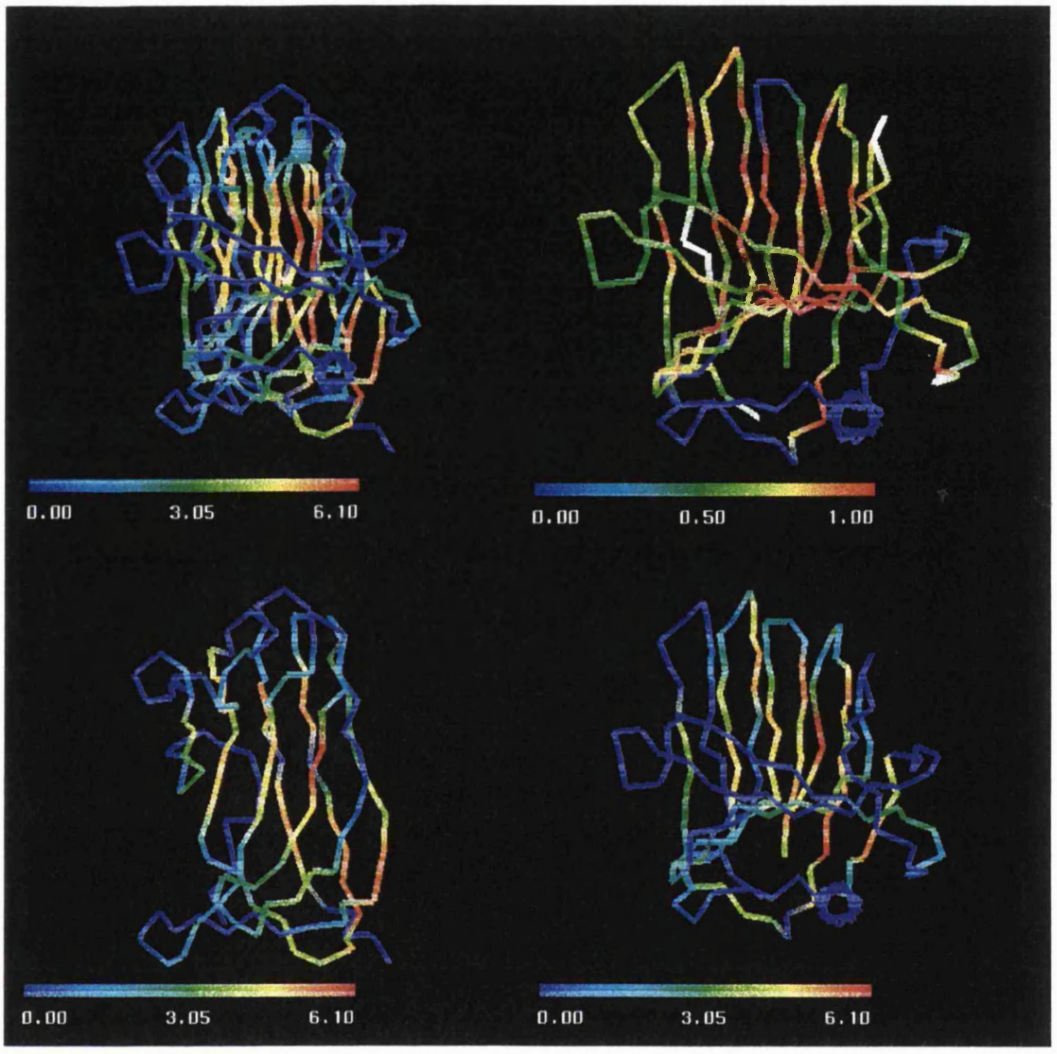


Figure 3.12: Superposition and threading of target 38 with 1xnb. Note the deleted helix in the threading (coloured blue), the strands crossing the front of the predicted fold do not occur in the structure (on the left).

3.3.11 Target T0038

Figure 3.12 shows the threading and superposition between target 38 and 1xnb. NB - a higher score was obtained for 4gcr (GAMMA-B CRYSTALLIN) score = 6621, however this was discounted due to the poor positions in the structure for the predicted disulphide bonds and the lack of duplication in the target sequence to fit with the double domained gamma crystallin. The fold was correctly identified.

3.4 Conclusion

Table 3.1 shows a summary of the predictions showing four out of eight predictions correct. Of the proteins with previously know folds, three out of four were correct. If the moderately easy threading target discussed in Chapter 4 is also included, then four out of five were correctly identified. Three folds were predicted, when there were no known folds, i.e. classified as impossible, it is these proteins which are particularly hard to know when not to predict a fold.

Fold recognition is a bit like secondary structure prediction: it is often quite accurate at finding the correct structure in roughly the right place, but where the ends of that structure lie is far harder to predict accurately. Similarly with fold recognition, a fold may be recognised due to a reasonable match with secondary structure, hydrophobic cores and whatever else that may make up the *potential* but when it comes to accurate alignment of the sequence on the structure then the register may be shifted: typically by the periodicity of the α-helix. In other cases the alignment of a sheet may be shifted. In many cases the alignment is just plainly wrong. Possible improvements to the gap function of the threading method used here are detailed in Chapter 6. The `MST` results make use of the *cones* algorithm discussed in Chapter 2, as one of the main driving forces for packing hydrophobic residues on the template structures. Where this has worked well is obvious from the coloured threadings showing amphipathic helices and the like.

Other groups who submitted to the CASP2 experiment also found similar problems

with accurate alignments. Other methods used included hidden Markov models, ID/3D fold recognition and mean force potential based threading. In general the groups who did well achieved four out of six correctly predicted folds, these are detailed in the CASP2 special issue of proteins (CASP, 1997). Of particular note was the Murzin group who correctly identified six folds by having a profound knowledge of the PDB and extensive use of the SCOP classification of structures.

It is worth noting that even with structural alignment programs such as those used in the CASP assessment (`VAST`, `DALI` and `SSAP`) do not agree with each other in the exact alignments. They do mostly agree, however, on which folds are similar.

3.5 Appendix: CASP2 alignments

Multiple alignment for T0005:

```
*USER>>P1;T0005:165 :      gamma fibrinogen C terminal (threading target 5)
*USER>>P1;FIBG_XENLA:330 : FIBRINOGEN GAMMA-CHAIN PRECURSOR - X.LAEVIS
*USER>>P1;FIBG_PETMA:329 : FIBRINOGEN GAMMA CHAIN  - P.MARINUS
*USER>>P1;JN0596:229 :     fibrinogen-related HFREP-1 precursor - human
*USER>>P1;FIB2_PETMA:551 : FIBRINOGEN ALPHA-2 CHAIN  - P.MARINUS
*USER>>F1;B44234:176 :     fibrinogen alpha chain, extended form - human
*USER>>P1;CHKFIBAB:655 :   CHKFIBAB extended fibrinogen - Gallus gallus
*USER>>P1;FIBB_BOVIN:377 : FIBRINOGEN BETA CHAIN. - BOS TAURU
*USER>>P1;FIBB_CHICK:373 : FIBRINOGEN BETA CHAIN - GALLUS GALLUS
*USER>>P1;RNU05675:388 :   RNU05675 NCBI gi: 455105 - Rattus norvegicus
*USER>>P1;FIBB_PETMA:388 : FIBRINOGEN BETA CHAIN - PETROMYZON MARINUS
*USER>>P1;FIBX_MOUSE:344 : FIBRINOGEN-LIKE PROTEIN  - MUS MUSCULUS
*USER>>P1;FIBA_PARPA:205 : FIBRINOGEN-LIKE PROTEIN A   - P.PARVIMENSIS
*USER>>P1;B47172:251 :     ficolin-beta - pig
*USER>>P1;FIBL_HUMAN:739 : FIBRINOGEN-LIKE PROTEIN - HOMO SAPIENS
*USER>>F1;S28170:139 :     tenascin homolog - pig
*USER>>P1;JH0675:1266 :    restrictin precursor - chicken
*USER>>P1;JQ1322:1935 :    tenascin precursor - mouse
*USER>>P1;SCA_DROME:643 :  SCABROUS PROTEIN PRECURSOR - D.MELANOGASTER

itgkdcqdiank  gakqsglyfikplkanqqflvyceidgsgngwtvfqkrldgsvdfkknwiqykegfghlsptgt
ftgkdcqevank  garlsglyyikplkakqqflvyceiepsgsawtviqrrldgsvnfhknwvqyregfgylspndk
itgkdcqqvvdn  ggkdsglyyikplkakqpflvfceie ngngwtviqhrhdgsvnftrdwvsyregfgylaptlt
rqyadcseifnd  gyklsgfykikplqspaefsvycdms dgggwtviqrrsdgsenfnrgwkdyengfgnfvqkhg
seyidcldvlqrrpggkasglyevrprgakraltvhceqdtdgggwtlvqqredgslnfnrsfsayregfgtvdgsgh
rpvrdcddvlqthpsgtqsgifniklpgsskifsvycdqetslggwlliqqrmdgslnfnrtwqdykrgfgslndege
yngkdcddirqkhtsgaksgifkikpegsnkvlsvycdqettlggwlliqqrmdgsvnfnrtwqdyrrgfgsvdgkgq
vsgkeJekiirn  egetsemyliqpedsskpyrvyJdmktekggwtviqnrqdgsldfgrkwdpykqgfgniatnaegk
vsgrecediyrk  ggetsemyiiqpdpfttpyrvycdmetdnggwtliqnrqdgsvnfgrawdeykrgfgniaksg gk
vsgkeceeiirk  ggetsemyliqpdtsskpyrvycdmktenggwtviqnrqdgsvdfgrkwdpykkgfgniatnedtk
vsgmhcediyrn  ggrtseayyiqpdlfsepykvfcdmeshgggwtvvqnrvdgssnfardwntykaefgnia fgngk
liykdcsdhyvl  grrssgayrvtpdhrnssfevycdmetmgggwtvlqarldgstnftrewkdykagfgnle
eyprdcydilqscsgqsppsgqyyiqpdggnlikvycdmetdeggwtvfqrridgtinfyrswsyyqtgfgnln
tgprtckelltrghfls   gwhtiylpdcqpltvlcdmdtdgggwtvfqrrsdgsvdfyrdwaaykrgfgsql
pfprdcgeemqngagasrtst iflngnrerplnvfcdmetdgggwlvfqrrmdgqtdfwrdwedyahgfgnis
pfprdcgeemqngvstsrttt iflngnrerplnvfcdmetdgggwlvfqrrmdgktdfwrdwedyahgfgnis
anpqdcaqhlmngdtlsgvyt isingdlsqrvqvfcdmstdgggwivfqrrqngltdffrkwadyrvgfgnle
pfprdcsqamlngdttsglyt iyingdktqalevycdmtsdgggwivflrrkngredfyrnwkayaagfgdrr
klphdcsevhtqtdglh     liapagqrhplmthctad    gwttvqrrfdgsadfnrswadyaqgfgapg
```

```
      tefwlgnekihlistqsaipyalrveledwngrtstadyamfkvgpeadkyrltyayfaggdagdafdgfdfgd
      tefwlgnekihllstqstipyvmrieledwsnqkstadystfrlgsekdnyrftyayfiggdagdafdgfdfgd
      tefwlgnekihlltgqqa  yrlridltdwenthryadyghfkltpesdeyrlfysmyldgdagnafdgfdfgd
       eywlgnknlhflttqed  ytlkidladfeknsryaqyknfkvgdeknfyelnigey    sgtagdslagnf
      gelwlgleamyllaheds   tmrvelqgwdgagahaey tvtlrddskgyalqvsdy    rgtagnalvsg
      gefwlgndylhlltqrgs   vlrveledwagneayaey hfrvgseaegyalqvssy    egtagdaliegs
      gelwlgnenihlltqndt   llrveledwdgnaayaey ivqvgteaegyaltvssy    egtagdalvagw
kyOgvpgeywlgndrisqltnmgpt  klliemedwkgdkvtalyegftvqneankyqlsvsky    kgtagnaliega
kycdtpgeywlgndkisqltkigpt  kvliemedwngdkvsalyggftihnegnkyqlsvsny    kgnagnalmega
kycglpgeywlgndkisqltrigpt  elliemedwkgdkvkahyggftvqteankyqvsvnky    kgtagnalmega
sicnipgeywlgtktvhqltkqhtq  qvlfdmsdwegssvyaqyasfrpeneaqgyrlwvedy    sgnagnallega
      refwlgndkihlltkskem  ilridledfngltlyalydqfyvaneflkyrlhigny    ngtagdalr
      tefwlgndnihyltsqgd  yelrvelnntlgnhyyakynkfrigdsfseyllvlgay    sgtagdsl
      gefwlgndhihaltaqgt  selrvdlvdfegnhqfakyrsfqvageaekyklvlggfl   egnagdsl
      gefwlgnealhsltqagd  ysirvdlrag deavfaqydsfhvdsaaeyyrlhlegy    hgtagdsm
      gefwlgnealhsltkagd  yslrvdlrag eeavfaqyesfqvdsaaehyrlhlegy    hgtagdsm
      defwlgldnihkitsqgr  yelridmrdg qeaayayydkfsvgdsrslyklrigdy    ngtsgdsl
      eefwlgldnlskitaqgq  yelrvdlqdh gesayavydrfsvgdaksryklkvegy    sgtagdsm
      gefwigneqlhhltldnc  srlqvqmqdiydnvwvaeykrfyissradgyrlhiaey    sgnasdal

dp   sdkfftsHNGMQFSTWDNDNDKF     EGNCAEQD   GSGWWMNKCHAGHLNGVYyqggtyska stpngydn
dp   sdkfytsHNGMQFSTFDKDNDKF     DGNCAEQD   GSGWWMNRCHAAHLNGKYyqggtyseadsgpsgydn
dp   qdkfyttHLGMLFSTPERDNDKY     EGSCAEQD   GSGWWMNRCHAGHLNGKYyfggnyrktdve fpydd
hp   evqwwasHQRMKFSTWDRDHDNY     EGNCAEED   QSGWWFNRCHSANLNGVYysgpytaktd       n
va   ddpeltsHGGMTFSTYDRDTDKW    SDGSCAEWY   GGGWWINACQAANLNGVYyqggpydprekppyeven
ve   egaeytsHNNMQFSTFDRDADQW    EE NCAEVY   GGGWWYNNCQAANLNGIYypggsydprnnspyeien
le   egseytsHAQMQFSTFDRDQDHW    EE SCAEVY   GGGWWYNSCQAANLNGIYypgghydprynvpyeien
sqlvgenrtmtiHNSMFFSTYDRDNDGWKTTDPRKQUSKED   GGGWWYNRUHAANPNGRYywggaytwdmakh gtdd
sqlygenrtmtiHNGMYFSTYDRDNDGWLTTDPRKQCSKED   GGGWWYNRCHAANPNGRYywggtyswdmakh gtdd
sqlvgenrtmtiHNGMFFSTYDRDNDGWVTTDPRKQCSKED   GGGWWYNRCHAANPNGRYywgglyswdmskh gtdd
tqlmgdnrtmtiHNGMQFSTFDRDNDNWNPGDPTKHCSRED   AGGWWYNRCHAANPNGRYywggiytkeqady gtdd
      fsrhynHDLRFFTTPDRDNDRY    PSGNCGLYY   SSGWWFDSCLSANLNGKYyhqkykgvrngifwgtwp
          ayHNTMRFSTYDNDNDVY     SINCASHSSYGRGAWWYKSCLLSNLNGQYy   dysga        p
          ssHRDQFFSTKDQDNDNH     SGNCAEQY   HGAWWYNACHSSNLNGRYlrglhtsya        n
          syHSGSVFSARDRDPNSL     LISCAVSY   RGAWWYRNCHYANLNGLYg   stvdh        q
          syHSGSVFSARDRDPNNL     LISCAVSY   RGAWWYRNCHYANLNGLYg   stvdh        q
          tyHQGRPFSTKDRDNDVA     VTNCAMSY   KGAWWYKNCHRTNLNGKYg   esrhs        q
          nyHNGRSFSTYDKDTDSA     ITNCALSY   KGAFWYKNCHRVNLMGRYg   dnnhs        q
          nyQQGMQFSAIDDDRDIS     QTHCAANY   EGGWWFSHCQHANLNGRYnlgltwfda        a
```

```
giiwatwk                        trwysmkkttmkiipf
giiwatwr                        rrwysmksvtmkimpl
giiwatwh                        drwyslkmttmkllpm
givwytwh                        gwwyslksvvmkirpn
gvvwatyr                        gsdyslkrtavrfrrv
gvvwvsfr                        gadyslravrmkir
gvvwipfr                        asdyslkvvrmkirpl
gvvwmnwq                        gswysmkkmsmkirpy
givwmnwk                        gswysmkkmsmkikpy
gvvwmnwk                        gswysmrrmsmkirpv
gvvwmnwk                        gswysmrqmamklrpk
ginqaqpg                        gykssfkqakmmirpk
siywsylp                        gdndqipfaemklrnr
gvnwrsgr                        gynysyqvsemkvrlt
gvswyhwk                        gfefsvpftemklrpr
gvswyywk                        gfefsvpftemklrpr
ginwyhwk                        ghefsipfvemkmrpy
gvnwfhwk                        gheysiqfaemklrps
rnewiavkssrmlvkrlpavecqanasasgafvsvsgsaadaaps
```

Multiple alignment for T0011:

```
     ..p- USER>>P1;HS83_LEIAM:145
     p-b- USER>>F1;JQ0129:160
     b--- USER>>P1;HS9B_MOUSE:154
815
     p--- USER>>P1;T0011:146
     b-p- USER>>P1;S49155:146
     ..b- USER>>P1;HS82_MAIZE:158


block 0 = 6 seqs
*USER>>P1;HS83_LEIAM:   HEAT SHOCK PROTEIN
*USER>>F1;JQ0129:       86K heat shock protein IV - Human
*USER>>P1;HS9B_MOUSE:   HEAT SHOCK PROTEIN HSP 84
*USER>>P1;T0011:        N-terminal part of heat shock protein HSP90
*USER>>P1;S49155:       heat-shock protein - Plasmodium falciparum
*USER>>P1;HS82_MAIZE:   HEAT SHOCK PROTEIN 82

               mtetfafqaeinqlmsliintfysnkeiflrdvisnasdacdkiryqsltdpsvlgdatrlcvrv
mpeetqtqdqpmeeeevetfafqaeiaqlmsliintfysnkeiflrelisnssdaldkiwyesltdpskldsgkelhinl
      peevhhgeeevetfafqaeiaqlmsliintfysnkeiflrelisnasdaldkiryesltdpskldsgkelkidi
              masetfefqaeitqlmsliintvysnkeiflrelisnasdaldkirykslsdpkqletepdlfiri
              mstetfafnadirqlmsliintfysnkeiflrelisnasdaldkiryesitdtqklsaepeffiri
   masadvhmaggaetetfafqaeinqllsliintfysnkeiflrelisnasdaldkirfesltdksnvnaqpelfirl

vpdkenktltvednggigmtkadlvnnlgtiarsgtkafmealeaga dmsmigqfgvgfysaylvadrvtvtsknnsdev
ipnkqdqtltivdtggigmtkadlinnlgtiaksgtkafmealqaga dismigqfgvsfysaylvaekvtvitkhnddeq
lpnpqertltlvdtggigmtkadlinnlgtiaksgtkafmealqaga dismigqfgvgfysaylvaekvvvitkhnddeq
tpkpeqkvleirdsggigmtkaelinnlgtiaksgtkafmealsaga dvsmigqfgvgfyslflvadrvqvisksnddeq
ipdktnntltiedsggigmtkndlinnlgtiarsgtkafmeaiqasg dismigqfgvgfysaylvadhvvviskdnddeq
vpdkasktlsiidsgvgmtksdlvnnlgtiarsgtkefmealaagatdvsmigqfgvgfysaylvadrvmvttkhnddeq

YVWESSAGGTFTitsa pesdmklparitlhlkedqleylearrlkelikkhsefigydielmvekttekevtded
YAWESSAGGSFTvrtd tgermgrgtkvilhlkedqteyleeqrikeivkkhsqligypitlfvekecdkevsdde
YAWESSAGGSFTvrad hgepigrgtkvilhlkedqteyleerrvkevvkkhsqfigypitlylekerekeisdde
YIWESNAGGSFTvtldevnerigrgtilrlflkddqleyleekrikevikrhsefvaypiqlvvtkevekevpip
YVWESAAGGSFTvtkdetneklgrgtkiilhlkedqleyleekrikdlvkkhsefisfpiklycerqnekeisas
YVWESQAGGSFTvthdttgeqlgrgtkitlflkddqleyleerrlkdlvkkhsefisypiylwtekttekeisdd
```

Multiple alignment for T0014:

```
     ..p- USER>>T0014;
     p-b- USER>>P1;AROD_SALTI
     b-p- USER>>P1;AROD_ECOLI
     ..b- USER>>P1ROD_BACSU
637
     p--- USER>>P1;ACU20284
     b--- USER>>P1;EBSD_ENTFA


*USER>>T0014; :          3-Dehydroquinase from Salmonella typhimurium (252a.a.)
*USER>>P1;AROD_SALTI :   3-DEHYDROQUINATE DEHYDRATASE - SALMONELLA TYPHI.
*USER>>P1;AROD_ECOLI :   3-DEHYDROQUINATE DEHYDRATASE - ESCHERICHIA COLI.
*USER>>P1ROD_BACSU :     3-DEHYDROQUINATE DEHYDRATASE - BACILLUS SUBTILIS.
*USER>>P1;ACU20284 :     ACU20284 NCBI gi: 644873 - Acinetobacter calcoaceticus
*USER>>P1;EBSD_ENTFA :   PROBABLE 3-DEHYDROQUINATE DEHYDRATASE - E. FAECALIS

MKTVTVKNLIIGEGMPKIIVSLMGRDINSVKAEALAYREA TFDILEWRVDHFMDIASTQSVLTAARVIRDAMPDIPLLF
MKTVTVKNLIIGEGMPKIIVSLMGRDINSVKAEALAYREA TFDILEWRVDHFMDIASTQSVLTAARVIRDAMPDIPLLF
MKTVTVKDLVIGTGAPKIIVSLMAKDIASVKSEALAYREA DFDILEWRVDHYADLSNVESVMAAAKILRETMPEKPLLF
MNVLTIKGVSIGEGMPKIIIPLMGKTEKQILNEAEAVKLL NPDIVEWRVDVFEKANDREAVTKLISKLRKSLEDKLFLF
KSTYVVKNLNIGDLPVKTLVPITAKTREQALAQAKVIAENKDADIAEFRIDLLEFASDTKKVIALGQELNQILKDKPLLA
MKPVIVKNVRIGEGNPKIVVPIVAPTAEDILAEATA SQTLDCDLVEWRLDYYENVADFSDVCNLSQQVMERLGQKPLLL

TFRSAKEGGEQTITTQHYLTLNRAAIDSGLVDMIDLELFTGDADVKATVDYAHAHNVYVVMSNHDFHQTPSAEEMVSRLR
TFRSAKEGGEQTITTQHYLTLNRAAIDSGLVDMIDLELFTGDADVKATVDYAHAHNVYVVMSNHDFHQTPSAEEMVSRLR
TFRSAKEGGEQAISTEAYIALNRAAIDSGLVDMIDLELFTGDDQVKETVAYAHAHDVKVVMSNHDFHKTPEAEEIIARLR
TFRTHKEGGSMEMDESSYLALLESAIQTKDIDLIDIELFSGDANVKALVSLAEENNVYVVMSNHDFEKTPVKDEIISRLR
TIRTSNEGGKLKVTDQEYEKIYSEYLKKPFMQLLDIEMFRDQAAVAKLTKLAHQKKVLVVMSNHDFDKTPSEQEIVSRLL
TFRTQKEGGEMAFSEENYFALYHELVKKGALDLLDIELFANPLAADTLIHEAKKAGIKIVLCNHDFQKTPSQEEIVARLR

KMQALGADIPKIAVMPQSKHDVLTLLTATLEMQQHYADRPVITMSMAKEGVISRLAGEVFGSAATFGAVKQASAPGQIAV
KMQALGADIPKIAVMPQSKHDVLTLLTATLEMQQHYADRPVITMSMAKEGVISRLAGEVFGSAATFGAVKQASAPGQIAV
KMQSFDADIPKIALMPQSTSDVLTLLAATLEMQEQYADRPIITMSMAKTGVISRLAGEVFGSAATFGAVKKASAPGQISV
KMQDLGAHIPKMAVMPNDTGDLLTLLDATYTMKTIYADRPIITMSMAATGLISRLSGEVFGSACTFGAGEEASAPGQIPV
KQDQMGADILKIAVMPKSKQDVFTLMNATLKVSEQ STKPLLTMSMGRLGTISRIATANMGGSLSFGMIGEASAPGQIDV
QMQMRQADICKIAVMPQDATDVLTLLSATNEMYTHYASVPIVTMSMGQLGMISRVTGQLFGSALTFGSAQQASAPGQLSV

NDLRSVLMILHNA
NDLRSVLMILHNA
NDLRTVLTILHQA
SELRSVLDILHKNTRG
TALKQFLKTVQPTP
QVLRNYLKTFEQNK
```

Multiple alignment for T0023:

```
     ..........p--- USER>1ald
     ........p-b--- USER>>P1;JX0233
     ......p-b----- USER>>P1;ALFH_CAEEL
     ....p-b-p----- USER>>P1;ALF1_PEA
     ....|...b----- USER>>P1;ALF_PLAFA
     ....b-p------- USER>>P1;ALFC_SPIOL
     ......b------- USER>>P1;ALF_TRYBB
165
     ............p- USER>>P1;KDSA_HAEIN
     ............b- USER>>T0023;


block 8 = 9 seqs
*USER>1ald :    ALDOLASE HUMAN SKELETAL MUSCLE: range=1-363, len=363
*USER>>P1;JX0233 :      Fructose-1,6-bisphosphate aldolase alpha - Fruit fly
*USER>>P1;ALFH_CAEEL :  PROBABLE FRUCTOSE-BISPHOSPHATE ALDOLASE - C. ELEGANS.
*USER>>P1;ALF1_PEA :    FRUCTOSE-BISPHOSPHATE ALDOLASE (GARDEN PEA).
*USER>>P1;ALF_PLAFA :   FRUCTOSE-BISPHOSPHATE ALDOLASE  - P.FALCIPARUM.
*USER>>P1;ALFC_SPIOL :  FRUCTOSE-BISPHOSPHATE ALDOLASE, CHLOROPLAST
*USER>>P1;ALF_TRYBB :   FRUCTOSE-BISPHOSPHATE ALDOLASE, GLYCOSOMAL
*USER>>P1;KDSA_HAEIN :  2-DEHYDRO-3-DEOXYPHOSPHOOCTONATE ALDOLASE
*USER>>T0023; : KDO 8-P Synthase from E.coli, 284a.a.

          PYQYPALTPEQKKELSDIAHRIVAPGKGILAADESTGSIAKRLQSIGTENTEENrRFYRQLLLTADDRVN
       MTTYFNYPSK  ELQDELREIAQKIVAPGKGILAADESGPTHGKRLQDIGVENTEDNRRAYRQLLFSTDPKLA
     MATVGGAFKDSLTQAQKDELHQIALKIVQDGKGILAADESTGTIGKRLDAINLENNETNRQKYRQLLFTT PNLN
          MSAFVGKYAD   ELIKNAKYIATPGKGILAADESTGTIGKRLASINVENIEANRQALRELLFTS PNAL
   MAHCTEYMNAPKKLPADVAEELATTAQKLVQAGKGILAADESTQTIKKRFDNIKLENTIENRASYRDLLFGT KGLG
VAGVRFTPSGSSSLTVRASSYADELVKTAKTVASPGRGILAMDESNATCGKRLASIGLENTEANRQAYRTLLISA PGLG
SKRVEVLLTQLPAYNRLKTPYEAELIETAKKMTAPGKGLLAADESTGSCSKRFAGIGLSNTAEHRRQYRALMLEC EGFE
                           MQNKIVKIGNIDVAND                                    KP
                           MKQKVVSIGDINVAND                                    LP

PCIGGVILFHETLYQKADDGRPFPQVIKSKGGVVGIKV  DKGVVPLA GTNGETTTQGLDGLSERCAQYKKDGADFAkW
ENISGVILFHETLYQKADDGTPFAEILKKKGIILGIKV  DKGVVPLF GSEDEVTTQGLDDLAARCAQYKKDGCDFAKW
QHISGVILYEETFHQSTDKGEKFTDLLIKQGIVPGIKL  DLGVVPLA GTIGEGTTQGLDKLAERAAAFKKGGCGFAKW
QYLSGVILFEETLYQKSSEGKPFVEILQENNVIPGIKV  DKGVVELA GTDGETTTQGFDSLGARCQQYYKAGARFAKW
KFISGAILFEETLFQKNEAGVPMVNLLHNENIIPGIKV  DKGLVNIP CTDEEKSTQGLDGLAERCKEYYKAGARFAKW
QYVSGAILFEETLYQSTTDGKKMVDVLIEQGIVPGIKV  DKGWLPLP GSNDESWCQGLDGLACRSAAYYQQGARFAKW
QYISGVILHDETVYQKAKTGETFPQYLRRRGVVPGIKT  DCGLEPLVEGAKGEQMTAGLDGYIKRAKKYYAMGCRFCKW
FVLFGGMNVLESRDMAMQVCEAYVKVTEKLGVPYVFKASFDKANRSSIHSYRGPGMEEGL     KIFQELKDTFGVKII
FVLLGGMNVLESRDLAMRICEHYVTVTQKLGIPYVFKASFDKANRSSIHSYRGPGLEEGM     KIFQELKQTFGVKII
```

```
RCVLKIGEHT  PSALAIMENANVLARYASICQQ   NGIVPIVEPEILPDGDHDLKRCQYVTEKVLAAVYKALSDHHIY
RCVLKIGKNT  PSYQSILENANVLARYASICQS   ERIVPIVEPEVLPDGDHDLDRAQKVTETVLAAVYKALSDHHVY
RCVLNIGTHT  PSHLGMLENANVLARYASICQA   NGLVPIVEPEVLCDGEHDLARAQKVTEQVLAFVYKALADHHVY
RAVLKIGPNE  PSELSIQQNAQGLARYAIICQE   NGLVLFVEPEILTDGSHDIAKCAAVTETVLAACYKALNDQHVL
RTVLVIDTAKGKPTDLSIHETAWGLARYASICQQ   NRLVPIVEPEILADGPHSIEVCAVVTQKVLSCVFKALQENGVL
RTVVSIPNGPS  ALAV KEAAWGLARYAAITQD   NGLDPILEPEIMLDGEHGIDRTFRVAQQVWAEVFFNLAENNVL
RNVYKIQNGTV  SEAVVRFNAETLARYAILSQL   CGLVPIVEPEVMIDGTHDIETCQRVSQHVWSEVVSALHRHGVV
TDVHEIYQCQPVADVVDIIQLPAFLARQTDLVEAMAKTGAVINVKKPQFLSPSQMGNIVEKIEECGNDKII  LCDRGTN
TDVHEPSQAQPVADVVDVIQLPAFLARQTDLVEAMAKTGAVINVKKPQFVSPGQMGNIVDKFKEGGNEKVI  LCDRGAN

LEGTLLkPNMVTPGHACTQKFSHEEIAMATVTALRRTV PPAVTGITFLSGGQSEEEASINLNAINKCPLLKPWALTFSY
LEGTLLKPNMVTAGQSAK KNTPEEIALATVQALRRTV PAAVTGVTFLSGGQSEEEATVNLSAINNVPLIRPWALTFSY
LEGTLLKPNMVTPGQSSASKASHEAIGLATVTALRRGV PAAVPGITFLSGGQSELDATANLNAINSVQLGKPWKLTFSY
LEGTLLKPNMVTPGSDSP KVSPEVIGEYTVNALRRTV PAAVPGIVFLSGGQSEEQATLNLNAMNKFDVVKPWTLSFSF
LEGALLKPNMVTAGYECTAKTTTQDVGFLTVRTLRRTV PPALPGVVFLSGGQSEEEASVNLNSINALGPH PWALTFSY
LEGSSLKPSMVGPGALSARKGPPEQVADYPLKLLHRRR GPVVPGIMVLSGGQSEVEATLNLNAMNQSPNP  WHVSFSY
WEGCLLKPNMVVPGAESGLKGHAEQVAEYTVKTLARVI PPALPGVTFLSGGLSEVMASEYLNAMNNCPLPRPWKLTFSY
FGYDNLIVDMLGFSVMKKASKGSPVIFDVTHSLQCRDPFGAASSGRRAQVTELARSGLAVGIAGLFLEAHPNPNQAKCDG
FGYDNLVVDMLGFSIMKKVSGNSPVIFDVTHALQCRDPFGAASGGRRAQVAELARAGMAVGLAGLFIEAHPDPEHAKCDG

GRALQASALKAWGGKKENLKAAQEEYVKRALANSLACQGKYTPSGQAGAAASES  LFVSNHAy
GRALQASVLRAWAGKKENIAAGQNELLKRAKANGEAACGNYTAGSVKGFAGKDT  LHVDDHRY
GRALQASVLKAWGGKDENIAAAQKTLLHRSKANGDASLGKYAGEDAAGAAA ES  LFVAKHSY
GRALQQSTLKTWSGKKENVGKAQDVFLARCKANSEATLGKYGGGSGTGLAS ES  LHVKDYKY
GRALQASVLNTWQGKKENVAKAREVLLQRAEANSLATYGKYKGGAGGENAG AS  LYEKKYVY
ARALQNTCLKTWVEGQENVKAQDFAC   AKSNSLAQLGKYTGEGESEERKKDMFVK  ATLTY
ARALQSSAIKRWGGKESGVEAGRRAFMHRAKMNSLAQLGKYNRADDD    KDSQSLYVAGNTY
PSALPLSAL              EGFVSQMKAIDDLVKSFPELDTSI
PSALPLAKL              EPFLKQMKAIDDLVKGFEELDTSK
```

Multiple alignment for T0031:

```
block 0 = 11 seqs
*USER>>T0031; : Exfoliative toxin A from Staphylococcus aureus, 242a.a.
*USER>>P1;ETB_STAAU:93 : EXFOLIATIVE TOXIN B PRECURSOR
*USER>>1gctA :  GAMMA-*CHYMOTRYPSIN *A BOVINE (BOS TAURUS) PANCREAS
*USER>>2cha :   CHYMOTRYPSIN A (TOSYLATED) COW (BOS TAURUS)
*USER>>1smfE :  TRYPSIN COMPLEXED WITH BOWMAN-BIRK INHIBITOR TRYPSIN
*USER>>1lmwB :  MOL_ID: 1; MOLECULE: UROKINASE-TYPE PLASMINOGEN ACTIVATOR
*USER>>1hf1 :   00 =  HANNUKA FACTOR (MODEL) HUMAN
*USER>>1thsH :  16 =  ALPHA-THROMBIN (E.C.3.4.21.5) COMPLEX
*USER>>1ppbB :  75 =  THROMBIN EC 3.4.21.5 IN COVALENT COMPLEX
*USER>>1etrH :  EPSILON-THROMBIN NON-COVALENT COMPLEX WITH MQPA BOVINE
*USER>>1hcgA :  BLOOD COAGULATION FACTOR XA HUMAN (HOMO SAPIENS)

EVSAEEIKKHEEKWNKYYGVN
eysaeeirklkqkfe

AFNLPKELFSKVDEKDRQKYPYNTIGNVFVK       GQTSATGVLIGKNTVLTNRHIAKFANGD
vpptdkelythitdnars  pynsvgtvfvk       gstlatgvligkntivTNYHVAREAAKN
           IVNGEEA VPGSWPWQVSLQ    DKTGFHFCGGSLINENWVVTAAHCG    VT
           IVNGEEA VPGSWPWQVSLQ    DKTGFHFCGGSLINENWVVTAAHCG    VT
           IVGGYTC GANTVPYQVSLN      SGYHFCGGSLINSQWVVSAAHCY    KS
           IIGGEFT TIENQPWFAAIYRRHRGGSVTYVCGGSLMSPCWVISATHCFID  YP
           IIGGNEV TPHSRPYMVLLS   LD  RKTICAGALIAKDWVLTAAHCNLNKRSQ
           IVEGSDA EIGMSPWQVMLF   RKSPQELLCGASLISDRWVLTAAHCLLYPPWD
           IVEGSDA EIGMSPWQVMLF   RKSPQELLCGASLISDRWVLTAAHCLLYPPWD
           IVEGQDA EVGLSPWQVMLF   RKSPQELLCGASLISDRWVLTAAHCLLYPPWD
           IVGGQEC KDGECPWQALLI   NEE NEGFCGGTILSEFYILTAAHCLYQAKRF

      PSKVSFRPSINTD DNGNT ETPYGEYEVKEILQEPFGAGVDLALIRLKPD     QNGVSLGDKISP
      PSNIIFTPAQNRD AEKNEFPTPYGKFEAEEIKESPYGQGLDLAIIKLKpn     ekgesagdliqp
T     SDVVVAGEFDQGSSSEKIQKLKIAKVFKNSKYNSLTI  NNDITLLKLSTA    ASFSQTVSAVCLP   SAS
T     SDVVVAGEFDQGSSSEKIQKLKIAKVFKNSKYNSLTI  NNDITLLKLSTA    ASFSQTVSAVCLP   SAS
      GIQVRLGEDNINVVEGNEQFISASKSIVHPSYNSNTL  NNDIMLIKLKSA    ASLNSRVASISLP   TSC
E     DYIVYLGRSRLNSNTQGEMKFEVENLILHKDYSADTLAHHNDIALLKIRSKEGRCAQPSRTIQTICLP   SMY
       KKVILGAHSITREEPT  KQIMLVKKEFPYPCYDPATREGDLKLLQLTEK    AKINKYVTILHLPKKGDDV
KNFTENDLLVRIGKHSRTRYERNIEKISMLEKIYIHPRYNWRENLDRDIALMKLKKP    VAFSDYIHPVCLPDRETAA
KNFTENDLLVRIGKHSRTRYERNIEKISMLEKIYIHPRYNWRENLDRDIALMKLKKP    VAFSDYIHPVCLPDRETAA
KNFTVDDLLVRIGKHSRTRYERKVEKISMLDKIYIHPRYNWKENLDRDIALLKLKRP    IELSDYIHPVCLPDKQTAA
        KVRVGDRN TEQEEGGEAVHEVEVVIKHNRFT KETYDFDIAVLRLKTP    ITFRMNVAPACLPERDWAE
```

```
   AKIGTSNDLKDGDK              LELIGYPFDHKVNQMH  RSE    IELTTLSRGL    RYYGFTVPG
   anipdhidiqkgdk              ysllgypynysaysly  qsq    iemfndsq       yfgytevg
DDFAAGTTCVTTGWGLTRYT         PDRLQQASLPLLSNTNCKK  YWGTK  IKDAMICAG  ASGV   SSCMG
DDFAAGTTCVTTGWGLTRYANT       PDRLQQASLPLLSNTNCKK  YWGTK  IKDAMICAG  ASGV   SSCMG
AS  AGTQCLISGWGNTKSSGTSY     PDVLKCLKAPILSDSSCKS  AYPGQ  ITSNMFCAGYLEGGK   DSCQG
NDPQFGTSCEITGFGKENSTDYLY     PEQLKMTVVKLISHRECQQPHYYGSE  VTTKMLCAADPQWKT   DSCQG
K   PGTMCQVAGWGRTHNSASWS      DTLREVNITIIDRKVCNDRNHYNFNPVIGMNMVCAGSLRGGR   DSCNG
SLLQAGYKGRVTGWGNLKET       GQPSVLQVVNLPIVERPVCKDSTRIR    ITDNMFCAGYKPDEGKRGDACEG
SLLQAGYKGRVTGWGNLKETWTANVGKGQPSVLQVVNLPIVERPVCKDSTRIR    ITDNMFCAGYKPDEGKRGDACEG
KLLHAGFKGRVTGWGNRRETWTTSVAEVQPSVLQVVNLPLVERPVCKASTRIR    ITDNMFCAGYKPGEGKRGDACEG
STLMTQKTGIVSGFGRTHSTR           LKMLEVPYVDRNSCKLSSSFI    ITQNMFCAGY   DTKQEDACQG

NSGSGIFNSNG     ELVGIHSSK VSH LDREHQINYGVGIGN   YVKRIINEKNE
nsgsgifnlkg     eligihsgkggqhnlpigvffnrkisslysvdntfgdtlgndlkkrakldk
DSGGPLVCKK  NGAWTLVGIVSWGSSTCSTSTPGVYARVTALVN   WVQQTLAAN
DSGGPLVCKK  NGAWTLVGIVSWGSSTCSTSTPGVYARVTALVN   WVQQTLAAN
DSGGPVVCSG      KLQGIVSWGSGCAQKNKPGVYTKVCNYVS   WIKQTIASN
DSGGPLVCSL  QGRMTLTGIVSWGRGCALKDKPGVYTRVSHFLP   WIRSHTKEE
DSGSPLLCEGV      FRGVTSFGLENKCGDPRGPGVYILLSKKHLNWIIMTIKGAV
DSGGPFVMKSPFNNRWYQMGIVSWGEGCDRDGKYGFYTHVFRLKK   WIQKVIDQFGE
DSGGPFVMKSPFNNRWYQMGIVSWGEGCDRDGKYGFYTHVFRLKK   WIQKVIDQFGE
DSGGPFVMKSPYNNRWYQMGIVSWGEGCDRDGKYGFYTHVFRLKK   WIQKVIDRLGS
DSGGPHVTR  FKDTYFVTGIVSWGEGCARKGKYGIYTKVTAFLK   WIDRSMKTRGLPKAK
```

Multiple alignment for T0037:

```
     ......p- USER>>P1;AAC1_HUMAN:176
     ....p-b- USER>>P1;AACT_DICDI:169
     ..p-b--- USER>>t0037
     p-b----- USER>>P1;S32565:153
     b-p----- USER>>P1;DMD_HUMAN:166
     ..b----- USER>>P1;GELA_DICDI:156
450
     -------- USER>>P1;HUMFLNG6PD:195


block 0 = 7 seqs
*USER>>P1;AAC1_HUMAN:176 :      ALPHA-ACTININ 1 (HUMAN).
*USER>>P1;AACT_DICDI:169 :      ALPHA-ACTININ (SLIME MOLD).
*USER>>t0037 :                  Calponin homology domain of beta-spectrin
*USER>>P1;S32565:153 :          actin-binding protein - slime mold
*USER>>P1;DMD_HUMAN:166 :       DYSTROPHIN. - (HUMAN).
*USER>>P1;GELA_DICDI:156 :      GELATION FACTOR (SLIME MOLD).
*USER>>P1;HUMFLNG6PD:195 :      HUMFLNG6PD NID: g1203968 - human.

 veetsakeglllwcqrktapyknvniqnfhiswkdGLGFCALIHRHRPELIDYGKLRKD DPLTNLNTAFDVAEKYLDI
 ieelsakealllwcqrktegydrvkvgnfhtsfqdGLAFCALIHKHRPDLINFDSLNKD DKAGNLQLAFDIAEKELDI
    KSAKDALLLWCQMKTAGYPNVNIHNFTTSWRDGMAFNALIHKHRPDLIDFDKLKKS NAHYNLQNAFNLAEQHLGL
 egdksseeglllwcknttagydgvdiksfktgfrdGHAFLALAHKYDPAQFNYDELNKL SPDQRLEKAFEIAEKTINI
lqqtnseki llswvrqstrnypqvnvinfttswsdGLALNALIHSHRPDLFDWNSVVCQQSATQRLEHAFNIARYQLGI
sesdnspkaallewvrkqvapy kvvvnnftdswcdGRVLSALTDSLKPGVREMSTLTGD  AVQDIDRSMDIALEEYEI
eakkqtpkqrllgwiqnk   lpqlpitnfsrdwqsGRALGALVDSCAPGLCPDWDSWDASKPVTNAREAMQQADDWLGI

PKMLDAEDIVGTARPDEKAIMTYvssfyhafsgaqkae
PKMLDVSDMLDVVRPDERSVMTYvaqyyhhfsasrkae
TKLLDPEDI SVDHPDEKSIITYVVTYYHYFSKMK
PKLLDVNEVMKGT ADERALILYtslffhafsaqaqar
EKLLDPEDV DTTYPDKKSILMYitslfqvlpqqvsie
PKIMDANDM NS LPDELSVITYvsyfrdyalnkekrd
PQVITPEEIVDPN VDEHSVMTYlsqfpkaklkpgapl
```

Multiple alignment for T0038:

```
     ....p--- USER>>P1;GUN1_STRRE:90
     ..p-b--- USER>>P1;THFENDOGLU:162
     p-b----- USER>>P1;MXEGLBG:180
     b-p----- USER>>P1;GUNC_CELFI:241
     ..b----- USER>>t0038
158
     ......p- USER>>P1;A36910:115
     ..p---b- USER>>P1;XYNA_RUMFL:113
     p-b----- USER>>P1;XYNC_FIBSU:400
     |.....p- USER>>1xnb :69
     b-p---b- USER>>P1;XYNB_STRLI:119
     ..b----- USER>>P1;XYNC_ASPAK:97


block 0 = 11 seqs
*USER>>P1;GUN1_STRRE:90 :  CELLULASE 1 PRECURSOR - STREPTOMYCES RETICULI.
*USER>>P1;THFENDOGLU:162 : THFENDOGLU NID: g310896 - Th.fusca (strain YX) DNA.
*USER>>P1;MXEGLBG:180 :    MXEGLBG NID: g895874 - Myxococcus xanthus.
*USER>>P1;GUNC_CELFI:241 : ENDOGLUCANASE C PRECURSOR - C.FIMI
*USER>>T0038 :             CBDN1 from Cellulomonas fimi, 152a.a.
*USER>>P1;A36910:115 :     xylanase, beta(1,3-1,4)-glucanase - R.flavefaciens
*USER>>P1;XYNA_RUMFL:113 : ENDO-1,4-BETA-XYLANASE XYLA PRECURSOR - R.FLAVEFACIENS
*USER>>P1;XYNC_FIBSU:400 : ENDO-1,4-BETA-XYLANASE C PRECURSOR - F.SUCCINOGENES
*USER>>1xnb :69 :          1XNB  =  XYLANASE  (BACILLUS CIRCULANS)
*USER>>P1;XYNB_STRLI:119 : ENDO-1,4-BETA-XYLANASE B PRECURSOR - S.LIVIDANS
*USER>>P1;XYNC_ASPAK:97 :  ENDO-1,4-BETA-XYLANASE C PRECURSOR - A.AWAMORI

             veqvrngtfdtttd pww tsnvtaglsdgrlcadvpggttn      rwdsaigqnditlv   kgetY
             vnqirngdfssgta pwwgteniqlnvtdgmlcvdvpggtvn      pwdviigqddipli   egesy
             telvsngtfnggtvspwwsgpntqsrvenarlrvdvgggtan      pwdaligqddiplv   ngray
                lphtsfaeslgpwslygtsepvfad grmcvdlpggqgn      pwdaglvyngvpvg   egesY
             ASPIGEGTFDDGPEGWVAYGTDGPLDTSTGALCVAVPAGSAQ       YGVGVVLNGVAIE   EGTTY
       wemwnqnytgtvsmnp gagsftcswsg      ienflarmgknyddqkknykafgdivltydv  eytprgnsy
       yemwnqngqgqasmnp gagsftcswsn      ienflarmgknydsqkknykafgnivltydv  eytprgnsy
       yeiwyqg  gnnsmtfydngtykaswng      tndflarvgfkyd ekhtyeelgpidayykwskqgsaggyny
   astdywqnwtdgggivnavngsggnysvnwsn      tgnfvvgkgwttgspfrtinynagvwa         pngngy
qegtnngyyysfwtdsqgtvsmnmgsggqystswrn      tgnfvagkgwanggr rtvqys gsfn         psgnay
   rsaginyvqnyngnladftydesagtfsmywedg    vssdfvvglgwttgss naisysaeysa         sgsssy
```

```
          (v)
RFSFHAsgipeghvravvglavspy dtwqeas pvlteadgsysytftapv        dttqgqvafqvgg
afsftasstvpvsiralvqepvepw ttqmder allgpeaetyefvftsnv        dwddaqvafqigg
tlsftasasvsttvrvtvqlesapy tapldrq itldgtsrrftfpftstl        atqagqvtfqmgg
VLSFTAsatpdmpvrvlvgegggay rtafeqgsapltgepatreyaftsnltfp  pdgdap gqvafhlgk
TLRYTATASTDVTVRALVGQNGAPY GTVLDTSPA LTSEPRQVTETFTASATYPATPAADDPEGQIAFQLGGF
mciYGWTRNPLMEYYivegwgdweppgndgvdnfgtttidgktykirksmrynqpsiegtkt fpqywsvrttsgsrnnt
mcvYGWTRNPLMEYYivegwgdwrppgndg evkgtvsangntydirktmrynqpsldgtat fpqywsvrqtsgsannq
igiYGWTVDPLVEYYivddwfn kpganllgqrkgeftvdgdtyeiwqntrvqqpsikgtqt fpqyfsvrksarsc
ltlYGWTRSPLIEYYvvdswgtyrp  t  gtykgtvksdggtydiytttrynapsidgdrttftqywsvrqskr
lalYGWTSNPLVEYYivdnwgtyrp  t  geykgtvtsdggtydiykttrvnkpsvegtrt fdqywsvrqskr
lavYGWVNYPQAEYYivedygdynpcss  atslgtvysdgstyqvctdtrtnepsitgtst ftqyfsvrestr

                          stdawrfcvddvsllggvpp
                          sdepWTFCLDDVALlgraep
                          ratgFSAFIDDISLttedgg
                          agayefcisqvslttsatp
                          SADAWTLCLDDVALDSEVEL
tnymkdqvsvtkhfdawskagldmsgtlyevslniegyrsngsanv
tnymkgtidvtkhfdawsaagldmsgtlyevslniegyrsngsanvk
     ghiditahmkkweelgmkm gkmyeakvlveagggsgsfdvtyfk
ptgsnatitftnhvnawkshgmnlgsnwayqvmategyqssgssnvtvw
tgg   tittgnhfdawaragmplgnfsyymimategyqssgtssinvgg
tsg   tvtvanhfnfwaqhgfg nsdfnyqvmaveawsgagsasvtiss
```

Chapter 4

Modelling by Multiple Sequence Threading and Distance Geometry

Here I describe a homology modelling prediction based on a scaffold identified by fold recognition and modelled using a distance regularisation algorithm for geometry optimisation, or DRAGON for short. The DRAGON modelling tool is based on distance geometry and relies on decreasing the dimensionality of the hierarchical projection of a simple model (C_α and C_β) into 3D and thus predicting its tertiary structure. For fold recognition I used a multiple sequence threading (MST) method (Taylor, 1997). Here I describe the use of MST to identify possible protein folds and from this construct several high resolution homology models by distance geometry of a CASP2

target. In this case I chose target T0004 (polyribonucleotide nucleotidyltransferase S1 motif (PNS1) from *Escherichia coli*) and built the sequence into several full atom representations. From the results of CASP2 it can be seen that one of these predicted models identified by `MST/DRAGON` was correct. The model created by `DRAGON` and the subsequent full atom representation compared well with the target PNS1. The model was better than the template with a C_α RMSD of 6.2Å compared with 6.4Å. Continuation of this work could lead to a better incorporation of distance geometry homology modelling with fold recognition in one fast automated procedure.

