Evolutionary Analysis of Protein Structural Families to Assist Comparative Modelling of Genome Sequences

Gabrielle Anne Reeves

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

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

August 2003

Abstract

The CATH domain database clusters closely related structures (>35% sequence identity) into families. More distant evolutionary links between these families are identified by common sequence patterns, functional and structural motifs in order to cluster further into homologous superfamilies. Relatives in these homologous superfamilies share core structural similarity. However, in some superfamilies extensive structural embellishments are observed. This thesis presents an analysis of the structural variability of the homologous superfamilies in the CATH database, focusing on the secondary structure embellishments present in many of the more variable families. It was found that secondary structure elements are inserted into a number of places in the peptide chain but are often co-located on the three-dimensional structure. Using this information, a protocol is developed to correlate the structural embellishments with the functional changes observed in three particularly variable families; the ATP-dependent carboxylase-amine/thiol ligase superfamily, the cupredoxin superfamily and the thioredoxin superfamily. A number of conclusions are drawn from this structural analysis, the embellishments often mediate the domain interfaces, illustrated in the cupredoxin and ATP-grasp superfamilies. Additionally, modifications to the active sites occur through the additions of secondary structure elements. In the ATP-grasp superfamily, a large embellishment encloses the active site in some members.

Experimental techniques for solving the three dimensional structure of a protein, primarily NMR and X-ray crystallography, are often hampered by technical limitations making them time consuming, and so comparative modelling techniques are being explored to create theoretical three-dimensional structural models. The second part of this thesis considers ways of modelling genome sequences, with assignments to CATH homologous superfamilies, by comparative modelling. An automatic comparative modelling pipeline has been developed where genome sequences are aligned and modelled using publicly available software in an optimised protocol (GenMod). GenMod was tested using a large dataset of 140 relatives from CATH superfamilies. Software to assess the quality of these models was selected and tested. One of the main areas reported to need improvement in current comparative modelling techniques is parent selection and here, a novel method is explored. Sets of parent structures have been created from structural sub-groups within each homologous superfamily. Regions from each of these parents were then selected by sequence similarity to create a final structural template. Results from the analysis showed that, below 30% none of the methods perfomed well, above 55% the closest relative is the best parent and between 30 and 55% the best method uses multiple parents.

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

Acknowledgements

First and foremost, thanks to Christine Orengo for the supervision and to Andrew Martin for making me feel so welcome at Reading and at his home for chapter chats as well as a wealth of other subjects.

UCL has meant two things, great work and great friends. To everyone in Janet Thornton and Christine Orengo's groups, your banter and bioinformatics have helped me enormously. But to mention a few heroes...

Ian Sillitoe and James Bray, who made me feel so welcome from my very first day. James for his good ideas and his endless pocket sized programs and Ian who has been so generous with his time answering my coding questions, of which there were many. Thanks also for the support you have given me throughout the write-up. Russell Marsden, a recent joiner to the lab hasn't had a moment's peace since he arrived. I would be some way from Solla Sollew without your suggestions, fantastic support and not to mention your very polite way of telling me to just do it! Annabel Todd for four years of friendship, chats, chapters comments, the placing of commas and other such fiddly things. Dan Buchan who arrived a week earlier than me, bagged the best desk and from then seemed to be just ahead of me at every turn. SWOT! Oliver Redfern for brightening up the long dark write-up with his endless hilarious stories of the Redfern family and latest updates on his poor sickly constitution. Stuart Rison for his great write up advice (more colour = less critical (I hope it works...)) and the LaTeX tips. Adrian Shepherd for the support and gossip even after he left for the dark side. Neil Stoker for asking how I was AND listening to the answer... Mark Dibley for perspective – I always knew who to come to when things were beginning to get out of hand. Stephen Campbell for being a great friend, Tina Eyre and Patricia Furtardo for the female support, Brian Ferguson for the Friday night fun in the JB and Thomas Kabir for sharing the write-up pain with me. Thanks also to Alistair Grant, Andrew Harrison, Frances Pearl, Mike Plevin, Sarah Addou, Juan Antonio, Chris Bennett, Ilhem Diboun, Adrian EkoUkeh, Dave Lee, Stefano Lise and Stathis Sideris who all made the lab an enjoyable place to be.

To those behind the scenes: I would not have managed without the help of Duncan McKenzie who laughed at my jokes, John Bouquiere who did not laugh when I broke two computers in my first week and the three newest arrivals, Jahid Ahmed and Donovan Binns and Jessie Oldershaw who cleared up after my blunders.

And of course not forgetting all those at the EBI who I still consider to be part of the BSM unit at UCL. We don't like to think of you as in any way inferior (even though you are). Most importantly to Gail Bartlett for keeping a foot in both camps and to Hugh Shanahan for the wonderful emails and his very generous help with the EBI software.

A very special thank you to Chris Taggart (and his credit card) who have been so unbelievably supportive whilst suffering bravely throughout many thesis rants and trips to Gap. You have kept me from the doors of insanity with your unique perspective on life.

Finally, to my family. To Rachel and Rob Greening for making me laugh and showing me that compared to them, I am sane. Finally, to my parents, Janet and Ken, I dedicate this thesis to them for their encouragement, patience, love, hugs, the confidence to start the PhD in the first place and the support to finish it.

I've bought a big bat. I'm all ready, you see.
Now my troubles are going to have troubles with me!

Contents

List of Figures

List of Tables

List of Abbreviations

Abbreviation	Details
2D	Two-dimensional
3D	Three-dimensional
Å	Angstrom
ATP	Adenosine triphosphate
BLAST	Basic Local Alignment Search Tool
BLOSUM	Blocks Substitution Matrices
C-terminal	Carboxy-terminal
CASP	Critical Assessment of Methods of Protein Structure Prediction
CATH	Class, Architecture, Topology, Homologous Superfamily
CORA	Conserved Residue Attributes
DDP	Double Dynamic Programming
DP	Dynamic Programming
EC	Enzyme Classification
EBI	European Bioinformatics Institute
GO	Gene Ontology
HMM	Hidden Markov Model
HOMSTRAD	Homologous Structure Alignment Database
HSP	High Scoring Segment Pairs
N-terminal	Amino-terminal
NAD	Nicotinamide Adenine Dinucleotide
NCBI	National Center for Biotechnology Information
NMR	Nuclear Magnetic Resonance
NRDB	Non-redundant Database
PAM	Point Accepted Mutation
PDB	Protein Data Bank
PSSM	Position Specific Score Matrices
PSI-BLAST	Position Specific Iterated-BLAST
RMSD	Root Mean Squared Deviations
RNA	Ribonucleic Acid
PDB	Protein Databank
SAM	Sequence Alignment and Modelling
SCOP	Structural Classification Of Proteins
SSAP	Sequential Structural Alignment Program
SSE	Secondary Structure Element
STAMP	Structural Alignment of Multiple Proteins
SSG	Structural Sub-Group
TIM	Triosephosphate Isomerase
VAST	Vector Alignment Search Tool
WWW	World Wide Web

List of Amino Acid Abbreviations

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

Chapter 1

Introduction

1.1 The Hierarchy of Protein Structure

In 1958 the first protein structure, myoglobin, was solved by X-ray crystallography by Kendrew *et al.* (1958). This provided the surprising result of a more complex, asymmetric structural arrangement than the simple and regular structure of the DNA double helix elucidated five years previously. Over forty years later, with many structures now solved and functionally characterised, it is possible to see that the structural irregularity is required in order that proteins can fulfil their diverse functional roles as they are integral to every biochemical process in life. They can serve as modules for building up large assemblies such as virus particles or muscle fibres, function as ion channels in the cell walls, as electron carriers in the respiratory chain or provide specific enzymatic activities within the cell dominating the biochemistry of our cells. They have evolved through selective pressure to perform specific functions and this function depends on their three-dimensional structure. In solution proteins often form globular structures. Protein structure can be explained by an underlying hierarchy that ranges from primary to quaternary structure.

1.1.1 Primary Structure

The primary structure describes the sequence of amino acids along the polypeptide chain. Amino acids have a central carbon atom (C_α) which is connected to a hydrogen atom, a carboxyl group (COOH) and an amino group (NH_2). The fourth valence is occupied by one of 20 'sidechains', varying in chemical properties. These properties can be grouped into three main classes (Branden, 1999): amino acids with strictly hydrophobic sidechains (Ala, Val, Leu, Ile, Pro, Phe and Met), those with charged sidechains (Asp, Glu, Arg and Lys), and the amino acids with polar sidechains (Ser, Thr, Cys, Asn, Gln, His, Tyr and Trp). Gly is an exception in that it contains only a hydrogen atom as its sidechain, and as such is either placed in its own class or considered as a hydrophobic amino acid. These twenty amino acids form the primary structure of a protein that in turn encodes its uniquely folded three-dimensional structure (Anfinsen, 1973), thus bestowing the huge variety of protein structures and functions in nature. The polypeptide chain is created by a condensation reaction between the carboxyl and amino groups forming a peptide bond.

1.1.2 Secondary Structure

The force that drives the folding of water-soluble globular proteins is the packing of hydrophobic sidechains into the interior of the molecule creating a hydrophobic core and a hydrophilic surface. However, the burial of hydrophobic sidechains is also accompanied by the burial of their main chain atoms which include polar N-H and C=O groups. These polar groups are neutralised by the formation of hydrogen bonds between them. This gives rise to regular patterns of hydrogen bonding or secondary structure elements. The route taken by the polypeptide chain in three-dimensional space to create the secondary structure can be described by the relative positions of three atoms linked within the backbone: the C-α (C_α), carbonyl carbon (C') and amide nitrogen (N) atoms. The relative positions or angles of rotation between these atoms are described as the ϕ angle (around the N-C-α bond), and the ψ angle (around the C-α-C' bond). Secondary structure is defined as two main types, the α-helix and the β-sheet but there are also a number of less stable secondary structures found in protein structure.

α-helix Here the C=O group (residue *i*) and the N-H group (residue *i+4*) hydrogen bond to form a cylindrical structure of the peptide chain, with approximately 3.6 residues per turn, corresponding to a distance of 5.4Å. The helix forms a right handed turn, with ψ and ϕ angles of -60^o and -50^o respectively.

β-sheet This secondary structure type is made up of two or more continuous regions of chain called β-strands, which are found in a fully extended conformation. β-strands line up in such a way as to allow hydrogen bond formation between adjacent C=O and N-H groups. β-sheets are built up of β-strands arranged parallel, anti-parallel, or a mixture of both. Parallel β-sheets have average ϕ, ψ angles of -119^o and 113^o respectively and anti-parallel β-strands, an average of between -139^o and 135^o.

3_{10} helix These helices are always short and frequently occur at the termini of regular α-helices. The name 3_{10} describes the three residues along the chain that form a pairwise hydrogen bond (C=O$_i$ to N-H$_{i+3}$) and the ten atoms that are enclosed in a ring formed by each hydrogen bond. The dipoles of the 3_{10}-helix are not so well aligned as in the α-helix, *i.e.* it is a less stable structure and sidechain packing is less favourable.

π-helix The π-helix is an extremely rare secondary structural element sometimes found on the ends of α-helices. Hydrogen bonds occur between $C{=}O_i$ and $N\text{-}H_{i+5}$. The ϕ and ψ angles of the pure π helix (-57.1, -69.7) lie at the very edge of an allowed minimum energy before unfavourable steric clashes occur.

β-turn The β-turn describes a turn of the peptide chain upon itself which is stabilised by hydrogen bonding. Usually, this region contains glycine, providing almost no steric hindrance, and proline in which the sidechain hydrogen bonds to its main chain N atom, forcing the bend in the chain.

1.1.3 Super-secondary Structure

Adjacent secondary structures can in turn assemble to form super-secondary structures or 'motifs'. Some of these motifs can be associated with a specific function and others form the building blocks of larger structural and functional assemblies.

β-hairpins β-hairpins are the simplest motif involving β-strands and consisting of two adjacent anti-parallel strands joined by a loop. This motif occurs very frequently either as an isolated ribbon or present as part of a more complex β-sheet. β-hairpin loops adopt specific conformations which depend on their lengths and sequences. Sibanda & Thornton (1985) have shown that 70% of β-hairpins are less than 7 residues in length with the two-residue turns forming the most noticeable component. Several consecutive anti-parallel β-strands form a super-secondary structure known as the β-meander.

Helix-turn-helix This is a functional motif also known as the EF hand. It was discovered that in parvalbumin two of the three helix-turn-helix motifs present in the structure are involved in binding calcium using the carboxyl sidechains and main chain carbonyl groups.

Helix-loop-helix motifs This motif was first observed in prokaryotic DNA binding proteins such as the cro repressor from phage lambda. This protein forms a dimer and each subunit consists of an anti-parallel three stranded β-sheet with three helical segments inserted sequentially between the first and second β-strands. The dimer forms so that the second helix from each monomer are located on one side of the sheet at the correct distance to fit into adjacent major groves in the DNA. Many other helix-turn-helix proteins with different folds exhibit essentially the same mode of binding to DNA.

β-α-β motifs Parallel β-strands are connected by longer regions of chain which cross the β-sheet and frequently contain α-helical segments. This motif is called the β-α-β motif and is found in most proteins that have a parallel β-sheet. The helix axis is roughly parallel with the β-strands and all three elements of secondary structure interact forming a hydrophobic core. In certain proteins the loop linking the carboxy terminal end of the first β-strand to the amino terminal end of the α-helix is involved in binding of ligands or substrates.

1.1.4 Tertiary Structure

The packing together of secondary structure elements or larger super-secondary motifs results in the tertiary structure of a protein which contains an individual hydrophobic core that is made up of secondary structure elements. This is known as a domain. Protein domains are often described as the fundamental units of protein structure, forming high-order building blocks of the protein polypeptide chain. The domain can be described as a semi-independent folding unit (Richardson, 1981), with a well packed hydrophobic core. Residue contacts between domains are less than those within domains and secondary structures are rarely shared between domains (Taylor, 1999). Secondary structure elements are connected by exposed loop regions that are usually much less conserved, unless involved in the function of the protein. Domains can be placed into different classes according to their secondary structure content. Four main classes were originally described by Levitt & Chothia (1976). All-α domains comprise mostly α-helices and are often small folds in which the α-helices are usually arranged in bundles packing against one another to form a globular core. All-β domains, comprise almost entirely of β-sheets normally in an anti-parallel arrangement within the domain core. β-sheets can pack against one another, with the hydrophobic sidechains located at the interface, forming β-sandwiches. $\alpha\beta$ domains are built up of a repeating β-α-β motifs that results in the outer layer of the structure being composed of amphipathic α-helices, that pack around the central core of β-sheets. α+β domains, like $\alpha\beta$ domains, contain α-helices and β-sheets, however the arrangement of these elements is mixed. The classification of these domains can be complicated by the fact that there are overlaps between this class and the $\alpha\beta$ class and so these classes are sometimes merged (Orengo *et al.*, 1997).

1.1.5 Quaternary Structure

Many proteins have a quaternary structure that is based on the association and interaction of two or more polypeptide chains that form an oligomeric complex. The formation of multi subunit complexes for the activity of the protein provides evolutionary advantages as the interactions of the chains can be more transient or reversible than the interactions between domains. This provides mechanisms such as allosteric control, higher active site concentrations, new active sites at subunit interfaces, and an economic way to produce protein interaction networks and molecular machines (Liu & Eisenberg, 2002).

1.2 Protein Domains

Domains are considered to be evolutionary units and sequence-based analyses have demonstrated that some domains have ancient origins because they are widespread in all three forms of cellular life, archaea, bacteria and eukarya whose common ancestor is thought to have existed over three billion years ago. This suggests that these domains are either susceptible to adaptation enabling them to fill many functional niches or that they fulfil essential functional processes. Many enzymatic domains of central metabolism $(\beta/\alpha)_8$ TIM barrels, flavoproteins and Rossman-like fold proteins appear to owe their heritage to ancestors that precede the last common ancestor of archaea, bacteria and eukarya (Doolittle & Brown, 1994). Very little is known about what the precursors to these structural units might have been. Lupas *et al.* (2001) examined the evolution of protein folds by locating structure/sequence similar motifs within proteins with different folds suggesting that the diversity of domain folds in existence today might have evolved from the conglomerates of peptide segments that are seen today as internal repeats and structure-integrated motifs.

1.3 Convergent and Divergent Evolution

It appears that during evolution many domains have diverged from a common ancestor to such an extent that they now show little sequence similarity and their structural similarity may be limited to the core structure only. By contrast, there are many cases of convergent evolution, where common structures have reinvented themselves. One of the most interesting examples of convergence is thermolysin and mitochondrial processing peptidase (Makarova & Grishin, 1999). These proteins show striking similarities in their active site residues and also in the arrangement and packing of the core secondary structure elements and therefore have the same architecture. However, the connectivity is completely different (and therefore do not share the same fold) making their evolution from a common ancestor unlikely. By contrast, structures such as the $\alpha\beta$-TIM-barrels which have the same connectivity between their secondary structures often share little or no significant sequence similarity. These structures were previously thought to be examples of convergent evolution but are now increasingly being revealed as homologues (Copley & Bork, 2000).

It is difficult to distinguish between those structures which have converged and those which have diverged. Domains with a similar structure, sequence and function are clear homologues, however in the case of proteins with a similar structure but very little sequence and functional similarity it not so easy to give a definitive answer.

In many cases, proteins have diverged beyond significant sequence similarity but retain close structural similarity. The most widely known family exhibiting high structural similarity with negligible sequence similarity are the globins (Aronson *et al.*, 1994). Cases of distantly related enzymes with very different functions are gradually being found. Murzin (1998) describes two enzymatic examples with very probable distant relationships including caseinolytic ClpP protease, an ATP-dependent protease responsible for protein degradation in *E. coli* which shares the barrel-shaped architecture of other ATP-dependent proteases. Within the active site, it shares the catalytic triad of chymotrypsin and three other structural superfamilies of serine proteases. However, the folding subunit shows no relationship with any protease of known structure. Instead it has a clear structural and very probable relationship with members of the crotonase family responsible for fatty acid metabolism: enoyl-CoA hydratase and 4-chlorobenzoyl-CoA dehalogenase.

Those proteins that recur in nature but have neither sequence nor functional similarity have been termed superfolds. Superfolds offer an interesting perspective on evolution as they appear to represent folds that have been reinvented many times, as such it is thought that such folds may offer favourable properties for folding stability (Orengo *et al.*, 1994).

1.3.1 Domain Evolution

Basic mechanisms of protein evolution include residue substitutions, deletions and insertions. Grishin (2001) has identified structural mechanisms which can change the fold of a protein. He described a number of mechanisms whereby homologous structures have been modified through evolution so that the fold has changed. Examples of the addition/subtraction or substitution of secondary structures have been revealed. The most dramatic of these is the evolutionary event which transformed bacterial luciferase, a complete $(\alpha\beta)_8$-barrel into non flourescent flavoprotein (NFP). NFP is a homologous relative with 30% sequence identity to bacterial luciferase, as identified by PSI-BLAST. NFP contains a 90 residue deletion, removing two complete $\alpha\beta$ units and an α-helix. The remaining parts of the barrel are connected with a single antiparallel β-strand. Grishin describes how it is possible for a shorter deletion in an α-helix to force an extension of the peptide to form a β-strand. The addition of extra β-strands is also considered, whilst extra β-strands inserted at the edge of a sheet would have little effect on the topology of a structural family, the addition of a β-strand into a β-barrel may warrant placement of structures into different fold groups. Grishin (2001) cites an example in which a β-hairpin addition into the 8 stranded β-barrel of retinoic acid binding protein results in the formation of 10 stranded β-barrel of retinol binding protein. Structural similarity between these proteins is pronounced, including conserved length and tilt of the β-strands. Function is also similar with both binding lipids inside the barrel.

Another example illustrates the invasion of a β-strand and an α-helix. The two folds of the K homology (KH) domain, a widespread RNA-binding motif with similarity to a number of other RNA binding proteins and a particularly strong similarity with ribosomal protein S3, part of the ribosome which translates mRNA. These two structures differ in the placement of an $\alpha\beta$-unit insertion changing the topology but nevertheless leaving the structures to converge on the same architecture. The two distinct topologies may have arisen from an ancestral KH-motif protein by N- and C-terminal extensions or one of the existing topologies may have evolved from the other by extension, displacement and deletion.