4.1 Introduction

Homology modelling is a technique where the conformation of a target protein is deduced from the known structures of homologous proteins. The approach can give a very accurate insight into the 3D structure of sequences with no solved crystal structure and the models produced can greatly aid in the understanding of detailed protein structure (Pearl and Taylor, 1987b; Sali *et al.*, 1990; Havel, 1993; Taylor, 1994).

This works well where there are similar proteins with a known fold in the database, but, unfortunately, fails when a close homologue is not available. Fold recognition can be used to identify structural similarity when this is not apparent in the sequence information (Bowie *et al.*, 1990; Jones *et al.*, 1992a; Bryant and Lawrence, 1993).

Overcoming this problem involved the combination of different techniques to predict protein structure, along with homology modelling, in the hope that this combination would enable us to see a wider picture than each individual part of the problem could convey. Specifically, I use distance geometry as a flexible method that can, simultaneously, combine the restraints needed for homology modelling along with less reliable constraints derived from secondary structure prediction and fold recognition. The combination of a fold recognition (or threading) method with a distance-geometry approach is potentially very powerful as it allows many a likely fold, or folds, to be identified and used as a rough template for the rapid generation and evaluation of the threading results as a full 3D model.

The fold recognition method I use incorporates multiple sequence alignments to compare sequences with structure and predict the most likely fold. The method has expanded on the more traditional threading methods which compare a single sequence with a single structure using pairwise potentials. Although the method can, in general, use a multiple alignment on both the sequence and structure side of the sequence/structure comparison problem, the studies presented below only have multiple data on the sequence side. Below, the method will be referred to as the Multiple Sequence Threading method (or `MST`) (Taylor, 1997).

The distance-geometry modelling program used was `DRAGON` (Aszódi and Taylor, 1996), which allows the fast generation of many simplified C_α models, as opposed to trying to model full atom models which is far more computationally expensive. The method also incorporates a randomisation algorithm for exploring the available

conformational space of alternate structures. Using a multiple alignment and the predicted fold alignments from MST, structural equivalences are extracted and used as weighted distance constraints in the modelling.

For the CASP2 experiment these methods were used on many of the available targets, ranging from close to distant similarity. However, as the current approach I use is directed towards very distant relationships. Of particular interest was the PNS1 sequence which had no known homologous structure but for which some candidate folds could be proposed by the MST method. In this chapter I describe these approaches in the prediction of the structure of PNS1 by homology modelling using templates determined by threading.

4.2 Methods

T0004 is a Polyribonucleotide Nucleotidyltransferase, S1 motif (from *Escherichia coli*) (Bycroft *et al.*, 1997). It is 84 amino acids in length with the following sequence:

```
AEIEVGRVYT GKVTRIVDFG AFVAIGGGKE GLVHISQIAD KRVEKVTDYL
QMGQEVPVKV LEVDRQGRIR LSIKEATEQS QPAA
```

4.2.1 Multiple alignment

The target sequence was aligned with the following homologous sequences (SwissProt id and accession number):

```
PNP_PHOLUP(P41121), PNP_HAEIN(P44584),  YABR_BACSU(P37560),
RS1_MYCLE(P46836),  RS1H_BACSU(P38494), YHGF_ECOLI(P46837)
```

using the multiple sequence alignment program `MULTAL` (Taylor, 1988), (see Figure 4.1). The sequences had been identified in the databanks using the programs `BLAST`, `BLITZ` and `FASTA`. `MULTAL` was used to align the sequences, typically, with a fixed (opening) gap penalty of between 15 and 20 and with an amino acid relatedness matrix composed of 30 percent multiples of the PAM120 values augmented by adding 7 to the diagonal (Dayhoff *et al.*, 1978).

```
      p- TARGET
    p-b- PNP_PHOLU
  p-b--- PNP_HAEIN
p-b----- YABR_BACSU
|
b-p----- RS1_MYCLE
  b-p--- RS1H_BACSU
    b--- YHGF_ECOLI
```

```
TARGET      --AEIEVGRVYTGKVTRIVDFGAFVAIGGGKEGLVHISQIADKRVEKVTD
PNP_PHOLU   --AEIEVGRIYAGKVTRIVDFGAFVAIGGGKEGLVHISQIADKRVEKVAD
PNP_HAEIN   --AEVEAGVIYKGKVTRLADFGAFVAIVGNKEGLVHISQIAEERVEKVSD
YABR_BACSU  --MSIEVGSKLQGKITGITNFGAFVELPGGSTGLVHISEVADNYVKDIND
RS1_MYCLE   FARTHAIGQIVPGKVTKLVPFGAFVRVEEGIEGLVHISELAERHVEVPDQ
RS1H_BACSU  ---KVKPGDVLEGTVQRLVSFGAFVEILPGVEGLVHISQISNKHIGTPHE
YHGF_ECOLI  ---DLQPGMILEGAVTNVTNFGAFVDIGVHQDGLVHISSLSNKFVEDPHT

TARGET      YLQMGQEVPVKVLEVDRQG-RIRLSIKEATEQSQPAA--
PNP_PHOLU   YLQVGQETSVKVLEIDRQG-RVRLSIKEATAGTAVEE--
PNP_HAEIN   YLQVGQEVNVKVVEIDRQG-RIRLTMKDLAPKQETEIN-
YABR_BACSU  HLKVGDQVEVKVINVEKDG-KIGLSIKKAKDRPQARPR-
RS1_MYCLE   VVAVGDDAMVKVIDIDLERRRISLSLKADQRGLHRGVR-
RS1H_BACSU  VLEEGQTVKVKVLDVNENEERISLSMRELEETPKA----
YHGF_ECOLI  VVKAGDIVKVKVLEVDLQRKRIALTMRLDEQPGETNARR
```

Figure 4.1: The multiple alignment of target T0004.

The resulting alignment was then examined by eye to find patterns of conserved positions that could provide the basis for a motif search. These were scanned across a non redundant PIR sequence database (OWL) using the `UNIX` pattern matching utility `regex`. This tool, implemented in a simple application program (D. Jones, W. Taylor) uses regular expression searches to scan the database and pull out any sequences which matched, including the flanking regions around the pattern that would be anticipated from a knowledge of the full alignment probe. Larger alignments were pruned by (recursively) removing one of the pair of most similar sequences. These related, but relatively distant proteins were threaded along with the target sequence in the `MST` program. In some cases the motif was refined and then used to research the database. Due to an overall lack of divergence among the sequences found, the alignment in Figure 4.1 was used in the modelling.

4.2.2 Secondary structure prediction

Secondary structure prediction techniques used included the predict protein server (Rost and Sander, 1993) and latterly DSC (King and Sternberg, 1996); other programs available on the web were used as an additional check. Where possible a multiple alignment was submitted to the secondary structure prediction program. In all the cases a MULTAL alignment was studied and a secondary structure prediction evaluated 'by eye'. Sequences of known structures could sometimes be aligned with relatively distant homologues of the target sequence. Many of these structures were first identified by threading methods.

4.2.3 Fold recognition

The PNS1 sequence was compared to a fold database using the THREADER program (Jones *et al.*, 1992a). The multiple sequence alignment of the target sequence and its homologues was then compared to the extended UCLA benchmark set of 319 structures (Fischer *et al.*, 1996a) using MST. The MST prediction uses a simple pairwise potential that favours the packing of conserved hydrophobics into the core along with the matching of predicted and observed secondary structure and predicted and observed solvent exposure. The protein structure with the highest threading score was chosen. The MST program automatically generates a model of the alignment (but with no attempt to model inserted regions) allowing the basic threaded structure to be vi-

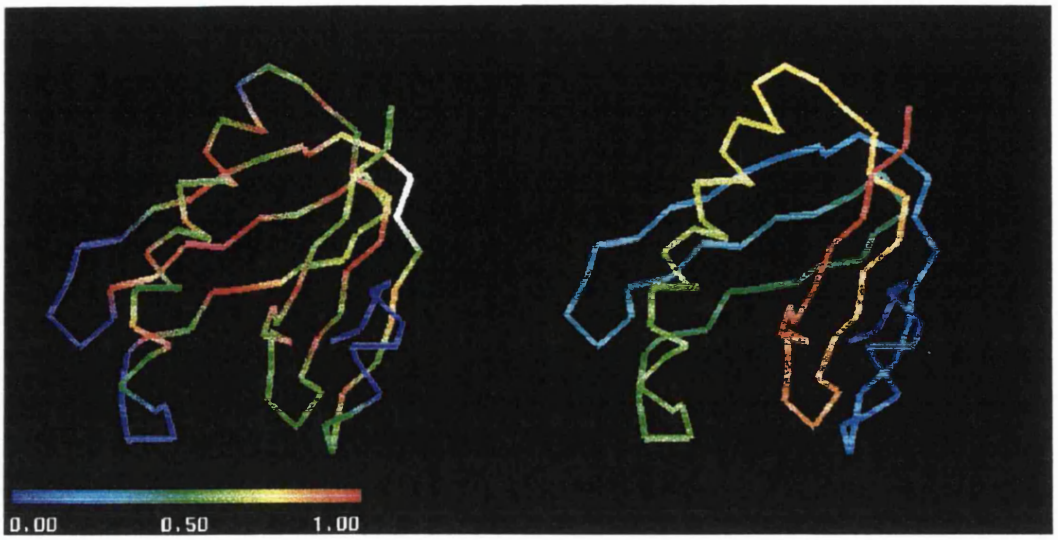


Figure 4.2: Threaded structure of 1LTS chain D. Where blue = deletions, white = inserts and red = hydrophobic. The right hand figure just shows the template coloured from the N to the C terminus (blue to red)

sualised. The hydrophobic core packing in each model was assessed especially in areas where insertion and deletion had occurred. To prevent a dislocated threading across two protein domains a modified version of MST was used to thread just one domain of given size. Three hits were selected from the MST results — chains: 1LTSD,1HRHA and 2SNS. These structures had the best scores considering both packing and secondary structure correspondence. See Table 4.1, Table 4.2 and Table 4.3 in results section for more details. All these structures were small $\beta + \alpha$ structures: 1LTSD is a 'classic' OB-fold (Murzin, 1994), 2SNS is a more elaborate OB-fold also, while 1HRHA is a ribonuclease RT domain. Figure 4.2 illustrates the threading on one of these structures (1LTS chain D).

4.2.4 Fold generation

Homology modelling was carried out using DRAGON (Aszódi and Taylor, 1996). The similarity between the unknown target structure and the scaffold proteins with known structures was described by mapping secondary structure assignments and specific distance restraints between C_α atoms on to the model through a multiple alignment. The results of the MULTAL alignments were submitted to DRAGON, version 4.16.1 containing the known structures (from MST) and the target sequence only. 50 models were generated for each of the three scaffold structures, using C_α:C_α distances shorter than 10Å to guide the folding process. Secondary structure assignments from the scaffolds were also mapped on to the target at 30% probability. The model based on 1LTS chain D can be seen in Figure 4.5. The average clustered results of the simplified chains were calculated and this backbone was then modelled in QUANTA (Molecular Simulations Inc.) and CHARMM (MSI) to produce a full coordinate homology model.

4.2.5 DRAGON methodology

Figure 1.3 in Chapter 1 shows information flow in DRAGON: all the different restraints which can be added or are built in to the program. DRAGON does not use a full atom representation, but instead works with a simplified protein chain. Figure 4.3 shows the simple model chain which DRAGON uses to model. From this simple representation full atom models can be built. This C_α chain with side chain centroids (SCC) speeds

up the modelling approach.

DRAGON has many features which make it a very useful modelling tool. It essentially works by initially generating a random distance matrix. The next generation of the distance matrix is then derived from a matrix with expected values for a given distance and a matrix with the confidence in those expectations.

For a current distance matrix, ${}^{k}d_{ij}$, an expected matrix P_{ij} and a confidence matrix S_{ij}; then the next matrix in the iterative procedure can be defined as:

$$ {}^{k+1}D_{ij} = S_{ij} \cdot P_{ij} + (1 - S_{ij}) \cdot {}^{k} D_{ij}. \tag{4.1} $$

Consequently anything to do with distance of atoms in a protein can be modelled with DRAGON. It makes assignments of additional distance restraints very easy. All distance based assignments in DRAGON have a confidence (or strictness) attached, where 1 is a distance with high confidence and 0 is a distance with no confidence. Due to a random seed matrix the program can generate different models and successfully sample the given conformational space. In this case due to the constrints imposed from the template structures, a much reduced space was sampled in the modelling. Chapter 5 has a diagram (Figure 5.1) showing how different the models can be when generated using this method. A non-random starting matrix will always produce identical results.

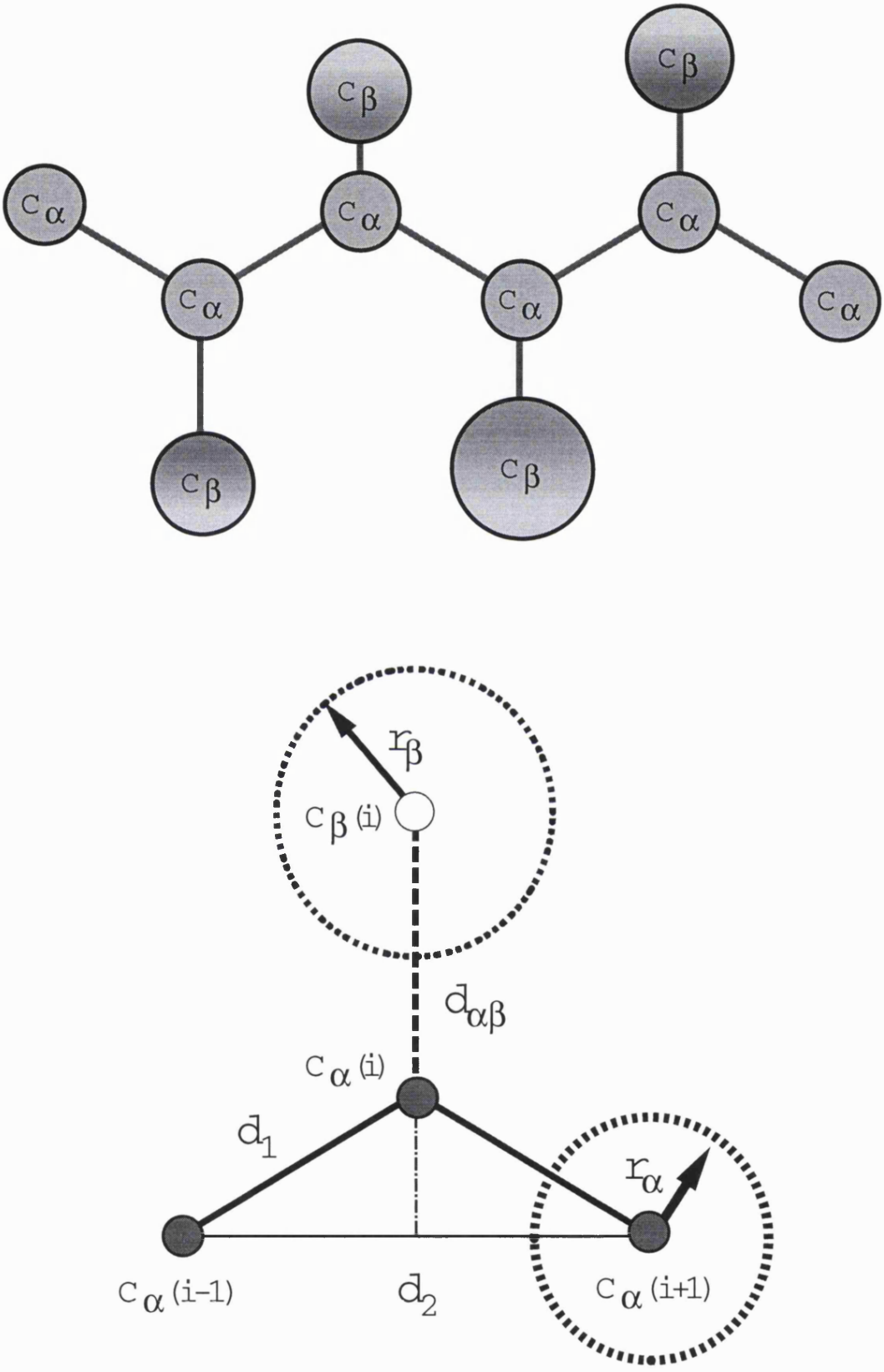


Figure 4.3: Simplified protein model chain.

4.2.6 Model building and refinement

The ten best-scoring DRAGON output structures were averaged for each scaffold. The missing atoms were added to the C_α average structures and the resulting full-atom structures were minimised by QUANTA version 4.1/CHARMM 23.1 (Brooks *et al.*, 1983).

4.2.7 Comparison

Once the NMR coordinates were obtained from the CASP2 organisers I superposed the known structure with the models to see how close the folds matched, as shown in Figure 4.4 and also illustrated in Figure 4.5. The template structures were also compared to the experimental structure although this could not be done by a straightforward rigid-body superposition as the template sequences were different from that of PNS1. SAP, a modification of the SSAP algorithm (Orengo and Taylor, 1996), was used to generate optimal correspondences between atoms for the superposition.

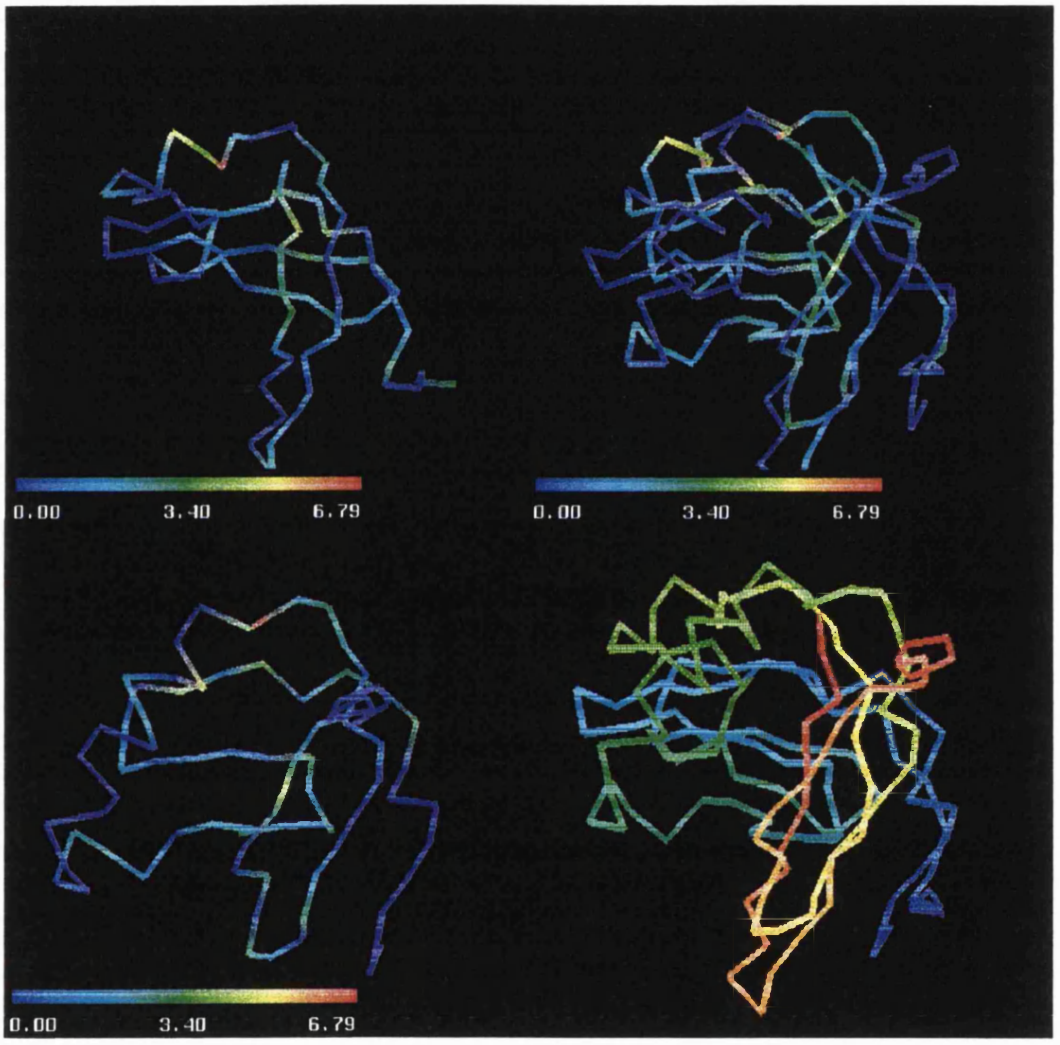


Figure 4.4: Superposition of T0004 model with structure. NMR structure on top-left, model bottom left. With superposition on the right. Colours indicate the weights in the alignment matrix, the "hotter" colour the better the structural alignment. Also shown the superposition coloured from N to C terminus, blue to red. N.B. colour bars only indicate the range of values in the alignment matrix, they have no relation to RMSD.

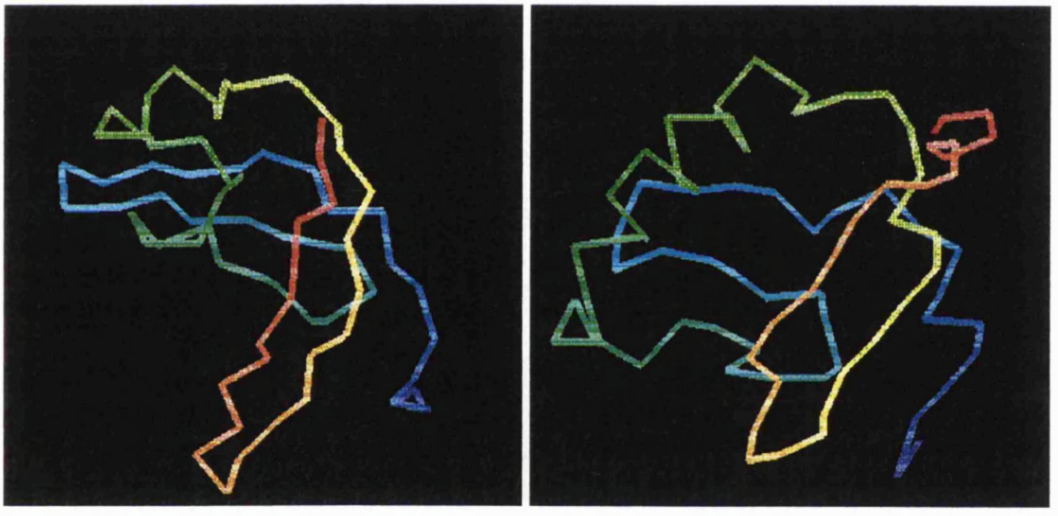


Figure 4.5: The NMR structure and the model based on the structure of 1ltsD. Coloured in blue from the N-terminus.

4.3 Results and discussion

Three potential folds were identified for target T0004 (polyribonucleotide nucleotidyltransferase S1 motif (PNS1) from *Escherichia coli*) by fold recognition. (1HRH chain A, 1LTS chain D and 2SNS). See Table 4.1, Table 4.2 and Table 4.3 for rankings used to determine these folds. The main interest from constructing models for these was to see if the more detailed refined models would provide a stronger basis on which to select one candidate of folds over the others. For this, I followed a path of increasing molecular detail. Firstly, constructing loops and refining the initial C_α-model with `DRAGON`, then adding full atomic detail in `CHARMM` and finally adding solvent.

The ten best scoring `DRAGON` models were averaged and converted into full atom representations, followed by a minimisation step. From the final full atom minimisations of the three models created it was not possible to tell with sufficient confidence which was the best model. By studying the `CHARMM` energies I saw that with an unsolvated minimised structure the 1HRHA template had the lowest energy, whereas with a solvated model the 2SNS template had the lowest energy.

The 1LTSD template, which turned out to have the correct fold, ranked second after all the minimisations. All three models after `DRAGON` homology modelling and full atom refinement were plausible – none of the modelled proteins had structures that provided any reason to reject them. The time spent in doing the solvated annealing made no improvement in any of the models.

Rank	Z score	%str aligned	%seq aligned	Chain	PDB
1	-2.53	87.8	77.4		1HOE
2	-2.43	85.7	100.0		7PCY
3	-1.99	69.5	86.9		1AAJ
4	-1.97	64.8	83.3	A	2TRX
5	-1.94	66.1	97.6	H	1FGV
6	-1.91	52.0	78.6		2FGF
7	-1.90	60.9	92.9		3CHY
8	-1.88	70.7	83.3		1PLC
9	-1.77	87.4	90.5		1PTF
10	-1.71	92.8	76.2	A	1BOV
11	-1.68	63.6	91.7	H	1FVW
12	-1.56	52.5	75.0	B	1FVC
13	-1.55	50.4	84.5		2SNS
14	-1.37	69.9	85.7	A	1SHA
15	-1.36	51.3	91.7		1F3G
16	-1.33	68.1	96.4	H	1FVB
.	..	..	..	.	..
.	..	..	..	.	..
29	-0.95	65.2	89.3	L	1IGM
30	-0.93	62.3	90.5		2APK
31	-0.88	56.5	88.1		1ALB
32	-0.87	62.8	96.4	H	1IGM
33	-0.86	54.5	100.0	A	1SRD
34	-0.86	70.9	86.9	D	1LTS
35	-0.82	75.7	96.4	L	1FGV
36	-0.82	50.7	91.7	B	2PKA
37	-0.81	72.8	98.8		2IMM
38	-0.81	59.2	91.7	A	1HRH
39	-0.74	96.8	72.6	I	2TEC
40	-0.73	65.1	82.1	L	1FVB

Table 4.1: THREADER score table for the target sequence. Ranked according to Z score. Also shown are the percentage of structure aligned and the percentage of sequence aligned. This can give an idea of how useful the threading may be.

rank	PDB	score	fold
1	1osa	3007	calmodulin
2	3inkC	2995	interleukin
3	1cd8	2974	ig-fold
4	2sns	2965	OB-fold
5	1paz	2954	ig-like
6	1rro	2888	calmodulin
7	2fx2	2871	flavo
8	1ltsD	2743	OB-fold
9	1mdc	2721	Ortho beta
10	1hrhA	2707	Ribonuc.RTdom.
11	3chy	2700	flavo
12	1opaA	2686	Ortho beta
13	1hli	2664	lectin (IgE rec)

Table 4.2: Showing the top MST scores and fold type.

rank	PDB	score	fold
1	1fus	n = 20	
2	1frrA	n = 22	
3	1hstA	n = 23	
4	1hrhA	n = 26	RT (10th above)
5	1ptf	n = 26	
6	1ltsD	n = 28	OB (8th above)
7	1paz	n = 30	IG (5th above)
8	2trxA	n = 35	
9	3chy	n = 35	flavo (11th above)
.	..	.	.
13	1noa	n = 44	
14	1osa	n = 46	calmodulin (1st above)
15	1stfI	n = 47	
16	3inkC	n = 56	interleukin (2nd above)
17	2sns	n = 74	OB (4th above)

Table 4.3: Secondary Structure agreement. A small value of n is best indicating a high secondary structure content agreement. n is the factor required to match the predicted and observed secondary structure. A value of $n = 0$ would indicate that the observed and predicted secondary structure are identical.

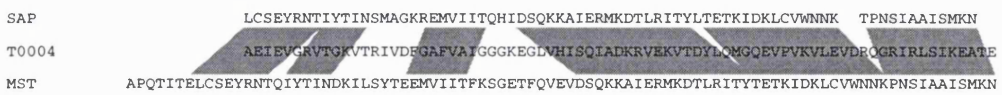


Figure 4.6: Alignment comparison of target 4 and 1ltsD. The target sequence (in this case T0004) is shown in the centre of the three sequences. Also shown is the mapping of 1ltsD by structural comparison (SAP) and by predicted threading (MST). Taking the SAP:1ltsD alignment as the best possible it is a simple case to compare how good the threading alignment is by showing how the sequence is shifted, or not. This is highlighted by the grey shading. A perfect threading alignment is boxed. The greater the shift in the alignment, the larger the slant of the grey shading. The less shift there is in the alignment, the better.

Comparing the model based on 1LTS chain D with the NMR data from the experimentalists it can be seen that the correct fold had been predicted. Superposing the C_α backbone of two structures gives an RMSD of 6.2Å (84 atoms) and a full atom comparison of 7.1Å (587 atoms). The alignment fares less well, this can be seen in Figure 4.6. In more detail: a helix which occurs in the model, between Arg-42 and Met-52, is only partially present in the NMR structure. Similarly a beta turn between Ile-25 and Lys-29 is out of position, as well as the loop between His-34 and Ala-40. The C_α backbone RMSD was not improved by any of the minimisations after the full atom model was generated with an RMSD of 6.2Å.

The rankings compared were: the `MST` threading scores, the pseudo-energy scores calculated by the MatchMaker program (supplied as part of the `SYBYL` molecular modelling package, version 6.3, Tripos Associates)(Godzik and Skolnick, 1992) (Provided by A. Aszódi) and the `CHARMM` potential energy values of the refined structures. The correct model based on `1LTSD` ranked second both by the MatchMaker and `CHARMM` energies, it also ranked second according to the `MST` threading score (Table 4.4). The energy was much higher on the NMR structure but it couldn't really put it through

same minimisation steps. MST score reflected badly as it is not yet normalised for size of protein - which is probably why the larger 2sns scored better. Perhaps by using an application of the *cones* method discussed in Chapter 2, then the models could be scored according to the proteinness of the buriedness factors. A filtering technique could perhaps be devised and a suitable rank of the models would be achieved.

Template	MST score	MatchMaker score [kT]	CHARMM energy [kcal/mole]	C_αRMSD [Å]
1LTSD	2743	-0.12	-4417	6.2
1HRHA	2707	-0.06	-4547	10.8
2SNS	2966	-0.14	-4403	11.0

Table 4.4: Model quality judged by various scores. The correct model based on the template structure 1LTSD comes second according to the MST, MatchMaker and CHARMM energy rankings. The RMS deviations of the models from the experimental structure (T0004) are shown in the last column for comparison.

The 2SNS fold identified also has the correct fold although there are much larger insertions in the structure. The best fit for 2SNS was 3.9Å (67 atoms) compared with 4.1Å (67 atoms). RMSD of the template structure 1LTS chain D is worse than the model's. The best C_α superposition of the template versus the NMR data was 6.4Å compared with the aforementioned 6.2Å, a comparison of the template structure versus the DRAGON model gave a C_α RMSD of 5.1Å. Although the fold is similar the proteins are not entirely homologous. The MST identified areas which were deleted from the template and were therefore not included in the modelled structure. This is shown in Figure 4.2 where the areas in blue on the template were deleted. The comparison between the NMR structure and the model is more obviously shown in Figure 4.5.

4.4 Conclusion

From this brief example of these methods it can be seen that the correct fold for T0004 has been predicted – as Figure 4.5 clearly shows. The superposition of the C_α chains are reasonably accurate and a similar comparison of the full-atom models gives an RMSd of 7.1Å (over 587 equivalent atoms).

While the correct overall fold for PNS1 was identified successfully, the atomic details of the model structure are not accurate enough. This is due to several factors. Firstly, threading-based methods cannot provide the large amount of high-quality structural information available in comparative modelling where the target and the templates are closely related both sequentially and structurally. Most participants at the CASP2 meeting agreed that model quality depends very much on the quality and quantity of *external* structural information supplied to the prediction algorithms. Second, it seems to be difficult to choose the appropriate level of resolution. In this case the low-resolution C_α:C_β model built by distance geometry appeared to be justified on grounds of efficiency and lack of detailed experimental information. Perhaps the method would have performed better if another refinement at intermediate resolution had been carried out before the full-atom modelling to improve the main-chain geometry. Finally, although the choice of detailed potential functions and sophisticated energy minimisation/refinement methods are important for the last stage of full-atom refinement, these cannot compensate for gross errors (such as misaligned residues in homology modelling) made earlier in the modelling process. Consider-

ing sub-optimal alignments under these circumstances might improve matters (Saqi *et al.*, 1992). Possible improvements to the approach should therefore include a careful choice of low-resolution interaction potentials and improved gap modelling (Taylor and Munro, 1997),

MST gave accurate predictions for many of the fold recognition targets, confirming that a well constructed multiple alignment can be a great aid in fold recognition. The approach, however, is sometimes limited by a lack of sufficient homologous sequences.

Combining threading with distance geometry modelling can be a useful way to construct a model for a protein. If a sequence has no known structural homologues then the sequence can be threaded to predict a likely scaffold on which to base the model. This method has several advantages over a pure *ab initio* prediction, where a fold is constructed using just secondary structure information. A threading alignment will be more accurate than just a sequence alignment, where there may only be 10% sequence homology.

Several points can be taken from the CASP2 experiment with respect to these methods. DRAGON performed best when working on *ab initio* targets or an example like this, where a target has possible template proteins identified from fold recognition and homology modelling is carried out based on them. As yet, the distance geometry method is not as good as the more classical homology modelling methods when dealing with closely homologous template structures, as these use a full atomic representation where DRAGON uses only C_α atoms. With distant homologous sequences the

distance geometry method may be as good as other homology modelling methods. It would be interesting to use MODELLER, for example and see if better models would be produced. A lot still depends on an accurate alignment with a good template, which proves hard in the current state of the field.

Chapter 5

Structure prediction of NK Lysin

5.1 Introduction

NK-lysin is a small protein of 78 amino acid residues (CASP target T0042). It is a membrane destabilising protein which has anti-bacterial activity and the capability to lyse tumour cells. There are three disulphide bonds and its secondary structure is composed of helices. It has homology to the family of saposin-like proteins. One of the interesting features of these proteins may be their ability to adopt the same or similar folds in both an aqueous or membrane environment. Studies of circular dicromism of NK-lysin show that its secondary structure does not change when in

these different environments. The pore forming activity and membrane binding of NK-lysin have been assessed by Ruysschaert *et al.* (Ruysschaert *et al.*, 1998). The structure of NK-lysin is quite compact and would be unable to bridge the width of a membrane and as such no obvious mechanism of action has been determined.

This chapter will concentrate on the prediction made for the CASP2 assessment and some of the ideas which have been developed since the results of the NMR structure were made available.

A distance geometry based modelling algorithm, DRAGON, has been developed for the prediction of protein structures. Here I show the prediction of a protein structure using this method. Incorporating a multiple alignment, secondary structure, disulphide bonding data as well as the built in restraints: simple low-resolution C_α and C_β models were constructed.

One such model was submitted for assessment to the CASP2 experiment. The analysis made by the CASP2 assessors and the subsequent comparison of the model is presented here. A slightly modified version of the SSAP algorithm was used to compare the model structures with the NMR coordinates. The effectiveness of the model scoring system is evaluated, as is a more up to date version of DRAGON. The correct fold was successfully identified and the models were found to be similar to the experimental structure.

An analysis of the modelling potential of DRAGON is presented here along with some further modelling where the structural coordinates of the target sequence are known.

Using the correct secondary structure can give models with RMSD of 4.8Å. The results show DRAGON as an efficient and reasonably accurate method for the *ab initio* prediction of tertiary structure. This distance geometry approach has the potential to provide models when there are no homologous models in the PDB and where no putative structure can be found by threading methods.

5.1.1 *Ab initio* modelling

Ab initio modelling (or *de novo* folding) is not based on any template structures, but rather on a secondary structure assignment and various sets of constraints, see introduction for more details.

An *ab initio* analysis using distance geometry was assessed by predicting a set of 8 helical proteins (Mumenthaler and Braun, 1995). One of these proteins contained disulphide bridges, so as a comparison DRAGON was used to model the protein in order to assess whether the connectivity between the cysteines could be determined. 1ERP a Pheremone (ER-10) contains a 3 helix packing motif (Brown *et al.*, 1993). A secondary structure was assigned using DSSP and incorporated into DRAGON along with a multiple sequence alignment, generated automatically by the PHD secondary structure prediction server. Using this easily obtained information is a bare minimum required for the modelling process.

5.1.2 DRAGON

The distance geometry approach to protein prediction has been used for many years (Mackay, 1974; Crippen and Havel, 1988; Kuntz *et al.*, 1989) but these techniques have rarely been applied to *ab initio* folding. Distance geometry is used more often in homology modelling (Havel and Snow, 1991; Havel, 1993; Srinivasan *et al.*, 1993; Sudarsanam *et al.*, 1994). Using `DRAGON` (Aszódi *et al.*, 1995a; Aszódi and Taylor, 1996), a simplified model chain is folded by projecting it into gradually decreasing dimensional spaces whilst subjecting it to a set of defined restraints, primarily secondary structure. In this way it is possible to explore the geometry space to produce a range of protein backbones that satisfy the restraints. See Figure 5.1 shows the range of different folds created by the distance geometry, clustered on an average structure using the `clumsy` program which is supplied in the `DRAGON` distribution.

The method generates many folds in a short time due to the efficient embedding algorithm incorporated into the program (Aszódi and Taylor, 1997). I applied `DRAGON` to one of the targets in the second meeting on the Critical Assessment of Techniques for Protein Structure Prediction (CASP2) and describe the results here along with a detailed post-analysis of the other models generated by `DRAGON`.

Other uses for `DRAGON` also include homology modelling (Aszódi and Taylor, 1996) and homology modelling with threading (Aszódi *et al.*, 1997a), in combination with a multiple sequence threading method (Taylor, 1997; Taylor and Munro, 1997) (see Chapter 4).

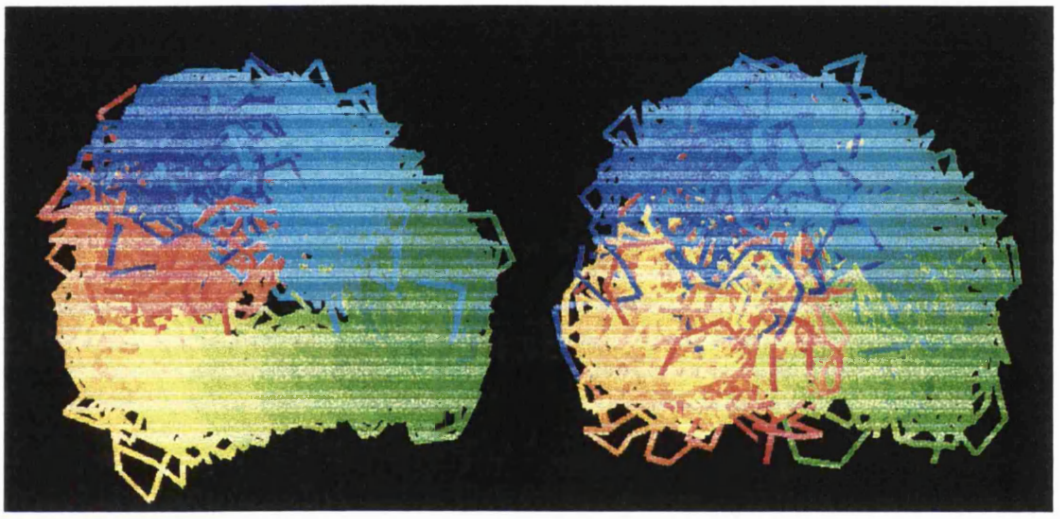


Figure 5.1: Shown here are two clusters of 60 models generated by DRAGON, superposed on an average backbone structure. On the left is the collection of models generated for the CASP2 with disulphide restraints, on the right are models where only the secondary structure has been specified, both with a stringency of 0.7. Both sets of models make use of the multiple alignment in Figure 5.2. Only C_α backbone models are shown, they are coloured according to residue position (N = blue, C= red). This amply demonstrates the way that the program samples the available conformational space.