Grishin also describes the mechanism of β-hairpin flip/swap. The lipocalins are used again as an example, this time the similarity between retinol-binding protein and triabin, the thrombin inhibitor, is described. Triabin shares significant sequence similarity with retinol-binding protein and is structurally very similar. However, the N-terminal regions of the structures are topologically distinct due to a flip of a β-hairpin through 180^o an as a result triabin and retinol binding protein are sometimes classified as different folds.

1.4 Domains and Function

1.4.1 Functional Classification Schemes

Genome sequencing projects have driven the development of functional classification schemes. Functional classification of proteins is essential for database annotations allowing the functional comparison of different proteins and organisms. However, functional classification is not trivial. Firstly, it is difficult to describe proteins as having a particular function on the basis of their domain composition. Domains can exhibit one function, such as ATP binding whereas the function of the whole protein may be controlled by a different domain. Secondly, function can be described physiologically, for example, 'cell regulation' or at the molecular level, for example, 'ion channel'. Two of the most widely used classification schemes are the Enzyme Commission (EC; NC-IUBMB, 1992) and the Gene Ontology or (GO) (Ashburner *et al.*, 2000) classification schemes.

1.4.1.1 Enzyme Commission Numbers

The hierarchical Enzyme Commission scheme categorises enzyme reactions. It is the best developed and the most widely used scheme, however, it is limited to enzymes and their biochemical functions. A reaction is assigned a four-digit EC number where the first digit describes the class of reaction the enzyme catalyses, (1, oxidoreductases; 2, transferases; 3, hydrolases; 4, lyases; 5, isomerases; 6, ligases). Further levels in the hierarchy depend upon the primary EC number. For example, for oxidoreductase reactions, the second level describes the substrate upon which the enzyme acts, but for the isomerases, it describes the type of reorganisation in isomerase reactions. Enzymes with more than one reaction can have more than one EC number assigned to them.

1.4.1.2 Gene Ontology

This classification scheme separates gene function into three independent ontologies: biological process, molecular function and cellular component. One particular protein may be described by many different categories within the classification scheme. For example, a particular protein may function in several different biological processes, contain domains with diverse molecular functions and participate in multiple interactions. Each ontology is therefore a network of nodes, able to handle data at different stages of completeness.

1.4.2 Methods for Predicting Function from Structure

With the explosion of sequence and structural data now available there is a need to define or predict functional mechanisms from a knowledge of the protein structure. In particular, a complete understanding of the complex relationships between protein sequence, structure and function is critical. Genome sequencing projects have driven the development of functional classification schemes. Among these is the **F**uzzy **F**unctional **F**orm (FFF) (Di Gennaro *et al.*, 2001) which is derived by the superposition of functionally significant residues in a few selected protein structures that have related functions. A descriptor of the active site is formed from the distance and angles between the C_α atoms and the sidechain centres of mass. This, in turn, is used to screen the structure for the presence of this functional site in other experimentally determined structures in the Protein Data Bank (PDB;(Bernstein *et al.*, 1977). **TE**mplate **S**earch **S**uperposition, (Wallace *et al.*, 1997) has been used for the detection of enzyme active sites by manually defining a template from the constellation of atoms important for the function of the protein. This constellation is then used to search for other PDBs with similar constellations. The **SP**atial **A**rrangements of **S**idechains and **M**ainchain (Jones *et al.*, 1991) program works in a similar way to TESS but identifies residues instead of atoms.

1.5 Pairwise Sequence Alignment

Aligning the sequences of proteins of known and unknown structures is integral to the process of inferring structural homology. Generally, as sequence identity decreases, the quality of the alignment also decreases (Martin *et al.*, 1997). In order to align protein sequences, it is necessary to have a method for scoring the similarity of the residues in protein sequences. Secondly, an optimisation method is needed for determining the best alignment of one sequence against the other. One method of scoring an alignment is by counting the number of identical residues. However, a more sensitive method is to account for the similarities between amino acids when comparing sequences.

1.5.1 Substitution Matrices

Substitution matrices are 20 by 20 matrices which score the similarity of pairs of amino acid residues. Scores can reflect similarity in the physiochemical properties, such as similar size or charge. Alternatively scores can reflect the frequency with which residues are found to exchange for one another in protein families.

1.5.1.1 Amino Acid Propensities

The 20 amino acid sidechains have distinct chemical and physical properties (Figure 1.1). This difference in amino acid physiochemical properties means that the substitution of like amino acids is more favourable than an unlike pair. For instance, the substitution of a negatively charged glutamic acid residue to a positively charged lysine residue is much less likely than to the similarly charged aspartic acid.

1.5.1.2 Observed Mutations

When comparing two proteins, residue substitution probabilities can be used as a more sensitive assessment of similarity than simply using amino acid identities. The likelihood of a given residue substitution can be quantified in a mutation data matrix (MDM). The 2D matrix provides a probability that each of the 20 amino acids could be substituted by

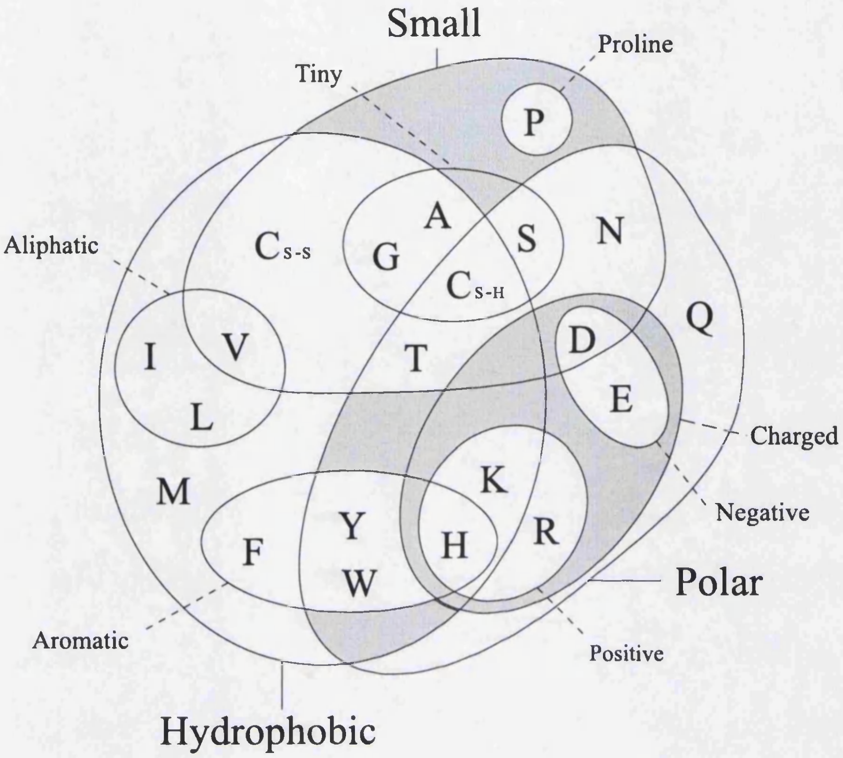


Figure 1.1: A Venn diagram describing the chemical and physical properties of amino acids (Taylor, 1986a). The residues are alanine (A), cysteine (C), aspartic acid (D), glutamic acid (E), phenylalanine (F), glycine (G), histidine (H), isoleucine (I), lysine (K), leucine (L), methionine (M), asparagine (N), proline (P), glutamine (Q), arginine (R), serine (S), threonine (T), valine (V), tryptophan (W) and tyrosine (Y).

any of the others. These probabilities are derived by examining closely related sequences and counting the occurrence of each amino acid at each position.

1.5.1.3 Dayhoff or Point Accepted Mutation (PAM)

The PAM (Point Accepted Mutation) matrices (Dayhoff, 1978) are derived by analysis of residues exchanging in close relatives. It is based on the concept of the PAM which describes the probability of each residue substitution occurring in an evolutionary period of 1 residue mutation every 100 residues. However considering substitution values, based on sequences where only 1 in 100 residues has mutated, will not provide much useful information on distant evolutionary relationships since the sequences would be almost identical. The PAM250 matrix gives similarity scores equivalent to 20% matches remaining between two sequences (the twilight zone) and is obtained by repeatedly multiplying the original matrix by itself. However, this means that these more distant relationships (at 20% sequence identity) are inferred as the original matrix is based on alignments with sequences at about 85% sequence identity.

1.5.1.4 BLOcks SUbstitution Matrix (BLOSUM)

Henikoff & Henikoff (1992) derived a substitution matrix from blocks of aligned sequences from the BLOCKS database. Sequences in the multiple alignment were clustered by sequence identity and the average contribution at each residue position is calculated. To represent the relationship between sequences at different sequence identities the sequences are clustered at different clustering percentages. Thus, sequences clustered at 80% sequence identity are used to generate the BLOSUM 80 matrix (BLOcks SUbstitution Matrix). The BLOSUM 62 matrix best represents the relationship at 20% sequence identity as the mutation rates in clusters at 62% provide the most satisfactory balance between information and amount of data. Clustering at lower sequence identities would cause the mutation rates to be averaged out within the clusters and the matrix would become less sensitive.

1.5.2 Insertions and Deletions

Evolutionary relationships can be identified most simply by aligning sequences in a pairwise manner then scoring the resulting alignments. As proteins become more distant, alignment programs must allow for insertions and deletions (indels) in addition to single site mutations (substitutions). In order to allow for these indels, gaps are introduced into the alignment. The positioning of these gaps is complex to compute accurately. Simply maximising the number of identical residue matches and providing no restrictions on the number and size of the gaps would achieve an optimum score but the alignment may be biologically meaningless. Instead, scoring penalties are introduced which minimise the number of gaps that appear. Affine gap penalties can also be applied. Rather than a fixed gap penalty, affine gap penalties encourage the extension of existing gaps rather than the opening of new gaps. This is achieved by applying a larger penalty for opening a gap in the alignment and a lower penalty for a gap extension.

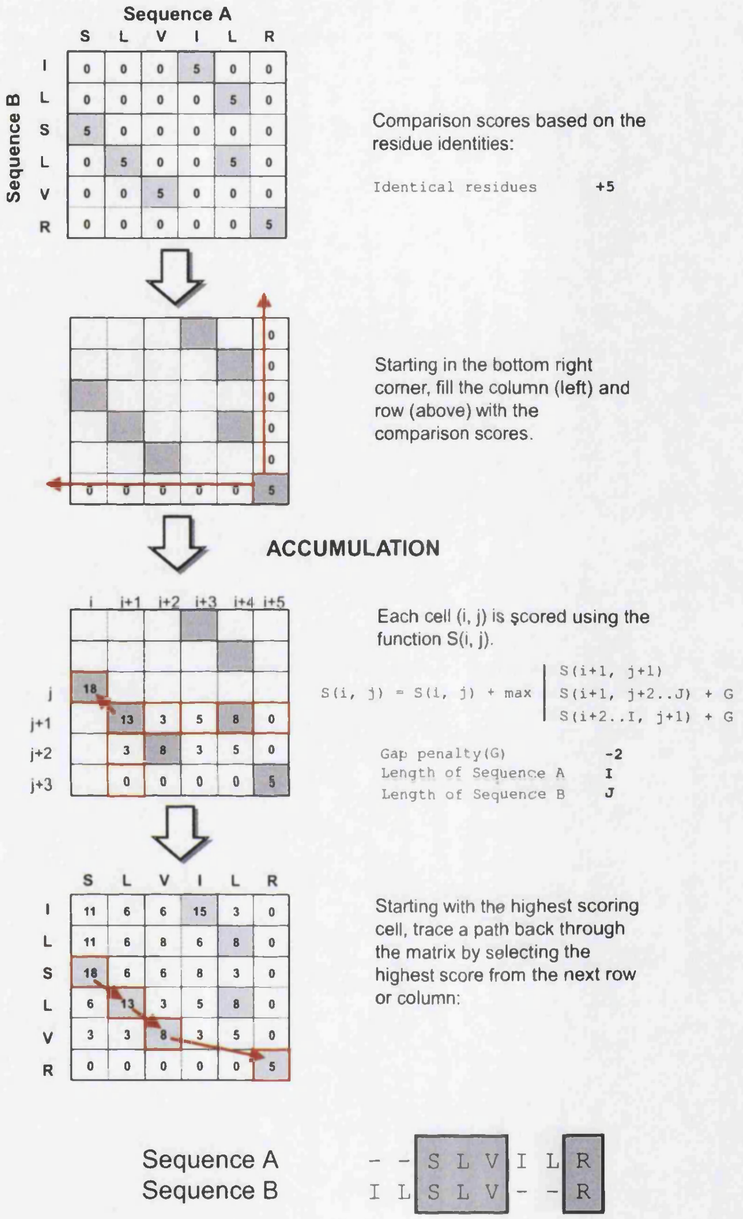
1.5.2.1 Local and Global Similarity

There are two general models for alignments: the global alignment which attempts to align two proteins along their entire lengths and the local alignment, which considers regions of similarity between specific parts of sequences. Local alignment algorithms allow the identification of localised regions of similarity which often convey functional information or domain similarities.

An example of an algorithm which is good at accommodating indels is the dynamic programming algorithm. This is a general mathematical procedure that provides an optimal alignment between two sets of data. This algorithm was first applied to sequence comparison by Needleman & Wunsch (1970). A 2D matrix is constructed from the sequences and the matrix is populated using a simple scoring scheme. In the example outlined in Figure 1.2, identical residues in the matrix score 5 and mismatches score 0. Working from bottom right of the matrix, to the top left, paths are drawn through the matrix summing the individual scores as shown in Figure 1.2. The scores are summed by using Equation 1.1 where the scoring function S(i,j) is added to the maximum score from the previous row or column in the 2D matrix.

$$S(i,j) = S(i,j) + max \begin{cases} S(i+1, j+1) \\ S(i+1, j+2..J) + G \\ S(i+2..I, j+1) + G \end{cases} \tag{1.1}$$

The score cell S(i,j) is added to the maximum score from the row and column below and to the right. If the value diagonally below S(i+1 ,j+1) is selected the value is added to the accumulating score. However, a higher value may be found in the row S(i+2...I, j+1) where I is the length of the row sequence, or the column S(i+1, j+2 ... J) where J is the length of the column sequence. Selecting from these values incurs a gap penalty (G) (shown on the example as a value of -2). A value in the row or column not diagonally below is only selected if it is still higher than the value in S(i+1, j+1) with the gap penalty. The alignment is generated by tracing back through the matrix, starting at the N-terminal end and following the pattern of high scoring elements.



Sequence B \ Sequence A	S	L	V	I	L	R
I	0	0	0	5	0	0
L	0	0	0	0	5	0
S	5	0	0	0	0	0
L	0	5	0	0	5	0
V	0	0	5	0	0	0
R	0	0	0	0	0	5

Comparison scores based on the residue identities:

Identical residues +5

Starting in the bottom right corner, fill the column (left) and row (above) with the comparison scores.

	i	i+1	i+2	i+3	i+4	i+5
j	18					
j+1		13	3	5	8	0
j+2		3	8	3	5	0
j+3		0	0	0	0	5

Each cell (i, j) is scored using the function S(i, j).

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

Gap penalty(G) -2
Length of Sequence A I
Length of Sequence B J

	S	L	V	I	L	R
I	11	6	6	15	3	0
L	11	6	8	6	8	0
S	18	6	6	8	3	0
L	6	13	3	5	8	0
V	3	3	8	3	5	0
R	0	0	0	0	0	5

Starting with the highest scoring cell, trace a path back through the matrix by selecting the highest score from the next row or column:

```
Sequence A    - - S L V I L R
Sequence B    I L S L V - - R
```

Figure 1.2: The Needleman and Wunsch dynamic programming algorithm. The sequences form a 2D matrix which is scored. Exact residue matches score 5 and mismatches score 0. Starting from the bottom right corner of the matrix the scores are summed. The maximum score from the previous row or column starting from the cell (i + 1, j + 1) can be inherited and added to the comparison score of the cell. A penalty is incurred from inheriting the score from any cell other than (i + 1, j + 1) as this corresponds to a gap in the alignment.

A modification of this algorithm was introduced by Smith & Waterman (1981) which focused on providing local, rather than global, alignments. This algorithm introduces negative scores for non-matching residue comparisons. When tracing back through the matrix the paths can start anywhere and is terminated when the score falls below zero.

1.5.3 Database Searching: FASTA and BLAST

Searching a database with a sequence can be conceived as an extension of the pairwise sequence alignment. Using a conventional dynamic programming method is time consuming and computationally expensive and so a great deal of effort has been put into making these searches as efficient as possible as the databases grow. The FASTA (Pearson & Lipman, 1988) and BLAST (Basic Local Alignment Search Tool, Altschul *et al.* (1990)) algorithms are both essentially local similarity searches which concentrate on finding short identical matches which may contribute to a total match.

FASTA

The FASTA algorithm searches for segments of one and two identical residues in pairs of sequences in the database. These possible matches are stored in a hash table along with all matched segments in the database (Figure 1.3). This data source is then used to screen database queries for likely matches. These possible matches are entered into a 2D matrix and the more rigorous dynamic programming algorithm strings the segments together for a global alignment. An E-Value, the expected number of times the observed value will occur by chance is calculated from the P-Value, which allows the user to assess the likelihood of the match being true. The P-Value is calculated from the raw score by summing matrix similarities of aligned residues and deducting gap penalties over the length. The P-Value (p) measures the probability of the query sequence matching a particular database sequence with a score of at least S; the corresponding E-Value (E) is the expected number of times that happened in a database of size N (Equation 1.2).

$$E = pN \tag{1.2}$$

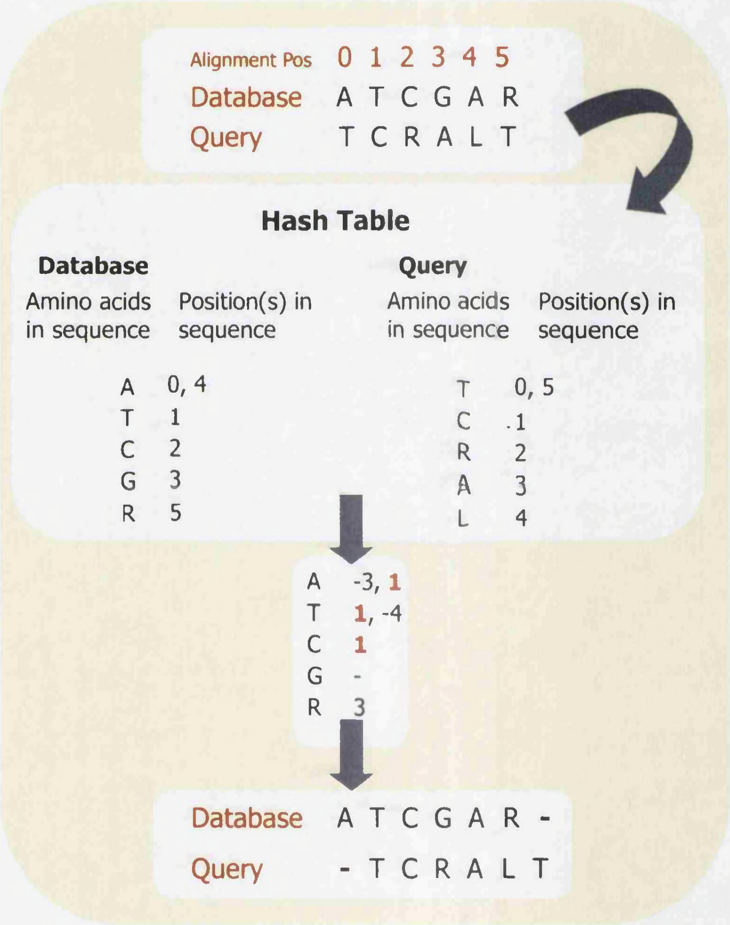


Figure 1.3: The FASTA algorithm builds a hash table of n identical letters, shown in this example as single letter words and their offsets in each database sequence. The offsets are then applied to the n-identical letters in the query to obtain the largest candidate alignment, the largest string with the same offset. In this an offset of 1 is obtained in the largest string.