5.2 Methods

A variety of methods were used to build up a picture about the target protein. This was achieved by examining the sequence, its predicted secondary structure, the evolutionary relationship to other proteins and the known function of the protein.

5.2.1 Sequence information

Protein: NK-Lysin (from pig)

Length: 78 amino acids

Sequence: GYFCESCRKI IQKLEDMVGP QPNEDTVTQA ASQVCDKLKI
LRGLCKKIMR SFLRRISWDI LTGKKPQAIC VDIKICKE

5.2.2 DRAGON *ab initio* model generation

Ab initio predictions consisted of a C_α and C_β chain modelled using secondary structure predictions (Aszódi *et al.*, 1995a). Multiple conformations were created using DRAGON-4. Where other constraints were used in the modelling, the best structure was chosen (which did not violate any of the pre-defined constraints). The program required a multiple alignment and a secondary structure alignment, plus any additional restraint information which in this case consisted of three disulphide bonds known to form in the protein (Andersson *et al.*, 1995).

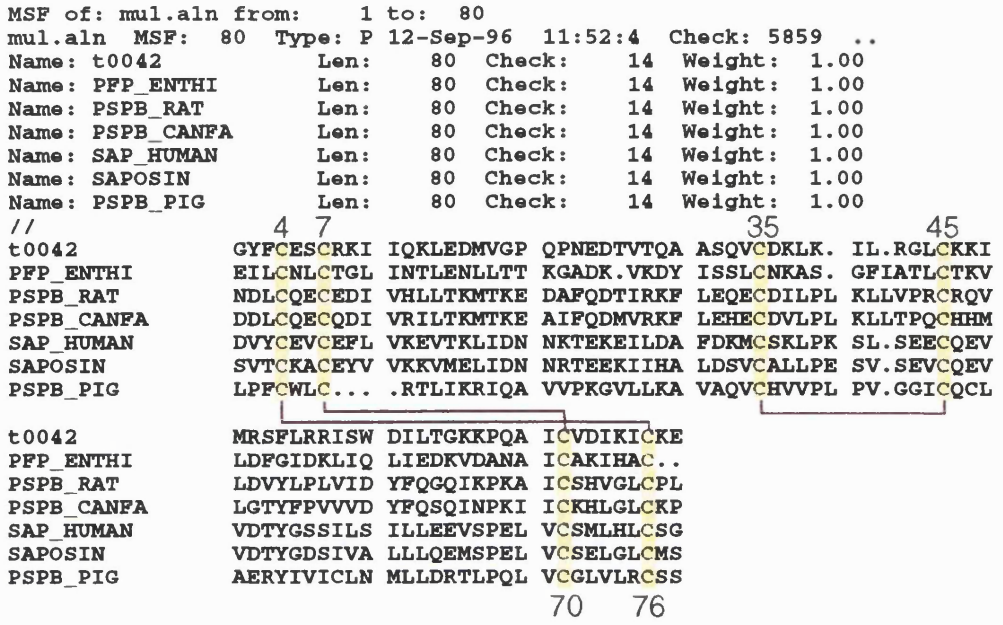


Figure 5.2: Multiple alignment of the target sequence, T0042. The cysteine pairs are highlighted.

Multiple Alignment The multiple alignments were built from sequences identified by BLAST (Altschul *et al.*, 1990) and BLITZ searches. The homologous sequences were aligned with the target sequence using MULTAL (Taylor, 1988). See Figure 5.2.

Consensus alignments can show those secondary structures which have been predicted most accurately by the programs. PHD (Rost and Sander, 1993) is particularly useful as it constructs a multiple alignment before using it to predict the secondary structure. MSF format multiple alignments can be submitted to PHD to give a SS prediction of any specific alignment.

MULTAL was used to align sequences which were related to the target sequence (including those found by BLAST, BLITZ and FASTA (Pearson and Lipman, 1988)) with a gap penalty of between 15 and 20 with a 30% PAM-120 matrix used (Dayhoff *et al.*,

1978). The resulting alignment was then examined by eye to find patterns of conserved residues which had aligned across the sequences. These motifs which were found were used to search an OWL database file for sequences containing that motif using the regex based programs such as PADGREP and HEADGREP (D. T. Jones and W. R. Taylor). These programs used regular expression searches to scan the database and pull out any sequences which matched. The larger alignments were pruned by removing sequences which were very similar.

Secondary Structure A program was written to convert the MULTAL alignment output into an MSF file which could be passed to the PHD program. This allowed a MULTAL alignment to be tweaked slightly by hand before a secondary structure prediction was re-assessed.

The multiple alignment was then sent to the predict protein PHD server (Rost and Sander, 1993) and the corresponding secondary structure prediction was incorporated into the DRAGON input. Further checks on the secondary structure prediction were performed using DSC (King and Sternberg, 1996) and SSPRED (Mehta *et al.*, 1995). All were in reasonable agreement with PHD, so just the PHD prediction was used, see Figure 5.3. Confidence in the prediction was supplied to DRAGON as a fixed value between zero and one. More sophisticated modelling might be achieved by assigning confidence to each of the segments based on the PHD scores (specified on the Rel Sec line of the program results). Most of the time a confidence of 0.7 or 70% was used in the DRAGON model building. This choice was based on the likely confidence in the

```
                  ....,....1....,....2....,....3....,....4....,....5....,....6....,....7....,...
Residue           GYFCESCRKIIQKLEDMVGPQPNEDTVTQAASQVCDKLKILRGLCKKIMRSFLRRISWDILTGKKPQAICVDIKICKE
-------------------------------------------------------------------------------------------------
DSSP               HHHHHHHHHHHHHHH     HHHHHHHHHHHHHH    HHHHHHHHH     HHHHH    HHHHHHHHH
built.dssp        SSSSSHHHHHHHHTS SSSSSSHHHHHHHHHHHTSTTTTTTHHHHHHHHHHHHTTTTS SSSSSSSTTTTTTSSS
main.dssp         SSSSSSHHHHHHHS SSSSSSHHHHHHHHHHHSSSSSSSSHHHHHHHHHHHHHHHHSSSSSSSSTTTTTTSSSS
minimised.dssp    SSSSSTTTTTTTTTSSSSSSSSHHHHHHHHHHHSSHHHHTTHHHHHHHHHHHHHTTS SSTTSSS  TTTTSSS
mod40.dssp.actual SHHHHHHHHHHHS SSSSSSHHHHHHHHHHHSSSSS HHHHHHHHS SSSSSSHHHHSSSSSHHHHHTTSS
PHD                HHHHHHHHHHHHHH     HHHHHHHHHHHH    HHHHHHHHHHHHHHHHHHHH    HHHHHHHHHH
PHD MSF              HHHHHHHHHHH     HHHHHHHHHHHH      HHHHHHHHHHHHHHHHHH     HHHHHHH
DSC MSF            HHHHHHHHHHHH     HHHHHHHHHHHH      HHHHHHHHHHHHHHHHHH     HHHHHHH
NNpredict            HHHHHHH          HHHHHHHHHHHHHHHHHHHHHHHHHHH E HH        EEE H
NNpredict (all a)  HHHHHHHHHHHH      HHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHH    HHHHHHHHHH
```

Figure 5.3: Secondary structure predictions/assignments. The DSSP classification of secondary structure is included for reference. Also shown are various DSSP assignments for different models of NK Lysin built from the DRAGON simple chain. `main.dssp` is a main chain representation of the DRAGON model. `built.dssp` is similar with a full side chain built on to the main chain frame. `minimised.dssp` this is the full atom model after 100 steps of steepest descents minimisation. `mod40.dssp.actual` is a different model where the assigned secondary structure before the DRAGON modelling directly corresponded to the DSSP secondary structure assignment. The other rows are different secondary structure prediction methods.

secondary structure prediction.

In total 60 models were created using DRAGON. Only one model, with the highest ranking restraint score, was submitted for the CASP2 assessment. Generating more models would tend to produce only more similar models and not add anything to the analysis.

Additional Restraints Restraints were imposed on the model to bring the disulphide bonds close together. The disulfide bridges are as follows: residues 4 and 76, 7 and 70, 35 and 45 (Andersson *et al.*, 1995). The assumption made when building this restraint information into DRAGON was that the distances between the C_α:C_α residues was 7.0Å and the distances between the corresponding side chain centroids was 4.5Å.

5.3 Results

5.3.1 Secondary structure accuracy

The secondary structure calculated by the predict protein server at EMBL, called PHD, was as follows: residue numbers 6-18, 24-36, 42-60 and 66-72 all predicted as helix. The DSSP classification of the actual coordinates (which were later available) was that: 3-17, 23-36, 40-51, 57-62, 66-73 are helical. Comparing the predicted and observed helix assignments gives an overall accuracy of 79.5%, 16 of the residues were incorrectly predicted. Using the multiple alignment, see Figure 5.2, there was an indication that there might be a break in the long third helix at position 55 – there is in fact a break in the helix in the structure of 5 residues (52-56) allowing a kink between two helices, where the predicted structure is an unbroken helix.

5.3.2 Handedness of models

Previous work with four α-helix bundles has shown that left-handed and right-handed structures occur with approximately equal frequencies (Presnell and Cohen, 1989). I found that of the 60 models, 55% were right-handed and 45% left-handed. Figure 5.4 illustrates the two potential folds a simple four helix bundle can form, either left or right handed.

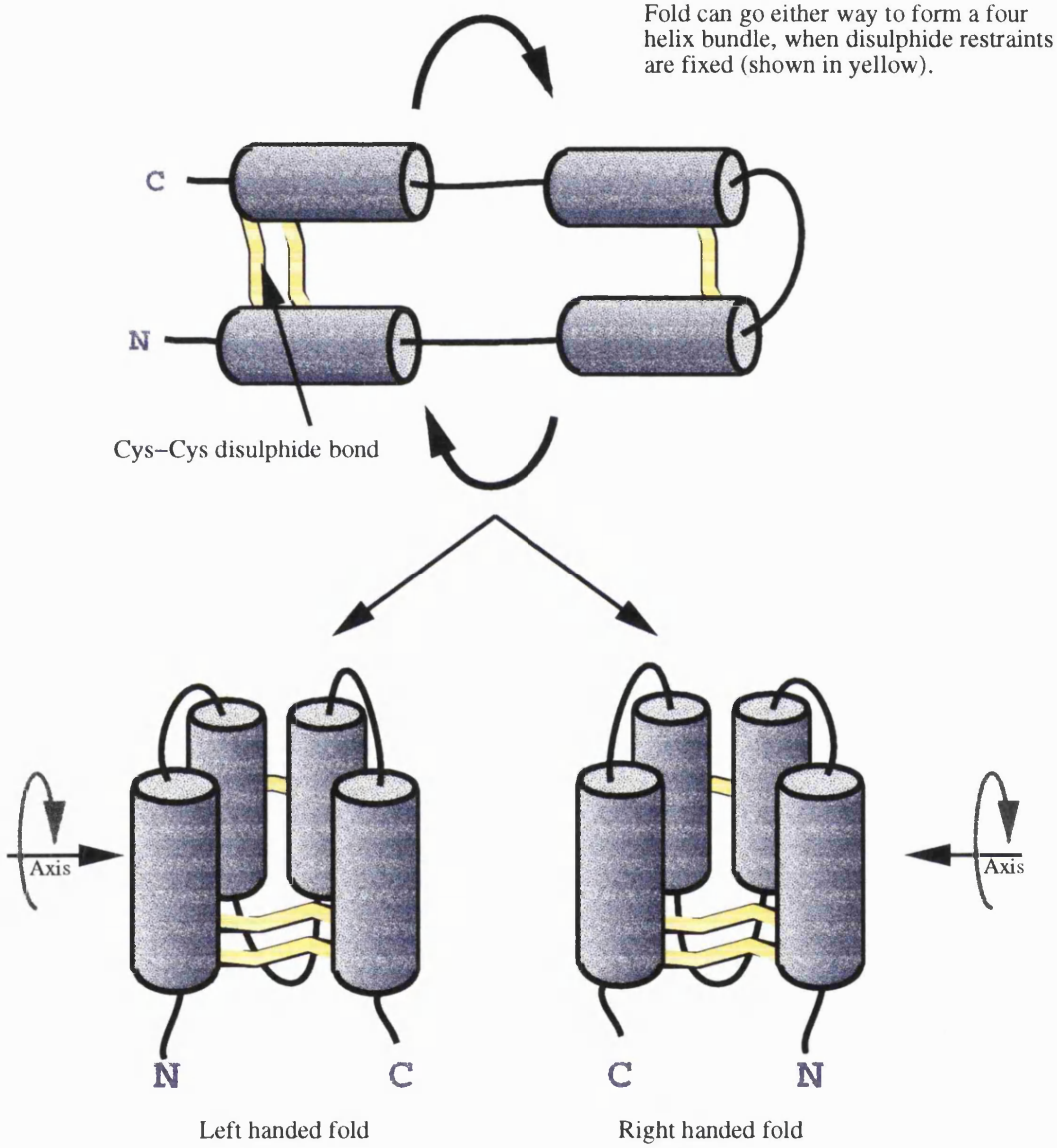


Figure 5.4: Illustration of the handedness of a four helix bundle. By looking down the two axes indicated, the left and right handedness of the models produced can be easily seen. The disulphide bonds shown in yellow indicate the way in which a fold of this type may be restricted into two roughly similar folds, with different handedness.

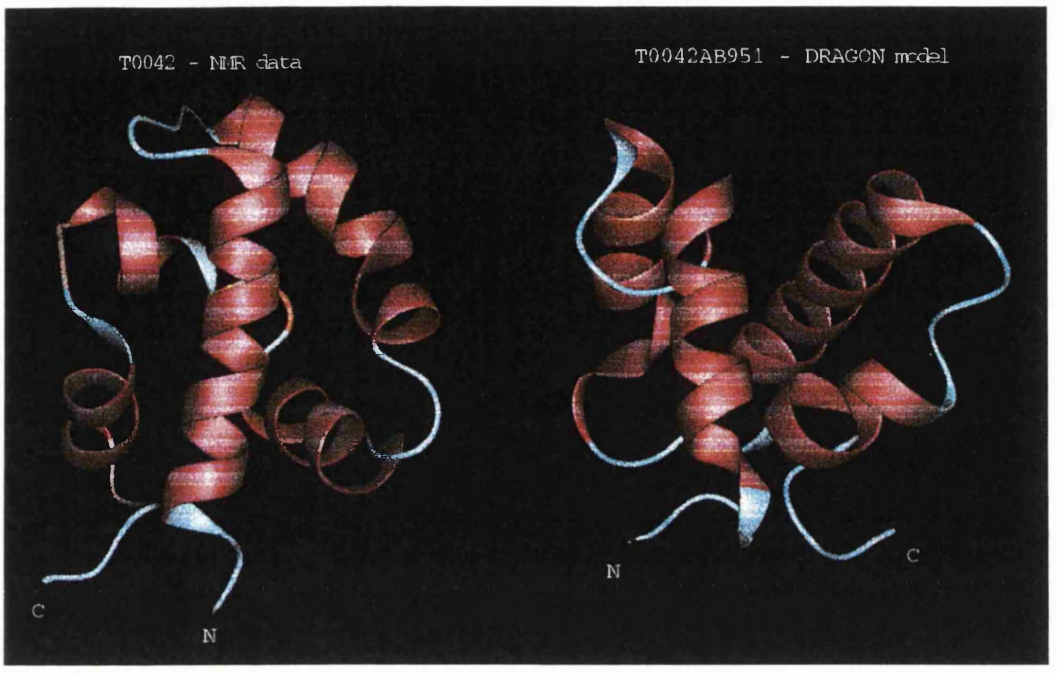


Figure 5.5: Side by side comparison. The target NMR structure with the model(2_12) submitted to CASP2.

5.3.3 Post CASP2 analysis

Several tables of results are presented showing SAP structural comparisons between models generated and the NMR structure. Table 5.1 shows models generated prior to CASP2 ranked according to the DRAGON restraint score. It was the first ranking model here (2_12) which was submitted for assessment by CASP2 (see chapter appendix). Table 5.2 similarly ranks the models but this time includes the bond score as an additional ranking factor. The bond score is the value reflecting the accuracy of the distance between the first and second neighbours. In hindsight this would have given a better model if the first ranked structure had been submitted to CASP2. Model 2_11 is ranked highest and would have been submitted to CASP2 had this ranking analysis been carried out. Table 5.4 illustrates the model ranked according to a score

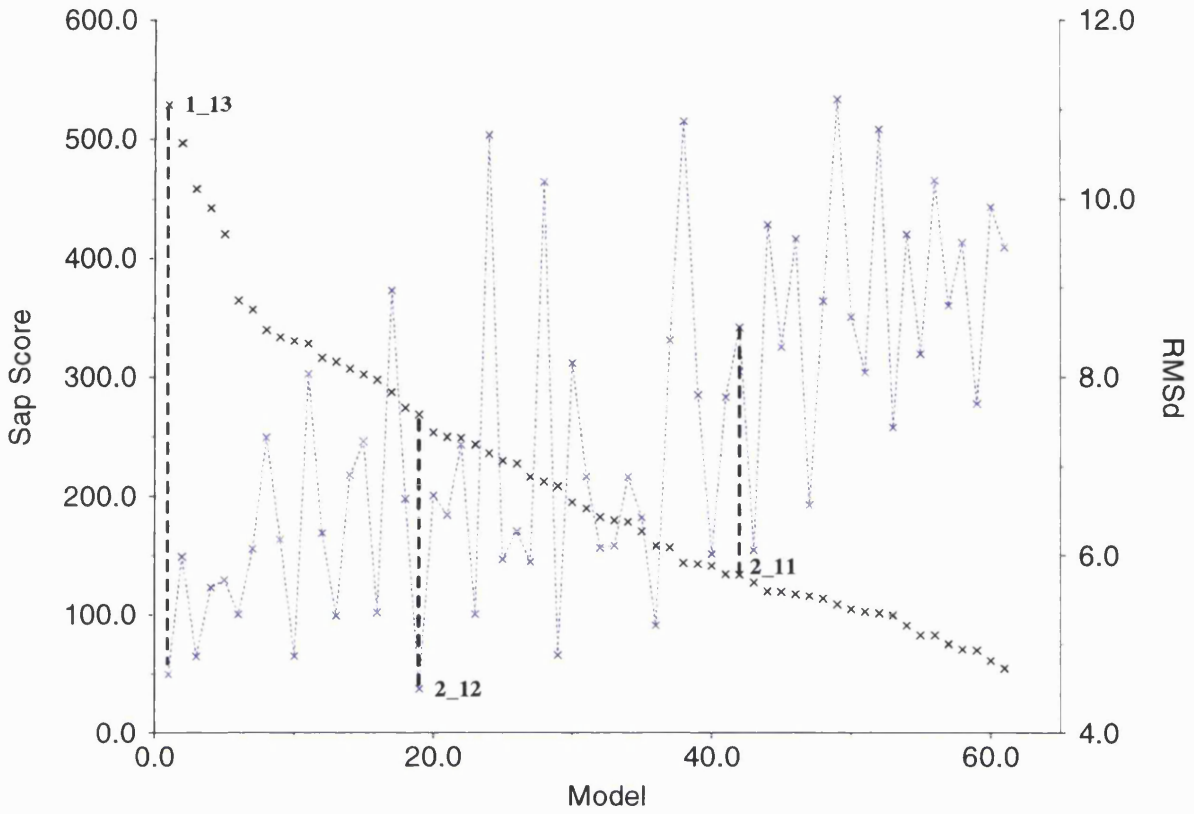


Figure 5.6: Models ranked according to RMSD. Of the 60 models it is clear to see the distribution of different RMSD's. Also shown are the corresponding SAP scores. Some of the models discussed in the text are highlighted on the graph.

produced by the SAP structure-structure comparison algorithm and the corresponding RMSD's. Here model 1_13 is considered the best model (RMSD=5.7). Table 5.5 lists the top ten SAP scores for the models generated prior to CASP2 and the top ten models ranked by each of the different RMSD calculations from SAP. Considering the RMSD's it appears that the best model is 2_40 which ranks top for all three measures. The SAP score for 2_40 is less good, placing it seventh. Table 5.6 gives details of the RMSD's obtained from model 2_11, when compared with the NMR data, before and after building the main chain into a full atom representation.

5.3.4 Building full side chain models

One of the better models (code: 2_11) from the first run was built into a full protein backbone. The resulting RMSD went from 5.9 to 6.0Å for all 78 C_α atoms. Table 5.7 shows the differences before and after the minimisation/MD. After building chains the energy was exceptionally high, this was lowered by a steepest descents minimisation (100 steps) – Lennard-Jones energy : 4.9×10^{11} to -1689. The energy went from -3147 to -3508 during the molecular dynamic step. 1.3Å difference before and after MD for all 78 C_α atoms - increasing the RMSD to 6.0Å. A 7.2Å RMSD for all 736 matched atoms between NMR structure and the full atom model. The backbone was not fixed at any point in the calculations. The brief use of molecular dynamics is justified in this case to sort out any bad contacts created by building the model into a full atom representation.

Combined Rank	model	Restraint score	SAP score	Weighted RMSD	Un-wtd RMSD	Un-wtd RMSD	Quanta RMSD	Hand
1	2_12	8.392e-02	106.6	7.0 [51]	7.7 [15]	8.2 [51]	11.9	Left
2	2_11	9.601e-02	181.9	4.0 [70]	4.1 [26]	5.6 [70]	5.9	Right
3	2_30	9.728e-02	47.2	10.6 [69]	10.3 [20]	11.9 [69]	11.4	Right *
4	2_24	1.013e-01	76.8	11.1 [57]	11.0 [21]	10.6 [57]	12.8	Left
5	2_33	1.030e-01	278.1	5.1 [62]	5.6 [23]	7.0 [62]	6.7	Right
6	2_29	1.049e-01	109.4	6.7 [74]	6.2 [28]	7.4 [74]	7.2	Right
7	2_25	1.104e-01	410.6	3.8 [78]	3.9 [27]	5.9 [78]	5.9	Right
8	1_7	1.136e-01	200.0	9.5 [75]	9.4 [28]	11.5 [75]	11.8	Left
9	2_31	1.137e-01	207.2	4.1 [53]	4.1 [21]	7.1 [53]	11.9	Left
10	1_11	1.188e-01	111.6	7.4 [67]	6.7 [22]	12.1 [67]	12.6	Left

Table 5.1: Models ranked according to the DRAGON restraint score. The figures in square brackets indicate the number of C_α atoms used in the RMSD calculation.
Note: the QUANTA RMSD use all equivalent atoms in the calculation.
* - indicates an unusual fold which would be discounted by a visual analysis.

Combined Rank	model	Bn Score	Bn Rank	Rs score	Rs Rank	SAP score	Weighted RMSD	Un-wtd RMSD	Un-wtd RMSD	Quanta RMSD	Hand
1	2_11	3.596e-03	10	9.601e-02	2	181.9	4.0 [70]	4.1 [26]	5.6 [70]	5.9	Right
2	2_30	4.096e-03	17	9.728e-02	3	47.2	10.6 [69]	10.3 [20]	11.9 [69]	11.4	Right *
3	2_28	3.354e-03	8	1.308e-01	15	267.6	3.8 [75]	3.5 [34]	5.1 [75]	5.0	Right
4	1_1	2.166e-03	1	1.531e-01	26	64.2	5.8 [75]	6.3 [21]	11.7 [75]	14.8	Left
5	2_33	4.344e-03	23	1.030e-01	5	278.1	5.1 [62]	5.6 [23]	7.0 [62]	6.7	Right
6	2_21	2.553e-03	2	1.633e-01	29	38.6	7.5 [64]	7.4 [13]	11.8 [64]	9.7	Right *
7	2_14	3.712e-03	12	1.434e-01	19	88.9	4.8 [77]	5.0 [25]	6.9 [77]	7.2	Right
8	2_40	2.935e-03	5	1.581e-01	27	302.1	3.1 [74]	3.0 [25]	5.0 [74]	5.4	Right
9	2_29	4.518e-03	26	1.049e-01	6	109.4	6.7 [74]	6.2 [28]	7.4 [74]	7.1	Right
10	2_24	4.846e-03	29	1.013e-01	4	76.8	11.2 [57]	11.0 [21]	10.6 [57]	15.2	Left

Table 5.2: Models ranked according to the **DRAGON** bond and restraint scores. The figures in square brackets indicate the number of C_α atoms used in the RMSD calculation.
Note: the **QUANTA** RMSD use all equivalent atoms in the calculation.
* - indicates an unusual fold which would be discounted by a visual analysis.

Model	E. after	RMSD before	RMSD after
casp	-2960	0.0	0.1
2_29	-2011	7.1	7.0
2_33	-1934	6.7	6.7
2_40	-1726	5.4	5.4
2_11	-1689	5.9	5.9
2_28	-1670	5.0	5.0
2_24	-1594	15.2	12.7
2_30	-1555	11.4	11.4
2_21	-1465	9.7	9.5
2_14	-1363	7.2	7.1
1_1	-1184	14.8	11.8

Table 5.3: Energy score for the DRAGON models. Lennard-Jones energy shown after a 100 steps steepest descent minimisation using CHARMM, models taken are the top ten predicted by DRAGON, see Table 5.2. The table is ranked according to lowest energy. The lower the energy the better the model should be, in theory. As can be seen, by comparing the RMSD, this is not exactly the case.

Rank	Model	Sap Score	Weighted RMSD	Un-wtd RMSD (best C_α)	Un-wtd RMSD	Quanta RMSD
1	1_13	491	3.9 [71]	4.1 [45]	5.0 [71]	5.7
2	2_1	440	4.1 [78]	4.2 [28]	5.7 [78]	5.7
3	2_25	410	3.8 [78]	3.9 [27]	5.9 [78]	5.9
4	2_26	363	3.7 [73]	3.8 [31]	6.0 [73]	7.0
5	2_2	323	4.1 [75]	4.0 [24]	5.5 [75]	5.7
6	1_3	304	4.9 [71]	4.7 [27]	5.6 [71]	5.7
7	2_40	302	3.1 [74]	3.0 [25]	5.0 [74]	5.4
8	1_4	297	4.9 [76]	5.3 [23]	7.4 [76]	7.7
9	1_5	294	4.8 [76]	4.4 [28]	6.4 [76]	6.4
10	2_18	283	4.1 [78]	4.3 [44]	6.0 [78]	6.0
11	1_20	278	5.2 [78]	4.9 [29]	8.6 [78]	8.7
12	2_33	278	5.1 [62]	5.6 [23]	7.0 [62]	6.7
13	1_17	275	7.2 [75]	6.8 [28]	8.8 [75]	9.1
14	2_28	267	3.8 [75]	3.5 [34]	5.1 [75]	5.0
15	2_22	258	5.7 [58]	6.2 [36]	6.7 [58]	6.4
16	1_19	222	7.4 [31]	6.9 [17]	6.3 [31]	14.5
17	2_31	207	4.1 [53]	4.1 [21]	7.1 [53]	11.9
18	1_7	199	9.5 [75]	9.4 [28]	11.5 [75]	11.7
19	2_8	194	3.9 [78]	4.2 [37]	6.7 [78]	6.7
20	1_9	193	4.4 [70]	4.2 [27]	6.5 [70]	6.7

Table 5.4: Ranking of models with highest scoring SAP and the corresponding RMSD. The models are ranked according to the SAP score and for comparison are also shown the calculated RMSD for the model compared to the NMR structure. The figures in square brackets indicate the number of C_α atoms used in the RMSD calculation. Note: the QUANTA RMSD use all equivalent atoms in the calculation.

Rank	model	Sap Score	model	Weighted RMSD	model	Un-wtd RMSD (best C_α)	model	Un-wtd RMSD (all matched)
1	1_13	491	2_40	3.1 [74]	2_40	3.0 [25]	2_40	5.0 [74]
2	2_1	440	1_16	3.6 [49]	1_15	3.3 [21]	1_13	5.0 [71]
3	2_25	410	1_15	3.6 [78]	2_28	3.5 [34]	2_28	5.1 [75]
4	2_26	363	2_26	3.7 [73]	2_26	3.8 [31]	2_32	5.1 [57]
5	2_2	323	2_25	3.8 [78]	2_25	3.9 [27]	2_2	5.5 [75]
6	1_3	304	2_28	3.8 [75]	2_2	4.0 [24]	2_4	5.6 [63]
7	2_40	302	1_13	3.9 [71]	2_11	4.1 [26]	2_11	5.6 [70]
8	1_4	297	2_8	3.9 [78]	1_13	4.1 [45]	1_3	5.6 [71]
9	1_5	294	2_11	4.0 [70]	2_31	4.1 [21]	2_1	5.7 [78]
10	2_18	283	2_31	4.1 [53]	1_9	4.2 [27]	1_16	5.8 [49]

Table 5.5: Ranking of the best SAP and RMSD measures These models were generated prior to CASP2, so these a really blind predictions.
The figures in square brackets indicate the number of C_α atoms used in the RMSD calculation.

	best SAP model	RMSD of best SAP model	best RMSD overall
Run 1	491	5.0 [71]	5.0 [74]
Run 2	204	5.1 [70]	5.1 [70]
Run 3	389	8.5 [77]	4.8 [71]
Run 4	808	6.3 [74]	4.7 [73]

Table 5.6: Simulation summary table.
Run 1: was the runs created before the CASP2 modelling assessment.
Run 2: involved the assignment of multiple helix stringency scores from PHD.
Run 3: was based on the correct secondary structure of NK-Lysin.
Run 4: was the same as run 3 but also incorporated the predicted Accessibility (SUB acc from the MSF submission to the PHD server). The figures in square brackets indicate the number of C_α atoms used in the RMSD calculation.

	Quanta RMSD (C_α)	Quanta RMSD (all atoms)	Weighted RMSD	Un-wtd RMSD (best C_α)	Un-wtd RMSD (all matched)
Before	5.9 [78]	n/a	4.0 [70]	4.0 [26]	5.6 [70]
After	6.0 [78]	7.2 [736]	4.1 [67]	4.5 [23]	5.7 [67]

Table 5.7: Differences before and after MD simulation for model 2_11. All figures are comparisons with the NMR structure. The figures in square brackets indicate the number of atoms used in the RMSD calculation.

5.3.5 Model comparison

One of the best methods for comparing protein structures is a combination of looking at the coordinates in 3D on a computer screen and structure-structure comparison methods such as SAP. Figure 5.7 to Figure 5.8 illustrate some of the results of these analyses: Figure 5.7 simply shows the structure of the "Answer", the NMR coordinates. It can be seen that the structure is an all alpha protein with 5 helical secondary structure elements. Figure 5.5 is a side by side comparison of the NMR

structure with the model generated with DRAGON and submitted to the CASP2 assessment. The handedness is different with a left-handed model being shown next to the right-handed NMR structure. The full atom representation of model 1_11 was built upon the DRAGON model, after a small amount of molecular dynamics using CHARMM. Figure 5.8 shows the best model so far generated using DRAGON, with a C_α RMSD of 4.6Å. As can be seen from the figure, the three main helical regions have been modelled reasonably well and it is only the C terminal helix which is less ordered. A stringency of 70% was applied to the model whereas a higher stringency may have improved results slightly.

Modelling with predicted accessibility: A further modification to modelling with DRAGON can be made by adding information about whether or not certain residues should be buried or exposed in the protein. The best model created had an RMSD of 4.6Å and is shown in Figure 5.8.

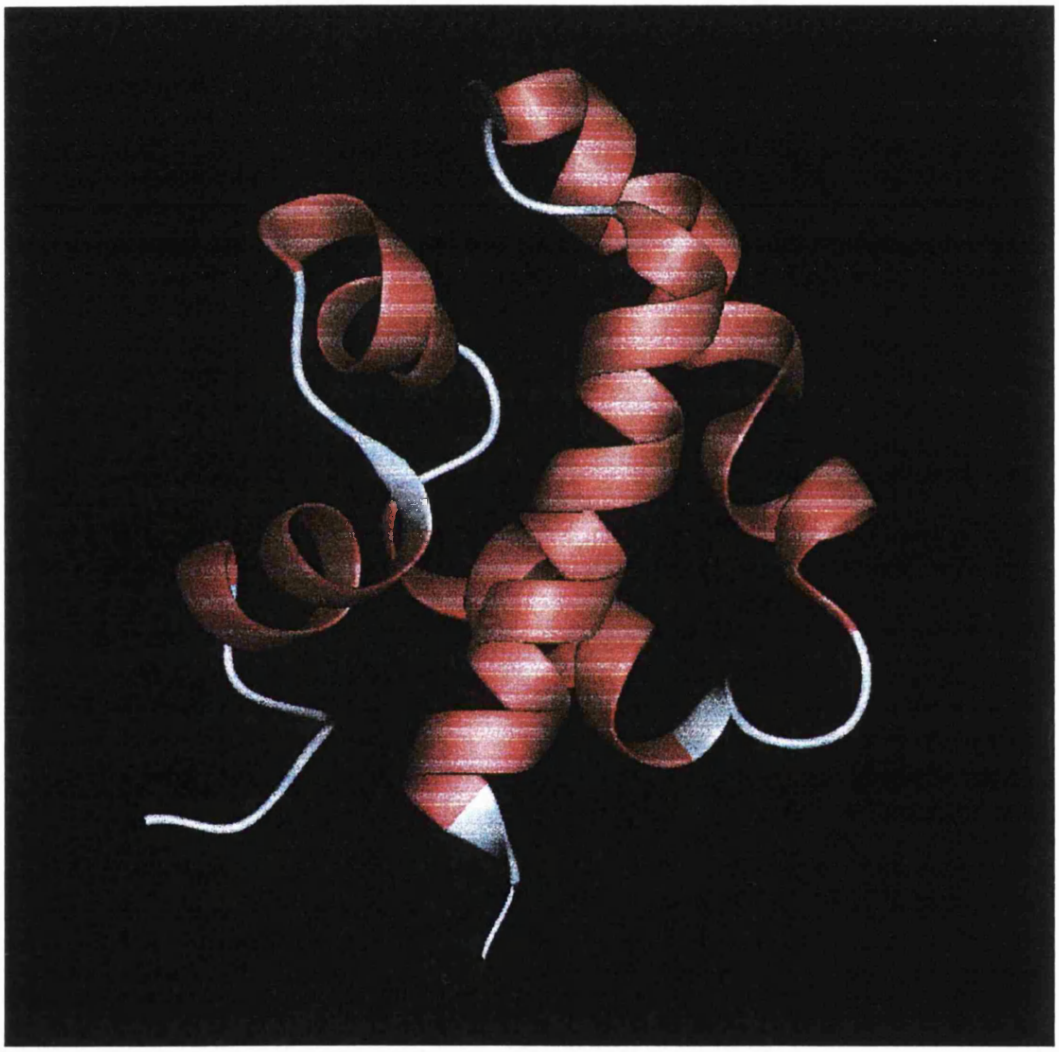


Figure 5.7: NMR structure of NK-Lysin.

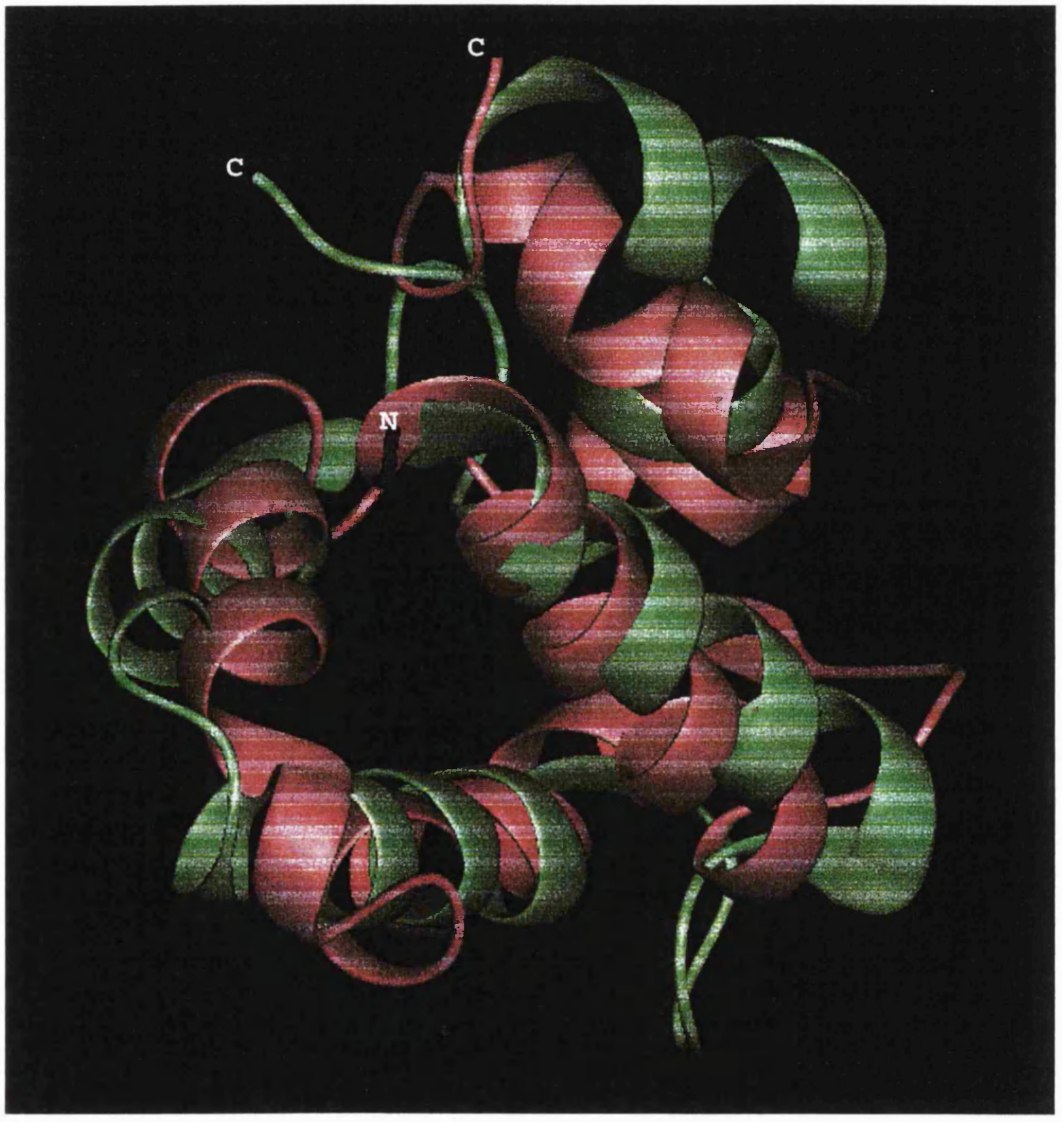


Figure 5.8: QUANTA superposition. A DRAGON model based on the correct secondary structure(green) with the NMR(red), giving a C_α overall superposition of 4.6Å.

5.4 Discussion

The result of submitting one model for the *ab initio* prediction assessment of CASP2 turned out to be a bad strategy. The highest ranked model using the restraints function score turned out to be close but had the incorrect handedness (see Figure 5.5). Using two measures, the restraints score and the bond score (see Table 5.2), a model with the correct handedness and 5.9Å RMSD from the NMR structure scores highest – 5.6Å RMSD over the best 70 C_α atoms. Furthermore models with visibly more plausible structures and the correct handedness occurred higher up the rankings when two scores were included to produce the ranks.

The highest scoring models when superposed using the most recent SAP algorithm (Taylor, 1999) were not easily identified by the DRAGON scoring methods alone, so future work would ideally bring some of the SAP measurement into play. However Table 5.6 shows that in most cases the highest SAP score does not have the best RMSD. The scores are also not easily comparable, varying between 200 and 800 for the highest score in any one run. Unless further work is carried out to try and normailse these scores then it is best to refer to the RMSD when comparing structures.

As the SAP scores clearly show in Table 5.4, Table 5.5 and Table 5.6 they cannot be used accurately to assess the models. In Table 5.5 model 2_40 does well in all but the SAP rankings. A high SAP score such as that obtained in one of the models generated in run 4 (Table 5.6) is not indicative of the best model with lowest RMSD. The RMSD measure is generally a much better assessment of overall model quality.

The SAP measure may score highly with a relatively poor RMSD if there are areas in the superposition with good local similarity.

The models created using the correct secondary structure look better by eye and generally give a lower RMSD but not by a large margin.

The models shown in Table 5.2 were ranked according to the DRAGON bond and restraint scores. Essentially the lower the number the better satisfied are the distance restraints. You would expect the best model, which satisfies the given restraints, to have a score close to zero. For example when the disulphide bonds were close enough to satisfy the distance restraints in the additional restraints file then the restraints may be satisfied and a low score produced for that model.

By analysis from the models it is clear that these scores do not go all the way to fulfilling a good method for the choice of *ab initio* models. The choice of the model for submission to the CASP2 was wholly on the external restraints score (in this case the relative positions of the 3 disulphide bonds) as in Table 5.1. As it turned out the model was of correct topology but incorrect handedness.

Variable secondary structure weights: Models created using DRAGON with variable strictness for the secondary structure according to PHD did not give good helical structures. This is because the helices were not as restrained and were treated more as several consecutive short helices rather than one longer helix.

Full atom representations: It is possible to take the simplified DRAGON models and build full atom representations based on the fold. Full atom models were not built from all the models but might have given more insight into a better fold with correct handedness. One model was chosen and converted to a full atom representation. It is possible that modelling with full disulphide geometry may improve the DRAGON models. The inability to resolve the handedness of the protein structures is a basic problem with a distance geometry based approach. Four-helix bundles are known to exist in a left handed form in nature, but most proteins adopt a right handed conformation.

Modelling with the observed secondary structure: A small amount of modelling was performed using DRAGON where the secondary structure assignments were taken from the NMR structure, post CASP2. The resulting models gave a slightly higher RMSD with the best of 4.8Å. This would be by no means the best achievable, further model runs would likely give better models.

5.4.1 Function

A knowledge of the function of the protein can be helpful, possibly classifying it into various families, of which there may have been known structure. Also by identifying any areas of known importance in the sequence of the protein the likely structure could also be predicted. For example, the location of disulphide bonds can give an easy idea of whether or not a protein structure is feasible with a given sequence, as has been seen here.

Investigations are under way to explore the ability for a more accurate incorporation of a secondary structure prediction into the program, making use of the strictness option in the assignment. Work to enhance the way in which the best model is chosen is also being considered. Without homology modelling restraints the models created by `DRAGON` are reasonably diverse so picking the right answer can be a difficult task. Future analyses might be able to incorporate information gained through correlated mutation studies and derive possible distance contacts from them.

5.5 Conclusion

An improvement in the DRAGON scoring function might make the choice of models simpler, with perhaps a type of energy function built in. Ideally the model with the best RMSD should be picked out *a priori*. DRAGON has been used to give an indication of the disulphide bonding possibilities in other proteins when this information is unknown. If models are generated where disulphide information is not included in the NK-Lysin models, would it then be possible to discern which disulphide bonds would form and are they the *correct* bonds? Chapter 7 details this idea and its application in other proteins with disulphide bonds.

Further modelling targets are being looked at in conjunction with various collaborators with a view to producing models before the crystallographic work is successfully completed. The prediction of the N-terminal domain of a GPCR is covered in Chapter 8.

The use of DRAGON as a method for building *ab initio* models shows considerable promise. As can be seen from the results, most models are plausible including many which are close to the correct fold. Taking all the ranking criteria into account one can come up with a close model to the NMR data supplied by the experimentalists at CASP2.

The efficacy of DRAGON's *ab initio* modelling can be increased with few additional restraints. With more in the way of NMR data and other biochemical information

it would be possible to get better and better models. It has already been shown that models can be created using DRAGON by homology modelling and also putative homologues identified by multiple sequence threading methods. A major point which came out of the CASP2 conference was the importance of gaining a good alignment between the homologous structure and the target sequence when homology modelling. For tertiary *ab initio* models it is vital to get a reasonable secondary structure prediction and a good multiple alignment. As the models generated were performed using a good secondary structure prediction, there is not much gain in the models when using the correct secondary structure. There were no models which were very close to the NMR structure so some work on the packing algorithms in the core and the driving force for more protein-like structures in DRAGON should be improved.

While the problem still remains of choosing the best model, without the benefit of hindsight, much has been gained in my ability to assess the performance of DRAGON and the role which distance geometry will have to play in the future of *ab initio* modelling.

A judgment "by eye" still plays an important part in the decision to accept or reject a model. Ideally, purely a computational decision would give a more reliable reason for accepting a model.