BLAST

The BLAST algorithm (BLASTP for proteins and BLASTN for nucleotides) separates the protein sequences into tripeptide fragments e.g. ACE. These tripeptide fragments are then expanded to include possible substitutions that could have occurred using the BLOSUM substitution matrix (section 1.5.1.4). For example, ACE is expanded so that it could match ACE, GCE, GME and AME. Database query sequences are then searched against the possible tripeptide fragments. When a matching tripeptide fragment is found it is extended in both directions along the segment as far as possible, creating the highest scoring segment pair from the two sequences. All matches are compared with each other in order to find the highest segment pair (HSP) which represents the highest scoring sequence match in the database. The overall scores assigned to the resulting sequence matches are based on E-Values. The E-Value in BLAST is calculated in a similar way to the E-Value in FASTA except they do not refer to the whole sequence but instead refer to the fragment of sequence that is in a given match (the HSP) The initial algorithm was gap-less, but has been refined to allow gaps by searching for two separate HSPs. Often several high scoring segments exist between homologous proteins. The gapped-BLAST implementation uses dynamic programming to link together these high scoring regions to obtain the final alignment.

1.5.4 Profile-based Sequence Comparison

One approach to improve the performance of these sequence comparison methods is to identify features that are conserved during the process of evolution by examining multiple sequence alignments of related protein sequences. The advantage of using this approach is that the variation of observed amino acids can then be modelled for each position in the alignment in a sequence 'profile'. A profile can assign significance to each alignment position based on the degree of conservation at that position, whereas a simple pairwise sequence comparison gives all positions in the alignment equal weighting. Emphasising the importance of highly conserved regions and reducing the importance of poorly conserved regions during the search procedure allows more accurate alignments and provides more discriminating scoring schemes (Barton & Sternberg, 1987; Taylor, 1987; Rice & Eisenberg, 1997; Park *et al.*, 1998; Kelley *et al.*, 2000).

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

1.5.4.1 PSI-BLAST

PSI-BLAST (Altschul *et al.*, 1997) uses an iterative approach that begins with a simple pairwise BLAST search of a sequence database. This identifies a set of close relatives from which a multiple sequence alignment is generated. Instead of searching with a single sequence, the database is now searched with a profile derived from the multiple sequence alignment (called the position specific substitution matrix, PSSM). This enables the identification of more distant homologues. The PSSM is generated using a substitution matrix and is refined at each iteration. The sequence information from distant relatives is therefore incorporated into the growing alignment and the process repeated until either no more sequences are found or a specified number of iterations has been reached. This approach relies heavily on a non-redundant database to avoid introducing bias into the relative importance of any of the residues.

1.5.4.2 Hidden Markov Models

Sequence profiles can be implemented using a statistical modelling technique known as a Hidden Markov Model (HMM) which allow sequences to be aligned against the model in a probabilistic manner. To use an analogy, HMMs can be considered as 'sequence generating factories' capable of producing many different sequences with different probabilities. Internally, the HMM works by representing each column in the multiple sequence alignment by three states; match, delete and insert (Figure 1.4). The match state models the distribution of residues allowed at a specific column of the sequence alignment, the delete state models having no residue at this column and the insert state models an insertion of one or more residues after this column. These states are connected by state-transition probabilities and a sequence of states is generated by moving from the start to the end point according to these probabilities. At each state, a residue is emitted according to the emission probability distribution and this creates an observable sequence of residues. The sequence of these internal states is hidden, hence the name hidden Markov models, therefore the most likely state sequence must be inferred from an alignment between the HMM and the query sequence (Eddy, 1996).

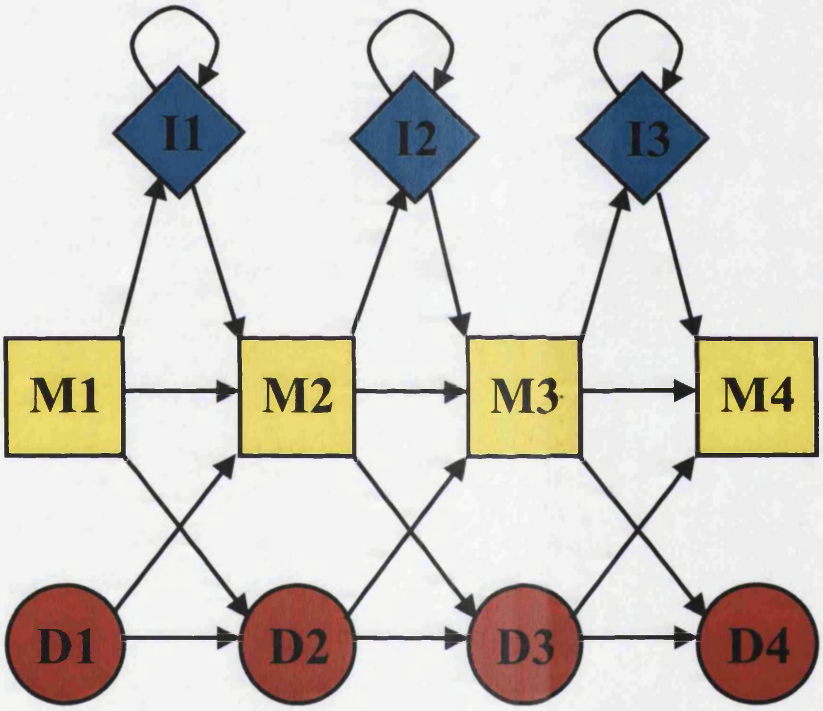


Figure 1.4: Overview of the profile Hidden Markov Hodel (HMM). This is characterised by its match (M), delete (D) and insert (I) states and the allowed transitions (arrows) between them

SAM-T99

The SAM-T99 method, recently updated to SAM-T02 and based on the earlier SAM-T98 protocol (Karplus *et al.*, 1998), builds a HMM from either a single seed sequence or a reliable seed alignment using a large sequence database such as the non-redundant translated GenBank sequence database (Benson *et al.*, 2000). After the initial scan of the sequence database, a model is generated from the alignment of these related sequences and this model is then used for a further database scan (Figure 1.5). Every added sequence provides the model with more detail on the acceptable sequence variability at each position within the sequence family. As a result, the method allows greater sensitivity for identifying more distantly related sequences.

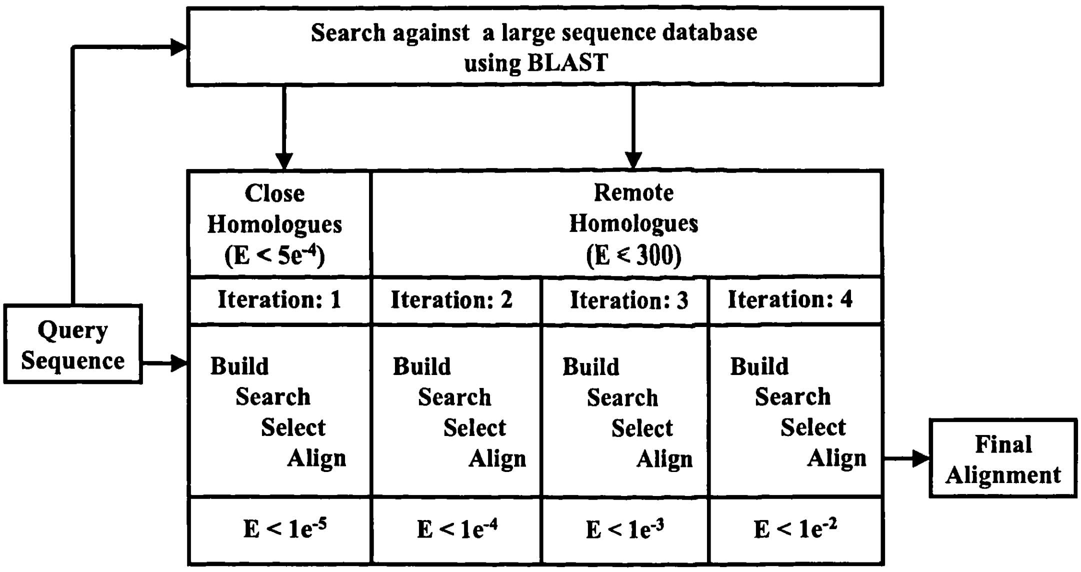

Search against a large sequence database using BLAST
Close Homologues ($E < 5e^{-4}$)
Remote Homologues ($E \leq 300$)
Iteration: 1
Iteration: 2
Iteration: 3
Iteration: 4
Query Sequence
Build Search Select Align
Build Search Select Align
Build Search Select Align
Build Search Select Align
Final Alignment
$E < 1e^{-5}$
$E < 1e^{-4}$
$E < 1e^{-3}$
$E < 1e^{-2}$

Figure 1.5: Overview of the SAM-T99 protocol for detecting remote homologues.

1.5.5 Structure Comparison

1.5.5.1 The Relationship Between Sequence and Structure

It is well established that evolution conserves protein structure more than protein sequence (Chothia & Lesk, 1986). The comparison of protein structures can lead to insights into the evolutionary relationships between them and the mechanisms by which the structures evolve from one another. Protein structure comparison can be used to detect similarity between homologous proteins when perhaps, sequence identity between homologous members is undetectable. Despite this, often more than half of the structure remains conserved, forming a structural fingerprint which can be used to identify other homologous proteins. Figure 1.6 shows the pairwise sequence identities and SSAP structural similarity scores for members of the structurally conserved but sequence diverse globin superfamily.

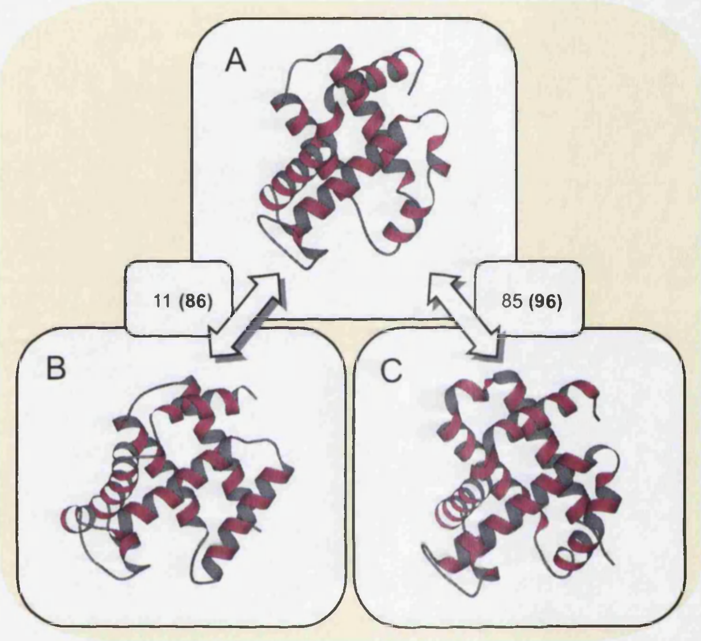


Figure 1.6: MOLSCRIPT (Kraulis, 1991) representations of relatives from the globin superfamily. Diagram from Orengo *et al.* (2003). Pairwise sequence identities are shown for each pair of structures and SSAP structural similarity scores are also given in parentheses. SSAP scores range from 0 up to 100 for identical structures (section 1.5.5.2).

As well as the identification of structural relatives, it is also possible to assess the tolerance to structural change of a given superfamily which in turn, could have impacts on the function.

Protein structures are compared, in most methods, by considering the properties and/or relationships of either the secondary structure elements or residues along the carbon backbone. This is usually done by comparing distances or vectors. However some structure comparison algorithms also take into account physiochemical properties such as accessibility, torsion angles or hydrophobicity. Methods for comparison of structure fall under two types:

1. Methods which superpose protein structures and measure **intermolecular** distances, minimising the distances between superposed positions

2. Methods which compare **intramolecular** distance vectors, comparing sets of internal distances between position to identify an alignment maximising the number of equivalent positions.

Intermolecular Methods

Most intermolecular methods employ rigid body superposition which was pioneered by Rossman and Argos in 1970s. Intermolecular methods must deal with insertions and deletions in the structures i.e. they must assign equivalences between residues. The algorithm superposes equivalent C_α onto each other and can be described in three steps:

1. Both proteins are translated into a common position in the co-ordinate frame of reference. Typically the centre of geometry is placed at the origin.
2. One protein is rotated onto the other protein through the orthogonal x, y and z axes.
3. The distances between equivalent atoms is measured.

Steps 2 and 3 are repeated until convergence on an optimum score is produced. The final score is the square root of the average squared distances between equivalent atoms. Additional information on **R**oot **M**ean **S**quare **D**eviation (RMSD) is provided in Chapter 5 where it is used to measure the quality of comparative models against their experimentally determined structures.

Intramolecular Methods

Intramolecular methods do not attempt to superpose structures but instead consider internal relationships within each structure such as distances or angles so that the most equivalent positions between the proteins can be determined. These methods can provide a set of equivalences for use by intermolecular methods. Allowing for insertions and deletions into the protein structures has been accomplished by a number of different techniques. Firstly, comparisons are made between the secondary structure elements only (section 1.5.5.3) and secondly, algorithms have been developed to divide the structure into fragments for example DALI (Holm & Sander, 1993). An approach that enables the accommodation of insertions and deletions is dynamic programming. This approach has been adopted by the SSAP structure comparison algorithm (Taylor & Orengo, 1989).

1.5.5.2 SSAP

The SSAP method employs dynamic programming at two levels to cope with the extensive indels between homologues. A structural environment, or vector view, is created from the C_β of a subject residue to all other C_β atoms in the structure (Figure 1.7). Vector views are calculated using the internal geometry of the residue, based on the C_α of the subject residue.

The first level of dynamic programming is applied to determine the equivalent residues in the two structures. For each pair of potentially equivalent residues (i.e. similar torsion angles, accessibility, secondary structure state), a two dimensional matrix is used to score the similarity in vectors between the residue in the first structure and the potentially equivalent residue in the second structure. The selection criteria identifying potentially equivalent residue pairs are based on the residues having similar physical and chemical properties, such as accessibility, secondary structure state and local conformation (measured by ϕ,ψ angles). The example in Figure 1.7 shows a residue level matrix for comparing the vector view from residue *a*in protein I with the vector view from K in protein II. Cells are scored depending on the similarity of the vector views then dynamic programming is used to find the optimal path through this matrix giving the best alignment of the residue views. This matrix is known as a residue level score matrix.

The alignment path from the residue level score matrix provides a score which represents the similarity between the structural environments in the two residues. If the residues are highly similar, i.e. provide a high scoring alignment path, then the scores from this alignment path are added to a summary score matrix. The best path through the summary matrix is then found using the second level of dynamic programming in order to find the best overall path and the structural alignment. A summary of the SSAP algorithm is shown in Figure 1.8.

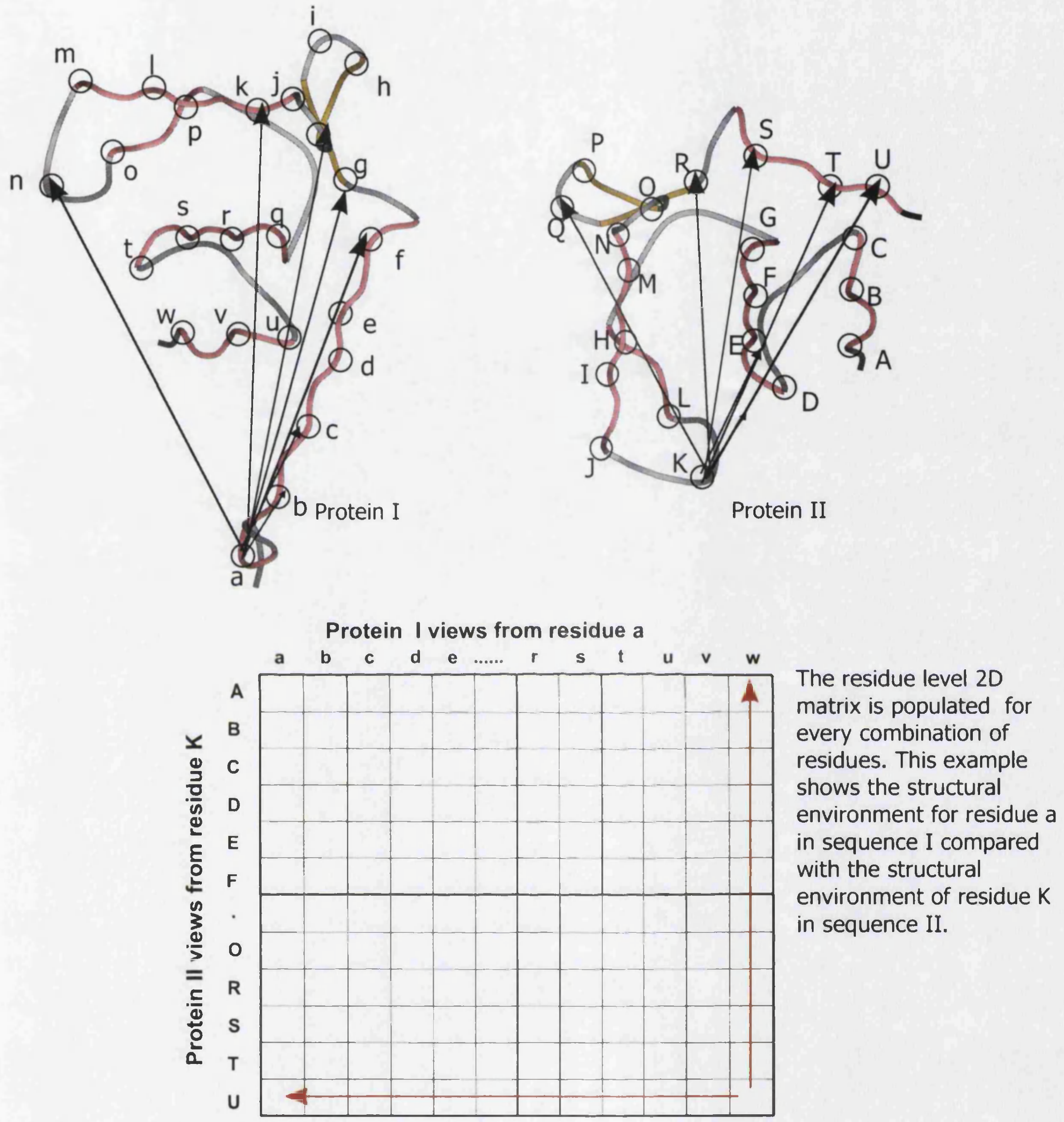
A simple outline of the SSAP algorithm can be summarised as follows:

1. Calculate the vector view for each residue in the two proteins, given by the set of vectors from the residue to all other residues in the protein.

2. For each potentially equivalent residue pair between the proteins, e.g. possessing similar torsional angles and accessible areas, compare vector views applying dynamic programming to find the best path through a residue level score matrix, scored according to the similarity of vectors.

3. For residue pairs scoring highly, add the scores along the optimal path obtained in step 2, to a two-dimensional summary score matrix.

4. Repeat steps 2 and 3 until all potentially equivalent pairs have been compared.


m
l
k
j
i
h
p
o
n
g
s
r
q
t
f
w
v
u
e
d
c
b
Protein I
a
P
S
R
T
U
O
Q
N
G
C
M
F
B
H
E
A
I
D
L
J
K
Protein II
Protein I views from residue a
a b c d e ...... r s t u v w
Protein II views from residue K
A
B
C
D
E
F
.
O
R
S
T
U

The residue level 2D matrix is populated for every combination of residues. This example shows the structural environment for residue a in sequence I compared with the structural environment of residue K in sequence II.

Figure 1.7: Illustration of a residue structural environment, or vector view, employed by the SSAP method (Taylor & Orengo, 1989) (Not all vectors are shown in the figure). The view for a given residue is taken as the set of vectors from the C_β atoms of this residue and all other residues in the protein.

5. Use dynamic programming again to determine the optimal path through the two-dimensional summary score matrix, giving the equivalent residues between the proteins and the global structural alignment.