5.6 Appendix: Example CASP2 submission

```
PFRMAT ABF1
TARGET T0042
AUTHOR 3272-9168-2129, Robin Munro, NIMR, r-munro@nimr.mrc.ac.uk
REMARK  Methods used secondary structure prediction methods PHD (1) and
REMARK  DSC (2).  Using disulphide bond constraints, a MULTAL (3) alignment
REMARK  and the secondary structure predictions a model was constructed
REMARK  using DRAGON (4,5) Distance Regularisation Algorithm for
REMARK  Geometry OptimisatioN, which generates C-Alpha traces. The
REMARK  model which satisfied the restraints best was chosen.
REMARK
REMARK  (1) Rost, B., Sander, C. and Schneider, R.  CABIOS 10:53-60 (1994)
REMARK  (2) King, R. D. and Sternberg, M. J. E.  Protein Science (in press)
REMARK  (3) Taylor, W. R.  J. Mol. Evol. 28:161-169 (1988)
REMARK  (4) Aszodi, A. et al.  J. Mol. Biol. 251:308-326 (1995)
REMARK  (5) Aszodi, A. and Taylor, W. R.  Folding & Design 1:325-334 (1996)
REMARK
REMARK
REMARK
BEGDAT 3.1  1  0.1
HEADER    PROTEIN MODEL                              04-OCT-96
COMPND    MODEL C-ALPHA:FAKE C-BETA CHAIN
SOURCE    DRAGON Version 4.16.1: compiled on Aug  2 1996, 19:01:16
EXPDTA    THEORETICAL MODEL
REMARK   1 NOT A GENUINE PDB ENTRY!
REMARK   2 RESOLUTION. NOT APPLICABLE.
REMARK   4 BOND SCORE: 5.831e-03
REMARK   4 BUMP SCORE: 0.000e+00
REMARK   4 EXTERNAL RESTRAINT SCORE: 8.392e-02
SEQRES   1     78  GLY TYR PHE CYS GLU SER CYS ARG LYS ILE ILE GLN LYS
SEQRES   2     78  LEU GLU ASP MET VAL GLY PRO GLN PRO ASN GLU ASP THR
SEQRES   3     78  VAL THR GLN ALA ALA SER GLN VAL CYS ASP LYS LEU LYS
SEQRES   4     78  ILE LEU ARG GLY LEU CYS LYS LYS ILE MET ARG SER PHE
SEQRES   5     78  LEU ARG ARG ILE SER TRP ASP ILE LEU THR GLY LYS LYS
SEQRES   6     78  PRO GLN ALA ILE CYS VAL ASP ILE LYS ILE CYS LYS GLU
HELIX    1  H1 SER      6  VAL     18  1
HELIX    2  H2 GLU     24  ASP     36  1
HELIX    3  H3 ARG     42  ILE     60  1
HELIX    4  H4 PRO     66  ASP     72  1
ATOM      1  N   GLY     1       8.226  -2.780  14.856  1.00  5.00
ATOM      2  CA  GLY     1       8.357  -3.775  13.742  1.00  5.00
ATOM      3  CA  TYR     2       5.709  -5.883  11.897  1.00  5.00
ATOM      4  CB  TYR     2       3.250  -8.802  13.367  1.00  5.00
ATOM      5  CA  PHE     3       5.955  -4.554   8.320  1.00  5.00
```

Chapter 6

Multiple Sequence Threading: gap placement

6.1 Introduction

A threading method by Taylor called `MST` has been described in the literature (Taylor, 1997) and it was this program which was primarily used on the CASP2 fold recognition targets, discussed in Chapter 3. An important problem in most, if not all, fold recognition (threading) methods is the inability correctly to predict the optimal sequence to structure alignment. It can be particularly difficult to get the placement

of insertions and deletions in structure or sequence correct. In this chapter it is considered whether it would be possible to build a weight into the alignment method to cope with the placement of gaps in a better fashion. For this the globin family is used as a suitable starting point to assess different measures for the placement of gaps in the multiple sequence threading alignment process. Based on a structure-structure comparison of two proteins in a sub-family the threadings can be compared, with multiple alignment information being used as a probe.

By using a multiple sequence alignment instead of a single sequence more information can be obtained. This may redress the problem where a two single structures may have a similar fold, but very remote sequence identity. It may be that a family of sequences can redress the imbalance and help with detection of similarity in this 'twilight-zone' (Taylor, 1995b).

This chapter breaks down the alignment problem into a series of measurements which will show whether the problem of gap placement can be incorporated into a type of gap penalty. Four different situations were considered: deleted structure, inserted sequence, gap ends in structure and broken ends in sequence. Each was analysed for exposure, occupancy and secondary structure. These measures should give some insight into the placement of gaps when using `MST`.

The predictions made by *ab initio* folding methods are often very hard to assess, although correct models can be generated as shown in Chapter 5. Conversely with a good sequence alignment to a known PDB structure good models can be obtained

far more easily from homology modelling. When sequence alignments fail to identify a protein with similar sequence then threading can often bridge the gap, as shown in Chapter 4.

Generally speaking when comparing protein sequences, the two sides to be aligned can both contain any number of pre-aligned sequences, some of which may be the sequences of proteins with known structure. In this chapter no comparisons are made when structural information is retained on both sides, as this is just a case of comparing 3D structures (Taylor and Orengo, 1989; Sali and Blundell, 1990). Conversely with only sequence knowledge, it is just a case of multiple sequence alignment (Taylor, 1988; Thompson *et al.*, 1994).

The `MST` method (Taylor, 1997) incorporates both a threading approach, where some position of the sequence on a structure gives good core packing, by using pairwise residue interaction preferences (Sippl, 1990; Jones *et al.*, 1992a; Bryant and Lawrence, 1993) and also a 3D/1D matching method, where characteristic states are measured from a structure and then compared against the predicted states of the sequence (Bowie *et al.*, 1990; Lüthy *et al.*, 1991; Russell *et al.*, 1996).

6.2 Methods

In this analysis test-data were constructed from the Globins of known structure by creating sub-families in which one structure was taken as 'known' and the structures of the others ignored. Pairs of sub-families were and then compared as a structure-sequence threading. The two sub-families to be aligned will be referred to as the *sequence-side* when no member has a structure and the *structure-side* when one or more known structure occurs in the alignment.

6.2.1 Burial of conserved hydrophobics

The hydrophobic effect has been much mentioned previously and dominates the folding process. The basic assumption that a globular protein should have a well packed hydrophobic core, is fundamental to the threading method. These buried positions in the sequence alignment appear as both hydrophobic and well conserved, or conphobic for short (Taylor and Aszódi, 1994a). Other functionally important areas may also be well conserved, but are rarely hydrophobic and are found near the surface of proteins and in loops. It is because of this that multipally aligned sequences with conphobics should try and pack them in the core of the protein, these score highly when this is the case.

As a measure of conphobicity the position in the multiple alignment is assessed for conservation and hydrophobicity and a product of these two measures taken. Con-

servation was measured as a pairwise sum of amino acid similarity (using a Dayhoff model) over the residues aligned at that point. Hydrophobicity was measured as the average over the aligned positions using a scale of hydrophobicity (Taylor and Aszódi, 1994b). Burial was measured as the sum of residue contacts with solvent molecules and their packing density, or neighbouring contacts.

The score used for aligning a conphobic position in the sequence with a buried position in the structure was the product of the measures described above. The product being used to give a high weight to a match of strong conphobics in deeply buried positions.

6.2.2 Matching of predicted and observed sec. str.

Due to the complexity of many of the methods now available for secondary structure prediction (see Introduction), it is easier to incorporate an older yet reasonably successful method into the program. The GOR method (Garnier *et al.*, 1978) using multiple sequences (Zvelebil *et al.*, 1987) was adopted as one still proving capable of producing reasonable results (Levine *et al.*, 1993).

A measure of secondary structure from coordinate sets was devised to coincide with a high propensity in the middle of a secondary structure region. This ties in well with the variable GOR propensity, allowing easy comparison between observed and predicted secondary structure.

6.2.3 Tertiary packing measure

So far all that has been mentioned constitutes a 3D/1D type fold recognition approach. Also incorporated in the MST program is a pairwise residue packing interaction. Where all pairs of residues are ranked by their C_β separation and then compared with each other to examine whether one pair is shielded from another. From this the packing of buried pairs can be calculated.

Further details of the method and the packing measure, all devised by W. R. Taylor can be found in a recent J. Mol. Biol. paper (Taylor, 1997).

6.2.4 Gap penalties

Dynamic programming is widely used in many alignment situations and can incorporate a gap penalty. Using an iterative version of this algorithm it is possible to allow gap weighting functions which are not restricted by the fact that once a gap has been created the process can't go back and change it. Thus, using double dynamic programming all gaps can be assessed together.

Here I investigate not only the gaps which occur in sequence alignment, but also the insertion and breakage of structure either side of the insert and in the inserted segment (Taylor, 1995a).

The following maxims outline the aims:

1. ends remaining after deletion of part of a structure should be close in space;
2. insertion of sequence should not occur between (adjacent) buried positions and preferably not in secondary structure;
3. strongly predicted and core secondary structures should be matched;
4. strongly predicted and deeply buried positions should be matched;
5. gapped regions are the preferred site of further insertion and deletion.

These rules define good molecular modelling practice and are commonly adhered to intuitively when 'modelling-by-homology' (Pearl and Taylor, 1987a). Rules 1 and 2 are usually applied quite strictly. Rules 3 and 4 will sometimes be broken, they emphasise that it is undesirable to omit conserved structures and positions from the final model. Rule 5 refers to the pre-existing gapped positions that occur when dealing with aligned sequences on one or both sides of the problem.

In the following sections, A designates the sub-family to be aligned that has a known structure (structure-side) and B the sub-family with no known structure (sequence-side). Penalties were developed for the point of insertion on both the sequence and structure sides and for the content of the inserted segments, again, both on the sequence and structure sides. For reference, it is convenient to represent the index of

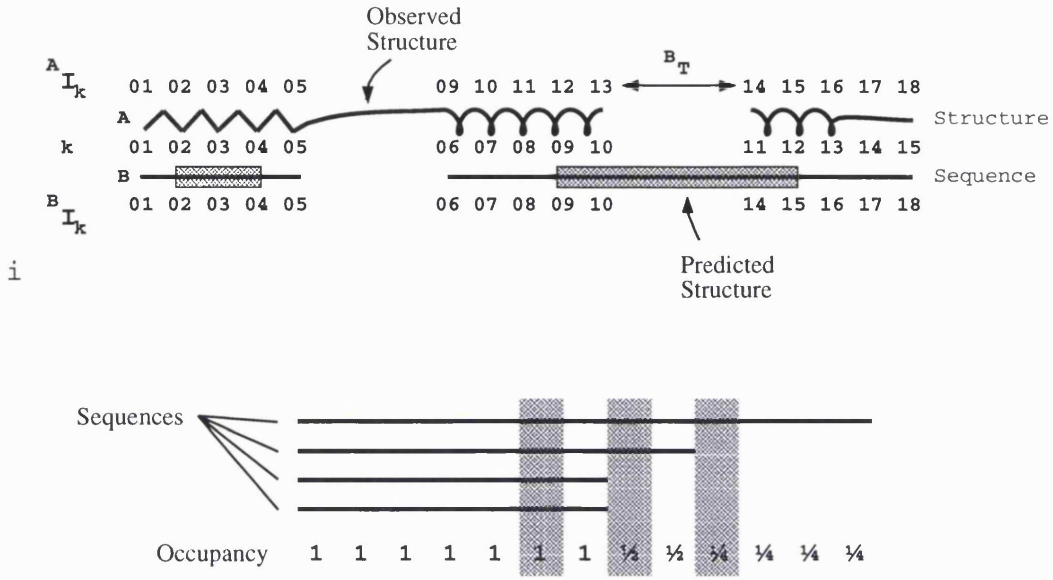


Figure 6.1: Sequence/Structure alignment. i) Illustration of the series representing the aligned elements in equations 1 and 2. — ii) the alignment occupancy is calculated as a proportion of residues found at any one position in an alignment.

the aligned elements of a sequence as a series (I), where:

$$^{A}I_k = \mathcal{P}(\mathbf{R}, A, k), \forall k = 1 \ldots K. \tag{6.1}$$

The function $\mathcal{P}(\mathbf{R}, A, k)$ returns the position in sequence A for the k^{th} match on the best path through the score matrix $\mathbf{R}$, for an alignment of K matches. Similarly for sequence B:

$$^{B}I_k = \mathcal{P}(\mathbf{R}, B, k), \forall k = 1 \ldots K. \tag{6.2}$$

(See Figure 6.1).

The terms: insertion and deletion will be used to refer to the final model. Thus, deleted structure will not form part of the final model whereas inserted sequence will.

${}^{B}O$	Occupancy, see Figure 6.1, ii
p	GOR prediction of secondary structure
c	Product of hydrophobicity and conservation
C	Number of residue contacts for a given position
e	Conic accessibility measure
s	Sec. Str. score based on deviation from an ideal structure

Table 6.1: Definitions of factors. The first three factors are predicted from the sequence alignment. The last three factors are observed from the structure side.

Deleted structure A gap on the sequence-side corresponds to a deletion of structure leaving two unconnected ends in the model see Figure 6.2. If these ends are taken as residue positions i and j, a score based on their α-carbon separation (d_{ij}) was devised thus:

$$^{A}S = \sum_{\text{gap } k} d_{ij}^2 \tag{6.3}$$

The sum ^{A}S was taken over all gaps with inserted structure between k and $k+1$ but for simplicity, i substitutes for $^{A}I_k$ and j for $^{A}I_{k+1}$. On the sequence-side, the alignment occupancy (^{B}O), predicted secondary structure state (p) and predicted degree of exposure (the conphobic score, c) of the broken ends should all affect gap placement. If an insert was placed between two residues, then the penalty for insertion should be high if both residues are buried or in a secondary structure or have full occupancy. This behaviour was achieved by taking the product of the properties on both ends, at the gap. As these components have differing numeric ranges and reliabilities they were retained as separate components for independent evaluation. These terms, designated respectively, T, P and Q were defined as follows:

$$^{A}T = \sum_{\text{gap } k} {}^{B}O_i \cdot {}^{B}O_{i+1}, \tag{6.4}$$

$$^{A}P = \sum_{\text{gap } k} {}^{z}p_i \cdot {}^{z}p_{i+1}, (z = \alpha, \beta), \tag{6.5}$$

$$^{A}Q = \sum_{\text{gap } k} (c_i + 1) \cdot (c_{i+1} + 1). \tag{6.6}$$

In each of the preceding equations the sum was taken over the two ends of all gaps with inserted structure between k and $k+1$ but for simplicity, i substitutes for $^{B}I_k$.

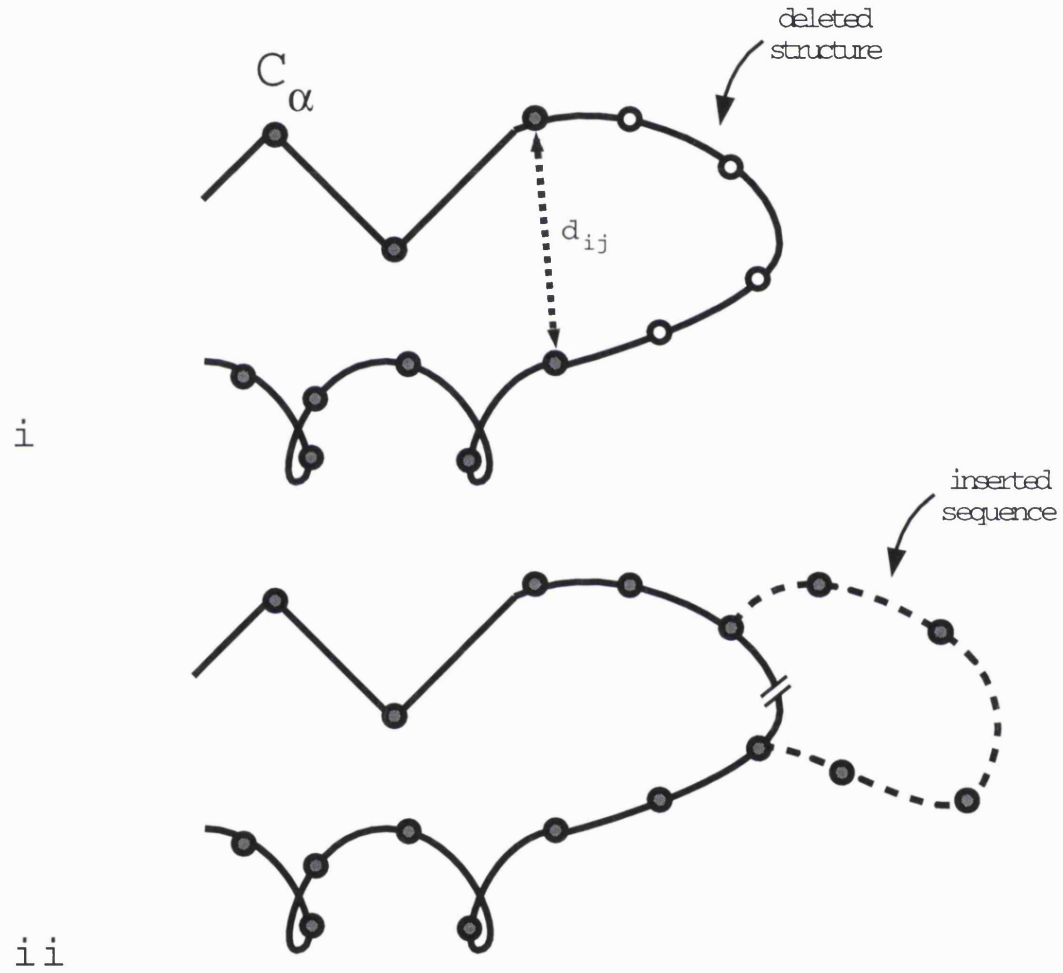


Figure 6.2: Insertions and deletions. i) deleted structure is represented by a gap on the sequence side — ii) inserted sequence into a structure.

Inserted sequence As exposed and variable loop regions in the structure are the preferred location for the insertion of sequence into a structure, a penalty was based on the number of existing gaps (occupancy, O), exposure and secondary structure state of the residues in the structure flanking the insertion. As above, an exposure measure was based on both the number of residues in contact (C) with a given position and the solvent exposure measure (e), as measured by conic accessibility, while the secondary state was measured by the score (s) based on deviation from an ideal secondary structure. Following the preceding formulation for deleted structure, the corresponding terms for inserted sequence are:

$$^{B}T = \sum_{\text{gap } k} {}^{A}O_i \cdot {}^{A}O_{i+1}, \tag{6.7}$$

$$^{B}P = \sum_{\text{gap } k} {}^{z}s_i \cdot {}^{z}s_{i+1}, (z = \alpha, \beta), \tag{6.8}$$

$$^{B}Q = \sum_{\text{gap } k} (C_i + e_i)(C_{i+1} + e_{i+1}). \tag{6.9}$$

In each of the preceding equations, the sum was taken over the two ends of all gaps with inserted sequence between k and $k+1$ and i is substituted for ${}^{A}I_k$.

Lost secondary structure: Unmatched secondary structure is effectively 'lost' from the final model. This was measured by calculating the fraction of both observed (${}^{A}V$) and predicted (${}^{B}V$) secondary structure matched in the sequence/structure alignment. For the observed structure measure (s), the fraction unmatched is then:

$$ {}^{A}V = 1 - \frac{\sum_{k=1}^{K} ({}^{z}s_{I_k}) {}^{A}O_{I_k}}{\sum_{k=I_1}^{I_K} ({}^{z}s_k) {}^{A}O_k}, (z = \alpha, \beta) \tag{6.10} $$

and similarly, for predicted structure (p):

$$ {}^{B}V = 1 - \frac{\sum_{k=1}^{K} ({}^{z}p_{I_k}) {}^{B}O_{I_k}}{\sum_{k=I_1}^{I_K} ({}^{z}p_k) {}^{B}O_k}, (z = \alpha, \beta). \tag{6.11} $$

In these fractions the numerator is the sum over the matched positions (I_k) which is normalised by the sum over all sequential positions (k) on the denominator. It is worth noting that this latter index runs from the first (I_1) to the last (I_K) match and therefore does not include terminal deletions. This form was chosen for consistency with earlier alignment algorithms (Taylor, 1988).

Lost burial: Fractional loss of exposure was calculated as above by summing exposure measures rather than secondary structure measures: substituting the observed exposure ($C+e$) for s and the predicted exposure ($c+1$) for p; giving equivalent terms: ${}^{A}U$ and ${}^{B}U$, respectively.

6.3 Results

Shown below is an initial characterisation of each of the measured factors, as described in the methods for the globin family.

A major problem in the proposed analysis is that even using a 'perfect' method for structure comparison, there will be variation in the exact placement of gaps when various members are compared within a family. To avoid this problem, I will consider families in which each member has a known structure and generate all pairs of structure comparisons. For each structure pair, the number of additional sequences aligned in each sub-family will directly affect the alignment occupancy and indirectly, the prediction of secondary structure and exposure on the sequence-side. To assess this effect, the remaining sequences could be allocated to the sub-families either combinatorially, randomly or phylogenetically. I have adopted the latter approach keeping each sub-family at all times composed of the sequences most related to the member that is used to determine the structure alignment. Structural alignments were generated using `SAP` (Taylor and Orengo, 1989) and phylogenetic relationships of sequences were taken from a `MULTAL` alignment (Taylor, 1988). The resulting overall sequence allocation scheme is shown in Figure 6.3.

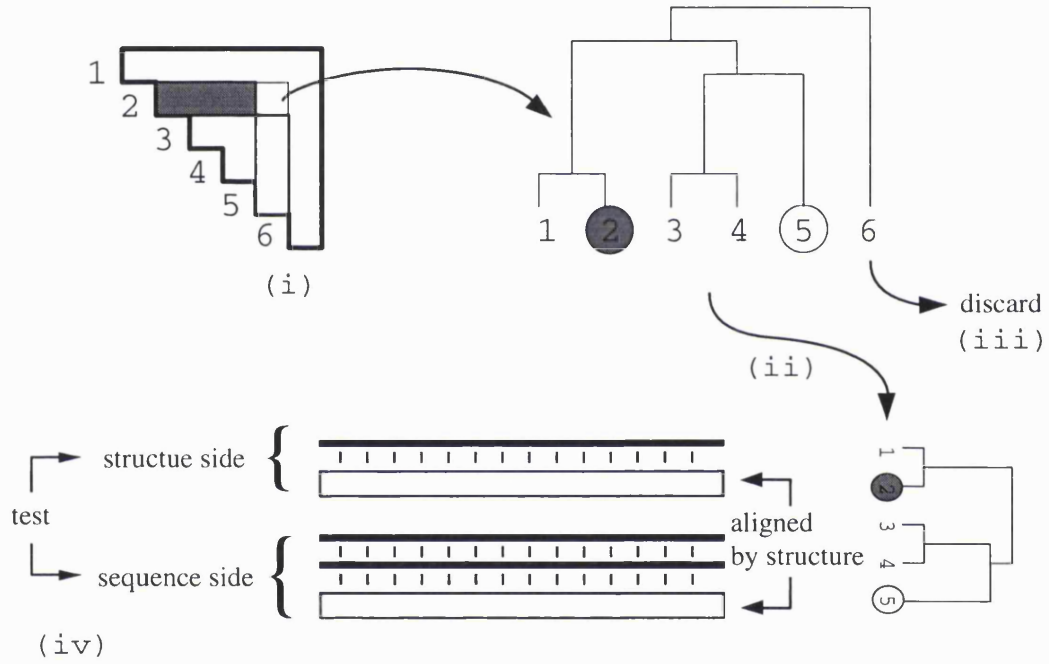


Figure 6.3: Schematic of the analysis of proteins using MST. In this case all pairs of Globins were compared (i). A pair of structures (2 and 5) were grouped into sub-families of homologous sequences (ii). Any sequences too remote to be included in either sub-family were not used for that pair of structures (iii). The two resulting sub-families were aligned by structure and tested, one on the structure side and the other on the sequence side (iv).

Twelve Globin sequences with known structure were chosen as a set for testing the placement of gaps. They are: `1gdj`, `1babA`, `1ash`, `1mba`, `1bvd`, `1hlb`, `3sdhB`, `2hbg`, `1eca`, `1flp`, `1ithA` and `2lhb` (where A or B is the chain identifier). A Threading of one sub-family on to a globin structure (`2lhb`) is shown in Figure 6.4.

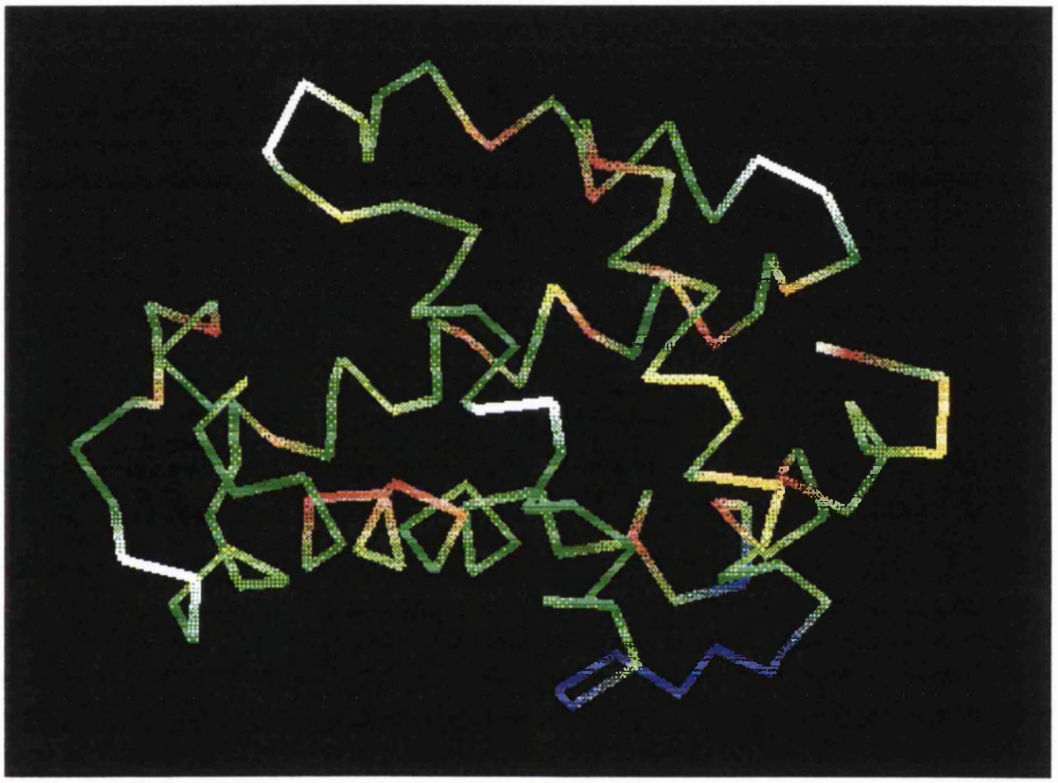


Figure 6.4: Threading of a sub-family on to a globin structure. α-carbon chain of the Multiple sequence threading of the globin sub-family `2lhb` on to the α-carbon chain of `1ithA` (based on the structure alignment of `1ithA` and `2lhb`) indicating areas of insertion in black and deletion in blue (figure iiia); red areas are hydrophobic.

6.3.1 Analysis of results

The structural alignment used to align the sub-families, one on the structure side and one on the sequence side are assumed to be the best possible alignments attainable by this approach. The multiple alignment step is open to some adjustment, but this is an inherent variable in the MST approach. All the factors (P, Q, S, T, U and V) were quantified by their mean and variance with a view that this might lead to a standard (inverse-variance) weighting scheme. However it was found that these values were very conservative estimates of the factors — being dominated by many alignments in which there were few gaps. Of greater interest to the future application of these quantities in alignment is the extreme upper values that can be found in structure alignments. These data were summarised by extracting the ten most extreme values for each factor, see Table 6.2, Table 6.3 and Table 6.4.

Gaps: (factors P and Q)

Secondary structure: Observed (${}^{A}P$) and predicted (${}^{B}P$) secondary structure were found to be almost universally broken by gaps in the in the worst cases. See Table 6.2.

Exposure: Sequence inserts in the structure (${}^{A}Q$) were found only between slightly buried positions which had a typical score around 12, Table 6.2. This value could be attained by two 'broken' ends that were 'half' buried ($e = 0.5$) with three neighbours each. (The maximum values for the most deeply buried residues would be an order of magnitude higher). The values for the predicted exposure also provide a strong constraint. As the conphobic measure (c) lies in the range -1:1, the maximum score

possible for ${}^{A}Q$ is 4, yet the worst observed values were typically 0.5 — corresponding to a gap between two relatively variable and hydrophilic positions.

			1	2	3	4	5	6	7	8	9	10
Sec	Obs	${}^{B}P$	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.981	0.970	0.961
	Pred	${}^{A}P$	1.000	0.999	0.999	0.993	0.983	0.983	0.981	0.979	0.952	0.934
Exp	Obs	${}^{B}Q$	13.543	13.071	11.749	11.662	11.534	10.703	10.529	10.377	10.334	10.281
	Pred	${}^{A}Q$	0.592	0.509	0.498	0.479	0.470	0.466	0.444	0.435	0.359	0.340

Table 6.2: Secondary structure and exposure state of the broken ends flanking gaps. The ten worst mean scores for pairwise comparisons of a set of 12 Globin sequences with known structure. *Sec* shows the breaks in predicted and observed secondary structure (${}^{A}P$ and ${}^{B}P$). *Exp* represents breaks between predicted buried residues (${}^{A}Q$) and observed buried positions (${}^{B}Q$). The values of ${}^{A}Q$ were not calculated for a single sequence.

Inserts: (factors U and V)

Secondary structure: The fraction of secondary structure found in (non-terminal) inserts is compiled in Table 6.3 for each of the ten most extreme alignments. These range from 4.5% to 6.8% for the observed structure (${}^{A}V$) and consistently about twice as much for the predicted structure (${}^{B}V$) (8.4%–11.7%). Any future weighting scheme could clearly reflect this two-fold difference.

Exposure: A similar, but less extreme trend was observed for the observed and predicted exposure (U).

Occupancy: (factor T) Many of the alignments generated did not contain gaps mainly due to the close similarity within the family. Hence any analysis of the extreme values for the occupancy would be trivial. Instead, the selection criterion for the ten examples was reversed. This would then extract the most gapped positions. Although a less informative statistic it does, nonetheless, indicate that gaps have been inserted into previously gapped positions. See Table 6.4.

End separation: (factor S) The end separation after the deletion of structure has a smallest distance of about 6Å, corresponding to the separation of two positions at i and $i+2$. The mean values found in the proteins with the biggest end separations were only slightly in excess of this — by typically, 3–4Å. However by looking at the individual gaps some end separations approach 15Å, which is undesirably large. See Table 6.4.

			1	2	3	4	5	6	7	8	9	10
Sec	Obs	${}^{A}V$	0.068	0.063	0.063	0.054	0.054	0.053	0.051	0.046	0.046	0.045
	Pred	${}^{B}V$	0.117	0.108	0.106	0.100	0.097	0.089	0.088	0.087	0.084	0.084
Exp	Obs	${}^{A}U$	0.072	0.056	0.047	0.043	0.037	0.037	0.037	0.036	0.035	0.033
	Pred	${}^{B}U$	0.076	0.068	0.065	0.060	0.059	0.056	0.054	0.054	0.052	0.052

Table 6.3: Secondary structure and exposure in inserts. Table showing observed and predicted lost secondary structure (${}^{A}V$ and ${}^{B}V$) and observed and predicted lost burial (${}^{A}U$ and ${}^{B}U$). N.B. These scores are normalised and can be compared directly.

			1	2	3	4	5	6	7	8	9	10
	gapSum	${}^{A}S$	353.857	334.378	318.415	295.708	268.174	258.847	252.792	248.408	244.714	243.127
	CadSum		40.077	37.768	36.832	36.376	36.018	35.844	35.304	34.303	33.874	33.578
	CadMean		13.145	12.727	9.867	9.805	9.645	8.645	8.635	8.424	8.348	8.289
	CadMax		14.490	13.842	13.794	13.535	13.459	13.396	13.368	13.145	13.059	12.899
Occ	Obs	${}^{B}T$	0.250	0.600	0.625	0.625	0.625	0.625	0.664	0.708	0.711	0.714
	Pred	${}^{A}T$	0.563	0.563	0.563	0.625	0.646	0.688	0.708	0.750	0.750	0.770

Table 6.4: End-point separation and occupancy of broken ends. The sum of squares of all end separations where structure is deleted (${}^{A}S$) is shown, along with the un-squared sum (cadSum), its mean (cadmu) and maximum value (cadMax). Also shown are the best 10 values for predicted and observed alignment occupancy (${}^{A}T$ and ${}^{B}T$, respectively). The worst values for these two measures is simply 1, in all sequence alignments without gaps.

6.3.2 Analysis of factors

Analysis of the factors measured in the above analysis was plotted on graphs for all points in the Globins analysed. This gave some idea of the differences between the observed and predicted distributions. Shown in Figure 6.5 are all the results for each factor. These graphs are effectively expansions of Table 6.2 to Table 6.4, showing all points rather than the ranked top 10. Also calculated, but not shown were the frequency distributions for each of the factors in figure 6.5. No reasonable weighting system for the gap penalty could be determined from these graphs.

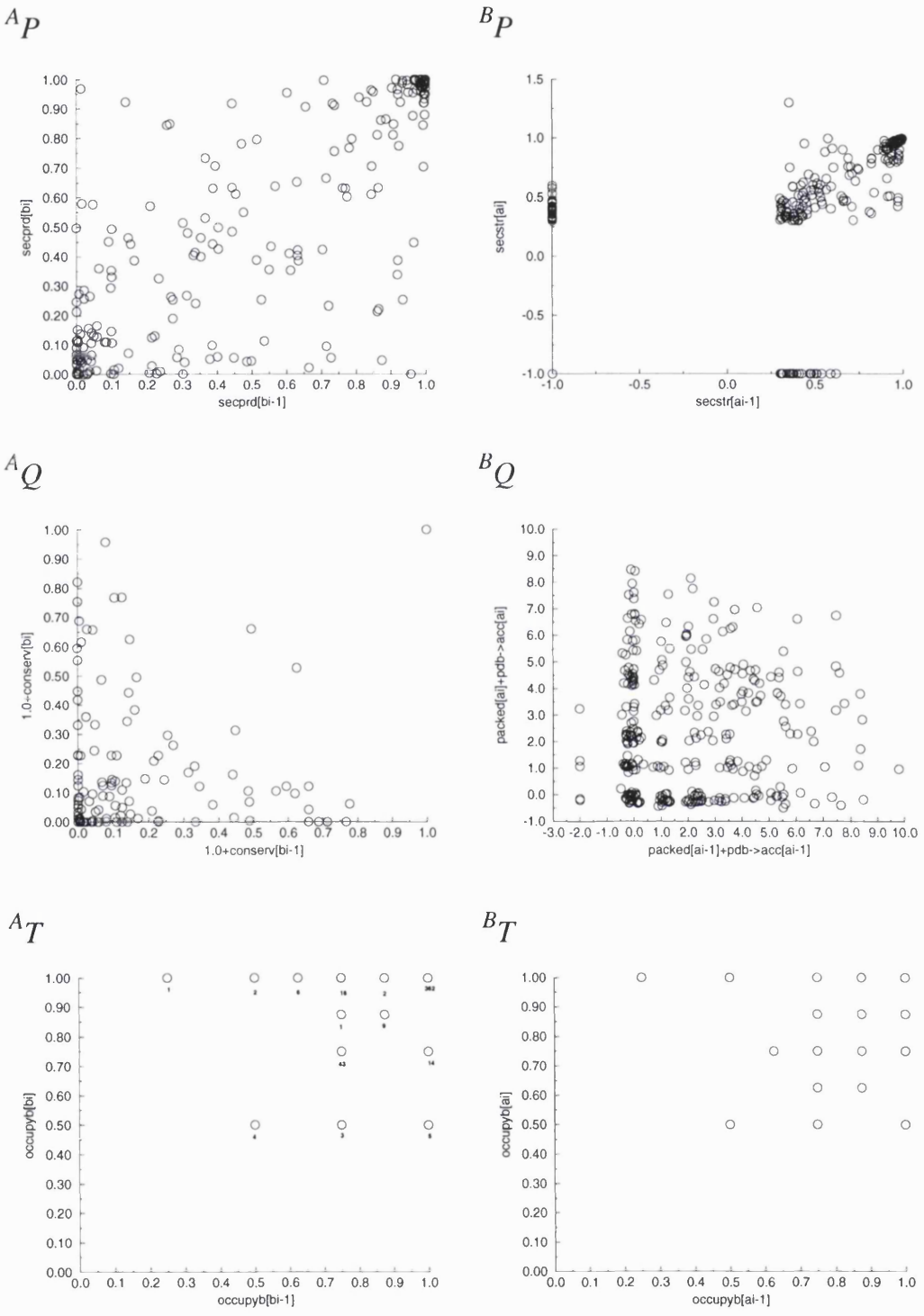


Figure 6.5: Plot of the measures given in equations.

6.4 Conclusion

These measures of gap placement were plotted (Figure 6.5) with the hope that they may give some insight into the overall building and refinement of a weight for the MST alignment process. Similarly, the most extreme situations were also analysed (Table 6.2 to Table 6.4). It can be seen that most secondary structures are broken by gaps. Sequence inserts in structure were never found in deeply buried positions. Where cuts have been made in the structure the ends were only separated by 3-4Å more than the minimum separation of 6Å. The maximum amount of observed secondary structure found in inserts was about half of the predicted structure (typically 5 and 10%, respectively). A similar trend occurred with observed and predicted exposure in inserts.

One of the more unexpected results was that the degree of exposure provides a stronger constraint on gap placement than the secondary structure state. Examples could easily be found where the insertions had occurred in secondary structure but none were found between buried positions. It is possible that the former measure is misleading, as it does not strongly distinguish gaps in the middle of secondary structure and at the ends. As would be expected, the predicted quantities provided less reliable constraints by around a factor of two.

The results of alignment occupancy (Table 6.4) were harder to evaluate. The main problem here was that the results were highly dependent on the number and similarity of the sequences used in the analysis. Perhaps a more rigorous test should be designed

with specifically controlled sequences.

The end separation remaining after deletion of structure is seldom violated to any great extent. This should provide a powerful constraint on gap-placement when encoded into a threading algorithm.

This is the first attempt to evaluate the contributions of the many factors involved in a generalised sequence/structure comparison in which multiple sequences appear on both sides being aligned. The main question arising from all this is the question of how all these mixed observed and predicted factors should be combined as a weight to improve the alignment algorithm. The problem has been well parameterised, but finding the best weights on the factors will be the harder problem which will require a wider test set.

Further analysis with other well characterised test sets such as immunoglobulins and flavodoxins will be required before a generic method can be applied to the more general threading problems. This is particularly so because the globin test set contains all alpha proteins, which may well bias the results so far attained. Closer analysis of the multiple alignments could also be made, but due to the reasonable sequence similarity of the proteins in this test set then the alignment is not a problem under these conditions. It is assumed that the SAP structural superposition was accurate and hence the sequence and structure sides were themselves well aligned (see Figure 6.3).

In principle, each could be weighted as a linear sum and the parameter space, defined by the weights, exhaustively explored. Even with fewer weights this can be a com-

putationally expensive procedure (Taylor, 1996), however, the search space can be narrowed by knowledge of the mean value and expected variation of the weighted factors. The current work has defined a protocol by which these values can be obtained from structure alignments. Future work should look at extending this approach using both more families whilst trying to incorporating a fuller analysis of the number and similarity of aligned sequences.

With the huge amount of interest in sequencing whole genomes, then being able to identify proteins which may have closely related structures will become more and more beneficial. When sequence alignment methods cannot detect any similarity, then a case identified by a fold recognition technique may be extremely valuable.

Chapter 7

Disulphide bond prediction

7.1 Introduction

Following on from the prediction of NK lysin, as shown in Chapter 5, several ideas arose. It was obvious from the folds generated for the NK lysin that the disulphide bonds played an important role in determining the fold of the protein. Without these additional restraints it would have been less easy to predict the correct fold. As knowledge of the disulphides were crucial for the NK lysin model, it was investigated to what extent disulphide bonds can generally be predicted in proteins and how the predictions can be used in modelling efforts. The following chapter looks at modelling

a protein with disulphide bonds with, as yet, unknown connectivity.

7.1.1 Disulphide bonds

Disulphides play a key role in stabilising folded proteins. When two cysteine residues are close in three dimensions, but not necessarily close in sequence, a disulphide bond is likely to form, by oxidation. Due to the requirement of oxidative conditions, intracellular proteins do not form disulphide bonds, even if the cysteines are in close proximity. A disulphide bonded pair of cysteines is commonly called cystine. Disulphide bonds have a hugely stabilising effect on proteins (Branden and Tooze, 1991) and may also be important in the formation of links between domains.

Thornton (Thornton, 1981), and more recently, Harrison and Sternberg have classified disulphide connectivity in proteins (Harrison and Sternberg, 1994) and went on to define regularities in small disulphide rich proteins (Harrison and Sternberg, 1996). They define a Disulphide β-cross motif, which they believe may be a core element in many small disulphide rich proteins.

Muskal *et al.* (Muskal *et al.*, 1990) trained a neural network to try and predict the disulphide bonding state of cysteines based on their flanking sequences. They achieved relatively high predictive power, indicating that the neighbouring amino acids do influence the formation of disulphide bonds. They also went some way in trying to predict disulphide forming pairs, but with only a few successes. Their

basis for prediction is to take the pairs with the most similar network properties and measure how close they were to each other in the sequence. Based on work by Wilmot and Thornton (Wilmot and Thornton, 1988), Muskal *et al.* suggest that a cysteine at position 0 is likely to bond to a cysteine at position +/- 5, if they are separated by residues with a high β-turn potential (e.g. Asp, Asn, Ser, Pro and Gly). Fiser *et al.* used a more simple method for distinguishing disulphide forming cysteines (half-cystines) (Fiser *et al.*, 1992), based also on the neighbouring sequence environment.

It was demonstrated using DRAGON that known disulphide bonds restrict the number of possible folds a protein can adopt (see Chapter 5). Without this information, the conformation space that needs to be searched increases dramatically. It would be a great step forward if the connectivity between cysteines in proteins could be predicted with greater accuracy.

As a result of the models produced in Chapter 5 for NK lysin, it was possible to compare the distances of all possible pairs of disulphides in many of the models generated with the DRAGON method. Assuming that DRAGON adequately samples the likely conformations which the protein could adopt and given a good secondary structure prediction (or even the correct secondary structure assignment, for the purposes of comparison) then all inter-cysteine distances can be monitored. For this purpose, a program called SSdist was written to calculate the distances associated with all possible cysteine pairings.

The SSdist program calculated both the C_α and C_β distances. For the purposes of this study an ideal disulphide bond is taken to have an approximate C_α:C_α distance of 7.8Å, with the side chain centroid (SCC) will be 3.5Å apart. If this were a true side chain representation, then the covalent sulphur:sulphur bond is 2Å. Disulphide bonds are rarely found with separation of less than two residues between the cysteines and normally with larger separation. Two examples with close cysteines are Deoxyribonuclease (DNASE I) where residues 101 and 104 form a bond; posterior pituitary peptide also has a similarly close disulphide. (For a review see ref. (Thornton, 1981)). It is worth mentioning that a value of 7.8Å does not always hold for a disulphide bond, for example the structure of CD4 has two types of disulphide bonds, one type which occurs between β-sheets and the other shorter bond within a β-sheet (Richardson, 1977). In this chapter only the major, dominant conformation is considered as a benchmark for future work.

Some DRAGON generated models may have closer cysteine pairs than others and it may be unlikely that one model will have all ideal disulphide bonds when no restraints are applied. By plotting the distances of all pairs of disulphide bonds in a group of models and by sorting them according to distance some inferences can be made as to the potential connectivity of cysteines in the models.

7.2 Method

The proposed method for the prediction of disulphide bond pairs is straightforward. Several steps were carried out, as follows:

1. Take one protein with cysteine residues known to form disulphide bonds as the target sequence.
2. This target sequence is compared to the many databases and homologues found.
3. A multiple alignment is constructed and checked for validity.
4. Secondary structure prediction is carried out based on the alignment.
5. Set up `DRAGON` to model the protein given its secondary structure or predicted SS, along with the appropriate multiple alignment.
6. Measure all possible pairs of cysteine:cysteine distances for C_α and optionally C_β atoms, this is not carried out in this analysis, but is discussed at the end of the chapter.
7. Rank by increasing inter-cysteine distance and plot all pairs on a graph.
8. Check secondary structure and disulphides, exclude pairs within a single helix or strand.
9. Plot sum of all combinations of pairs of cysteine distances and assess lowest scores.

Models can also be constructed where all permutations of disulphide bond are constrained and then the quality of the models assessed. Computationally this is more expensive as a greater number of models have to be produced. This would preclude some pairs of potential cystine bonds as the models produced would be unfeasible. The more straight forward protocol described above is much faster and the unlikely pairs should be consistently further apart.

NOTE: Model ranks in the following analyses relate to the individual cysteine pair distances, they do not refer to any one model, unless stated (as in Figure 7.11 and Figure 7.12). Consequently a monotonic increase will be observed for all curves. The reasoning behind this is that it is easier to gain a more "averaged" view over all the models, so a pair which is generally lower than the other pairs in all models, may well be a disulphide bond. Models which have, say, three ideal pairs would be picked up in a figure such as Figure 7.11. When the correct disulphide pairs are unknown (as they would be in reality) Figure 7.11 would be very hard to assess, unlike Figure 7.16. The pairs in the figures are sorted with a rank of 0 as the lowest distance.

7.3 Results

7.3.1 NK lysin

Figure 7.1 shows the distance plot of the disulphide bonds in NK lysin. Residue pairs 4+7 and 70+76 were excluded as they occur too close to each other in the sequence to form disulphide bonds and are consequently close in distance uniformly across most of the models. The next closest pairs are 35+45 followed by 4+76 and then 7+76. The first two pairs are observed in the NMR structure of NK lysin, the third is not. However with only three pairs of bonds if the first two are taken as the prediction then by inference the last pair must be 7+70 occurring about fourth in the chart, over all 60 models. In fact, looking more closely at the first 10 ranks, the three best pairs appear to be the correct disulphide bonding pairs, see Figure 7.2.

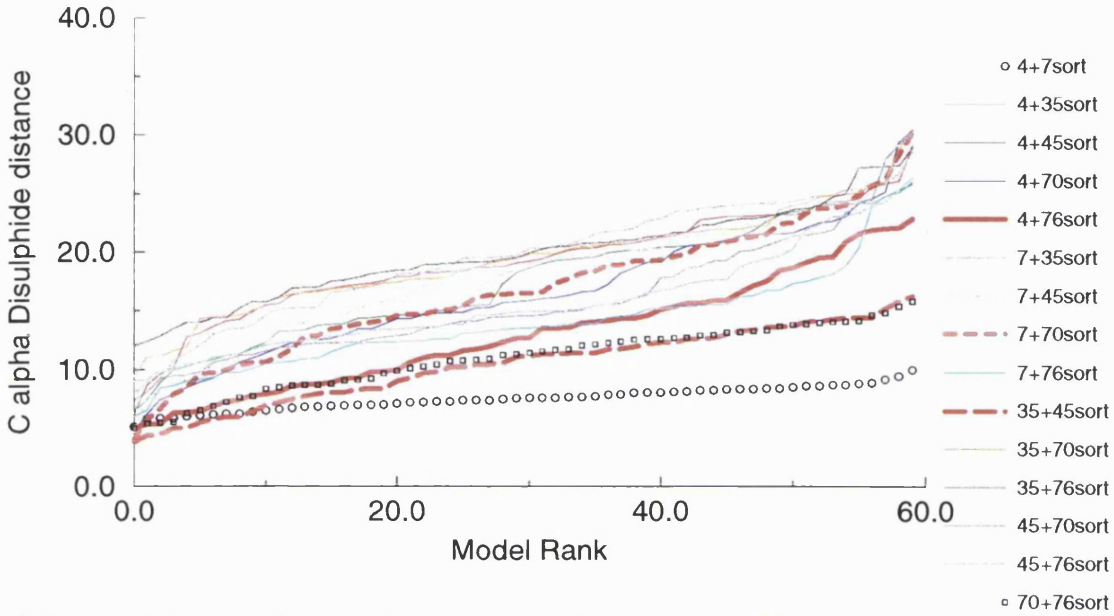


Figure 7.1: Disulphide analysis of the unconstrained NK Lysin DRAGON models. All 60 models with only 40% stringency and predicted secondary structure from PHD (see Chapter 5).

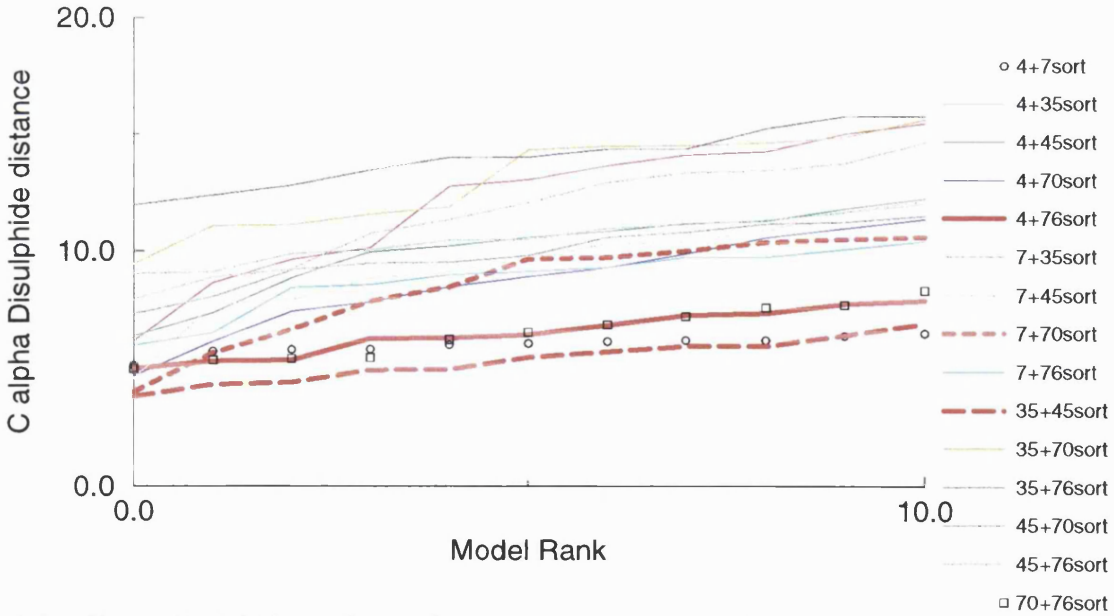


Figure 7.2: Closer disulphide analysis of the unconstrained NK Lysin models. Shown are the top ten disulphides with only 40% stringency and predicted secondary structure.

While the method seems hopeful for this protein, I examine in this Chapter how effectively the principle can be applied to other proteins. The prediction of C_α and C_β inter-cysteine distances is used in Chapter 8.

7.3.2 Test proteins

The following small cysteine rich disulphide bonding proteins were considered as they can be very quickly modelled in DRAGON. As a further complication, not all the cysteines which occur in the sequences are involved in disulphide bond formation. Disulphide bridging was modelled in the following proteins:

1occ(Tsukihara *et al.*, 1996), chain H has 75 residues. The structure was solved to 2.8Å by x-ray diffraction and contains three helices and two disulphide bonds (29+64, 39+53, pdb file commences at residue 11).

2crd(Bontems *et al.*, 1991) is a protein with 37 residues. It has a helix and a sheet made up of two strands, anti-parallel. It was solved by NMR, contains 12 models and has three disulphides (7+28, 13+33, 17+35). The average structure was used to define the secondary structure. It contains a Disulphide β-cross motif (Harrison and Sternberg, 1996).

1ehs(Sukumar *et al.*, 1995) is deposited as a single NMR structure with two helices and two disulphide bonds (10+48, 21+36); its length is 48 residues.

1sis(Lomize *et al.*, 1991) is also an NMR model, with ten sets of coordinates. It has 36 residues and four disulphide bridges (2+19, 5+26, 16+31, 20+33). The structure has one helix and an anti-parallel sheet of three strands.

1ps2(Polshakov *et al.*, 1997) 60 residues long, this NMR model has one sheet and two strands forming a small anti-parallel sheet. There are three disulphides (7+33, 17+32, 27+44).

1vib(Barnham *et al.*, 1997) the NMR structure (20 models) has 55 residues and two helices, with four disulphide bonds (12+52, 16+48, 23+41, 26+37, residue 10 is hypothetical). The turn between the helices is sometimes classed as one turn of a helix, but in this case the DSSP classification was followed, of just two helices.

1erc(Mronga *et al.*, 1994) a 40 residue long NMR structure (20 conformers) with three helices and three disulphide bonds (3+19, 10+36, 15+28).

1hyp(Baud *et al.*, 1993) is an all alpha protein, solved to 1.8Å. It has four helices and four disulphide bonds (8+43, 14+28, 29+67, 45+77 pdb file commences at residue 6), with two cysteines which do not form cross-links. The protein is 75 residues long.

1kjs(Zhang *et al.*, 1997) an all alpha NMR solved protein structure with five helices and three disulphides (21+47, 22+54, 34+55). It has also has a non disulphide forming cysteine, it is 74 residues long.

Also examined was the NK lysin protein from Chapter 5, which inspired the work. In the next chapter the analysis is also used on a protein where only the fact that there

may be disulphide bonds present is known.

Another protein small enough to be used for disulphide matching (Nerve growth factor (NGF) (McDonald *et al.*, 1991)) was not considered because it is an all beta protein and using DRAGON with the correct secondary structure would have given the model too many correct restraints. The other possibility would be to predict the secondary structure and then perform a combinatorial analysis on the different arrangements of the strands.

7.3.3 Interpretation of results

In the figures, the pairs of disulphides which form in the 'real' structure are highlighted in bold red lines. If the DRAGON models are reasonably correct then the disulphide bonds most likely to form would be closer together in space. If this were consistent throughout all the models then the disulphides would be the lowest lines on the plot (i.e. – the smallest C_α:C_α cysteine distance). Due to the inherent variation in the DRAGON modelling, when based on such few external restraints, the sorted potential disulphide distances increase, some faster than others. Theoretically, consistently lower pair distances will more likely be the actual disulphides in the DRAGON models. Reasoning that the core elements are more uniformly modelled throughout, then the core disulphides should be more consistently placed in the model, in this case the models may have good disulphides. It is worth bearing in mind that the pairs are all sorted by distance and consequently do not correspond to one particular model.

Although, Figure 7.11 and Figure 7.12 show plots where each cystine pair correspond to one model, i.e. unsorted by distance, but sorted by DRAGON score.

It is not likely for disulphide bonds to form between cysteines within the same secondary structure element, i.e. a single helix or strand. Cysteines less than five residues apart are also highly unlikely to pair with each other (Muskal *et al.*, 1990). Taking these factors into account can help to eliminate some pairs from the analysis.

7.3.4 NK Lysin analysis

When examining the deposited highly refined NMR coordinates in 1NKL, all pairs of cysteines have conserved distance, see Figure 7.3. The models are fairly similar, in other words. The three pairs of disulphides are the smallest distance apart, excluding atom pairs 4+7 and 70+76 which are close in the amino acid sequence.

A similar analysis of DRAGON models is shown in Figure 7.4 where the disulphide restraints are imposed on NK lysin (as explained in Chapter 5). This gives a similar, albeit far less well conserved plot as that seen in Figure 7.3. This is the case for the three restrained pairs, all the others are still variable. The disulphide restraints impose much fewer conformations on the modelling.

Some of the more rigid restraints seem to prevent the fold increasing in accuracy, this is exemplified by Figure 7.4 and Figure 7.1 where the former should have a better secondary structure prediction and therefore better prediction of the disulphides.

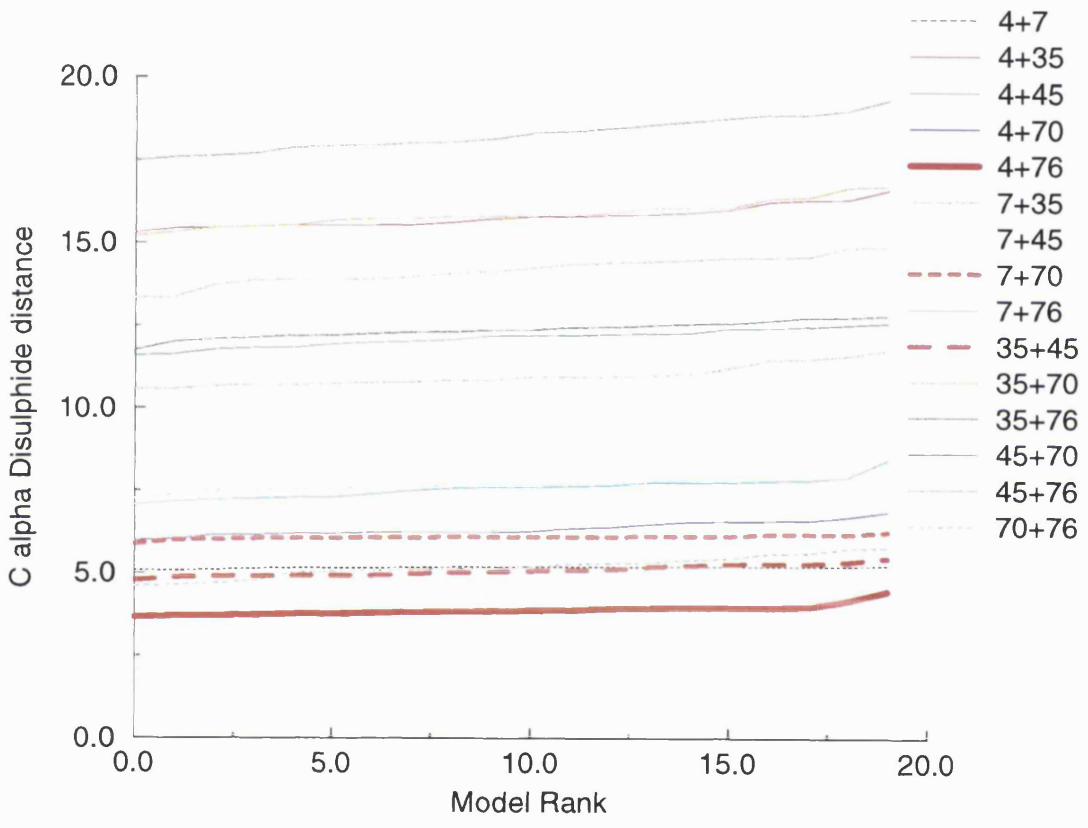


Figure 7.3: Disulphide analysis of the 20 NMR models from 1nkl. DRAGON models are analysed for this protein in the previous three figures.

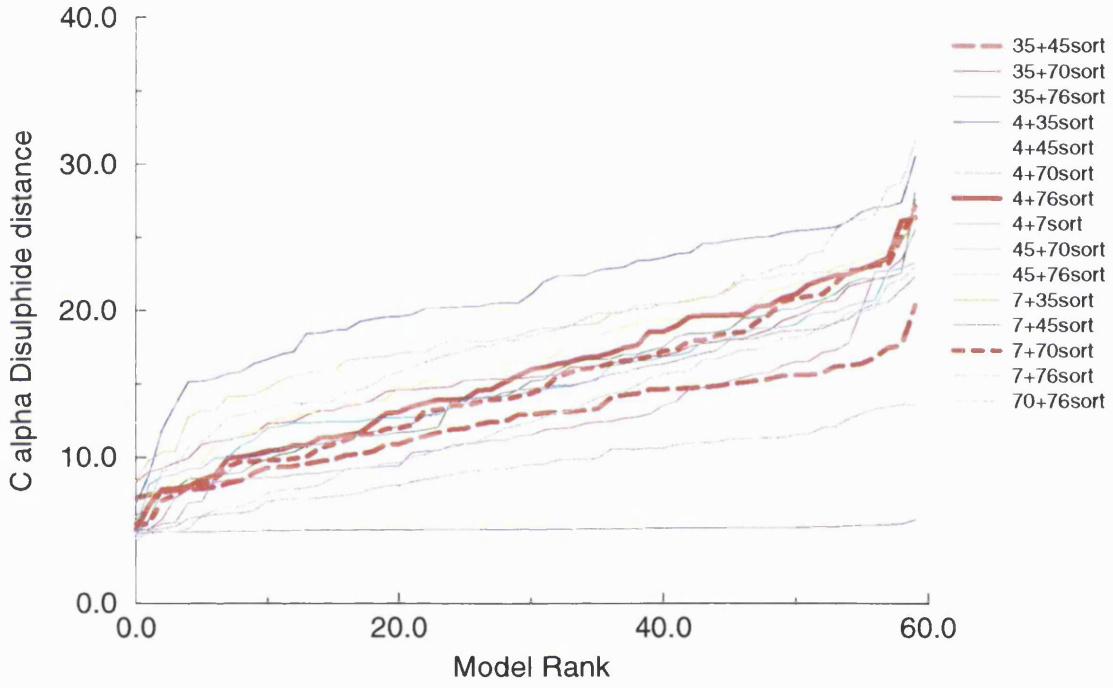


Figure 7.4: Disulphide analysis of the unconstrained DRAGON models. Models are of NK-lysin and in this case the secondary structure is fixed to the actual (from the NMR coordinates) with 100% stringency.

This is not the case, but is due to a combination of the protein, secondary structure prediction and DRAGON. The analysis in Figure 7.1 and more closely in Figure 7.2 had a four helix bundle prediction, whereas the correct secondary structure has one of these helices split into two. The real secondary structure, therefore, has more different possible outcomes. This may also be due to the fact that if the secondary structure assignments are fixed at high stringency, then the packing can be affected if the loops between secondary structure elements are too short.

7.3.5 Number of models

All the examples have been generated from 150 DRAGON models of each protein. In fact the number of models generated could have been less as Figure 7.5 shows for pairs 10+37. Between 50 and 500 models all produce roughly similar results for the disulphide pairs, as shown for 10+37 in the figure. Runs were performed to generate 50, 100, 250 and 500 models. It appears that just 50 models are still a reasonable representation of the combinatorics of the modelling. 150 models were chosen as a representative number of models. More than 150 would have produced redundant models. Due to the speed with which models can be generated, 150 was kept as a representative number of models, although reducing the number of models to 50 or 100 would not appear to have a detrimental effect.

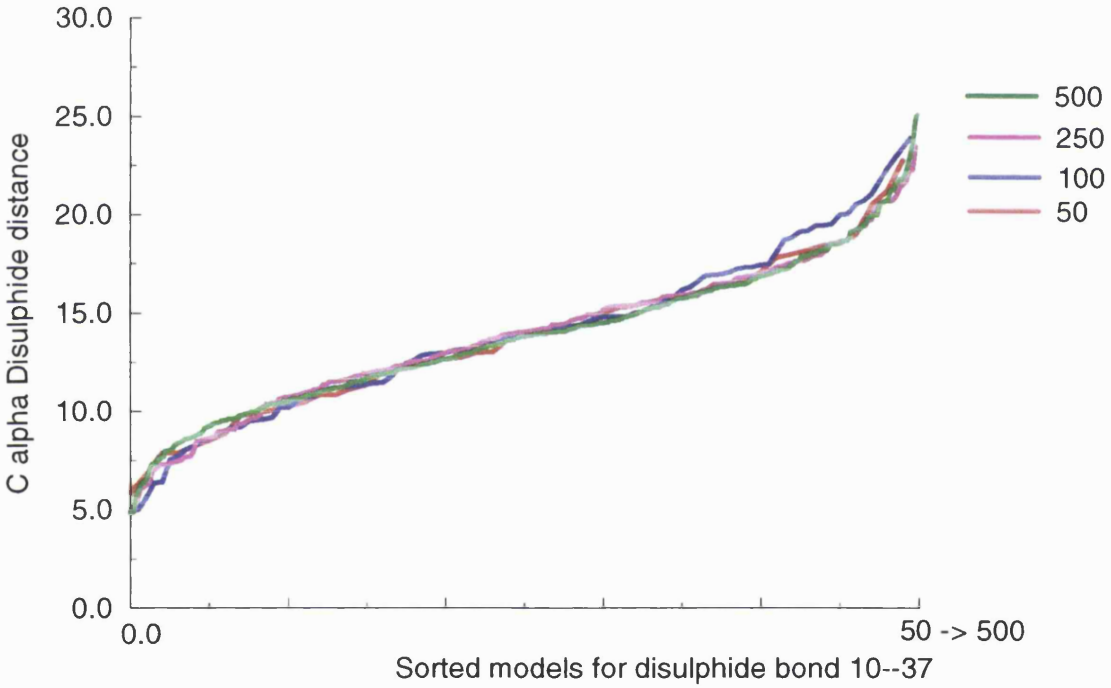


Figure 7.5: Disulphide analysis of compareRUN.pdb.

Example 1: 1occH The results for 1occH are all fairly similar. This is partly due to the fact that there are only two disulphides and therefore only six possible pairs, or 3 combinations. Residues 19+29 occur moderately close in sequence and are in fact in a helix. They are uniformly close, as indicated in Figure 7.6 by an almost straight line. Apparently the best pair is 29+54, followed by 29+43. These could not be the correct pairs as residue 29 occurs in both (an obvious criterion for an incorrect selection). The next best pair is 19+54. Of these three pairs, the only two which are mutually exclusive are 29+43 and 19+54. These are in fact the disulphide bonding residues.

This is easy to do with hindsight, but looking at the models as a whole over all 150 it would seem better to choose 29+54, which would leave only 19+43. This pair is the worst predicted as the closest distance between this pair only gets the C_α atoms about 14Å apart (the PDB structure has these at 15.7Å apart). The next best overall line is either 29+43 or 43+54 depending whether you consider the earlier or latter part of the curve.

In this example, as with any other protein with only four cysteines, the fact that once you choose one pair, by default you have the other, so that 4-cysteine proteins are not optimal as test cases for this method.

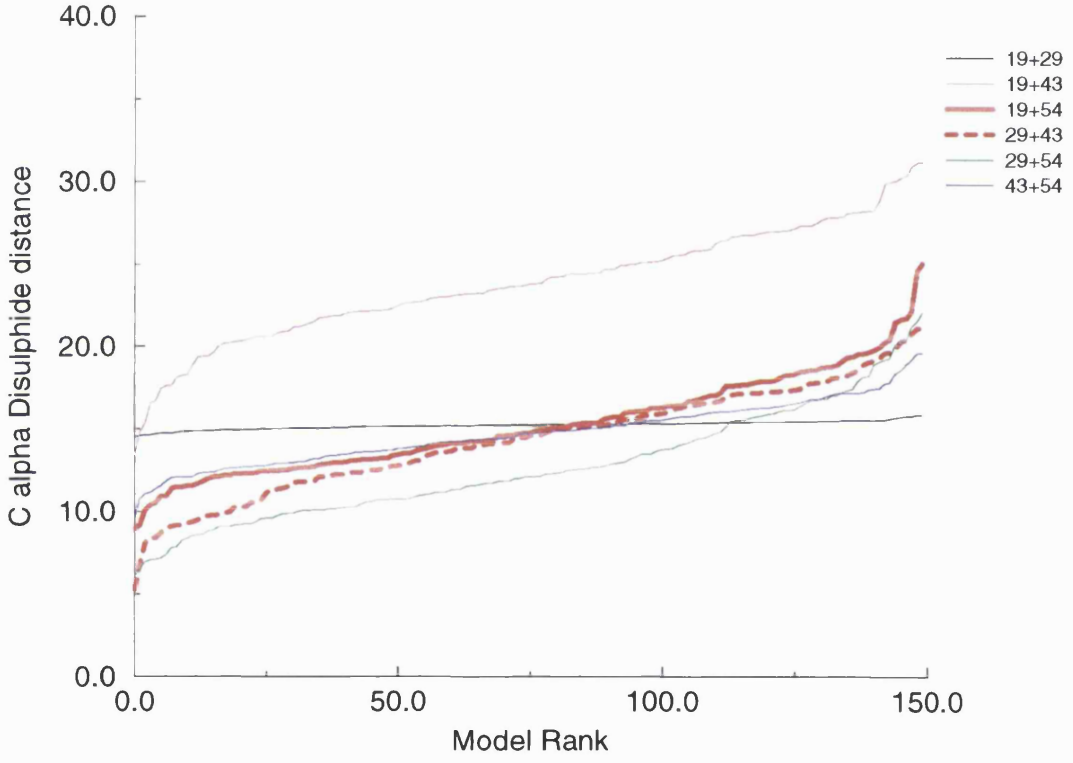


Figure 7.6: Disulphide analysis of 1occH.

Example 2: 2crd It is apparent when first looking at all possible pairs of cysteines in Figure 7.7 that some are very close in sequence and therefore highly unlikely to bond with each other – pairs 33+35 and 13+17 in particular – and a case could be made to also exclude 7+13 and 28+33. By excluding these two pairs, this would reduce the number of possible combinations from 15 to 10 (in a case such as this with 6 cysteines). If all four pairs were excluded then only eight combinations could occur. One factor which can also be taken into consideration is whether or not the cysteines occur in the same predicted secondary structure elements. For example, if pair 7+13 (which are five residues apart) both occur in a single strand, then in this model they could not bond. In fact this is not the case, so perhaps pairs 7+13 could pair if residues 8 to 12 formed a loop. It is unlikely from this case that the correct pairs would be predicted. The pairs which would be predicted (incorrectly) are 7+33, 13+35 and 17+28. Or perhaps 7+35, 13+33 and 17+28.

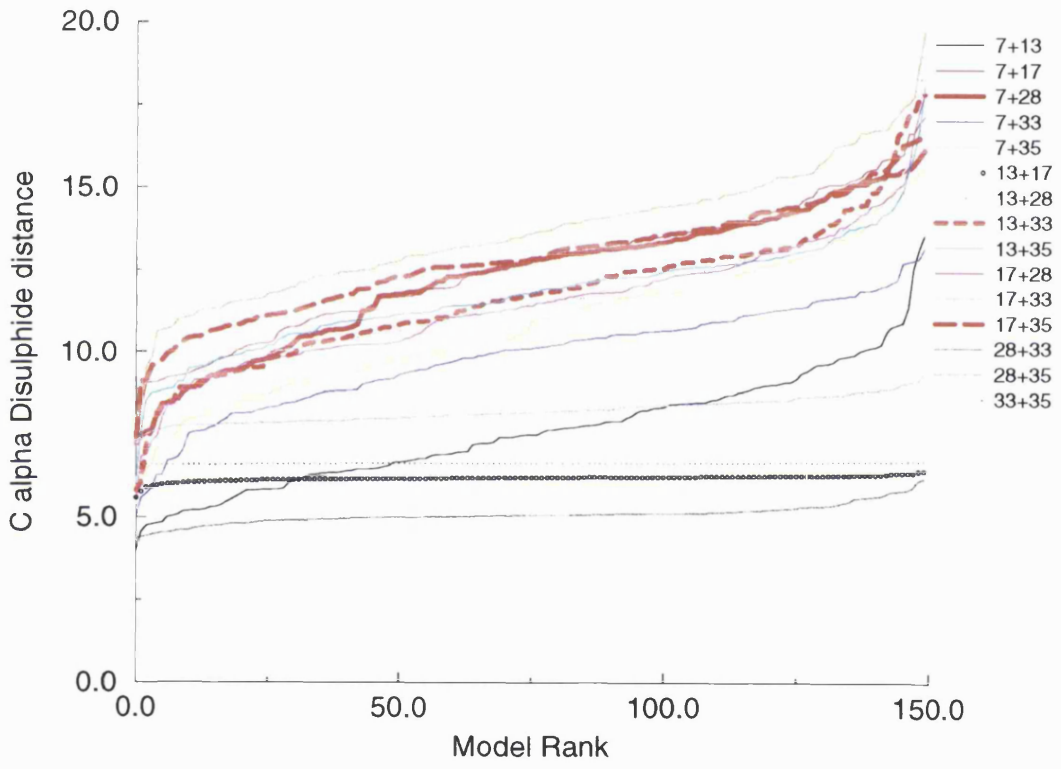


Figure 7.7: Disulphide analysis of 2crd.

Example 3: 1ehs The best pair from the graph in Figure 7.8 is 36+48, which means that 10+21 must be the other, by elimination. This latter pair, however, seem to be limited to a very constrained 15.5Å. This pair should be excluded, as it is in a helix, explaining its constant distance. The only other pair combinations are 10+36 with 21+48, both quite poor compared to 10+48 and 21+36. This latter combination is the correct answer, as highlighted in bold red curves.

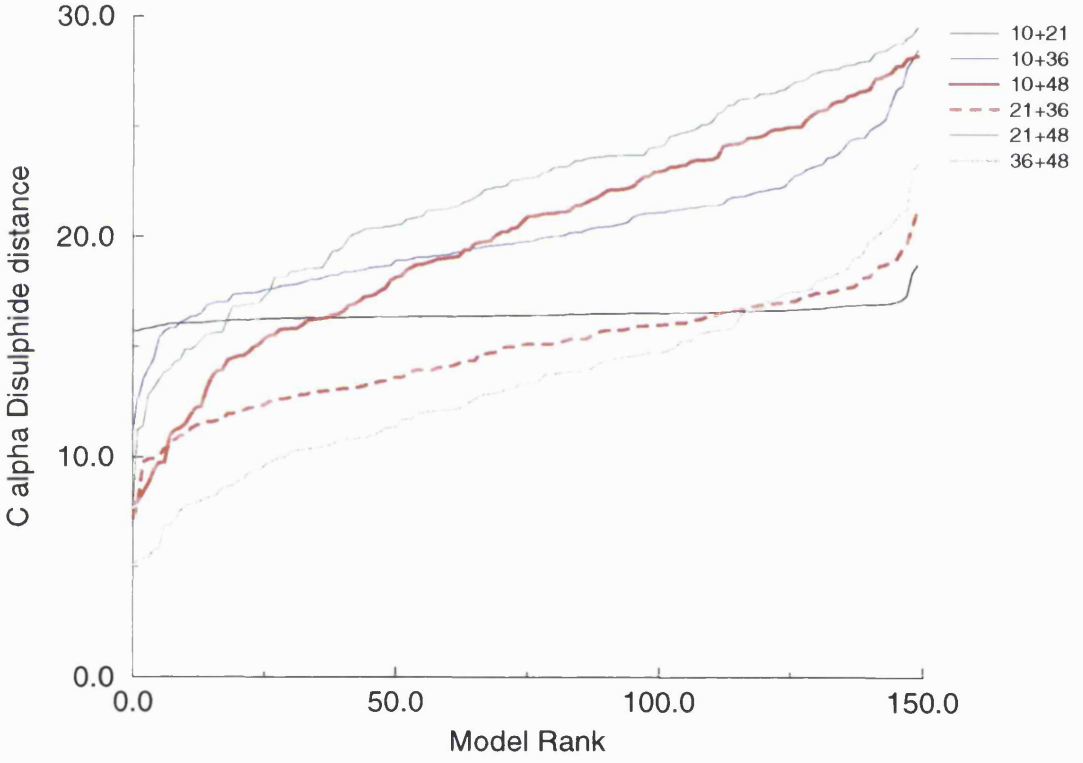


Figure 7.8: Disulphide analysis of 1ehs.

Example 4: 1sis Looking at this graph (Figure 7.9) the first impression is that there is no way of predicting the correct pairs, which is probably the case. Pairs 16+19, 16+20, 19+20, 2+5 and 31+33 can all be excluded as being too close.

Because there are so many cysteines in such close proximity, even looking at the correct (α-carbon) structure would not give a correct indication of which pairs pair. By eliminating the above mentioned pairs, in the NMR structure the 10 closest pairs are: 26+31, 5+31, 20+33, 2+33, 16+31, 2+31, 5+26, 2+19, 5+16 and 2+20. The disulphide binding pairs are ranked 3rd, 5th, 7th and 8th. Although in this instance the method is unsuitable for telling which pairs form disulphides, it can of course tell you those which are not.

In the structure three of the cystines (3+31, 19+24, 20+33) occur between the helix and the sheet. The other (5+26) occurs between two strands in the sheet.

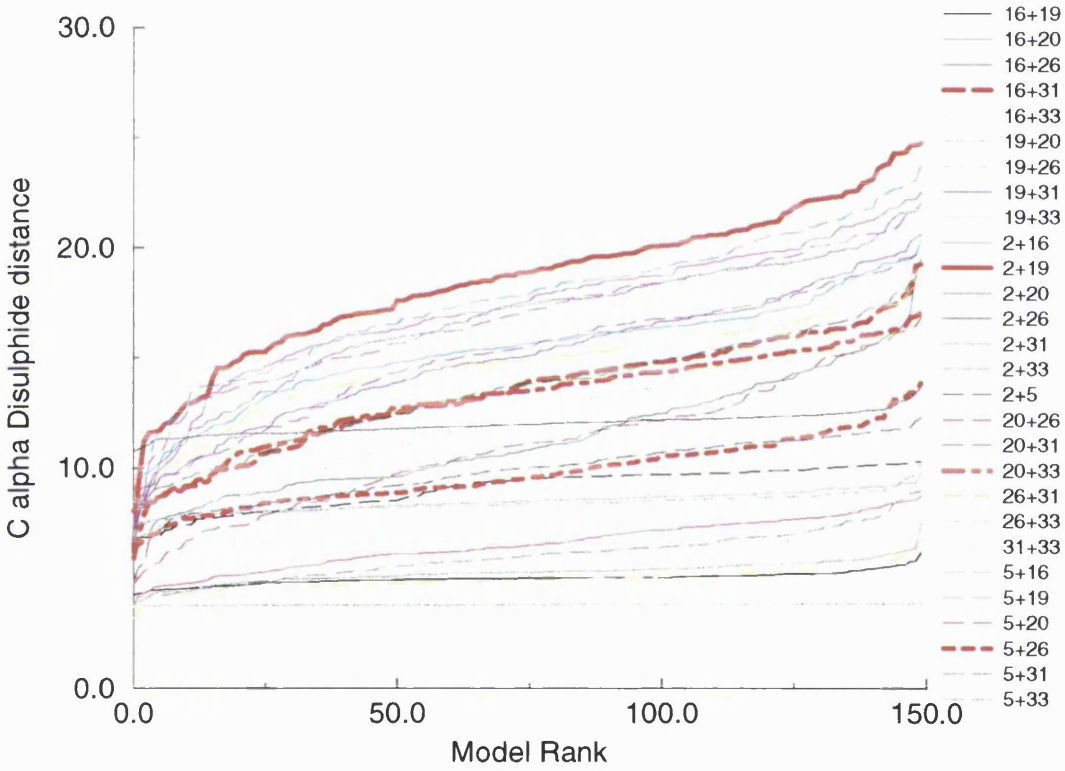


Figure 7.9: Disulphide analysis of 1sis.

Example 5: 1ps2 The 1ps2 protein results are shown in Figure 7.10. A pair of cysteines can be discarded from the analysis because they occur consecutively in the sequence (32+33). This leaves 12 possible combinations of pairs. By the numerical analysis described in the next section, set 11 is the correct answer and it was not predicted correctly by the method. Set 13 was the best prediction according to the DRAGON models, see Table 7.1 for a description of the sets.

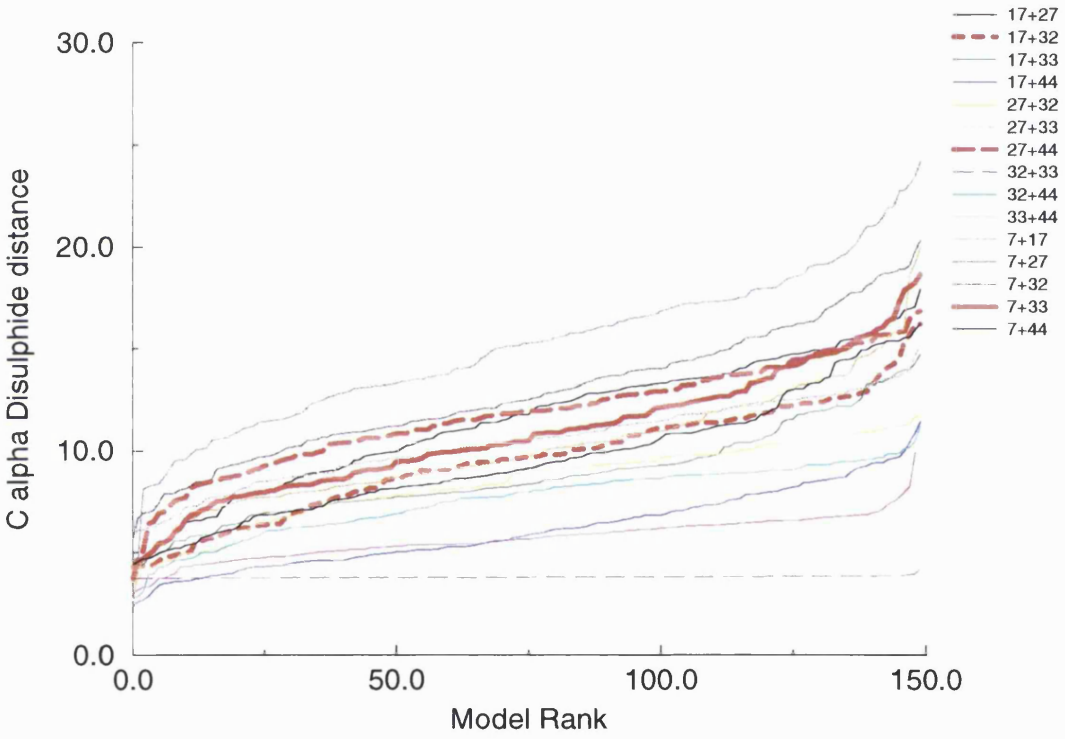


Figure 7.10: Disulphide analysis of 1ps2.

Examples 6-9 More examples of the plots can be found in the appendix at the end of this chapter. Also in the appendix are the multiple alignments and secondary structure assignment files used in the analysis. NB: where a sequence has no homology to any other proteins then only that single sequence occurs in the MSF.

Ranking the models

Figure 7.11 and Figure 7.12 show the ranked models according to DRAGON's criteria for model quality, by assessing contacts, bonds, accessibility and secondary structure. As Table 5.3 shows the rank does not perform as well as may be expected. A better scenario would be to build full models and then energy minimise them and take the lowest energy structure as the best and rank the models accordingly. It is more likely that the models with lower restraint violations will have the correct disulphides closer together.

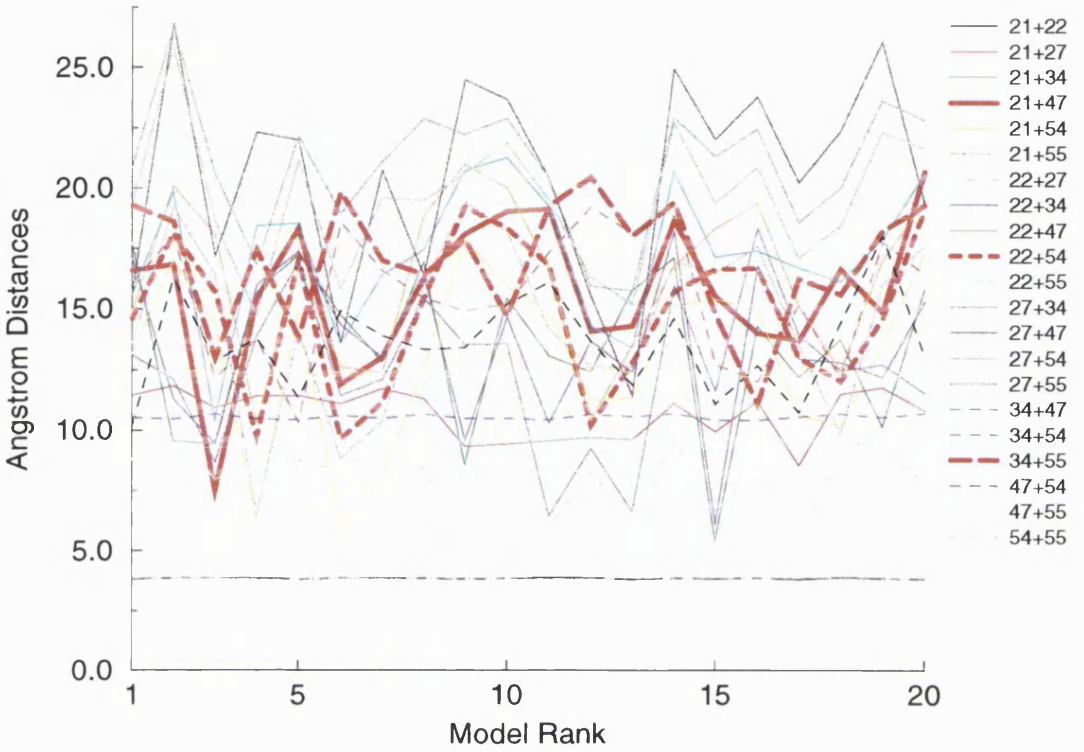


Figure 7.11: Disulphide analysis of the top 20 1kjs models.

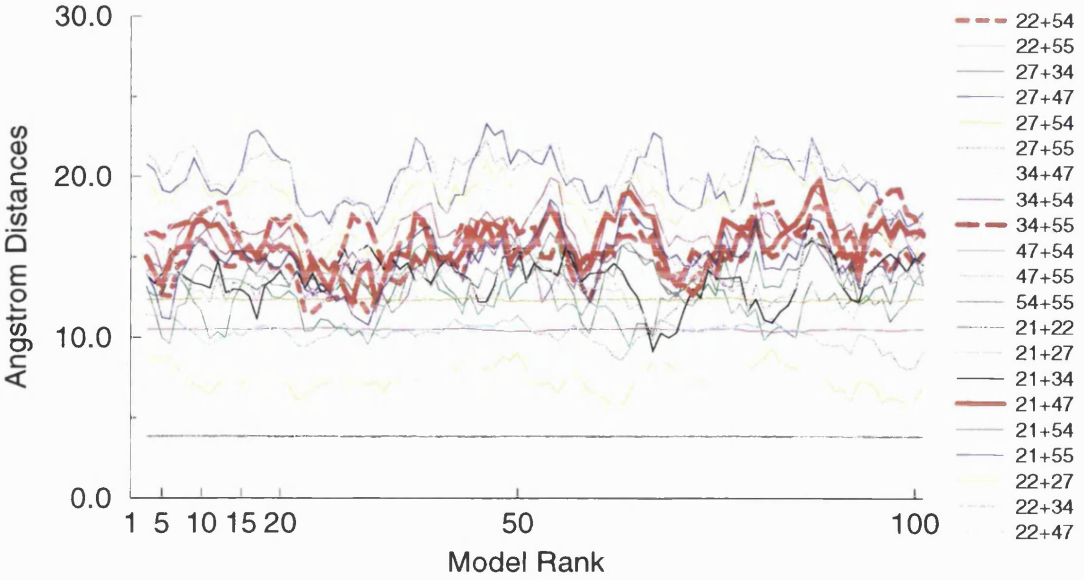


Figure 7.12: Disulphide analysis of the top 100 1kjs models.

7.3.6 Summing disulphides among models

For a protein which is thought to have, say, three disulphide bonds then there are 15 possible pairs. These pairs are dependent on each other. For example, if the first two cysteines bonded (lets call them 1+2), then the only three combinations of the others: 3+4 with 5+6, 3+5 with 4+6, or 3+6 with 4+5. So for all possible pairs of three there are 15 different sets of disulphides which can form (see Table 7.1). All the method has to do is decide which pairs are most likely to form. This makes the assuming that the closest pairs of disulphides will form. By calculating all pairs and finding the three which have the smallest sum of inter cysteine distances, should give the best disulphide prediction.

Pairs	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1+2	•	•	•												
1+3				•	•	•									
1+4							•	•	•						
1+5										•	•	•			
1+6													•	•	•
2+3							•			•			•		
2+4				•							•			•	
2+5					•			•							•
2+6						•			•			•			
3+4	•											•			•
3+5		•							•					•	
3+6			•					•			•				
4+5			•			•							•		
4+6		•			•					•					
5+6	•			•			•								

Table 7.1: Possible disulphide pairs for six cysteines forming three disulphide bonds.

It is a simple step to produce an algorithm to sum all possible sets of three. If there is an obvious trend among the models, where a particular set of disulphides have a low total distance, then those may be the most likely to form for that given structure. With a selection of models from DRAGON all based on the same structure, a trend could be observed of which pairs are most likely to form. The situation becomes somewhat more complicated when there are an odd number of disulphides. There is usually only one cysteine which does not form a bond; three free cysteines are not observed (Thornton, 1981). Often the single free cysteine will be involved in some functional aspect of the protein; so it might be identified in a model (or set of models) if it is positioned consistently at the protein surface.

For the NK-lysin model the results of this analysis are shown in Table 7.2. Pairs 1+2, i.e. residues 4+7 are very close in sequence and occur in the same helix. Eliminating

these pairs effectively eliminates sets 1 to 3. The next best set is set 15, which corresponds with pairs 1+6, 2+5 and 3+4 from Table 7.1. Taking the six known disulphide forming cysteines in NK-lysin then these pairs are the correct bonding pairs, i.e. 4+76, 7+70 and 35+45.