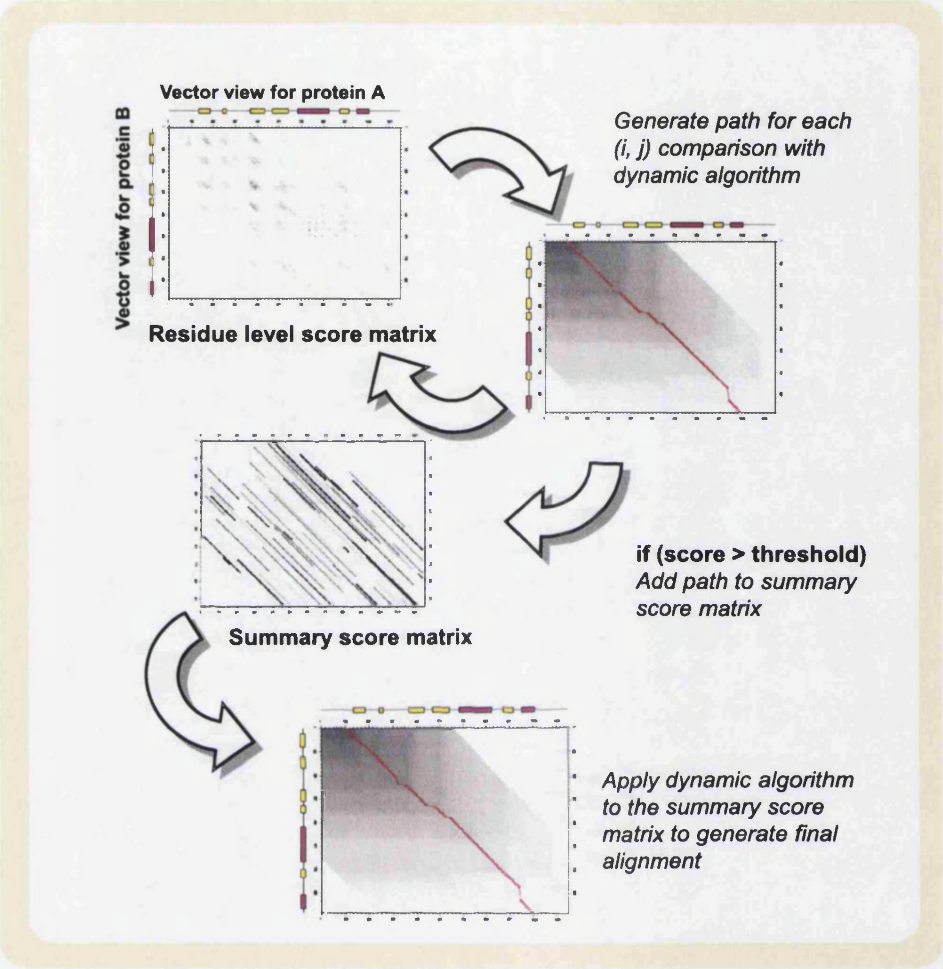


Figure 1.8: Flowchart describing the SSAP protocol. DP is applied on two levels. First to find the optimal alignment based on the comparison of structural environments of two residues (scored in the residue-level score matrix). If the residues are deemed sufficiently similar (i.e. the alignment path scores greater than a given threshold), the scores from this path are added to a summary score matrix. DP is then used again to find the optimal global alignment through the summary score matrix to identify equivalent residues between the two proteins.

1.5.5.3 GRATH

One way of dealing with insertions and deletions is to make comparisons between the secondary structure elements. A procedure which is carried out in a number of algorithms by a mathematical technique is known as graph theory. One such method which considers proteins in terms of their secondary structure elements is GRATH (Harrison *et al.*, 2002).

GRATH describes each secondary structure element as a node which is labelled according to the secondary structure type. The structural relationship between each of the secondary structure elements are then described by measuring the distance and the angle between them. Graph theory then searches for the most equivalent nodes, or clique, between two graphs.

1.5.6 Structural Classification

As the number of experimentally solved structures has increased from more than 1,000 structures in the 1990s to 17,200 in 2003, databases of structural classification have been developed. As structure is much more conserved than sequence through evolution it is possible to detect more distant evolutionary relationships through the comparison of structures. In turn, these evolutionary relationships can elucidate evolutionary mechanisms, the recurrence of structural motifs in families and the evolution of function. There are two major databases of protein structure domains, CATH (Orengo *et al.*, 1997; Pearl *et al.*, 2001) and SCOP (Murzin *et al.*, 1995; Lo Conte *et al.*, 2000).

1.5.6.1 CATH

CATH, a protein domain database, classifies over 34,300 domains (Version 2.4) which clusters protein domains in a hierarchy according to their Class (C), Architecture (A), Topology (T) and Homologous superfamily (H). Domain boundaries are defined using a consensus approach which seeks agreement between three domain assignment algorithms (Jones *et al.*, 1998). The class, the first level of the hierarchical classification, describes the proportion of α-helices and β-strands which divides domains into three major categories: Mainly α, mainly β, and $\alpha\beta$. The domains in each class are then divided into architecture which describes the spatial arrangement of the secondary structures in three-dimensions. Figure 1.9 summarises the main architectures. There are 37 distinct architectures in CATH Version 2.4. The topology, or fold level distinguishes the different connectivities of the secondary structure elements, giving a total of 775 different fold groups for CATH Version 2.4. Proteins are only grouped at the last level of the hierarchy, homologous superfamilies or H-level, if there is sufficient evidence that proteins are related by evolution. Domains are defined as homologous if two of the three criteria described below are met:

1. Similar structures (SSAP score > 80 or significant CATHEDRAL E-Value)

2. Similar sequence (>35% identity or significant PSI-BLAST E-Value)

3. Functional Similarity

Much of the CATH classification is carried out automatically, with manual inspection to assess some domain boundaries and homologous relationships. Structural similarity is measured by SSAP (section 1.5.5.2) or CATHEDRAL, which identifies equivalent secondary structure elements using the GRATH software (section 1.5.5.3) which are then used by SSAP for the structural comparison. Protein families in the CATH database are given a unique identification number. For example, the $\alpha\beta$-hydrolase superfamily is 3.40.50.950 describing this superfamily as the $\alpha\beta$ class (C identifier 3), the three layer ($\alpha\beta\alpha$) architecture (A identifier 40), and the Rossmann fold (T identifier 50).

A further level in the CATH database clusters homologous superfamilies into sequence clusters at 35% sequence identity using single linkage clustering (S35 families). Structures in the same sequence family are considered structurally very similar and the extent of structural variability in a superfamily can be assessed by taking a representative from each of these S35 clusters in a homologous superfamily.

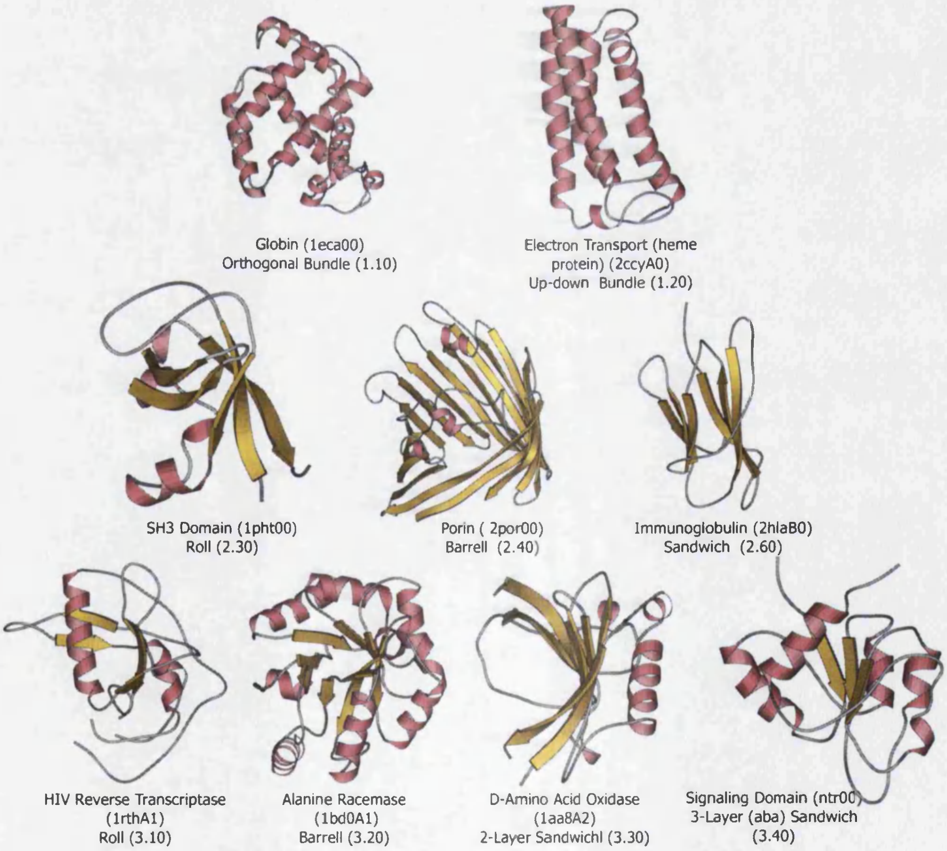


Figure 1.9: Illustrating the main architectures in CATH.

1.5.7 SCOP

The protein domain database, SCOP (Structural Classification of Proteins) is constructed from mainly manual methods but like CATH, classifies domains into a hierarchy of structural relationships. SCOP defines four levels of classification; class, common fold, superfamily and family in descending hierarchical order. The definition of class is similar to that of CATH although the $\alpha\beta$ class is split into α/β, where α-helices and β-strands are interspersed, and $\alpha+\beta$, where the regions of α-helices and β-strands are segregated. The common fold is equivalent to the topology (T) level in CATH and the superfamily level classifies at the same level as the homologous superfamily (H). The family level clusters structures at greater than 30% sequence identity or if they have a close structural or functional similarity.

1.5.8 Other Structure Classification Databases

A variety of structural databases have been constructed with contrasting levels of manual and automated intervention. Table 1.1 summarises some of the more commonly used structure databases. With the exception of SCOP (section 1.5.7), all the databases discussed here use an automated method for protein structure comparison at some point in the classification procedure. Rather than generate strict hierarchical boundaries, some resources, such as FSSP (Holm & Sander, 1998) and Entrez (Madej *et al.*, 1995), provide lists of domains with similar structures. These lists, also known as nearest neighbour lists, describe a model of protein folding space that resembles a continuum rather than a series of discrete structural clusters.