Set1		Set6		Set11	
20.867	model_38.pdb	33.385	model_17.pdb	26.709	model_8.pdb
21.256	model_29.pdb	34.475	model_35.pdb	30.046	model_12.pdb
21.858	model_20.pdb	36.926	model_8.pdb	32.990	model_36.pdb
21.937	model_19.pdb	37.255	model_9.pdb	35.891	model_30.pdb
22.423	model_4.pdb	40.038	model_24.pdb	37.943	model_19.pdb
Set2		Set7		Set12	
18.428	model_10.pdb	29.474	model_9.pdb	24.686	model_1.pdb
21.428	model_13.pdb	30.715	model_59.pdb	25.754	model_8.pdb
22.602	model_50.pdb	30.882	model_35.pdb	26.676	model_2.pdb
23.735	model_44.pdb	32.895	model_57.pdb	29.575	model_27.pdb
25.447	model_47.pdb	32.917	model_38.pdb	31.674	model_43.pdb
Set3		Set8		Set13	
22.100	model_30.pdb	29.189	model_12.pdb	35.263	model_17.pdb
23.135	model_13.pdb	32.016	model_8.pdb	37.039	model_8.pdb
24.506	model_44.pdb	32.073	model_36.pdb	37.896	model_28.pdb
25.676	model_24.pdb	36.579	model_24.pdb	37.944	model_35.pdb
25.879	model_32.pdb	36.915	model_19.pdb	38.706	model_51.pdb
Set4		Set9		Set14	
24.305	model_9.pdb	29.204	model_21.pdb	29.475	model_21.pdb
29.000	model_59.pdb	33.122	model_8.pdb	32.039	model_8.pdb
29.937	model_35.pdb	33.451	model_47.pdb	32.640	model_47.pdb
30.331	model_57.pdb	34.255	model_15.pdb	34.563	model_51.pdb
31.788	model_54.pdb	35.744	model_51.pdb	35.127	model_15.pdb
Set5		Set10		Set15	
35.349	model_8.pdb	31.238	model_8.pdb	22.188	model_2.pdb
36.600	model_17.pdb	36.424	model_5.pdb	26.862	model_1.pdb
36.830	model_9.pdb	36.977	model_1.pdb	29.978	model_8.pdb
37.024	model_5.pdb	37.194	model_36.pdb	31.020	model_33.pdb
37.039	model_52.pdb	38.324	model_52.pdb	31.205	model_27.pdb

Table 7.2: Lowest scoring total distance for each possible disulphide set of pairs for the six cysteines forming three disulphide bonds. See Table 7.1 for description of the sets.

A further analysis like this of example 2, 2crd backed up the results, with the correct disulphide bonding pattern having a high overall total in even the best models. This proteins disulphide bonding would have been incorrectly predicted.

The N-terminal domain of GPCR also has three disulphide bonds and using some of the methods so far described in this thesis a model has been constructed. This is explained in the following chapter.

7.4 Discussion

Some proteins have more cysteines than just twice their number of disulphide bonds. This can make prediction particularly difficult as all possible pairs have to be considered. It would also be difficult to know that not all the cysteines actually form disulphide links, although some which may be functional might be identified and eliminated from the analysis. So for example a protein with five cysteines may have only two disulphides, but a higher number of combinations of possible pairs.

The analyses here show that in all the cases tested, the disulphides are never the consistently furthest distance pairs of cysteines in the models.

With only the minimal information included in the modelling (a multiple sequence alignment and secondary structure – see Appendix), particularly when the secondary structure is not 100% correct (unless it is assigned the `DSSP` prediction), `DRAGON` does not produce many highly accurate models. The analysis would benefit by a pre-screening to try and only analyse correct or nearly correct structures. This leaves the problem of how to identify the best models for an analysis. At the moment the best thing to do, without building full models, is to use the rank program which checks on how well the restraints have been satisfied in the modelling process. Models which have disulphides which are far apart are filtered by the numerical analysis.

The C_α distance calculation was more beneficial as the `DRAGON` folds would not always have the pseudo-side chain atoms (the C_β's) pointing in a direction that would form a

disulphide, although in theory the immediate conformation could be changed latterly to constrain the side chain atoms into a more likely bond forming distance.

Further improvements could include the analysis of close C_α's and then check to see if the C_β's were pointing towards each other. Ambiguous distance restraints have aided in the assignment of disulphide connectivities from NOE data (Nilges, 1995). So perhaps a similar clustering algorithm as that used by Nilges, might be useful when trying to predict the disulphide pairs from DRAGON models, although much would depend on the accuracy of the models produced.

It would be interesting to investigate the gradient of the lines and perhaps to concentrate only on the top 10 or 20 scores. The relationship of the different gradients of the curve (which in many cases is sigmoidal) between ranked models 1 to 20, 21 to 120 and 121 to 150 may show some trend which might be more useful in prediction. It may be that some parts of the curves have better predictive powers than other parts.

Obviously the major limiting factor is the accuracy of the modelling. In this case DRAGON has been given virtually correct assignments of secondary structure, based on DSSP, with very rigid constraints on the length of the secondary structure elements. Other than this and an elementary multiple sequence alignment, the modelling has been left without any additional restraints. These results would be consistent with a good secondary structure prediction, as in the NK lysin case. Potential developments in DRAGON and our understanding of *ab initio* prediction may well improve the ideas described here.

None of the models were examined for "compactness" as this was designed to be an automated procedure. Clearer results might be obtained if some filtering methods were employed to remove models which were badly packed.

More work should be carried out to incorporate the measurement of C_β distances, so that a predictive score could take into account cysteine pairs which may have suitably close C_α atoms but the side chains would be pointing in the wrong direction. By applying the predicted disulphides to the models and iterating the DRAGON modelling process the predictions may be improved. Using a method such as this would show more clearly if the predicted disulphides could form protein like models.

With two pairs of possible disulphides and four cysteines a correct prediction could be made from these graphs, assuming that the secondary structure assignment is approximately correct. With three pairs of disulphides the problem becomes far more difficult, but the method seems to hold up for some cases moderately well. However, once the disulphides increase to three or more pairs, with more than six cysteines then the likelihood of making correct predictions is greatly reduced.

It is clear that the prediction of disulphide bonds is not easy, but perhaps this semi-empiracle method will aid the expertise required by a successful molecular modeller.

7.5 Appendix: Other examples

Below are some more examples of the plots from the disulphide analysis for some small disulphide rich proteins. More about these proteins can be found in section 7.3.2.

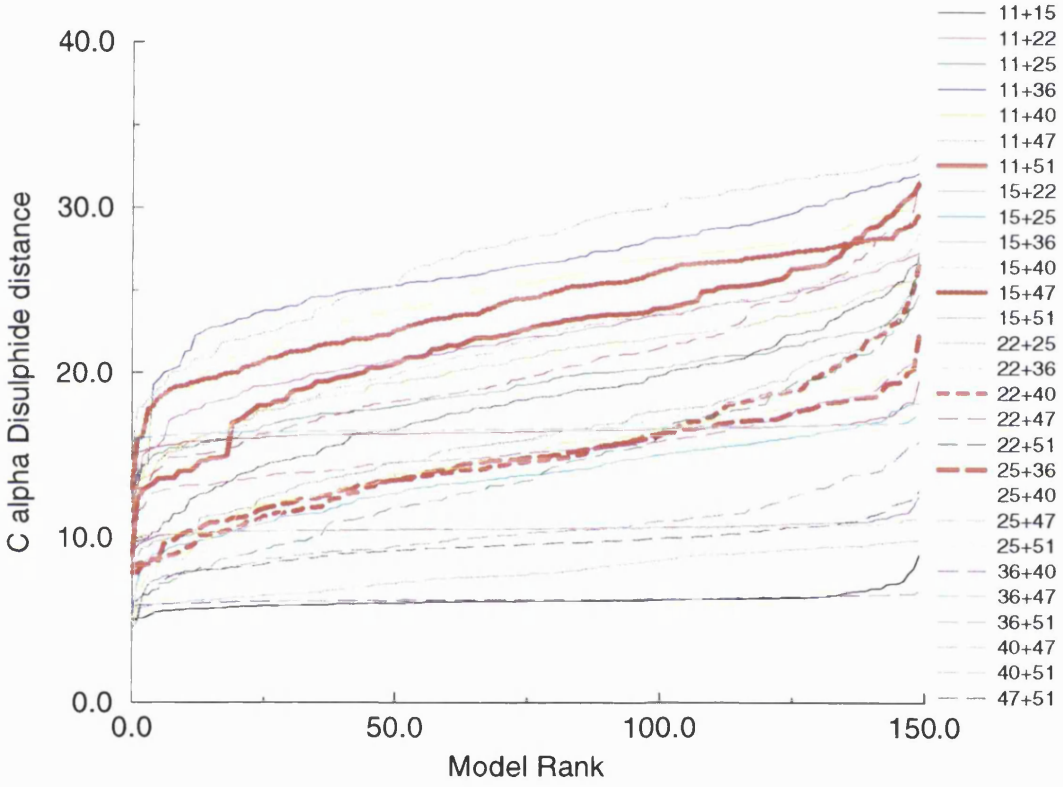


Figure 7.13: Disulphide analysis of 1vib.

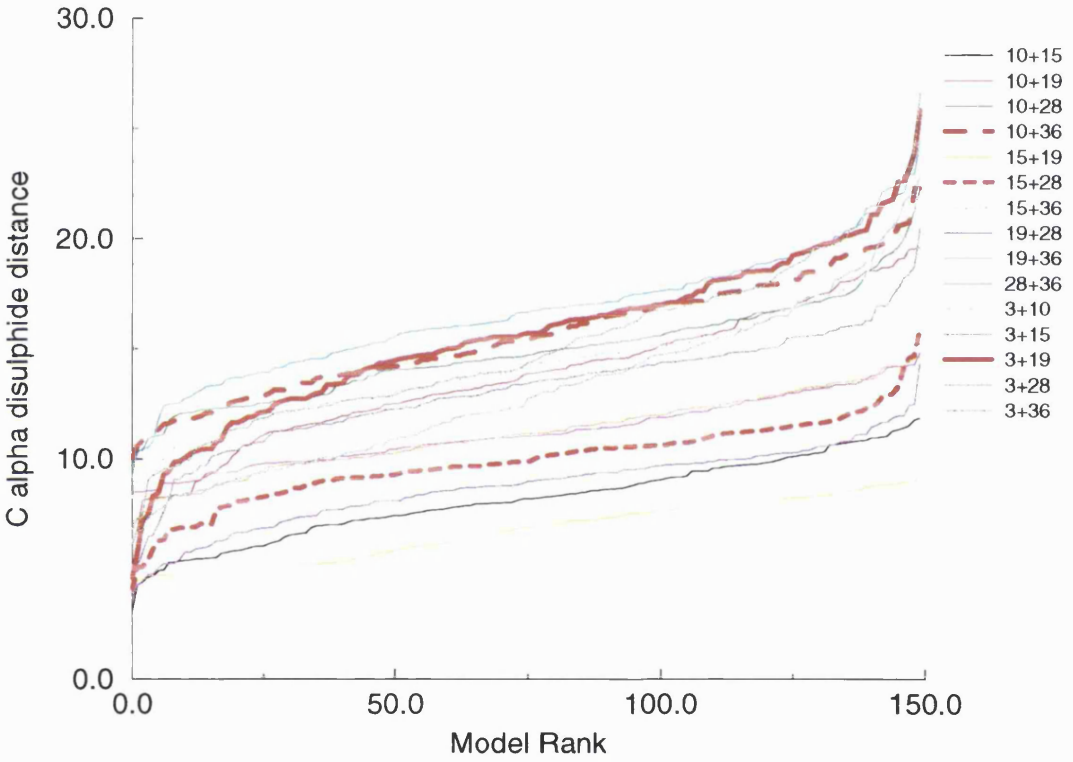


Figure 7.14: Disulphide analysis of 1erc.

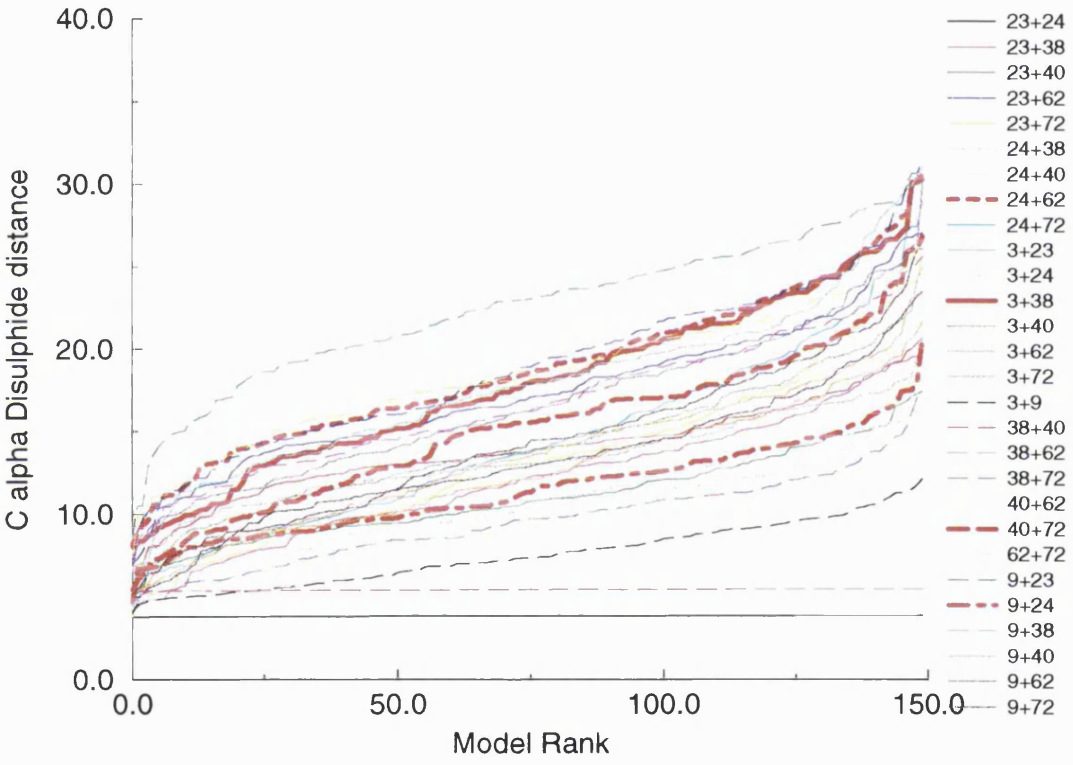


Figure 7.15: Disulphide analysis of 1hyp.

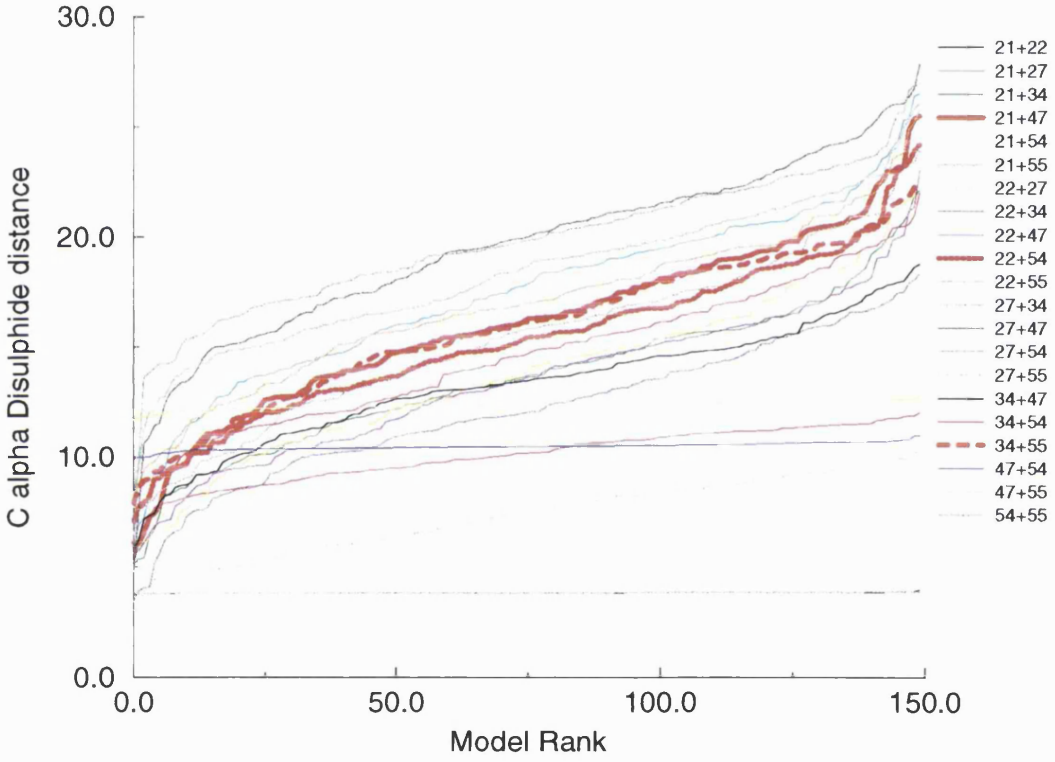


Figure 7.16: Disulphide analysis of 1kjs.

7.6 Appendix: SSdist Files

Multiple alignments and secondary structure assignments
1ehs

```
MSF:   48  Type: P 11-Oct-98  17:14:2  Check: 1384  ..

 Name: 1ehs.seq        Len:    48  Check: 5692  Weight:  1.00

//

              1                                                  48
1ehs.seq      STQSNKKDLC EHYRQIAKES CKKGFLGVRD GTAGACFGAQ IMVAAKGC


# Secstr assignment of 1ehs
# based on dssp
ALPHA 10 22 1.0
ALPHA 40 45 1.0

--------------------------------------


MSF:   40  Type: P 11-Oct-98  17:13:5  Check: 7134  ..

 Name: mer1_eupra      Len:    40  Check: 3567  Weight:  1.00

//

              1                                       40
mer1_eupra    DACEQAAIQC VESACESLCT EGEDRTGCYM YIYSNCPPYV


# Secstr assignment of 1erc
# based on dssp
ALPHA 3 8 1.0
ALPHA 12 18 1.0
ALPHA 24 35 1.0
```

1erp

```
MSF:   38  Type: P 1-Oct-98
21:32:1  Check: 1198  ..


 Name: merx_eupra      Len:    38  Check: 5599  Weight:  1.00

//


              1                                       38
merx_eupra    DLCEQSALQC NEQGCHNFCS PEDKPGCLGM VWNPELCP


# Secstr assignment of 1erp
# based on dssp
ALPHA 2 8 1.0
ALPHA 12 17 1.0
ALPHA 23 31 1.0
```

1hyp

```
MSF:   75  Type: P 11-Oct-98  17:13:5  Check: 2330  ..

 Name: hpse_soybn      Len:    75  Check: 6612  Weight:  1.00
 Name: ccdp_maize      Len:    75  Check: 6698  Weight:  1.00
 Name: 14kd_dauca      Len:    75  Check: 5715  Weight:  1.00
 Name: prf1_lyces      Len:    75  Check: 6693  Weight:  1.00

//

              1                                                  50
hpse_soybn    PSCPDLSICL NILGGSLGTV DDCCALIGGL GDIEAIVCLC IQLRALGILN
ccdp_maize    LKLKVCAKVL GLVKVGLPQY EQCCPLLEGL VDLDAALCLC TAIKANVllN
14kd_dauca    GVCADVLNLV HNVVIGSPPT LPCCSLLEGL VNLEAAVCLC TAIKANIllN
prf1_lyces    GACVDVLGGL IHIGIGGSAK QTCCPLLGGL VDLDAAICLC TTIRLKLliI

              51                        75
hpse_soybn    LNRNLQLILN SCGRSYPSNA TCPRT
ccdp_maize    VPLSLNFILN NCGRICPEDF TCPN.
14kd_dauca    LPIALSLVLN NCGKQVPNGF ECT..
prf1_lyces    LPIALQVLID DCGKYPPKDF KCPST


# Secstr assignment of 1hyp
# based on dssp
ALPHA 6 14 1.0
ALPHA 20 28 1.0
ALPHA 32 46 1.0
ALPHA 51 61 1.0
```

1kjs

```
MSF:   74  Type: P 11-Oct-98  17:13:3  Check: 5582  ..

 Name: co5_human       Len:    74  Check: 5017  Weight:  1.00
 Name: co5a_bovin      Len:    74  Check:  933  Weight:  1.00
 Name: co5a_pig        Len:    74  Check: 3200  Weight:  1.00
 Name: co5_mouse       Len:    74  Check: 5801  Weight:  1.00
 Name: co5a_rat        Len:    74  Check: 5354  Weight:  1.00
 Name: co4a_rat        Len:    74  Check: 4553  Weight:  1.00
 Name: co4_mouse       Len:    74  Check: 4658  Weight:  1.00
 Name: co4_bovin       Len:    74  Check: 4762  Weight:  1.00
 Name: co4_human       Len:    74  Check: 5126  Weight:  1.00
 Name: co3_cavpo       Len:    74  Check: 5987  Weight:  1.00
 Name: co3_oncmy       Len:    74  Check: 5784  Weight:  1.00
 Name: co3_eptbu       Len:    74  Check: 2716  Weight:  1.00
 Name: co3_human       Len:    74  Check: 4840  Weight:  1.00
 Name: co3_najna       Len:    74  Check: 1841  Weight:  1.00

//

               1                                                  50
co5_human      MLQKKIEEIA AKYKHSVVKK CCYDGACVNN DETCEQRAAR ISLGPRCIKA
co5a_bovin     MLKKKIEEEA AKYRNAWVKK CCYDGAHRND DETCEERAAR IAIGPECIKA
co5a_pig       MLQKKIEEEA AKYKYAMLKK CCYDGAYRND DETCEERAAR IKIGPKCVKA
co5_mouse      LLRQKIEEQA AKYKHSVPKK CCYDGARVNF YETCEERVAR VTIGPLCIRA
co5a_rat       LLHQKVEEQA AKYKHRVPKK CCYDGARENK YETCEQRVAR VTIGPHCIRA
co4a_rat       NFQKAISEKL GQYSSPDTKR CCQDGMtlPM ARTCEQRAAR V.PQPACREP
co4_mouse      NFQKAVSEKL GQYSSPDAKR CCQDGMtlPM KRTCEQRAAR V.PQQACREP
co4_bovin      NFQKAIHEKL GQYTSPVAKR CCQDGLtlPM ARTCEQRAAR V.QQPACREP
co4_human      NFQKAINEKL GQYASPTAKR CCQDGVtlPM MRSCEQRAAR V.QQPDCREP
co3_cavpo      QLMERRMDKA GKYKSKELRR CCEDGMREnm QFSCQRRARY VSLGEACVKA
co3_oncmy      TISDVITSMA SKYHG.LAKE CCVDGMRDNt gYTCDRRAQY ISDGDVCVQA
co3_eptbu      ELVLEIAIEK ASTYPAELRK CCRDAAIESP llSCEERTKH IheGEGCQET
co3_human      QLTEKRMDKV GKYPKE.LRK CCEDGMRENP mfSCQRRTRF ISLGEACKKV
co3_najna      LLLDSKASKA AQFQDQGLRK CCEDGMHEnm GYTCEKRAKY IQEGDACKAA

               51                    74
co5_human      FTECCVVASQ LRANISHKDM QLGR
co5a_bovin     FKSCCAIASQ FRADEHHKNM QLGR
co5a_pig       FKDCCYIANQ VRAEQSHKNI QLGR
co5_mouse      FNECCTIANK IRKESPHKPV QLGR
co5a_rat       FKECCTI.DP IRKNQSHKGM LLGR
co4a_rat       FLSCCKFAED LRRNQTRSQA GLAR
co4_mouse      FLSCCKFAED LRRNQTRSQA HLAR
co4_bovin      FLSCCQFAES LRKKARTRGQ VGLA
co4_human      FLSCCQFAES LRKKSRDKGQ AGLQ
co3_cavpo      FLDCCTYMAQ LRQQHRREQN LGLA
co3_oncmy      FLVCCTEMAS KKIESKQDAL LLSR
co3_eptbu      FLECCkvEEE LLIAMEEEDE DLGR
co3_human      FLDCCNYITE LRRQHARASH LGLA
co3_najna      FLECCHYIKG IRDENQRESE LFLA
```

```
# Secstr assignment of 1kjs
# based on dssp
ALPHA 5 11 1.0
ALPHA 16 26 1.0
ALPHA 34 38 1.0
ALPHA 45 62 1.0
ALPHA 68 71 1.0
```

1occ

```
MSF:   75  Type: P 11-Oct-98  17:27:1  Check: 9794  ..

 Name: coxg_bovin      Len:    75  Check:  963  Weight:  1.00
 Name: coxg_mouse      Len:    75  Check: 1403  Weight:  1.00
 Name: coxg_human      Len:    75  Check: 1604  Weight:  1.00
 Name: coxg_yeast      Len:    75  Check: 4861  Weight:  1.00

//

              1                                                  50
coxg_bovin    YQTAPFDSRF PNQNQTRNCW QNYLDFHRCE KAMTAKGGDV SVCEWYRRVY
coxg_mouse    YKTAPFDSRF PNQNQTKNCW QNYLDFHRCE KAMTAKGGDV SVCEWYRRVY
coxg_human    YKTAPFDSRF PNQNQTRNCW QNYLDFHRCQ KAMTAKGGDI SVCEWYQRVY
coxg_yeast    LHTVGFDARF PQQNQTKHCW QSYVDYHKCV N...MKGEDF APCKVFWKTY

              51                       75
coxg_bovin    KSLCPISWVS TWDDRRAEGT FPGKI
coxg_mouse    KSLCPVSWVS AWDDRIAEGT FPGKI
coxg_human    QSLCPTSWVT DWDEQRAEGT FPGKI
coxg_yeast    NALCPLDWIE KWDDQREKGI FAGDI


# Secstr assignment of 1occ
# based on dssp
ALPHA 15 36 1.0
ALPHA 44 53 1.0
ALPHA 56 69 1.0
```

1ps2

```
MSF of: mul.aln from:    1 to:  80
mul.aln  MSF:  80  Type: P 12-Sep-96  11:52:4  Check: 5859  ..
Name: ITF_HUMAN:42    Len:    80  Check:    25  Weight:  1.00
Name: ITF_RAT:43      Len:    80  Check:    25  Weight:  1.00
Name: MUA1_XENLA:33   Len:    80  Check:    25  Weight:  1.00
Name: MUA1_XENLA:84   Len:    80  Check:    25  Weight:  1.00
Name: MUA1_XENLA:36   Len:    80  Check:    25  Weight:  1.00
Name: MUC1_XENLA:31   Len:    80  Check:    25  Weight:  1.00
Name: MUC1_XENLA:36   Len:    80  Check:    25  Weight:  1.00
Name: MUC1_XENLA:53   Len:    80  Check:    25  Weight:  1.00
Name: MUC1_XENLA:58   Len:    80  Check:    25  Weight:  1.00
Name: MUC1_XENLA:63   Len:    80  Check:    25  Weight:  1.00
Name: PS2_HUMAN:41    Len:    80  Check:    25  Weight:  1.00
Name: PS2_MOUSE:44    Len:    80  Check:    25  Weight:  1.00
Name: SP_HUMAN:43     Len:    80  Check:    25  Weight:  1.00
Name: SP_HUMAN:92     Len:    80  Check:    25  Weight:  1.00
Name: SP_MOUSE:41     Len:    80  Check:    25  Weight:  1.00
Name: SP_MOUSE:90     Len:    80  Check:    25  Weight:  1.00
Name: SP_PIG:40       Len:    80  Check:    25  Weight:  1.00
Name: SP_PIG:89       Len:    80  Check:    25  Weight:  1.00
Name: SP_RAT:42       Len:    80  Check:    25  Weight:  1.00
Name: SP_RAT:91       Len:    80  Check:    25  Weight:  1.00
Name: XP1_XENLA:42    Len:    80  Check:    25  Weight:  1.00
Name: XP2_XENLA:361   Len:    80  Check:    25  Weight:  1.00
Name: XP2_XENLA:408   Len:    80  Check:    25  Weight:  1.00
Name: XP4_XENLA:86    Len:    80  Check:    25  Weight:  1.00
Name: XP4_XENLA:185   Len:    80  Check:    25  Weight:  1.00
//
ITF_HUMAN:42    .....ssssa eeyvglsanq ..ca....vp akdrvdCGYP HVTPKECNNR
ITF_RAT:43      .....sscka qefvglspsq ..cm....vp anvrvdCGYP TVTSEQCNNR
MUA1_XENLA:33   ......ihfl lmvvglgqaq .dcs....va pnmrvnCGYP TVTEADCRAV
MUA1_XENLA:84   ......wcfy natagpikkl .ecs....gd ptkridCGFP RITEKQCILR
MUA1_XENLA:36   ......wcfy teaeaparka .ect....vd psvrtdCGYP GITDKECREK
MUC1_XENLA:31   ..tttkattt tttts.g... .eck....me pskredCGYS GITESQCRTK
MUC1_XENLA:36   ..pqtkwcfy tlsqv.a... .dck....ve psqrvdCGFR GITADQCRQK
MUC1_XENLA:53   ..tttttkat tttts.g... .eck....me pskradCGYP GITESQCRSK
MUC1_XENLA:58   ..pqtkwcfy slpqv.a... .dck....va pssrvdCGFG GITADQCRQR
MUC1_XENLA:63   ...gtkwcfy stsqgna... .mcs....gp ptkrrdCGYP GISSSVCINR
PS2_HUMAN:41    .......... .......eaq tetct...va prerqnCGFP GVTPSQCANK
PS2_MOUSE:44    ........sl aqaqaqaqaq eetci...ma prerinCGFP GVTAQQCTER
SP_HUMAN:43     ........am clagsekpsp cqcsr...ls phnrtnCGFP GITSDQCFDN
SP_HUMAN:92     ....vpwcfh plpkqesd.. .qcv....me vsdrrnCGYP GISPEECASR
SP_MOUSE:41     ........lh alvegekpsp crcsr...lt phnrknCGFP GITSEQCFDL
SP_MOUSE:90     ....vpwcfh plpnqese.. .qcv....me vsarknCGYP GISPEDCASR
SP_PIG:40       ........av cltgaqkpaa crcsr...qd pknrvnCGFP GITSDQCFTS
SP_PIG:89       .......... .vpwcfkplp aqeseecvme vsarknCGYP GISPEDCARR
SP_RAT:42       ........lh alaegekpsp crcsr...mt psnrknCGFP GITSDQCFNL
SP_RAT:91       ....vpwcfh plpnqas... eqc.v...me vsarenCGYP GISPEDCASR
XP1_XENLA:42    .......sig sangqaafte eqc.s...ve rlarvnCGYS GITPQECTKQ
XP2_XENLA:361   apapapvevg pkte...... .dc.k...gd pfkrtdCGYP GITEGQCKAK
XP2_XENLA:408   ..vgvkwcff prtara.... .qc.l...fs pgdredCGYS SITPMECMKR
```

```
XP4_XENLA:86    ....siwcyt p.wkfed... ticnp...ae pkarvnCGYP GITSQDCDKK
XP4_XENLA:185   ....vpwcfk peikkel... lqc.a...vl pkarinCGYP DITMDQCYKK

ITF_HUMAN:42    GCCFdsripg vpwcf..... kplqeaectf
ITF_RAT:43      GCCFdssipn vpwcf..... kplqetectf
MUA1_XENLA:33   GCCFdssiln tkwcfynata gpikklecs.
MUA1_XENLA:84   GCCFdssisg vkwcyartvi ttpapdttt.
MUA1_XENLA:36   GCCYdecipd viwcfekav. ..pvvns...
MUC1_XENLA:31   GCCFdssipq tkwcfytlsq vadckveps.
MUC1_XENLA:36   NCCFdssisg tkwcfystsq vaatktttt.
MUC1_XENLA:53   GCCFdssipq tkwcfyslpq vadckvaps.
MUC1_XENLA:58   NCCFdssisg tkwcfystsq gnamcsgpp.
MUC1_XENLA:63   GCCWdnsvmn vpwcfyrt.. ..........
PS2_HUMAN:41    GCCFddtvrg vpwcfypnti dvppeeesef
PS2_MOUSE:44    GCCFddsvrg fpwcfhpmai entqeeecp.
SP_HUMAN:43     GCCFdssvtg vpwcfhplpk qesdqcvme.
SP_HUMAN:92     KCCFsnfife vpwcffp... nsvedchy..
SP_MOUSE:41     GCCFdssvag vpwcfhplpn qeseqcvme.
SP_MOUSE:90     NCCFsnlife vpwcffp... qsvedchy..
SP_PIG:40       GCCFdsqvpg vpwcfkplpa qeseecvme.
SP_PIG:89       NCCFsdtipe vpwcffpms. ..vedchy..
SP_RAT:42       GCCFdssvag vpwcfhplpn qaseqcvme.
SP_RAT:91       HCCFsnlife vpwcffpqsv ddchy.....
XP1_XENLA:42    GCCFdstiqd apwcfyprat pey.......
XP2_XENLA:361   GCCFdssivg vkwcffprta raqclfspg.
XP2_XENLA:408   GCCFdasitg vkwcfhqk.. ..........
XP4_XENLA:86    GCCFndtipn vvwcyqpiie averdcsav.
XP4_XENLA:185   GCCYdssesd siwcfypdie dvtiie....


# Secstr assignment of 1ps2
# based on actual
ALPHA 24 30 1.0
SHEET 1.0
STRAND 3 7
STRAND 47 52 ANTI  52 3
END
SHEET 1.0
STRAND 33 34
STRAND 45 44 ANTI 44 34
STRAND 16 15 ANTI 16 45
END
```

1sis

```
MSF:   35  Type: P 11-Oct-98  17:16:0  Check: 2960  ..

 Name: scx5_buteu      Len:    35  Check: 6332  Weight:  1.00
 Name: scxl_leiqu      Len:    35  Check: 5465  Weight:  1.00
 Name: scx1_butsi      Len:    35  Check: 4405  Weight:  1.00
 Name: scx1_buteu      Len:    35  Check: 6651  Weight:  1.00
 Name: scxp_andma      Len:    35  Check: 7175  Weight:  1.00
 Name: scx8_leiqh      Len:    35  Check: 6600  Weight:  1.00

//

              1                                   35
scx5_buteu    MCMPCFTTDP NMAKKCRDCC GGNGKCFGPQ CLCNR
scxl_leiqu    MCMPCFTTDH QMARKCDDCC GGkgKCYGPQ CLCR.
scx1_butsi    RCKPCFTTDP QMSKKCADCC GGkgKCYGPQ CLC..
scx1_buteu    MCMPCFTTRP DMAQQCRACC KGRGKCFGPQ CLCGY
scxp_andma    .CGPCFTTDP YTESKCATCC GGRGKCVGPQ CLCNR
scx8_leiqh    RCSPCFTTDQ QMTKKCYDCC GGkgKCYGPQ CICAP


# Secstr assignment of 1sis
# based on dssp
ALPHA 12 19 1.0
#
SHEET 1.0
STRAND 2 4
STRAND 30 33 ANTI 33 2
STRAND 24 27 ANTI 24 33
END
```

1vib

```
MSF:   54  Type: P 11-Oct-98  17:14:4  Check: 7559  ..

 Name: nxb4_cerla      Len:    54  Check: 8758  Weight:  1.00
 Name: nxb2_cerla      Len:    54  Check:   43  Weight:  1.00

//

              1                                                    50
nxb4_cerla    ASATWGAAya CENNCRKKYD LCIRCQGKWA GKRGKCAAHC IIQKNNCKGK
nxb2_cerla    ASSTWGGSya CENNCRKQYD DCIKCQGKWA GKRGKCAAHC AVQTTSCNDK

                54
nxb4_cerla    CKKE
nxb2_cerla    CKKH


# Secstr assignment of 1vib
# based on dssp
ALPHA 11 23 1.0
ALPHA 34 48 1.0
```

2crd

```
MSF of: 2crd.mul from:    1 to:  40
2crd.mul  MSF:  40  Type: P 12-Sep-96  11:52:4  Check: 5859  ..
Name: SCKC_LEIQH      Len:    40  Check:    17  Weight:  1.00
Name: SCK2_LEIQH      Len:    40  Check:    17  Weight:  1.00
Name: SCKI_MESTA      Len:    40  Check:    17  Weight:  1.00
Name: SCK3_LEIQH      Len:    40  Check:    17  Weight:  1.00
Name: SCA2_LEIQH      Len:    40  Check:    17  Weight:  1.00
Name: SCKA_TITSE      Len:    40  Check:    17  Weight:  1.00
Name: SCK2_ANDMA      Len:    40  Check:    17  Weight:  1.00
Name: SCA3_LEIQH      Len:    40  Check:    17  Weight:  1.00
Name: SCK1_ANDMA      Len:    40  Check:    17  Weight:  1.00
Name: SCK1_ORTSC      Len:    40  Check:    17  Weight:  1.00
Name: SCA1_LEIQH      Len:    40  Check:    17  Weight:  1.00
Name: SCKM_CENMA      Len:    40  Check:    17  Weight:  1.00
Name: SCK1_CENLL      Len:    40  Check:    17  Weight:  1.00
Name: SCKB_PANIM      Len:    40  Check:    17  Weight:  1.00
Name: SCKA_PANIM      Len:    40  Check:    17  Weight:  1.00
Name: SCXM_SCOMA      Len:    40  Check:    17  Weight:  1.00
Name: SCKG_PANIM      Len:    40  Check:    17  Weight:  1.00
//
SCKC_LEIQH      .QFTNVSCTT SKECWSVCQR LHN.TSRGKC MNKKCRCYS.
SCK2_LEIQH      .QFTQESCTA SNQCWSICKR LHN.TNRGKC MNKKCRCYS.
SCKI_MESTA      .QFTDVDCSV SKECWSVCKD LFG.VDRGKC MGKKCRCYQ.
SCK3_LEIQH      .GLIDVRCYD SRQCWIACKK VTG.STQGKC QNKQCRCY..
SCA2_LEIQH      GVPINVSCTG SPQCIKPCKD A.G.MRFGKC MNRKCHCTPK
SCKA_TITSE      .VFINAKCRG SPECLPKCKE AIG.KAAGKC MNGKCKCYP.
SCK2_ANDMA      AVRIPVSCKH SGQCLKPCKD A.G.MRFGKC MNGKCDCTPK
SCA3_LEIQH      GVPINVPCTG SPQCIKPCKD A.G.MRFGKC MNRKCHCTPK
SCK1_ANDMA      GVEINVKCSG SPQCLKPCKD A.G.MRFGKC MNRKCHCTPK
SCK1_ORTSC      GVIINVKCKI SRQCLEPCKK A.G.MRFGKC MNGKCHCTPK
SCA1_LEIQH      GVPINVKCTG SPQCLKPCKD A.G.MRFGKC INGKCHCTPK
SCKM_CENMA      .TIINVKCTS PKQCLPPCKA QFGQSAGAKC MNGKCKCYPH
SCK1_CENLL      .ITINVKCTS PQQCLRPCKD RFGQHAGGKC INGKCKCYP.
SCKB_PANIM      ....TISCTN EKQCYPHCKK ETG.YPNAKC MNRKCKCFGR
SCKA_PANIM      ....TISCTN PKQCYPHCKK ETG.YPNAKC MNRKCKCFGR
SCXM_SCOMA      .....VSCTG SKDCYAPCRK QTG.CPNAKC INKSCKCYGC
SCKG_PANIM      ....LVKCRG TSDCGRPCQQ QTG.CPNSKC INRMCKCYGC


# Secstr assignment of 2crd
# based on actual
ALPHA 11 19 1.0
#beta sheet
SHEET 1.0
STRAND 1 3
STRAND 32 36 ANTI 35 1
STRAND 25 29 ANTI 29 32
END
```

Chapter 8

The N-terminus of the glucagon-like-peptide-1 receptor

G-protein coupled receptors of the secretin family are activated by peptide hormones of about 30 residues in length. There is considerable sequence homology within both the hormone and receptor families. The receptors possess in addition to their integral membrane domain a characteristic extracellular domain of about 120 residues in length. The latter have conserved cysteine residues, which are presumably involved in disulphide bridge formation and tryptophans, which have been shown to be critical for hormone binding.

This extracellular domain does not have appreciable homology to any known protein

fold. In order to be able to propose a structure for this domain I have used computational tools for predicting secondary structure and accessibility, ligand binding and mutational data and used this information as input data to the *ab initio* protein folding program DRAGON. The calculations were carried out in a combinatorial manner in order to explore different permutations of disulphide bond connectivity, tryptophan side chain position and chain topology.

8.1 Introduction

The superfamily of G protein-coupled receptors (GPCRs) are membrane proteins present in all higher animals where they perform vital signalling functions between the external environment (vision, olfaction) and the organism's nervous system and between cells in the body (neuromuscular, endocrinological and metabolic control and CNS functioning). Dysfunction of these receptors due to mutation or interference with the normal mechanism of agonist action will be harmful to the organism; many diseases arise directly from these kinds of molecular lesion.

For this reason there has been a continued activity over many years in the study of these receptors at the physiological, pharmacological and biochemical levels. In particular, many of these receptors have been cloned and sequenced. There are, in fact, more sequences for GPCRs (over 800 at the present time (Horn *et al.*, 1998b)) than for any other protein family. Despite the very intensive level of research, but

as a direct consequence of the membrane location of these receptors, there exists no high-resolution structure for any GPCR. The difficulties of obtaining sufficiently high expression levels and of extracting, purifying, reconstituting and crystallising these proteins have so far proved to be insurmountable. A considerable degree of success in obtaining lower resolution structural data from electron crystallography experiments has however been reported (Unger *et al.*, 1997; Krebs *et al.*, 1998).

The proteins themselves form a superfamily in which members *within* each of the constituent families have considerable sequence similarity. Reliable alignments have been carried out by several authors (Oliveira *et al.*, 1994; Taylor and Jones, 1995) and a continuously updated database of sequences and alignments is maintained at a web-site at the EMBL, Heidelberg, Germany (Horn *et al.*, 1998b). There is much lower sequence similarity *between* the families and alignments of the families has only recently been accomplished (Frimurer *et al.*, 1999) The various families of the GPCR superfamily differ not only in the degree of similarity of their seven transmembrane helical domains ('7TM'). In particular, the family of the secretin receptors (otherwise referred to as 'Class B') is characterised by having an extracellular domain of about 120 residues attached to the N-terminus of the integral membrane domain. This extracellular domain has been shown to be critical for binding of the hormones that activate these receptors and for overall function (DeAlmeida and Mayo, 1998; Eyll *et al.*, 1996; Graziano *et al.*, 1996; Vilardaga *et al.*, 1997; Vilardaga *et al.*, 1995; Wilmen *et al.*, 1997; Wilmen *et al.*, 1996). It is therefore critical to obtain structural information for N-terminus as well as for 7TM. Although it has been possible to

isolate the N-terminus from the glucagon-like-peptide-1 (GLP1) receptor (Wilmen *et al.*, 1996) and to show that it alone is responsible for much of the binding energy for the hormone, this domain has so far not been crystallised. Structural biology plays an important role in furthering the understanding of the function of biological molecules and in stimulating the design of new biochemical experiments. Until crystal structures become available, carefully constructed models can serve as a very useful substitute. Attempts to construct a model for N-terminus has been made more difficult by the lack of homology to any known protein structure. I have used `DRAGON` in an attempt to build a model for the N-terminal GPCR domain.

8.1.1 GPCR's

GPCR's are a widely studied family of proteins which, as mentioned, play an important role in the function of many aspects of cellular function. Consequently they have been much studied. There has been only modest success in predicting these proteins. Models of the 7TM domains are based on the seven helix structure of bacteriorhodopsin and are regarded as being irrelevant as models for GPCRs (Soppa., 1994). Currently, GPCR modellers either tend to favour the model proposed by Baldwin (Baldwin *et al.*, 1997) based on the low resolution electron crystal structure of Unger et al. (Unger *et al.*, 1997) or take a more *ab initio* approach (Herzyk and Hubbard., 1995; Perez *et al.*, 1998; Taylor and Jones, 1995).

GPCRs all possess a 7TM domain, but in addition may possess other extracellular

domains, which has led to a system of classification based on these topological details and on mutual sequence similarity (Horn *et al.*, 1998b). The largest family, those GPCRs which are homologous to rhodopsin do not usually possess large domains beyond the obligatory 7TM domain, but two other families possess large N-terminal domains, the secretin family (c.120 residues), the Ca^{2+}/metabotrope glutamate family (c.580 residues) and the family with a 'frizzled' cysteine-rich N-terminal domain (Horn *et al.*, 1998b).

Of particular interest is the receptor for glucagon and GLP1 and their interactions with their respective hormones. Both are members of the secretin family and it has been shown for several members that the N-terminal domain is critical for binding the hormone (DeAlmeida and Mayo, 1998; Eyll *et al.*, 1996; Graziano *et al.*, 1996; Vilardaga *et al.*, 1997; Vilardaga *et al.*, 1995; Wilmen *et al.*, 1997; Wilmen *et al.*, 1996), in conjunction with some of the extracellular loops between the transmembrane helices (Buggy *et al.*, 1995; Paolo *et al.*, 1998). See Figure 8.1 for a schematic representation of the receptor and hormone.

As the two hormone/receptor systems are very similar (48.3% sequence identity for the hormones, 47.8% for the receptors) I, in collaboration with R. Bywater, only performed folding studies on the GLP1 receptor. The aim was to predict the structure of the protein and this is described in the remainder of this chapter. It is then hoped that some of the possible interactions between the N terminus and the hormone and loops between the 7TM segments may be determined.

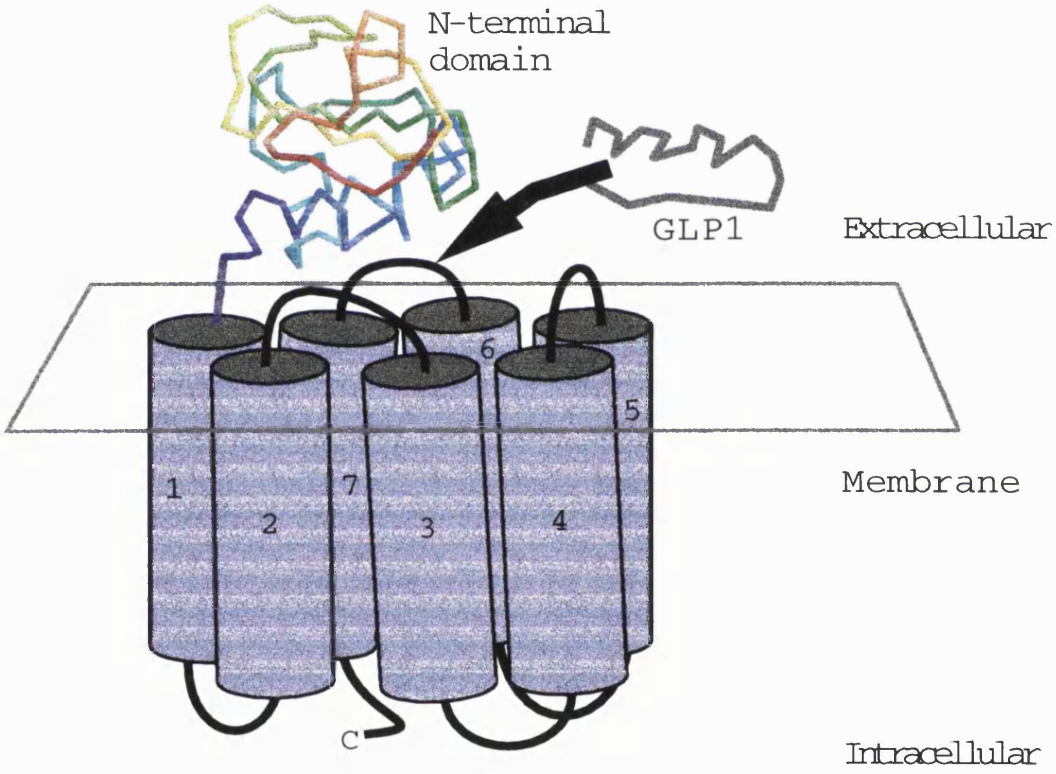


Figure 8.1: Shown is a representation of the GLP-1 receptor and the large N-terminal domain, along with the GLP-1 hormone. It is proposed that the hormone binds partly to the N-terminal domain and some of the loops on the extracellular side of the receptor.

8.1.2 Glucagon

Like insulin, glucagon is involved in regulating blood glucose, but has the opposite effect. In the liver, it activates its receptor which in turn stimulates the gluconeogenesis cascade. Under normal circumstances glucagon will be secreted in response to low blood glucose, and inhibited when blood sugar levels are high.

The N-terminus has six cysteine residues with unknown connectivity. I have tried to predict a likely scenario for the formation of these bridges based on a secondary structure prediction and models constructed using `DRAGON`.

8.2 Methods

A variety of methods have been employed to find out as much additional information as possible about the N-terminus of the receptor.

8.2.1 Sequence alignment and secondary structure.

Many different search methods were used to detect sequence similarities, to the target protein. These included `BLAST` (Altschul *et al.*, 1990), `BLITZ` (Smith and Waterman, 1981a), `FASTA` (Lipman and Pearson, 1985), `PSI-BLAST` (Altschul *et al.*, 1997) and an early version of `QUEST` (Taylor, 1998a), all of which were used as described in Chapter 3.

I employed several programs in order to determine the secondary structure of the protein: `PHD` (Rost and Sander, 1993), `DSC` (King and Sternberg, 1996), NNPRED (Kneller *et al.*, 1990) and NNSSP (Salamov and Solovyev, 1995). Also predicted by `PHD` was information on the accessibility of the N-terminus.

A great deal of time was spent at the multiple alignment stage, examining the sequences "by eye" to try and gain a better idea of the secondary structure prediction.

8.2.2 Correlated mutation analysis

The correlated mutation analysis (CMA) was carried out by R. Bywater.

Conservation in multiple alignments are known to be important in predicting protein structure. It is postulated that mutations which occur in important sites in a sequence may have to be balanced by other correlated mutations to preserve some aspect of structure or function. Although the original hopes that this method would assist in predicting the fold of proteins (Shindyalov *et al.*, 1994; Taylor and Hatrick, 1994) have not been fulfilled (Taylor and Hatrick, 1994; Pollock and Taylor, 1997), it has been shown (Singer *et al.*, 1995; Pazos *et al.*, 1997; Horn *et al.*, 1998a; Gouldson *et al.*, 1998) that sites at domain and protein-protein interfaces can be successfully predicted as overall function is preserved. If the correlated mutations (CMs) that are observed in the secretin family (Horn *et al.*, 1998a) do correspond to close contacts within the receptors and receptor hormone complexes, then this distance information could be built in to `DRAGON` as a distance constraint.

8.2.3 Fold recognition

Both `MST` (Taylor, 1997) and `THREADER2` (Jones *et al.*, 1992a) were run with the sequence across a UCLA benchmark fold database (Fischer *et al.*, 1996a), in the first case, and a CATH (Orengo *et al.*, 1997) based set supplied with the `THREADER2` program. A method by Fischer and Eisenberg which accepts requests via the `WWW` was

also used (Fischer and Eisenberg, 1996).

8.2.4 Disulphide pairing analysis

For some of the initial modelling runs, with an anti-parallel and parallel sheets both the C_α and C_β potential disulphide distances were measured. See Chapter 7 for more details on the method.

8.2.5 Folding by distance geometry

The distance geometry based method devised by Aszódi and Taylor was used extensively (Aszódi *et al.*, 1995a; Aszódi and Taylor, 1996). The method, called `DRAGON`, has been tested in the CASP assessment with good results (Aszódi *et al.*, 1997a; Aszódi *et al.*, 1997b) (see also Chapter 4). It has also performed well on an *ab initio* prediction of CASP2 target 42 (see Chapter 5). `DRAGON` was used for the majority of the fold generation in the prediction of the N-terminus. All `DRAGON` modelling was based on sequence number seven in Figure 8.3. This is the human GLP1 receptor N-terminal region, which is most relevant to the pharmaceutical industry.

The two most important sets of information which `DRAGON` uses when constructing models is the multiple alignment input file and the secondary structure assignment file.

Combinatorial models

An additional aid in generating models was to look at all the possible combinations of interconnectivity between the predicted secondary structure elements on a scaffold structure. Using a method by Taylor (Taylor, 1993) all possible combinations of fold were created for the given secondary structure prediction. Based on the secondary structure prediction and the scaffold, 60 different fold conformations were built using this method.

Glycosylation sites

There are several sites in the protein sequence which would undergo glycosylation. These sites are characterised by this sequence: NX(S/T) i.e. an asparagine followed by anything and then serine or threonine and will most probably be close to the surface of the protein. This factor can be built into the DRAGON modelling. Particular note was made of glycosylation sites which were conserved across the family of sequences and were thus especially likely to appear on the surface.

Disulphide bonds

Disulphide bonds were predicted from the disulphide pairing analysis discussed in Chapter 7. Using DRAGON it is possible to build in these pairs as additional restraints to try and improve on the fold prediction.

Sheet geometry

Several different possibilities were available to us when predicting the topology of the sheet. With four strands (three pairs), there are nine combinations of parallel/anti-parallel arrangements. Connectivity with loops make a further 12 combinations of each arrangement. Models were generated for all possible combinations and then analysed by `WHATIF` (Vriend, 1990), to see if a particular fold would be more or less likely to occur in a real protein.

Tryptophan positions

Wilmen *et al.* (Wilmen *et al.*, 1997) showed that five out of the six tryptophans in the rat N-terminal GLP1 receptor region are essential to allow the receptor to bind the glucagon like peptide. It is likely, therefore, that one or more of these five Trp's would be on the surface of the domain and may well be situated relatively close together, if they are involved in the binding of the ligand. By using the buried/exposed functionality in `DRAGON` I was able to constrain the trp residues so that they would be always exposed.

8.2.6 DRAGON usage

Some examples of DRAGON files are shown here. The DRAGON command file is designed to take the DRAGON parameter file *drag.par* as input and then run DRAGON 30 times to produce 30 models and then quit.

```
# Run command file
p drag.par
r 30
q
```

DRAGON can be told of specific residues which should be positioned at the surface of a protein, or buried in that protein. Particular interest would be in the glycosylation sites or the tryptophans thought to be important in ligand binding. This is accomplished with a very simple file such as follows:

```
# drag.acc
# Residues thought to be on the surface
# due to N glycosylation sites Nx(S/T)
S 27 32 33 41 56 73 93 95 100
S 17 50 69 88 98
# Residues thought to be buried
B
```

When there are some additional constraints to be applied to a structure then these can be specified in a suitable file. These restraints can be between any atom in the model and they can be constrained to a lower and upper limit, which will preferably not be exceeded, depending on a strictness value, or confidence in the restraint (between 0

and 1). The file takes the form of:

reside1 residue2 mindistance maxdidtance stringency atomtype1 atomtype2

e.g. below are three putative disulphide restraints,

```
# drag.restr
#
#Disulphide restraints
107 85 6.80 7.00 1.00 CA CA
107 85 3.80 4.50 1.00 SCC SCC
66 43  5.80 6.00 1.00 CA CA
66 43  3.80 4.50 1.00 SCC SCC
27 52  5.80 6.00 1.00 CA CA
27 52  3.80 4.50 1.00 SCC SCC
```

Probably the most important file after the multiple alignment is the secondary structure prediction assignment. In this example the strands within the sheet are two pairs anti-parallel and the last pair parallel with each other.

```
# drag.str
# Secstr assignment of combinatorial ss
#
ALPHA 8 20 1.0
ALPHA 23 34 1.0
ALPHA 68 71 1.0
ALPHA 90 93 1.0
#pred beta
SHEET 1.0
STRAND 46 53
STRAND 60 64 ANTI 64 46
STRAND 80 86 ANTI 86 60
STRAND 104 107 PAR 104 80
END
```

8.2.7 Glucagon like peptide (GLP) model

A model for GLP1 was carried out by R. Bywater. It was computed on the basis of NMR data, secondary structure prediction, mutations including 'ala' scans and CM analysis which had been proposed (Frimurer *et al.*, 1999). Independently, a set of *ab initio* structures for GLP1 was calculated using the ECEPP force field with the application of a smoothing algorithm for searching for relevant potential minima. The structure first proposed fell exactly in the middle of the envelope of the superimposed *ab initio* structures (Frimurer *et al.*, 1999) which added credibility to the original structure. This model for GLP1 will be used in conjunction with the model for the N-terminus modelled here to examine the possible mode of binding of GLP1 to the N-terminus, in the future.

The sequence of the Glucagon Precursor, Human is shown below.

```
HAEGTFTSDV SSYLEGQAAK EFIAWLVKGR
```

8.2.8 Analysis

Analysis of models was being carried out, by R. Bywater, at the time of writing by using the WHATIF suite of programs (Vriend, 1990). The quality check programs QUACHK (checks core-packing quality) and BBCCHK (checks backbone conformation normality) are being used. The quality scores in the output files will be used to

sort and select the best quality structures.

A further analysis may be implemented to take the multiple alignment conservation and hydrophobicity information and build it on to all the models produced in DRAGON. It is then a simple matter to check for a model which has good conserved hydrophobic packing according to our alignment. In theory DRAGON will try and do this when modelling, but undoubtedly some models will be better than others in this respect.

8.3 Results and discussion

An early version of QUEST, an iterated search program (Taylor, 1998a) was run to identify any further sequences which had not been picked up by the other commonly available similarity search methods. From this sequence search the multiple alignment was made, all sequence search tools were in agreement with each other and no unexpected outliers were identified using any of the more advanced methods such as QUEST and PSI-BLAST.

A working copy of the multiple alignment is shown in Figure 8.2. This was coloured by hand to highlight the features of interest. The final alignment was converted into MSF format and is shown in Figure 8.3. From this we were able to be more confident about the secondary structure predictions, which when possible used this multiple alignment as the basis for predicting the secondary structure. The same alignment was used by DRAGON in its derivation of distance restraints.

The results of the secondary structure prediction were not conclusive, see Figure 8.4 for a predict protein (PHD) secondary structure prediction. A multiple alignment was constructed using MULTAL, with some manual corrections. From the alignment secondary structure predictions were made using PHD and DSC. In a combination with these methods and studying the patterns of conserved residues in the alignment we came up with a secondary assignment which we were confident about.

The secondary structure assignment shown in Figure 8.4 is not always very well in

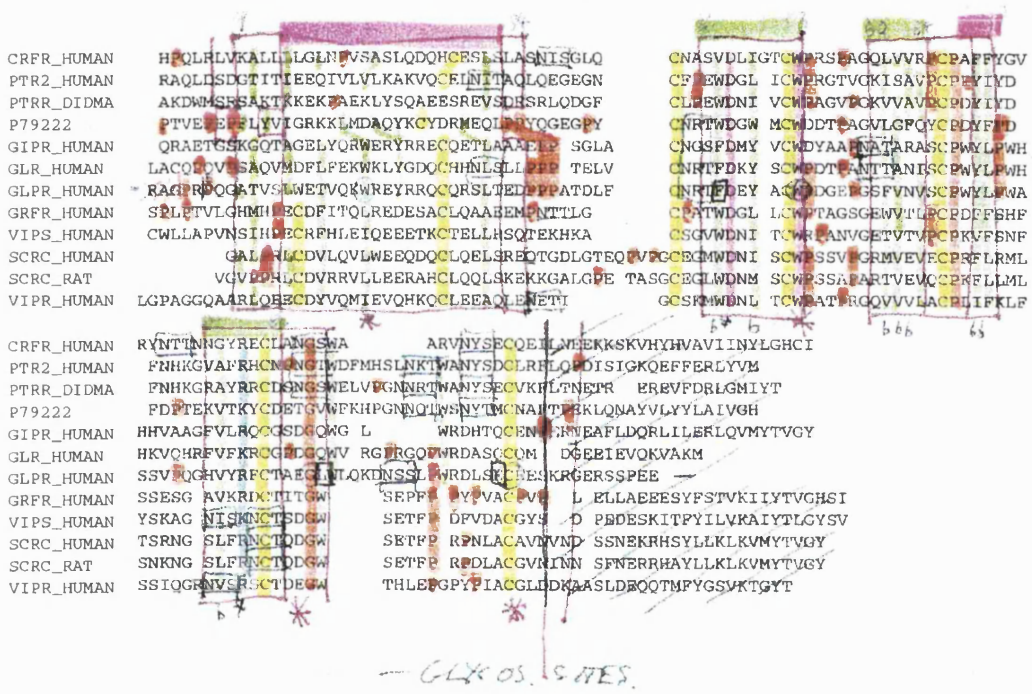


Figure 8.2: Working alignment with important features highlighted. Conserved cysteines are highlighted, along with other conserved residues, including some patterns of conserved hydrophobicity. Glycosylation sites are boxed in green and prolines highlighted in orange. Correlated mutations are in black boxes. This sort of hand editing can aid a great deal in the prediction of secondary structure and acts as an aid when predicting and modelling proteins.

```
MSF of: drag.mul from:    1 to:  117
drag.mul  MSF:  117  Type: P 12-Sep-96  11:52:4  Check: 5859  ..
    . . . . . .
    .deleted headers.
    . . . . . .
//
CRFR_HUMAN:44   ..HPQLRLVK ALLLLGLNPV SASLQDQHCE SLSLASNISG LQ......CN
PTR2_HUMAN:63   ..RAQLDSDG TITIEEQIVL VLKAKVQCEL NITAQLQEGE GN......CF
PTRR_DIDMA:10   ..AKDWMSRS AKTKKEKPAE KLYSQAEESR EVSDRSRLQD GF......CL
P79222:72       .....PTVEP EPFLYVIGRK KLMDAQYKCY DRMEQLPPYQ GEGPY...CN
GIPR_HUMAN:61   ..QRAETGSK GQTAGELYQR WERYRRECQE TLAAAEPP.S GLA.....CN
GLR_HUMAN       .LACQPQVPS AQVMDFLFEK WKLYGDQCHH NLSLLPPP.T ELV.....CN
**GLPR_HUMAN**  .RAGPRPQGA TVSLWETVQK WREYRRQCQR SLTEDPPPAT DLF.....CN
GRFR_HUMAN:55   .SPLPTVLGH MHPECDFITQ LREDESACLQ AAEEMPNTTL G.......CP
VIPS_HUMAN:52   .CWLLAPVNS IHPECRFHLE IQEEETKCTE LLRSQTEKHK A.......CS
SCRC_HUMAN:66   ........GA LPRLCDVLQV LWEEQDQCLQ ELSREQTGDL GTEQPVPGCE
SCRC_RAT:66     .......VGV PPRLCDVRRV LLEERAHCLQ QLSKEKKGAL GPE.TASGCE
VIPR_HUMAN:63   LGPAGGQAAR LQEECDYVQM IEVQHKQCLE EAQLENETI. .......GCS

CRFR_HUMAN:44   ASVDLIGTCW PRSPAGQLVV RPCPAFFYGV RYNTTNNGYR ECLANGSWA.
PTR2_HUMAN:63   PEWDGL.ICW PRGTVGKISA VPCPPYIYD. .FNHKGVAFR HCNPNGTWDF
PTRR_DIDMA:10   PEWDNI.VCW PAGVPGKVVA VPCPDYIYD. .FNHKGRAYR RCDSNGSWEL
P79222:72       RTWDGW.MCW DDTPAGVLGF QYCPDYFPD. .FDPTEKVTK YCDETGVWFK
GIPR_HUMAN:61   GSFDMY.VCW DYAAPNATAR ASCPWYLPWH HHVAAGFVLR QCGSDGQWG.
GLR_HUMAN       RTFDKY.SCW PDTPANTTAN ISCPWYLPWH HKVQHRFVFK RCGPDGQWV.
**GLPR_HUMAN**  RTFDEY.ACW PDGEPGSFVN VSCPWYLPWA SSVPQGHVYR FCTAEGLWLQ
GRFR_HUMAN:55   ATWDGL.LCW PTAGSGEWVT LPCPDFFSHF SSESG.AVKR DCTITGW...
VIPS_HUMAN:52   GVWDNI.TCW RPANVGETVT VPCPKVFSNF YSKAG.NISK NCTSDGW...
SCRC_HUMAN:66   GMWDNI.SCW PSSVPGRMVE VECPRFLRML TSRNG.SLFR NCTQDGW...
SCRC_RAT:66     GLWDNM.SCW PSSAPARTVE VQCPKFLLML SNKNG.SLFR NCTQDGW...
VIPR_HUMAN:63   KMWDNL.TCW PATPRGQVVV LACPLIFKLF SSIQGRNVSR SCTDEGW...

CRFR_HUMAN:44   ......ARVN YSECQEI
PTR2_HUMAN:63   MHSLNKTWAN YSDCLRF
PTRR_DIDMA:10   VPGNNRTWAN YSECVKF
P79222:72       HPGNNQTWSN YTMCNAF
GIPR_HUMAN:61   L......WRD HTQCENP
GLR_HUMAN       RGPRGQPWRD ASQCQM.
**GLPR_HUMAN**  KDNSSLPWRD LSECEES
GRFR_HUMAN:55   ..SEPFP.PY PVACPVP
VIPS_HUMAN:52   ..SETFP.DF VDACGYS
SCRC_HUMAN:66   ..SETFP.RP NLACAVN
SCRC_RAT:66     ..SETFP.RP DLACGVN
VIPR_HUMAN:63   ..THLEPGPY PIACGLD
```

Figure 8.3: MSF of alignment of N-terminal region.

```
          ....,....1....,....2....,....3....,....4....,....5....,....6
                                    #               #       #
AA        RAGPRPQGATVSLWETVQKWREYRRQCQRSLTEDPPPATDLFCNRTFDEYACWPDGEPGS
PHD                    HHHHHHHHHHHHHHHHHHH           E      EE
DSC                 EEEEEEHHHHHHHHHHHHHH                     EEE
nnpred               E  HHHHHH HHHHHH    HHH HHHH    E
nnssp                  HHHHHHHHHHHHHH
DRAGON            HHHHHHHHHHHHH  HHHHHHHHHHHH           EEEEEEEE      E

          ....,....7....,....8....,....9....,....10...,....11...,....12
               #                   #                     #
AA        FVNVSCPWYLPWASSVPQGHVYRFCTAEGLWLQKDNSSLPWRDLSECEESKRGERSSPEEQ
PHD        EEEE       HHHEEEEEEEEE     EEE
DSC       EEEE                EEEEE                  HHHHHH
nnpred    EEE     E     HH    EEEE  HH HEE
nnssp      EE                 EEEE      EE
DRAGON    EEEE   HHHH        EEEEEEE   HHHH          EEEE
```

Figure 8.4: PHD secondary structure prediction of the N-terminal region. This prediction was based on the multiple sequence alignment, where applicable.

agreement, so making a consensus is not easy. The secondary structure prediction in this instance is not a clear cut scenario. The secondary structure assignment which was used for the modelling is shown in Figure 8.4 called DRAGON. This was based on the program predictions and discussions with Taylor over the alignment in Figure 8.2. The most different prediction made in the DRAGON models is that of residues 91 and 92, they were assigned as a helix for the purposes of the modelling rather than as part of a beta sheet. The reason for this was the inability to make a reasonable fold which was compatible with the rest of the predicted secondary structure, as well as a lack of confidence in that region of the multiple alignment used for the secondary structure prediction.

From the results of the fold recognition, I was unable to find a clear link to any particular fold type. All the results were not significant as the Z-scores were below

the confidence limits for the programs. Additionally, there was no apparent pattern in the identified folds. Either the methods are not sensitive enough to identify a correct fold or this is a novel fold, not yet in the database. No useful information could be obtained by fold recognition, so a purely *ab initio* model was constructed.

For model building DRAGON was the main program used. Another method based on a combinatorial connectivity assignment of large secondary structure elements was used to build models based on the secondary structure (Taylor, 1991). In combining these approaches it was hoped that a model of reasonable quality could be produced.

Many models were generated using DRAGON, but even if the approximate fold may be correct there is still uncertainty about the arrangements of the strands within the predicted sheet. The subsequent analysis which was carried out generated all possible combinations of sheet organisation. The results of these models were analysed by WHATIF and the most likely combination was chosen.

Based on the disulphide analysis explained earlier in this chapter, three disulphide bond pairs were calculated.

Other pairs may well be possible and different restraints could be applied to different models in an attempt to classify the likely models given a specific set of disulphide restraints. It may be that the elimination of some pairs could be carried out in this manner.

The plots of C_α and C_β potential disulphide bonds were calculated and compared.

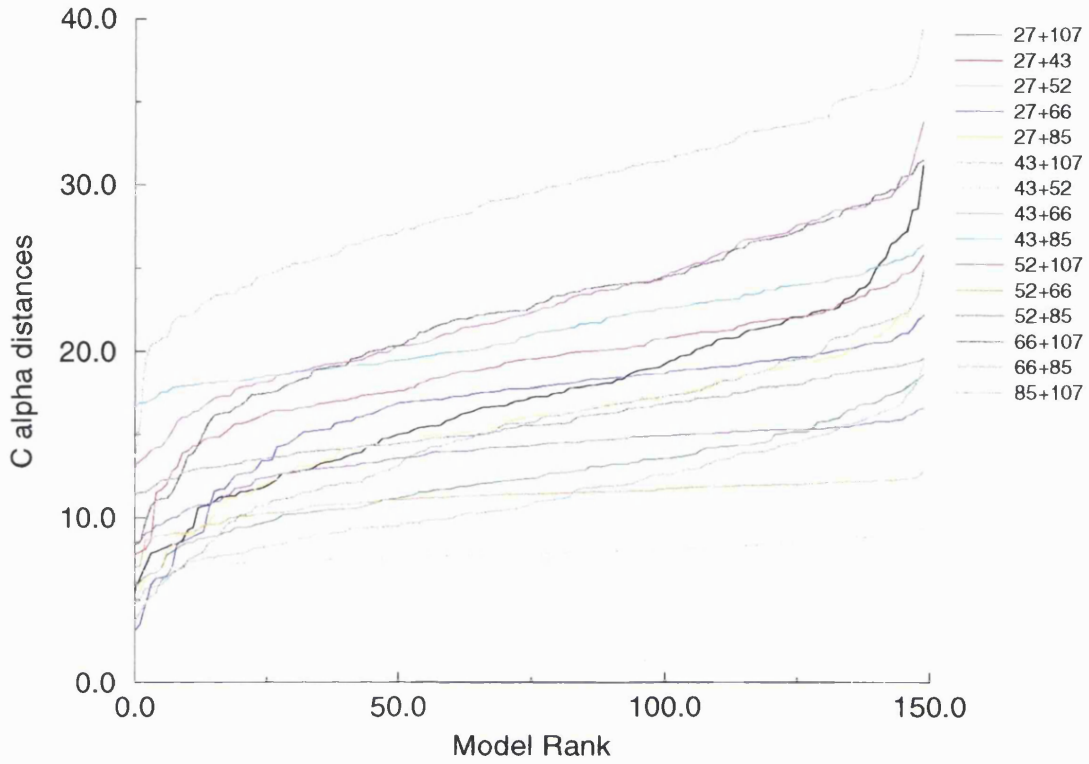


Figure 8.5: C_α disulphide analysis of DRAGON model of N-terminal GLP1 receptor.

Figure 8.5 and Figure 8.6 were overlaid on acetate to make for easy comparison of the differences between the C_α and C_β distances. The main graph to consider is the C_α plot, as minor differences in the side chain centroid (SCC) orientation can severely effect the distances, which are subsequently remedied when restraints are applied. The comparison of the two gives an idea of which pairs of side chains are generally pointing towards each other or away from each other. For example if the average C_β:C_β distance is less than the average C_α:C_α distance then the two SCC's will be pointing towards each other in a disulphide like forming position.

Once a set of models are generated which I am confident in they can then be built into

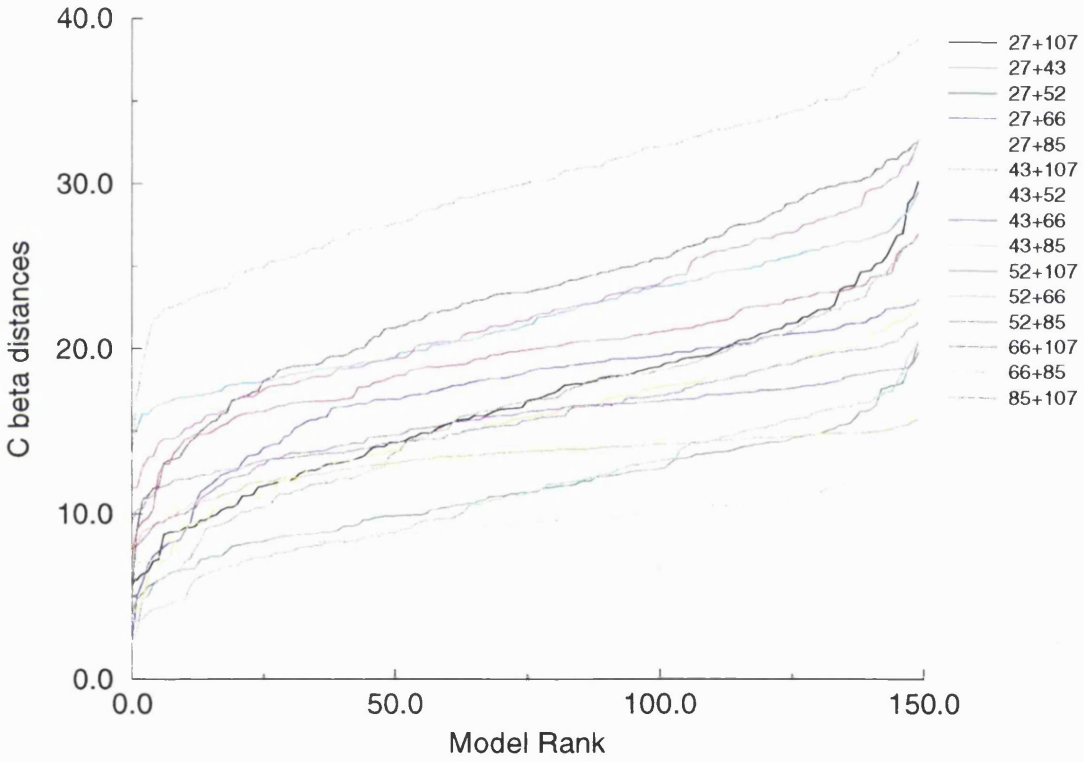


Figure 8.6: C_β disulphide analysis of DRAGON model of N-terminal GLP1 receptor.

full atom proteins and minimised. Then an energy based ranking can be imposed and the disulphides recalculated. Then the models can be re-run with the new restraints, to see if the models improve.

Using the novel disulphide prediction method sets 7 followed by set 13 were identified as having the closest models. The second set is in direct agreement with a by-eye judgment prediction. See Table 8.1 for results of the disulphide analysis.

Set 7 Predicted Disulphides: 27-66, 43-52, 85-107.
Set 13 predicted and also identified by-eye: 66-85, 43-52, 27-107.

Table 8.1 also shows model_51.pdb as having a consistently high disulphide bonding pattern, even amongst different pairs. This is probably due to a significant clustering of the cysteines together, or perhaps two disulphides which are particularly well formed and biasing the other. More detailed analysis based on these figures can give a lot of insight into the possible connectivity of the bonds in the models.

Based upon earlier CM analysis (Horn *et al.*, 1998a) the following sites in N-terminus were suggested to be important for hormone binding and/or contacts with the 7TM domain or within the N-terminus itself. In the list below are shown some of the original correlation networks (clusters of residues that are correlated). 'L' denotes residues in the hormone (ligand), 'N' denotes residues in N-terminus while integers without a preceding denote positions in the 7TM domain.

Set1		Set6		Set11	
26.718	model_51.pdb	38.353	model_31.pdb	28.452	model_102.pdb
29.295	model_80.pdb	40.010	model_149.pdb	33.638	model_52.pdb
29.434	model_31.pdb	41.079	model_80.pdb	34.153	model_106.pdb
29.483	model_135.pdb	41.393	model_123.pdb	36.456	model_72.pdb
32.681	model_112.pdb	41.514	model_72.pdb	37.155	model_86.pdb
Set2		Set7		Set12	
34.512	model_135.pdb	16.733	model_51.pdb	29.111	model_102.pdb
40.428	model_31.pdb	18.941	model_59.pdb	36.886	model_52.pdb
41.072	model_80.pdb	22.556	model_120.pdb	39.590	model_106.pdb
41.572	model_107.pdb	24.013	model_112.pdb	40.861	model_72.pdb
41.577	model_95.pdb	28.897	model_135.pdb	43.033	model_114.pdb
Set3		Set8		Set13	
33.014	model_31.pdb	41.316	model_51.pdb	22.269	model_38.pdb
34.221	model_123.pdb	43.302	model_136.pdb	23.446	model_97.pdb
37.745	model_135.pdb	43.678	model_72.pdb	23.608	model_72.pdb
38.041	model_136.pdb	46.99	model_59.pdb	24.718	model_32.pdb
38.826	model_149.pdb	47.791	model_70.pdb	25.351	model_117.pdb
Set4		Set9		Set14	
25.382	model_51.pdb	42.208	model_51.pdb	30.341	model_38.pdb
27.566	model_36.pdb	43.929	model_136.pdb	32.347	model_50.pdb
27.711	model_80.pdb	45.032	model_114.pdb	33.714	model_68.pdb
29.518	model_112.pdb	45.267	model_59.pdb	34.028	model_86.pdb
30.142	model_26.pdb	45.997	model_72.pdb	35.448	model_106.pdb
Set5		Set10		Set15	
40.491	model_95.pdb	24.841	model_102.pdb	33.464	model_38.pdb
42.448	model_135.pdb	29.092	model_106.pdb	35.641	model_50.pdb
42.857	model_143.pdb	29.568	model_76.pdb	38.012	model_32.pdb
42.931	model_80.pdb	30.634	model_99.pdb	38.106	model_117.pdb
43.245	model_110.pdb	31.404	model_95.pdb	39.461	model_86.pdb

Table 8.1: Lowest scoring total distance for each possible disulphide set of pairs for the six cysteines forming three disulphide bonds in the N-terminus of GLP1 receptor. See Table 7.1 for description of the sets.

```
                  10        20
           12345678901234567890123
  3 (  L3) QQQEEDDDDDDDDDDDDDDEE
104 ( N71) FFFFFWWWWWWWWWWWWWFF
147 (N114) QQQILWWWWWWWWWWWWWQQ
163 (N130) QQQEEAAAAAAAAAAAAAAQQ

                  10        20
           12345678901234567890123
  5 (  L5) TTTTTIIIIIIIITTVVVVVTT
101 ( N68) NNNNNPPPPPPPPEESSSSSSNN

                  10        20
           12345678901234567890123
 16 ( L16) SSSGGQQQQQQQQSGQQQQQQKK
 21 ( L21) DDDEEKKKKKKKKRRKKKKKDD
130 ( N97) YCHAAFFFFFFFLLFFFFFYH

                  10        20
           12345678901234567890123
 17 ( L17) RRRQQLLLLMMMMAAMMMILIII
144 (N111) PPPAATIIIEEEEQQDEDSSSSS


                  10        20
           12345678901234567890123
 25 ( L25) WWWWWDEDDAAAAGGSSSSSSWW
 73 ( N40) KKKKKQQQQIIIIVVMLMEEERR
 74 ( N41) WWWWWLLLLFFFFLLIIIIIIWW
137 (N104) LLFHHAAAFVVVVSSNNNNNNFF

                  10        20
           12345678901234567890123
 77 ( N44) YYYYYDDDDEEEEEEQQQEEEYY
107 ( N74) YYYYYLLLLIIIIMILLLIIIYY
```

Those CM clusters that contain more than one mention of N-ter allow a distance limit between the corresponding residues to be assigned or at least postulated, while those mentioning both 'L' and 'N' will be used in the future to attempt to 'dock' the hormone to the N-terminus.

The CMA calculation highlighted three residues which were likely to all fall in the same vicinity in the model. These were: 47-106, 47-90, 90-106. These residue pairs could be introduced into the modelling as restraints, but due to the tenuous theory

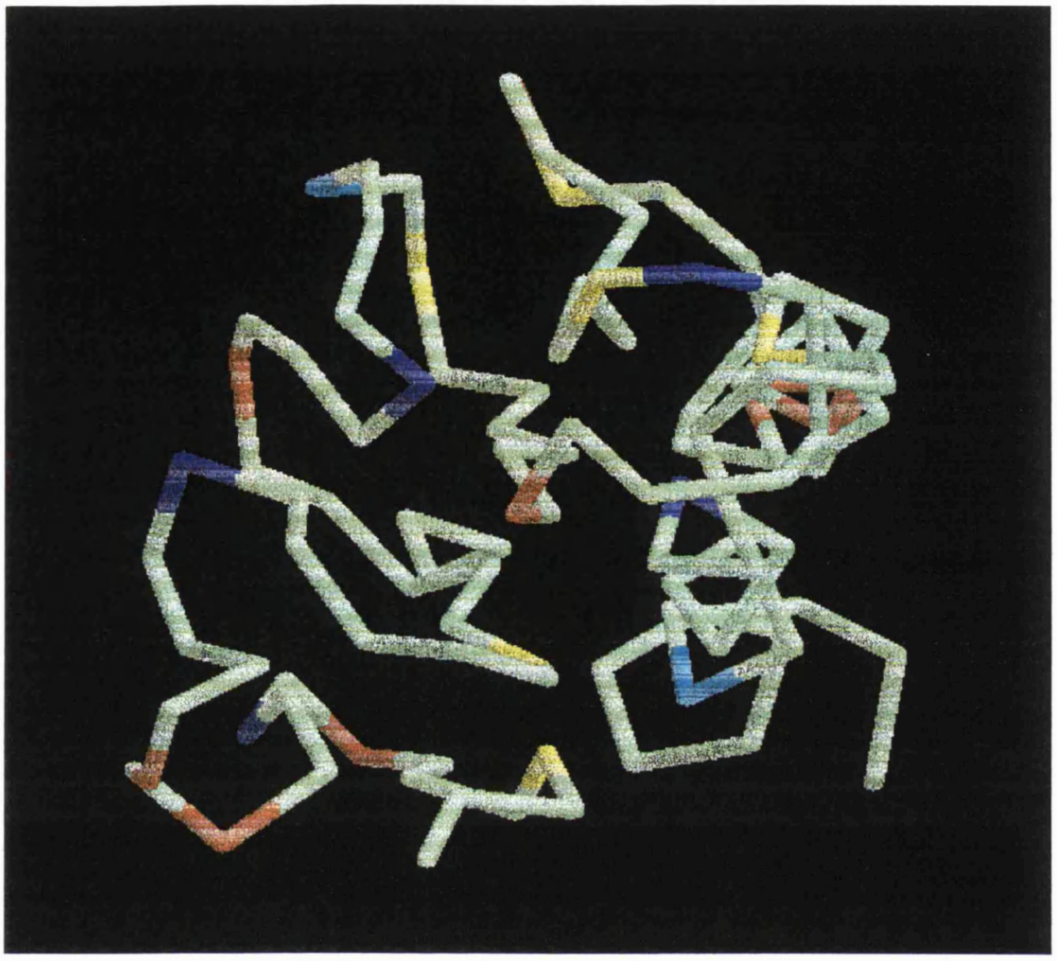


Figure 8.7: A DRAGON model of the N-terminus. Cysteines are shown in yellow, the dark blue indicates 5 of the tryptophans thought to be on the surface, the light-blue are other tryptophans, the orange residues indicate glycosylation sites found in the multiple alignment.

behind the idea of CMA's they were not built in as actual distance restraints. Distances of CMA's were measured after modelling in a similar way to the disulphide bond prediction method, to see if the pairs were in roughly the same vicinity.

Figure 8.7 shows one of the models produced by DRAGON.

8.4 Conclusion

Using a combination of specialised sequence alignment tools, secondary structure and accessibility prediction, correlated mutation information and the distance geometry program DRAGON a potential solution was derived for the structure of the amino-terminal domain of the glucagon receptor.

Many hundreds of models can be generated with this distance geometry approach, using different combinations of restraints of what is know or predicted about the protein. The hardest part, as with any modelling, is to make a meaningful analysis of the models and sort them into some order of confidence. Using the many algorithms built into the WHATIF suite of programs, a systematic analysis is being carried out to find the models which were most protein like, given the different restraints. Although a very difficult problem such as this may generate more ideas than solutions it is hoped that the analyses present herein may bridge some of the gap which still exists in our understanding of the human GLP-1 receptor.

As yet there is no crystal structure for the N-terminal domain, although work is being carried out to try and find out as much as possible about the biochemistry and structure of the protein. It is hoped that some of the models generated for the prediction of this protein will aid the development of new ideas and eventually result in drugs for patients with type II diabetes.

Using all the many methods available to the "Bioinformaticist" can give huge amounts

of information about protein sequences. However in many cases, such as this, there is still a large amount of uncertainty as to the quality of the information and understanding of what makes a protein fold the way it does. Building a good *ab initio* model which factors in all of what is though to be known about a protein, is far from a trivial matter.

Even if the models turn out to be incorrect, which would not be too surprising given the current state of the art, they may still have promoted some further ideas and experiments and thus increase our understanding of the system at hand.

Chapter 9

Conclusions

CASP2 is now almost super by CASP3 and it will be interesting to see if any of the new methods in the community prove successful, more so than two years ago. The results of CASP2 generally considered that Human knowledge based predictions are better than those carried out using computer methods only.

The central prediction method for this thesis, DRAGON, saw its first real test case with the protein sequences of CASP2. DRAGON was used for homology modelling (Aszódi *et al.*, 1997a), as well as some threading based modelling which has been described in Chapter 4. DRAGON's first real test came with the *ab initio* prediction on NK-lysin. It turned out to be quite successful. Subsequent analysis has shown that many of the models produced were accurate in the placement of core secondary structure elements.

It appears that the method for predicting the connectivity between cysteines in proteins falls some way short of the mark for being useful. In some cases the predictions are accurate, but in the majority of cases it is hard to gain any really useful information. With more restraints built into the DRAGON models then perhaps more useful results would be obtained. It would not be an overly complex matter to try restraining the more likely cysteines into close proximity and then analysing the subsequent models for "proteiness". The motto of protein structure prediction could be "That the more information you have the better – and better still the X-ray structure".

In a recent review by Westhead and Thornton (Westhead and Thornton, 1998) they conclude that the best way to improve the sequence alignments in threading is to use a multiple sequence threading approach. If the problem of gap placement could be more easily quantified, then perhaps a better analysis could be adopted and better quality alignments achieved.

One of the major areas for improvements in DRAGON could come from the inclusion of an energy based potential within the program. At the moment the folds are driven by the hydrophobic packing and the confidence to which the secondary structure is assigned, along with other external restraints. Perhaps some sort of Sippl like potential may enable DRAGON to make less stark decisions when generating the models.

As far as the GPCR model is concerned, we will just have to sit back and wait to see if any more information shows my model to be correct, or rather to see if one of them is close.

Due to the phenomenal speed at which models can now be produced, I am convinced that more predictive methods could be based on simple chain models. For example, multiple models could be used to predict secondary structure. By building models and analysing the forms they take, some preferences for secondary structure can be found, so if the unconstrained models all give a helix like structure in one part of a protein, then that could give confidence to a more 1D secondary structure prediction approach.

In time, more and more folds will be solved, eventually making the *ab initio* predictioner redundant. There will be a growth in sequence analysis to identify models and, as is already proving to be the case, a much larger emphasis is being placed on genomic data as it becomes more and more readily available. There is a great debate as to whether there is a finite number of potential protein folds out there. Obviously there must be a finite number, as there is a finite number of proteins, but it is still the case that more and more new folds are being uncovered. The definition of a fold is also very subjective and beyond the scope of this thesis to examine, but once a representative genome has been sequenced and all the representative proteins solved, then a better idea of this problem should be one step nearer to being solved. Perhaps it would be possible to generate many different fold types and test the possible overlap with other folds to determine an accurate prediction of the number of possible protein folds. By building many random proteins more could be understood about packing of residues. With the work of characterising the amino acids with the *cones* algorithm it may be possible to analyse models for their overall conformation to the

patterns obtained from the amino acid burial. Further use of the *cones* algorithm and the profiles of where amino acids occur within proteins could be made. They could be used as a method of calculating how good models are and producing a ranked output of models produced. Until we understand more about the way proteins behave in real systems, we will always be playing catch-up to Mother Nature's solution to the folding problem.

The limiting factor in much of the modelling is the quality of the alignment. With obvious sequence homologues and homology modelling, the models generated can be accurate, but homologues are not always to hand. The correct alignment of sequence with structure is still a major failing of all fold recognition methods.

In the meantime, protein structure prediction offers insight into the whole field of molecular biology and protein chemistry. Because of the lengthy X-ray crystallographic process, new methods to predict protein structure are essential if we are to keep up with the huge influx of sequence data and the ever increasing need to solve the tertiary structure of proteins.

Combining techniques to predict protein structure enables us to see a wider picture than each individual part of the problem could possibly convey. Homology modelling gives very accurate insight into the 3D structure of sequences with no solved crystal structure. The models produced can greatly aid in the understanding of detailed protein structure. This works well where there are similar proteins with known fold in the database. Unfortunately this approach falls over when a close homologue is

not available. By using a threading method a likely fold, or folds, can be identified and used as a template for the modelling. Failing to identify a potential template structure leads to *ab initio* prediction and modelling. Perhaps when all the possible fold types have been recognized there will be no need for purely *ab initio* prediction. Current and future methods are likely to be concerned with more and more genomic data thus facilitating pharmaceutical companies to determine and design better and better solutions for the myriad of diseases affecting our lives.

Far more important for the future of Bioinformatics will be the ability to analyse the large volumes of data which are being produced experimentally. For instance, the wealth of information being generated by genome sequencing means that it is essential to have the ability to analyse complete genomes. It is also important to be able to predict protein:protein interaction, the role that metabolites play in systems and the function of proteins from sequence.

As a general overview of this thesis, I think that the application of multiple sequence threading will become a useful addition to protein prediction and should improve the alignment accuracy of threading. The disulphide prediction ideas, with some more work, should be a useful indicator in the goal of protein prediction and modelling.

References

Al-Karadaghi, S., Hansson, M., Nikonov, S., Jonsson, B., and Hederstedt, L. (1997). Crystal structure of ferrochelatase: the terminal enzyme in heme biosynthesis. *Structure*, 5:1501–1510.

Altschul, S. F. and Erickson, B. W. (1986). Optimal sequence alignment using affine gap costs. *Bull. Math. Biol.*, 48:603–616.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.*, 214:403–410.

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nuc. Acids Res.*, 25:3389–3402.

Andersson, M., Gunne, H., Agerberth, B., Boman, A., Bergman, T., Olsson, B., Dagerlind, A., Wigzell, H., Boman, H. G., and Gudmundsson, G. H. (1995). NK-lysin, a novel effector peptide of cytotoxic t-cells and nk-cells – structure and cDNA cloning of the porcine form, induction by interleukin-2, antibacterial and antitumor-activity. *EMBO J.*, 14:1615–1625.

Aszódi, A. and Taylor, W. R. (1994). Secondary structure formation in model polypeptide chains. *Prot. Engng.*, 7:633–644.

Aszódi, A. and Taylor, W. R. (1996). Homology modelling by distance geometry. *Folding & Design*, 1:325–334.

Aszódi, A. and Taylor, W. R. (1997). Hierarchical inertial projection: A fast distance matrix embedding algorithm. *Computers Chem.*, 21:13–23.

Aszódi, A., Gradwell, M. J., and Taylor, W. R. (1995a). Global fold determination from a small number of distance restraints. *J. Mol. Biol.*, 251:308–326.

Aszódi, A., Gradwell, M. J., and Taylor, W. R. (1995b). Protein fold determination using a small number of distance restraints. In Bohr, H. and Brunak, S., editors, *Protein Folds: A Distance Based Approach*, pages 85–97. CRC Press Inc., Boca Raton, Florida, U.S.A.

Aszódi, A., Munro, R. E. J., and Taylor, W. R. (1997a). Distance based comparative modelling. *Folding & Design*, 2, suppl.:S3–S6.

Aszódi, A., Munro, R. E. J., and Taylor, W. R. (1997b). Protein modelling by multiple sequence threading and distance geometry. *Prot. Struct. Funct. Genet.*, suppl. 1:38–42.

Bairoch, A. (1990). *The SwissProt protein sequence databank user manual.* EMBL Data Library, Heidelberg, FRG., 14 edition.

Baldwin, J. M., Schertler, G. F., and Unger, V. M. (1997). An alpha-carbon template for the transmembrane helices in the rhodopsin family of G-protein-coupled receptors. *J. Mol. Biol.*, 272:144–164.

Barnham, K. J., Dyke, T. R., Kem, W. R., and Norton, R. S. (1997). Structure of neurotoxin B-iv from the marine worm cerebratulus lacteus: a helical hairpin cross-linked by disulphide bonding. *J. Mol. Biol.*, 268:886–902.

Baud, F., Pebay-Peyroula, E., Cohen-Addad, C., Odani, S., and Lehmann, M. S. (1993). Crystal structure of hydrophobic protein from soybean a member of a new cysteine-rich family. *J. Mol. Biol.*, 231:877–887.

Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T., and Tasumi, M. (1977). The protein databank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.*, 112:535–542.

Blundell, T. L., Sibanda, B. L., Sternberg, M. J. E., and Thornton, J. M. (1987). Knowledge-based prediction of protein structures and the design of novel molecules. *Nature*, 326:347–352.

Blundell, T. L., Carney, D., Gardner, S., Hayes, F., Howlin, B., Hubbard, T., Overington, J., Singh, D. A., Sibanda, B. L., and Sutcliffe, M. J. (1988). Knowledge-based protein modelling and design. *Eur. J. Biochem.*, 172:513–520.

Bontems, F., Roumestand, C., Gilquin, B., Menez, A., and Toma, F. (1991). Refined structure of charybdotoxin: common motifs in scorpion toxins and insect defensins. *Science*, 254:1521–1523.

Bowie, J. U., Clarke, N. D., Pabo, C. O., and Sauer, R. T. (1990). Identification of protein folds: matching hydrophobicity patterns of sequence sets with solvent accessibility patterns of known structures. *Prot. Struct. Funct. Genet.*, 7:257–264.

Branden, C. I. and Tooze, J. (1991). *Introduction to Protein Structure.* Garland, New York.

Brocklehurst, S. M. and Perham, R. N. (1993). Prediction of the three-dimensional structures of the biotinylated domain from yeast pyruvate carboxylase and of the lipoylated h-protein from the pea leaf glycine cleavage system: A new automated method for the prediction of protein tertiary structure. *Prot. Sci.*, 2:626–639.

Brooks, B. R., Bruccoleri, R. E., Olafson, B. D., States, D. J., Swaminathan, S., and Karplus, M. (1983). CHARMM a program for macromolecular energy, minimisation, and dynamics calculations. *J. Comp. Chem.*, 4:187–217.

Brown, L. R., Mronga, S., Bradshaw, R. A., Ortenzi, C., Luporini, P., and Wuthric, K. (1993). Nuclear magnetic resonance solution structure of the pheromone Er-10 from the ciliated protozoan Euplotes raikovi. *J. Mol. Biol.*, 231:800–816.

Browne, W. J., North, A. C. T., Phillips, D. C., Brew, K., Vanaman, T. C., and Hill, R. L. (1969). A possible three-dimensional structure of bovine alpha-lactalbumin based on that of hen's egg-white lysozyme. *J. Mol. Biol.*, 42:65–86.

Bryant, S. H. and Lawrence, C. E. (1993). An empirical energy function for threading protein-sequence through the folding motif. *Prot. Struct. Funct. Genet.*, 16:92–112.

Buggy, J. J., Livingston, J. N., Rabin, D. U., and Yoo-Warren, H. (1995). Glucagon.glucagon-like peptide I receptor chimeras reveal domains that determine specificity of glucagon binding. *J. Biol. Chem.*, 270:7474–7478.

Bycroft, M., Hubbard, T. J., Proctor, M., Freund, S. M., and Murzin, A. G. (1997). The solution structure of the S1 RNA binding domain: a member of an ancient nucleic acid-binding fold. *Cell*, 88:235–242.

Carugo, K. D., Banuelos, S., and Saraste, M. (1997). Crystal structure of a calponin homology domain. *Nature, Struct. Biol.*, 4:175–179.

CASP, . (1995). CASP special issue. *Prot. Struct. Funct. Genet.*, 23(3).

CASP, . (1997). CASP2 special issue. *Prot. Struct. Funct. Genet.*, suppl. 1.

Chiu, T. L. and Goldstein, R. A. (1998). Optimizing energy potentials for success in protein tertiary structure prediction. *Folding & Design*, 3:223–228.

Chothia, C. (1976). The nature of the accessible and buried surfaces in proteins. *J. Mol. Biol.*, 105:1–12.

Chothia, C. (1984). Principles that determine the structure of proteins. *Ann. Rev. Biochem.*, 53:537–572.

Chothia, C. (1992). Proteins - 1000 families for the molecular biologist. *Nature*, 357:543–544.

Chou, P. Y. and Fasman, G. D. (1974). Prediction of protein conformation. *Biochemistry*, 13:222–245.

Chou, P. Y. and Fasman, G. D. (1978). Prediction of the secondary structure of proteins from their amino acid sequence. *Adv. Enzymol.*, 47:45–148.

Cohen, F. E., Sternberg, M. J. E., and Taylor, W. R. (1980). Analysis and prediction of protein β-sheet structures by a combinatorial approach. *Nature*, 285:378–382.

Collins, J. F. and Coulson, A. F. W. (1990). Significance of protein sequence similarities. In Doolittle, R. F., editor, *Methods Enzymol.*, volume 183 of *Methods Enzymol.*, pages 474–487. Academic Press, Inc.

Covell, D. G. (1992). Folding protein α-carbon chains into compact forms by Monte Carlo methods. *Prot. Struct. Funct. Genet.*, 14:409–420.

Crippen, G. M. and Havel, T. F. (1988). *Distance Geometry and Molecular Conformation.* Chemometrics Research Studies Press, Wiley, New York.

Dandekar, T. and Leippe, M. (1997). Molecular modeling of amoebapore and NK-lysin: A four-alpha-helix bundle motif of cytolytic peptides from distantly related organisms. *Folding & Design*, 2:47–52.

Dayhoff, M. O., Schwartz, R. M., and Orcutt, B. C. (1978). A model of evolutionary change in proteins. In Dayhoff, M. O., editor, *Atlas of Protein Sequence and Structure*, pages 345–352. Nat. Biomed. Res. Foundation, Washington D.C., USA. Volume 5, Supplement 3.

DeAlmeida, V. I. and Mayo, K. E. (1998). Identification of binding domains of the growth hormone-releasing hormone receptor by analysis of mutant and chimeric receptor proteins. *Mol. Endocrinol.*, 12:750–765.

Dill, K. A. (1985). Theory for the folding and stability of globular proteins. *Biochemistry*, 24:1501–1509.

Doolittle, R. F., editor (1990). *Molecular Evolution: computer analysis of protein and nucleic acid sequences*, volume 183 of *Meth. Enzymol.* Academic Press, San Diego, CA, USA.

Eisenberg, D. and McLachlan, A. D. (1986). Solvation energy in protein folding and binding. *Nature*, 319:199–203.

Eyll, B. V., Goke, B., Wilmen, A., and Goke, R. (1996). Exchange of W39 by A within the N-terminal extracellular domain of the GLP-1 receptor results in a loss of receptor function. *Peptides*, 17:565–570.

Feng, D. F. and Doolittle, R. F. (1987). Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *J. Mol. Evol.*, 25(4):351–360.

Finney, J. L. (1978). Volume occupation, environment, and accessibility in proteins: environment and molecular area of RNase-S. *J. Mol. Biol.*, 119:415–441.

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

Fischer, D., Elofsson, A., Rice, D., and Eisenberg, D. S. (1996a). Assessing the performance of fold recognition methods by means of a comprehensive benchmark. In *Proceedings of the Pacific Symposium on Biocomputing, Hawaii*, pages 300–318.

Fischer, D., Rice, D., Bowie, J. U., and Eisenberg, D. (1996b). Assigning amino-acid-sequences to 3-dimensional protein folds. *Faseb J.*, 10:126–136.

Fiser, A., Cserzo, M., Tudos, E., and Simon, I. (1992). Different sequence environments of cysteines and half cystines in proteins. application to predict disulfide forming residues. *FEBS Lett.*, 302:117–120.

Flores, T. P., Moss, D. S., and Thornton, J. M. (1994). An algorithm for automatically generating protein topology cartoons. *Prot. Engng.*, 7:31–37.

Francesco, V. D., Geetha, V., Garnier, J., and Munson, P. J. (1997). Fold recognition using predicted secondary structure sequences and hidden markov models of protein folds. *Prot. Struct. Funct. Genet.*, suppl. 1:123–128.

Frimurer, T. M., Kostrowicki, J., and Bywater., R. P. (1999). Modelling the GLP1 hormone-receptor complex: Proposed conformation of the hormone and its interaction with the membrane domain of the receptor. in press.

Garnier, J., Osguthorpe, D. J., and 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.

Geourjon, C. and Deleage, G. (1995). SOPMA: significant improvements in protein secondary structure prediction by consensus prediction from multiple alignments. *Comput. Appl. Biosci.*, 11:681–684.

Gerstein, M., Tsai, J., and Levitt, M. (1995). The volume of atoms on the protein surface: calculated from simulation, using voronoi polyhedra. *J. Mol. Biol.*, 249:955–966.

Gibrat, J.-F., Garnier, J., and Robson, B. (1987). Further developments of protein secondary structure prediction using information theory – new parameters and consideration of residue pairs. *J. Mol. Biol.*, 198:425–443.

Godzik, A. and Skolnick, J. (1992). Sequence structure matching in globular-proteins - application to supersecondary and tertiary structure determination. *Proc. Natl. Acad. Sci. USA*, 89:12098–12102.

Gonnet, G. H., Cohen, M. A., and Benner, S. A. (1992). Exhaustive matching of the entire protein-sequence database. *Science*, 256:1443–1445.

Gotoh, O. (1982). An improved algorithm for matching biological sequences. *J. Mol. Biol.*, 162:705–708.

Gouldson, P. R., Snell, C. R., Bywater, R. P., Higgs, C., and Reynolds., C. A. (1998). Domain swapping in G-protein coupled receptor dimers. *Prot. Engng.*, 11.

Graziano, M. P., Hey, P. J., and Strader., C. D. (1996). The amino terminal domain of the glucagon-like peptide-1 receptor is a critical determinant of subtype specificity. *Receptors and Channels*, 4:9–17.

Greer, J. (1981). Comparative model-building of the mammalian serine proteases. *J. Mol. Biol.*, 153:1027–1042.

Greer, J. (1991). Comparative modeling of homologous proteins. *Methods Enzymol.*, 202:239–252.

Guex, N. and Peitsch, M. C. (1997). SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. *Electrophoresis*, 18:2714–2723.

Harrison, P. M. and Sternberg, M. J. (1994). Analysis and classification of disulphide connectivity in proteins. the entropic effect of cross-linkage. *J. Mol. Biol.*, 244:448–463.

Harrison, P. M. and Sternberg, M. J. (1996). The disulphide beta-cross: from cystine geometry and clustering to classification of small disulphide-rich protein folds. *J. Mol. Biol.*, 264:603–623.

Havel, T. F. and Snow, M. E. (1991). A new method for building protein conformations from sequence alignments with homologs of known structure. *J. Mol. Biol.*, 217:1–7.

Havel, T. F. (1993). Predicting the structure of the flavodoxin from escherichia-coli by homology modeling, distance geometry and molecular-dynamics. *Mol. Simulation*, 10:175–210.

Hendlich, M., Lackner, P., Weitckus, S., Floeckner, H., Froschauer, R., Gottsbacher, K., Casari, G., and Sippl, M. J. (1990). Identification of native protein folds amongst a large number of incorrect models – the calculation of low-energy conformations from potentials of mean force. *J. Mol. Biol.*, 216:167–180.

Henikoff, S. and Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. USA*, 89:10915–10919.

Herzyk, P. and Hubbard., R. E. (1995). Automated method for modeling seven-helix transmembrane receptors from experimental data. *Biophys. J.*, 69:2419–2442.

Higgins, D. G., Bleasby, A. J., and Fuchs, R. (1992). CLUSTAL V: improved software for multiple sequence alignment. *CABIOS*, 8:189–191.

Hinds, D. A. and Levitt, M. (1992). A lattice model for protein-structure prediction at low resolution. *Proc. National Academy Sciences United StatesAmerica*, 89:2536–2540.

Holbrook, S. R., Muskal, S. M., and Kim, S.-H. (1990). Predicting surface exposure of amino acids from protein sequence. *Prot. Engng.*, 3:659–665.

Holliger, P. and Riechmann, L. (1997). A conserved infection pathway for filamentous bacteriophages is suggested by the structure of the membrane penetration domain of the minor coat protein g3p from phage fd. *Structure*, 5:265–275.

Holm, L. and Sander, C. (1998). Touring protein fold space with Dali/FSSP. *Nuc. Acids Res.*, 26:316–319.

Horn, F., Bywater, R., Krause, G., Kuipers, W., Oliveira, L., Paiva, A. C. M., Sander., C., and Vriend, G. (1998a). The interaction of class B G-protein-coupled receptors with their hormones. *Receptors and Channels*, 5:305–314.

Horn, F., Weare, J., Beukers, M. W., Horsch, S., Bairoch, A., Chen, W., Edvardsen, O., Campagne, F., and Vriend, G. (1998b). GPCRDB: an information system for G protein-coupled receptors. *Nuc. Acids Res.*, 26:275–279.

Huang, E. S., Subbiah, S., Tsai, J., and Levitt, M. (1996). Using a hydrophobic contact potential to evaluate native and near-native folds generated by molecular dynamics simulations. *J. Mol. Biol.*, 257:716–725.

Hubbard, S. J. and Thornton, J. M. (1993). NACCESS computer program.

Hubbard, T. J. P., Murzin, A. G., Brenner, S. E., and Chothia, C. (1997). SCOP: a structural classification of proteins database. *Nuc. Acids Res.*, 25:236–239.

Huber, T. and Torda, A. E. (1998). Protein fold recognition without Boltzmann statistics or explicit physical basis. *Prot. Sci.*, 7:142–149.

Islam, S. A. and Weaver, D. L. (1990). Molecular interactions in protein crystals: solvent accessible surface and stability. *Prot. Struct. Funct. Genet.*, 8:1–5.

J., S. M., M., K. R., and D., G. P. (1998). Practical evaluation of comparative modelling and threading methods. *Comput. Chem.*, 22:369–375.

Jaroszewski, L., Rychlewski, L., Zhang, B., and Godzik, A. (1998). Fold prediction by a hierarchy of sequence, threading, and modeling methods. *Prot. Sci.*, 7:1431–1440.

Jeanmougin, F., Thompson, J. D., Gouy, M., Higgins, D. G., and Gibson, T. J. (1998). Multiple sequence alignment with CLUSTAL X. *TIBS*, 23:403–405.

Johnson, P. E., Joshi, M. D., Tomme, P., Kilburn, D. G., and McIntosh., L. P. (1996). Structure of the N-terminal cellulose-binding domain of cellulomonas fimi CenC determined by nuclear magnetic resonance spectroscopy. *Biochemistry*, 35:14381–14394.

Jones, T. A. and Thirup, S. (1986). Using known substructures in protein model building and crystallography. *EMBO J.*, 5:819–822.

Jones, D. T. and Thornton, J. M. (1993). Protein fold recognition. *J. Comp. Aided Mol. Desig.*, 7:439–456.

Jones, D. T. and Thornton, J. M. (1996). Potential energy functions for threading. *Curr. Opin. Struct. Biol.*, 6:210–216.

Jones, D. T., Taylor, W. R., and Thornton, J. M. (1992a). A new approach to protein fold recognition. *Nature*, 358:86–89.

Jones, D. T., Taylor, W. R., and Thornton, J. M. (1992b). The rapid generation of mutation data matrices from protein sequences. *CABIOS*, 8:275–282.

Jones, D. T., Orengo, C. A., Taylor, W. R., and Thornton, J. M. (1993). Progress towards recognising protein folds from amino acid sequence. *Prot. Engng.*, 6 (supplement):124. (abstract).

Jones, D. T. (1997). Successful *ab initio* prediction of the tertiary structure of NK-lysin using multiple sequences and recognized supersecondary structural motifs. *Prot. Struct. Funct. Genet.*, suppl. 1:185–191.

Kabsch, W. and Sander, C. (1983). Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22:2577–2637.

Kanaoka, M., Kishimoto, F., Ueki, Y., and Umeyama, H. (1989). Alignment of protein sequences using the hydrophobic core scores. *Prot. Engng.*, 2:347–351.

Karplus, K., Sjolander, K., Barrett, C., Cline, M., Haussler, D., Hughey, R., Holm, L., and Sander, C. (1997). Predicting protein structure using hidden markov models. *Prot. Struct. Funct. Genet.*, suppl. 1:134–139.

Kellis, J. T., Nyberg, K., Sali, D., and Fersht, A. R. (1988). Contribution of hydrophobic interactions to protein stability. *Nature*, 333:784–786.

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

Kim, P. S. and Baldwin, R. L. (1982). Specific intermediates in the folding reactions of small proteins and the mechanism of protein folding. *Ann. Rev. Biochem.*, 51:459–489.

King, R. D. and Sternberg, M. J. E. (1996). Identification and application of the concepts important for accurate and reliable protein secondary structure prediction. *Prot. Sci.*, 5:2298–2310.

Kneller, D. G., Cohen, F. E., and Langridge, R. (1990). Improvements in protein secondary structure prediction by an enhanced neural network. *J. Mol. Biol.*, 214:171–182.

Kolinski, A. and Skolnick, J. (1994). Monte-carlo simulations of protein-folding .1. lattice model and interaction scheme. *Prot. Struct. Funct. Genet.*, 18:338–352.

Krebs, A., Villa, C., Edwards, P. C., and Schertler, G. F. X. (1998). Characterisation of an improved two-dimensional p22121 crystal from bovine rhodopsin. *J. Mol. Biol.*, 282:991–1003.

Kuntz, I. D., Thomason, J. F., and Oshiro, C. M. (1989). Distance geometry. *Meth. Enzymology*, 177:159–204.

Kyte, J. and Doolittle, R. F. (1982). A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.*, 157:105–132.

Lawrence, C. E. and Bryant, S. H. (1991). Hydrophobic potentials from statistical-analysis of protein structures. *Methods Enzymol.*, 202:20–31.

Lee, B. and Richards, F. M. (1971). The interpretation of protein structures: estimation of static accessibility. *J. Mol. Biol.*, 55:379–400.

Lesser, G. J. and Rose, G. D. (1990). Hydrophobicity of amino acid subgroups in proteins. *Prot. Struct. Funct. Genet.*, 8:6–13.

Levine, J. M., Pascarella, S., Argos, P., and Garnier, J. (1993). Quantification of secondary structure prediction improvement using multiple alignments. *Prot. Engng.*, 6:849–854.

Levitt, M. and Chothia, C. (1976). Structural patterns in globular proteins. *Nature*, 261:552–558.

Levitt, M. (1976a). A simplified representation of protein conformations for rapid simulation of protein folding. *J. Mol. Biol.*, 104:59–107.

Levitt, M. (1976b). A simplified representation of protein conformations for rapid simulation of protein folding. *J. Mol. Biol.*, 104:59–107.

Levitt, M. (1997). Competitive assessment of protein fold recognition and alignment accuracy. *Prot. Struct. Funct. Genet.*, suppl. 1:92–104.

Liepinsh, E., Andersson, M., Ruysschaert, J. M., and Otting, G. (1997). Saposin fold revealed by the NMR structure of NK-lysin. *Nature, Struct. Biol.*, 4:793–795.

Lipman, D. J. and Pearson, W. R. (1985). Rapid and sensitive protein similarity searches. *Science*, 227:1435–1441.

Lipman, D. J., Altschul, S. F., and Kececioglu, J. D. (1989). A tool for multiple sequence alignment. *Proc. Natl. Acad. Sci. USA*, 86:4412–4415.

Lomize, A. L., Maiorov, V. N., and Arsen'ev, A. S. (1991). Determination of the spatial structure of insectotoxin 15A from buthus erpeus by (1)H-NMR spectroscopy data. *Bioorg Khim*, 17:1613–1632.

Lüthy, R., Mclachlan, A. D., and Eisenberg, D. (1991). Secondary structure-based profiles - use of structure-conserving scoring tables in searching protein-sequence databases for structural similarities. *Prot. Struct. Funct. Genet.*, 10:229–239.

Mackay, A. L. (1974). Generalized structural geometry. *Acta Cryst.*, A30:440–447.

Marchler-Bauer, A. and Bryant, S. H. (1997). Measures of threading specificity and accuracy. *Prot. Struct. Funct. Genet.*, suppl. 1:74–82.

Marchler-Bauer, A., Levitt, M., and Bryant, S. H. (1997). A retrospective analysis of CASP2 threading predictions. *Prot. Struct. Funct. Genet.*, suppl. 1:83–91.

Mason, S. F. (1984). Origins of biomolecular handedness. *Nature*, 311:19–23.

McDonald, N. Q., Lapatto, R., Murray-Rust, J., Gunning, J., Wlodawer, A., and Blundell, T. L. (1991). New protein fold revealed by a 2.3-Å resolution crystal structure of nerve growth factor. *Nature*, 354:411–414.

Mehta, P. K., Heringa, J., and Argos, P. (1995). A simple and fast apprach to prediction of protein secondary structure from multiply aligned sequences with accuracy above 70%. *Prot. Sci.*, 4:2517–2525.

Miller, W. and Myers, E. W. (1989). Sequence comparison with concave weighting functions. *Bull. Math. Biol.*, 50:97–120.

Miller, S., Janin, J., Lesk, A. M., and Chothia, C. (1987). Interior and surface of monomeric proteins. *J. Mol. Biol.*, 196:641–656.

Mronga, S., Luginbuhl, P., Brown, L. R., Ortenzi, C., Luporini, P., Bradshaw, R. A., and Wuthrich, K. (1994). The NMR solution structure of the pheromone Er-1 from the ciliated protozoan euplotes raikovi. *Prot. Sci.*, 3:1527–1536.

Mumenthaler, C. and Braun, W. (1995). Predicting the helix packing of globular proteins by self-correcting distance geometry. *Prot. Sci.*, 4:863–871.

Murzin, A. G., Brenner, S. E., Hubbard, T., and Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, 247:536–540.

Murzin, A. G. (1994). New protein folds. *Curr. Opin. Struct. Biol.*, 4:441–449.

Muskal, S. M., Holbrook, S. R., and Kim, S.-H. (1990). Prediction of the disulfide-bonding state of cysteine in proteins. *Prot. Engng.*, 3:667–672.

Needleman, S. B. and Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.*, 48:443–453.

Neuwald, A. F., Liu, J. S., Lipman, D. J., and Lawrence, C. E. (1997). Extracting protein alignment models from the sequence database. *Nuc. Acids Res.*, 25:1665–1677.

Nilges, M. (1995). Calculation of protein structures with ambiguous distance restraints. automated assignment of ambiguous NOE crosspeaks and disulphide connectivities. *J. Mol. Biol.*, 245:645–660.

Nishikawa, K. and Matsuo, Y. (1993). Development fo pseudoenergy potentials for assessing protein 3D-1D compatability and detecting weak homologies. *Prot. Engng.*, 6:811–820.

Notredame, C. and Higgins, D. G. (1996). SAGA: sequence alignment by genetic algorithm. *Nuc. Acids Res.*, 24:1515–1524.

Oliveira, L., Paiva, A. C., Sander, C., and Vriend, G. (1994). A common step for signal transduction in G protein-coupled receptors. *Trends Pharmacol. Sci.*, 15:170–172.

Orengo, C. A. and Taylor, W. R. (1996). SSAP: sequential structure alignment program for protein structure comparison. In Doolittle, R. F., editor, *Computer methods for macromolecular sequence analysis*, volume 266 of *Meth. Enzymol.*, pages 617–635. Academic Press, Orlando, FA, USA.

Orengo, C. A., Michie, A. D., Jones, S., Jones, D. T., Swindells, M. B., and Thornton, J. M. (1997). CATH – A hierarchic classification of protein domain structures. *Structure*, 5:1093–1108.

Orengo, C. A. (1994a). Classification of protein folds. *Curr. Opin. Struct. Biol.*, 4:429–440.

Orengo, C. A. (1994b). Test list of fold families. *By communication.*

Paolo, E. D., Neef, P. D., Moguilevsky, N., Petry, H., Bollen, A., Waelbroeck, M., and Robberecht., P. (1998). Contribution of the second transmembrane helix of the secretin receptor to the positioning of secretin. *FEBS Lett.*, 424:207–210.

Pascarella, S. and Argos, P. (1992). Analysis of insertions/deletions in protein structures. *J. Mol. Biol.*, 224:461–471.

Pascarella, S., Persio, R. D., Bossa, F., and Argos, P. (1998). Easy method to predict solvent accessibility from multiple protein sequence alignments. *Prot. Struct. Funct. Genet.*, 32:190–199.

Pazos, F., Helmer-Citterich, M., Ausiello, G., and Valencia., A. (1997). Correlated mutations contain information about protein-protein interactions. *J. Mol. Biol.*, 271:511–523.

Pearl, L. H. and Taylor, W. R. (1987a). Sequence specificity of retroviral proteases. *Nature*, 328:482. (Communication).

Pearl, L. H. and Taylor, W. R. (1987b). A structural model for the retroviral proteases. *Nature*, 329:351–354.

Pearson, W. R. and Lipman, D. J. (1988). Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. USA*, 85:2444–2448.

Pearson, W. R. and Miller, W. (1992). Dynamic programming algorithms for biological sequence comparison. In Brand, L. and Johnson, M. L., editors, *Numerical Computer Methods*, volume 210 of *Methods Enzymol.*, chapter 27, pages 575–601. Academic Press Inc., N.Y.

Pearson, W. R. (1990). Rapid and sensitive sequence comparison with FASTP and FASTA. In Doolittle, R. F., editor, *Molecular Evolution: computer analysis of protein and nucleic acid sequences*, volume 183 of *Methods Enzymol.*, chapter 5, pages 63–98. Academic Press, Inc.

Peitsch, M. C. (1996). ProMod and Swiss-Model: Internet-based tools for automated comparative protein modelling. *Biochem. Soc. Trans.*, 24:274–279.

Perez, J. J., Filizola, M., and Cariteni-Farina., M. (1998). A general procedure for building the transmembrane domains of G-protein coupled receptors. *J. Math. Chem.*, 23:229–238.

Pollock, D. and Taylor, W. R. (1997). Effectiveness of correlation analysis in identifying protein residues undergoing correlated evolution. *Prot. Engng.*, 10:647–657.

Polshakov, V. I., Williams, M. A., Gargaro, A. R., Frenkiel, T. A., Westley, B. R., Chadwick, M. P., May, F. E., and Feeney, J. (1997). High-resolution solution structure of human pNR-2/ps2: a single trefoil motif protein. *J. Mol. Biol.*, 267:418–32.

Presnell, S. R. and Cohen, F. E. (1989). Topological distribution of four-α-helical bundles. *Proc. Natl. Acad. Sci. USA*, 86:6592–6596.

Qian, N. and Sejnowski, T. J. (1988). Predicting the secondary structure of globular proteins using neural network models. *J. Mol. Biol.*, 202:865–884.

Reinert, K., Lenhof, H.-P., Mutzel, P., Mehlhorn, K., and Kececioglu, J. D. (1997). A branch-and-cut algorithm for multiple sequence alignment. *Recomb. 97*, pages 241–249.

Reva, B. A., Finkelstein, A. V., Sanner, M., Olson, A. J., and Skolnick, J. (1997). Recognition of protein structure on coarse lattices with residue–residue energy functions. *Prot. Engng.*, 10:1123–1130.

Reva, B., Rykunov, D., Finkelstein, A. V., and Skolnick, J. (1998). Optimization of protein structure on lattices using a self-consistent field approach. *J. Comput. Biol.*, 5:531–538.

Rice, D. and Eisenberg, D. (1997). A 3D-1D substitution matrix for protein fold recognition that includes predicted secondary structure of the sequence. *J. Mol. Biol.*, 267:1026–1038.

Richardson, J. S. and Richardson, D. C. (1988). Amino acid preferences for specific locations at the ends of alpha helices. *Science*, 240:1648–1652.

Richardson, J. S. (1977). β-Sheet topology and the relatedness of proteins. *Nature*, 268:495–500.

Rooman, M. and Gilis, D. (1998). Different derivations of knowledge–based potentials and analysis of their robustness and context–dependent predictive power. *Eur. J. Biochem.*, 254:135–143.

Rose, G. D., Geselowitz, A. R., Lesser, G. J., Lee, R. H., and Zehfus, M. H. (1985). Hydrophobicity of amino acid residues in globular proteins. *Science*, 229:834–838.

Rost, B. and Sander, C. (1993). Prediction of protein secondary structure at better than 70-percent accuracy. *J. Mol. Biol.*, 232:584–599.

Rost, B. and Sander, C. (1994). Conservation and prediction of solvent accessibility in protein families. *Prot. Struct. Funct. Genet.*, 20:216–226.

Rost, B. (1995). TOPITS: Threading one-dimensional predictions into three-dimensional structures. In Rawlings, C., Clark, D., Altman, R., Hunter, L. Lengauer, T., and Wodak, S., editors, *The third international conference on Intelligent Systems for Molecular Biology (ISMB)*, pages 314–321. AAAI Press, Menlo Park, CA, USA. Cambridge, U.K., Jul 16-19.

Russell, R. B., Copley, R. R., and Barton, G. J. (1996). Protein fold recognition by mapping predicted secondary structures. *J. Mol. Biol.*, 259:349–365.

Russell, R. B., Saqi, M. A. S., Bates, P. A., Sayle, R. A., and Sternberg, M. J. E. (1998). Recognition of analogous and homologous protein folds – assessment of prediction success and associated alignment accuracy using empirical substitution matrices. *Prot. Engng.*, 11:1–9.

Ruysschaert, J. M., Goormaghtigh, E., Homble, F., Andersson, M., Liepinsh, E., and Otting, G. (1998). Lipid membrane binding of NK-lysin. *FEBS Lett.*, 425:341–344.

Saitou, N. and Nei, M. (1987). The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.*, 4:406–425.

Salamov, A. A. and Solovyev, V. V. (1995). Prediction of protein secondary structure by combining nearest-neighbor algorithms and multiple sequence alignments. *J. Mol. Biol.*, 247:11–15.

Sali, A. and Blundell, T. L. (1990). Definition of general topological equivalence in protein structures: a procedure involving comparison of properties and relationship through simulated annealing and dynamic programming. *J. Mol. Biol.*, 212:403–428.

Sali, A. and Blundell, T. L. (1993). Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.*, 234:779–815.

Sali, A., Overington, J. P., Johnson, M. S., and Blundell, T. L. (1990). From comparisons of protein sequences and structures to protein modeling and design. *TIBS*, 15:235–240.

Sali, A. (1995). Modeling mutations and homologous proteins. *Curr. Opin. Biotechnol.*, 6:437–451.

Sanchez, R. and Sali, A. (1998). Large-scale protein structure modeling of the saccharomyces cerevisiae genome. *Proc. Natl. Acad. Sci. USA*, 95:13597–13602.

Saqi, M. A. S., Bates, P. A., and Sternberg, M. J. E. (1992). Towards an automatic method of predicting protein-structure by homology - an evaluation of suboptimal sequence alignments. *Prot. Engng.*, 5:305–311.

Saraste, M., Sibbald, P. R., and Wittinghofer, A. (1990). The p-loop – a common motif in ATP- and GTP-binding proteins. *TIBS*, 15:430–434.

Sellers, P. H. (1974). On the theory and computation of evolutionary distances. *SIAM J. Appl. Math.*, 26:787–793.

Shindyalov, I. N., Kolchanov, N. A., and Sander, C. (1994). Can three-dimensional contacts in protein structures be predicted by analysis of correlated mutations? *Prot. Engng.*, 7:349–358.

Shrive, A. K., Polikarpov, I., Krell, T., Coulson, A., Coggins, J. R., Hawkins, A. R., and Sawyer, L. Three-dimensional structure of type I dehydroquinase - an enzyme recruited for a role in eukaryotic transcription regulation. *unpublished results.*

Singer, M. S., Oliveira, L., Vriend., G., and Shepherd, G. (1995). Potential ligand-binding residues in rat olfactory receptors identified by correlated mutation analysis. *Receptors and Channels*, 3:89–95.

Sippl, M. J. and Weitckus, S. (1992). Detection of native-like models for amino-acid-sequences of unknown 3-dimensional structure in a data-base of known protein conformations. *Prot. Struct. Funct. Genet.*, 13:258–271.

Sippl, M. J. (1990). Calculation of conformational ensembles from potentials of mean force. an approach to the knowledge-based prediction of local structures in globular proteins. *J. Mol. Biol.*, 213:859–883.

Skolnick, J. and Kolinski, A. (1991). Dynamic monte-carlo simulations of a new lattice model of globular protein folding, structure and dynamics. *J. Mol. Biol.*, 221:499–531.

Smith, T. F. and Waterman, M. S. (1981a). Comparison of bio-sequences. *Adv. Appl. Math.*, 2:482–489.

Smith, T. F. and Waterman, M. S. (1981b). Identification of common molecular subsequences. *J. Mol. Biol.*, 147:195–197.

Soppa., J. (1994). Sequence comparison does not support an evolutionary link between halobacterial retinal proteins including bacteriorhodopsin and eukaryotic G-protein-coupled receptors. *Febs. Lett.*, 342:7–11.

Srinivasan, S., March, C. J., and Sudarsanam, S. (1993). An automated method of modeling proteins on known templates using distance geometry. *Prot. Sci.*, 2:277–289.

Sternberg, M. J. E., Cohen, F. E., and Taylor, W. R. (1982). A combinatorial approach to the prediction of the tertiary fold of globular proteins. *Biochem. Soc. Trans.*, 10.

Sudarsanam, S., March, C. J., and Srinivasan, S. (1994). Homology modeling of divergent proteins. *J. Mol. Biol.*, 241:143–149.

Sukumar, M., Rizo, J., Wall, M., Dreyfus, L. A., Kupersztoch, Y. M., and Gierasch, L. M. (1995). The structure of escherichia coli heat-stable enterotoxin b by nuclear magnetic resonance and circular dichroism. *Prot. Sci.*, 4:1718–1729.

Sutcliffe, M. J., Haneef, I., Carney, D., and Blundell, T. L. (1987a). Knowledge based modelling of homologous proteins: I. 3-dimensional frameworks derived from the simultaneous superposition of multiple structures. *Prot. Engng.*, 1:377–384.

Sutcliffe, M. J., Hayes, F. R. F., and Blundell, T. L. (1987b). Knowledge based modelling of homologous proteins: II. rules for the conformations of substituted side-chains. *Prot. Engng.*, 1:385–392.

Swindells, M. B. (1995). A procedure for the automatic determination of hydrophobic cores in protein structures. *Prot. Sci.*, 4:93–102.

Tanford, C. (1997). How protein chemists learned about the hydrophobic factor. *Prot. Sci.*, 6:1358–1366.

Taylor, W. R. and Aszódi, A. (1994a). Building protein folds using distance geometry: towards a general modeling and prediction method. In Merz, K. M. and LeGrand, S. M., editors, *The protein folding problem and tertiary structure prediction*, chapter 6, pages 165–192. Birkhäuser (Springer-Verlag), Boston, MA., USA.

Taylor, W. R. and Aszódi, A. (1994b). Modelling and predicting protein structure using distance geometry. In Bohr, H. and Brunak, S., editors, *Distance based approaches to protein structure determination*, pages 213–221. IOS press, Elsevier (Amsterdam).

Taylor, W. R. and Hatrick, K. (1994). Compensating changes in protein multiple sequence alignments. *Prot. Engng.*, 7:341–348.

Taylor, W. R. and Jones, D. T. (1993). Deriving an amino acid distance matrix. *J. Theor. Biol.*, 164:65–83.

Taylor, W. R. and Jones, D. T. (1995). Modelling and predicting α-helical transmembrane structures. In Bohr, H. and Brunak, S., editors, *Protein Folds: A Distance Based Approach*, pages 283–293. CRC Press Inc., Boca Raton, Florida, U.S.A.

Taylor, W. R. and Munro, R. E. J. (1997). Multiple sequence threading: Conditional gap placement. *Folding & Design*, 2, suppl.:S33–S39.

Taylor, W. R. and Orengo, C. A. (1989). Protein structure alignment. *J. Mol. Biol.*, 208:1–22.

Taylor, W. R. (1986). The classification of amino acid conservation. *J. Theor. Biol.*, 119:205–218.

Taylor, W. R. (1988). A flexible method to align large numbers of biological sequences. *J. Mol. Evol.*, 28:161–169.

Taylor, W. R. (1991). Towards protein tertiary fold prediction using distance and motif constraints. *Prot. Engng.*, 4:853–870.

Taylor, W. R. (1993). Protein fold refinement: building models from idealised folds using motif constraints and multiple sequence data. *Prot. Engng.*, 6:593–604.

Taylor, W. R. (1994). Protein structure modelling from remote sequence similarity. *J. Biotechnol.*, 12:281–291. Special issue: *Genome Research/Molecular Biotechnology*, Part 1. ed., Helmut Blöcker.

Taylor, W. R. (1995a). An investigation of conservation-biased gap-penalties for multiple protein sequence alignment. *Gene*, 165:GC27–35. Internet journal Gene Combis: `http://www.elsevier.nl/locate/genecombis`.

Taylor, W. R. (1995b). Sequence alignment of proteins and nucleic acids. In Meyers, R. A., editor, *Molecular Biology and Biotechnology: comprehensive desk reference*, pages 856–859. VCH, New York, USA.

Taylor, W. R. (1996). Multiple protein sequence alignment: algorithms for gap insertion. In Doolittle, R. F., editor, *Computer methods for macromolecular sequence analysis*, volume 266 of *Meth. Enzymol.*, pages 343–367. Academic Press, Orlando, FA, USA.

Taylor, W. R. (1997). Multiple sequence threading: an analysis of alignment quality and stability. *J. Mol. Biol.*, 269:902–943.

Taylor, W. R. (1998a). Dynamic sequence databank searching with templates and multiple alignment. *J. Mol. Biol.*, 280:375–406.

Taylor, W. R. (1998b). Protein structural domain identification. *Prot. Engng.*, in press.

Taylor, W. R. (1999). Protein structure comparison using iterated double dynamic programming. *Prot. Sci.* in press.

Teller, D. C. (1976). Accessible area, packing volumes and interaction surfaces of globular proteins. *Nature*, 260:729–731.

Thompson, M. J. and Goldstein, R. A. (1996). Predicting solvent accessibility: higher accuracy using bayesian statistics and optimized residue substitution classes. *Prot. Struct. Funct. Genet.*, 25:38–47.

Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994). CLUSTAL-W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nuc. Acids Res.*, 22:4673–4680.

Thompson, J. D., Gibson, T., Plewniak, F., Jeanmougin, F., and Higgins, D. G. (1997). CLUSTAL-X windows interface: Flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nuc. Acids Res.*, 25:4876–4882.

Thornton, J. M. (1981). Disulphide bridges in globular proteins. *J. Mol. Biol.*, 151:261–287.

Torda, A. E. (1997). Perspectives in protein-fold recognition. *Curr. Opin. Struct. Biol.*, 7:200–205.

Tsukihara, T., Aoyama, H., Yamashita, E., Tomizaki, T., Yamaguchi, H., Shinzawa-Itoh, K., Nakashima, R., Yaono, R., and Yoshikawa, S. (1996). The whole structure of the 13-subunit oxidized cytochrome c oxidase at 2.8å˚ *Science*, 272:1136–1144.

Unger, V. M., Hargrave, P. A., Baldwin, J. M., and Schertler, G. F. X. (1997). Arrangement of rhodopsin transmembrane alpha-helices. *Nature*, 389:203–206.

Vath, G. M., Earhart, C. A., Rago, J. V., Kim, M. H., Bohach, G. A., Schlievert, P. M., and Ohlendorf, D. H. (1997). The structure of the superantigen exfoliative toxin A suggests a novel regulation as a serine protease. *Biochemistry*, 36:1559–1566.

Vilardaga, J. P., Neef, P. D., Paolo, E. D., Bollen, A., Waelbroeck, M., and Robberecht., P. (1995). Properties of chimeric secretin and VIP receptor proteins indicate the importance of the N-terminal domain for ligand discrimination. *Biochem. Biophys. Res. Commun.*, 211:885–891.

Vilardaga, J. P., Paolo, E. D., Bialek, C., Neef, P. D., Waelbroeck, M., Bollen, A., and Robberecht., P. (1997). Mutational analysis of extracellular cysteine residues of rat secretin receptor shows that disulfide bridges are essential for receptor function. *Eur. J. Biochem.*, 246:173–180.

Vingron, M. and Waterman, M. S. (1994). Sequence alignment and penalty choice: review of concepts, case-studies and implications. *J. Mol. Biol.*, 235:1–12.

Vriend, G. (1990). WHAT IF: a molecular modeling and drug design program. *J. Mol. Graph.*, 8:52–56.

Wako, H. and Blundell, T. L. (1994a). Use of amino acid environment-dependent substitution tables and conformational propensities in structure prediction from aligned sequences of homologous proteins. i. solvent accessibility classes. *J. Mol. Biol.*, 238:682–692.

Wako, H. and Blundell, T. L. (1994b). Use of amino acid environment-dependent substitution tables and conformational propensities in structure prediction from aligned sequences of homologous proteins. ii. secondary structures. *J. Mol. Biol.*, 238:693–708.

Weber, I. T. (1990). Evaluation of homology modeling of HIV preotease. *Prot. Struct. Funct. Genet.*, 7:172–184.

Westhead, D. R. and Thornton, J. M. (1998). Protein structure prediction. *Curr. Opin. Biotechnol.*, 9:383–389.

Wilmen, A., Goke, B., and Goke, R. (1996). The isolated N-terminal extracellular domain of the glucagon-like peptide-1 (GLP)-1 receptor has intrinsic binding activity. *FEBS Lett.*, 398:43–47.

Wilmen, A., Eyll, B. V., Goke, B., and Goke, R. (1997). Five out of six tryptophan residues in the N-terminal extracellular domain of the rat GLP-1 receptor are essential for its ability to bind GLP-1. *Peptides*, 18:301–305.

Wilmot, C. M. and Thornton, J. M. (1988). Analysis and prediction of the different types of beta-turn in proteins. *J. Mol. Biol.*, 203:221–232.

Yee, V. C., Pratt, K. P., Cote, H. C., Trong, I. L., Chung, D. W., Davie, E. W., Stenkamp, R. E., and Teller, D. C. (1997). Crystal structure of a 30 kDa C-terminal fragment from the gamma chain of human fibrinogen. *Structure*, 5:125–38.

Zhang, X., Boyar, W., Toth, M. J., Wennogle, L., and Gonnella, N. C. (1997). Structural definition of the C5a C terminus by two-dimensional nuclear magnetic resonance spectroscopy. *Prot. Struct. Funct. Genet.*, 28:261–267.

Zimmer, R., Wohler, M., and Thiele, R. (1998). New scoring schemes for protein fold recognition based on voronoi contacts. *Bioinformatics*, 14:295–308.

Zvelebil, M. J., Barton, G. J., Taylor, W. R., and Sternberg, M. J. E. (1987). Prediction of protein secondary structure and active sites using the alignment of homologous sequences. *J. Mol. Biol.*, 195:957–961.