Database	**Location**	**Structure Comparison Method**	**Description**
CATH	UCL, London, UK	SSAP (Taylor & Orengo, 1989)	Class Architecture Topology Homology. Semi-automated hierarchical classification of structural domains.
SCOP	MRC LMB, Cambridge, UK	Manual (Murzin *et al.*, 1995)	Structural Classification Of Proteins. A mainly manually maintained hierarchical classification.
3Dee	EBI, Cambridge, UK	STAMP (Russell & Barton, 1992)	Fully automated, multi-hierarchical classification with some class and fold definitions taken from SCOP.
DDD	EBI, Cambridge, UK	DALI (Holm & Sander, 1993)	Dali Domain Dictionary. Fully automated structural classification of recurring protein domains.
ENTREZ/ MMDB	NCBI, Bethesda, MD, USA	VAST (Madej *et al.*, 1995)	Fully automated structural descriptions using pre-calculated nearest neighbour lists.
FSSP	EBI, Cambridge, UK	DALI (Holm & Sander, 1993, 1998)	Fold classification based on Structure-Structure alignment of Proteins. Fully automated structural descriptions using nearest neighbour lists.
HOMSTRAD	Cambridge University, UK	COMPARER (Sali & Blundell, 1990)	HOMologous STRructure ALignment Database. Manual classification using information from SCOP, and various sequence family databases.

Table 1.1: Summary of structure classification databases.

1.6 Protein Structure Prediction

The knowledge of a proteins structure is of great importance as the structure of a protein can reveal essential insights into its function. However, traditional methods for determining protein structure (X-ray crystallography and nuclear magnetic resonance (NMR)) are limited by inherent difficulties in their application, and as such are time-consuming. The ability to predict protein structure has therefore received great interest in recent years, especially in light of whole-genome sequencing projects. Nevertheless, the accurate prediction of a proteins structure still remains a challenge, as it is related to the 'protein folding problem', the central dogma of structural biology. In general, the prediction of protein structure can be viewed at two levels; secondary structure prediction and tertiary structure prediction.

1.6.1 Secondary Structure Prediction

Early secondary prediction methods were based on empirical statistics, which described the intrinsic properties of secondary structures, obtained from the analysis of known structures. Such methods include those by Chou & Fasman (1974) and Garnier *et al.* (1978). More recently, prediction methods that exploit evolutionary relationships held within a PSI-BLAST sequence alignment profile have been developed. Two such methods, PHD (Rost *et al.*, 1994) and PSIPRED (Jones, 1999) use artificial neural networks to interpret PSI-BLAST profiles built for a target sequence, routinely giving prediction accuracies of over 77%.

1.6.2 Tertiary Structure Prediction

The prediction of tertiary structure is often classed into three areas, *ab initio* , fold recognition and comparative modelling.

Ab initio methods aim to predict protein structure from first principles, that is, they aim to generate a protein chain conformation using a potential function that is based on observed interactions within proteins of known structure. Even with a simplified representation of physio-chemical forces, or the use of a lattice model, this is an incredibly complex undertaking, and has generally met with little overall success.

Fold recognition methods were developed to exploit the fact that there may be a limited number of naturally occurring protein folds. Methods of fold recognition attempt to detect similarities between protein three-dimensional structure that are not necessarily accompanied by any significant sequence similarity. Such methods are often referred to as 'threading' methods, a term first described by (Jones *et al.*, 1992), as it can be imagined that the target sequence is threaded onto a library of representative protein folds. A global energy term is then used to assign the most likely fold for the target sequence. Fold recognition methods therefore aim to avoid the computational expense that is often encountered with the use of *ab initio* methods.

1.6.2.1 Comparative Modelling

Comparative modelling uses the information from homologous parents to infer the structure of the target sequence. The traditional approach outlined below is based on the findings of Greer (1981) which suggested that insertions are generally small local changes in the structurally variable regions leaving the rest of the structure, the structurally conserved regions, largely unchanged. The traditional method can be carried out manually, but has also been automated in software such as COMPOSER (Sutcliffe *et al.*, 1987a,b) and Swiss-Model (Schwede *et al.*, 2003).

1. Select a suitable template.
2. The target sequence is aligned to the template so that equivalent residues are identified.
3. Model structurally conserved regions.
4. The structurally variable regions are built, but with lower accuracy.
5. Sidechain modelling.
6. Energy minimisation may be performed.
7. Assess the quality of the structure.
8. Some of the steps above are iterated. For example, after the quality of the model is assessed, the alignment is often adjusted and a new model made.

Comparative modelling is discussed in more detail in Chapter 5.

1.7 Aims of the Thesis

With the recent success of the genome projects the number of protein sequences (presently 1, 497, 800 in GenPept) determined far exceeds the number of protein structures solved by NMR and X-ray crystallography (presently 17, 200 in the PDB). In order to understand the functions of these sequences it is important to know their structure. Bioinformatics techniques for the prediction of structure from sequences are being developed (section 1.6). The most accurate methods are those which infer the structure of a closely related homologue (>35% sequence identity) onto the target sequence, i.e. comparative modelling. However, in order to improve comparative modelling a number of areas must be addressed: The quality of the alignment, the selection of the best parents and the prediction of regions which deviate from the parent structure (Tramontano *et al.*, 2001). In order to provide better methods for the prediction of regions which deviate from the parent structures, more must be understood about how the homologous structures within a superfamily vary amongst themselves.

In Chapter 2, a method is developed for studying secondary structural variation in superfamilies. Homologous domains share an invariant core arrangement and connection of secondary structures but some superfamilies also contain peripheral secondary structures. 2DSEC identifies a profile core and embellished secondary structures from a CORA multiple structural alignment. Chapter 3 aims to study the mechanisms whereby structural change occurs between homologous structures. The change in structure as proteins become more distantly related is measured using the SSAP structural alignment and percentage sequence identity. Particular interest is shown in the region above 35% sequence identity, the region which is most reliable for comparative modelling. Protein evolution is then analysed in terms of indels. The average length of residue indels are examined and then superfamilies with indels comprising whole secondary structure elements are identified using 2DSEC. The impact of the secondary structural embellishments in three superfamilies are examined in more detail in Chapter 4.

Finally, Chapter 5 describes the development of an automatic comparative modelling pipeline GenMod. A CATH resource aligns target sequences to a structural alignment (Sillitoe, 2002) of the parents using the SAM-T99 software (Karplus *et al.*, 1998). Models are created using the MODELLER software (Sali & Blundell, 1993). A method for selecting the best regions from a set of multiple parents (Mosaic) is described and the quality of the final models are assessed using three methods, RMSD (section 5.1.3.1), LGScore (Cristobal *et al.*, 2001) and ProsaII (Sippl, 1993).

Chapter 2

2DSEC: A Two-Dimensional Representation of Three-Dimensional Structure

2.1 Introduction

2.1.1 Background

Databases such as CATH (Orengo *et al.*, 1997) and SCOP (Lo Conte *et al.*, 2000) group protein domains by sequence and structure (section 1.5.6). The fourth level of the CATH classification hierarchy groups domains sharing significant sequence, functional and/or structural similarity into homologous superfamilies (H level). Whilst these homologous domains share an invariant core arrangement and connection of secondary structure elements, in some superfamilies they may also contain additions or subtractions of peripheral secondary structures. These variations in structure between domains may provide a useful insight into their differing functional roles or protein-protein interactions. Thus, a method for identifying those core secondary structures and those secondary structures which are only present in one or a few superfamily members is key to the elucidation of the structural evolution of the superfamily.

Determining the structural core or invariant region from the variable region has been carried out in a number of different studies. The structural core of the protein can be defined as the hydrophobic centre and is considered to be the invariant scaffold on which the more variable regions are hung. A method for defining this hydrophobic region was proposed by Swindells (1995). Residues are considered to be part of the hydrophobic core if they are in a regular secondary structure conformation, their sidechains are buried and sidechain–sidechain contacts are mainly non-polar. However, the core of the protein can also be considered to be the regions of the protein which show the least structural variation between homologous members. Kelley & Sutcliffe (1997) provide an automatic method for defining core atoms or atoms with low spatial variance, in structures determined by NMR, and using this to cluster into ensembles of conformationally related sub-families. These sub-families are stored in an **O**n **L**ine **D**atabase of **E**nsemble **R**epresentatives **A**nd **DO**mains, OLDERADO. Similarly, a core finding algorithm was developed by Gerstein & Altman (1995a). They measure the spatial variation between equivalent atoms in a structural alignment. Those atoms with a low structural variation are determined as the core of the structure. This summarises the commonalities and differences within a family and has been compiled into an internet library of protein family core structures (Schmidt *et al.*, 1997). In addition to defining the core, the method creates an ordered list of atoms, ranked by their structural variation. In applying the core-finding procedure to the globins (Gerstein & Altman, 1995b) it was discovered that four of the helices form a structural core with low variance. The same α-helices were subsequently identified as having conserved, structurally important residues (Ptitsyn & Ting, 1999). The definition of core and variable regions of a set of homologous proteins has implications for comparative modelling. Simply by defining how much the structure varies between homologous proteins can provide the modeller with an idea of the quality of the model. With a set of particularly variable structures it may be expected that the true structure of the target sequence may vary just as much. In addition, the definition of core and variable region may provide an insight into regions of structure which will be modelled with greater accuracy than others.

Structural variation in the number of secondary structure elements can be viewed in terms of embellishment. The increase in complexity of organisms through evolution, from single cell to multi cellular and the increase in complexity of biochemical pathways and functions has lead to an increase in domain complexity through structural embellishment. However, variation may have occurred also due to truncation.

2.1.2 Aims of the Chapter

This chapter introduces 2DSEC, a tool for identifying and visualising protein domains within a given protein superfamily in terms of their secondary structure elements (SSEs). This enables a distinction to be made between core SSEs and those which are present only in a few superfamily members (embellishments). 2DSEC aims to calculate this secondary structure profile using structural alignment files generated by CORA (Orengo, 1999) which calculates a multiple structural alignment using a double dynamic programming method (section 2.2.2) (Taylor & Orengo, 1989). As such, CORA is a powerful approach for identifying core structural regions of a superfamily and 2DSEC provides a way of measuring this. 2DSEC provides a valuable tool to visualise and quantify the level of secondary structural variation present within a CATH superfamily allowing the identification and characterisation of variable superfamilies based on this measurement alone. The identification of core secondary structure and secondary structure embellishments enables an analysis of the mechanisms which have driven the evolution within the protein superfamilies - as such a number of analyses using this program are presented in Chapters 3 and 4 of this thesis.

2.2 Methods

2.2.1 Secondary Structure Assignment

Secondary structure assignments for each chain were calculated by the DSSP program (Kabsch & Sander, 1983) and the DSSP derived secondary structure is used by the CORA program. The two most regular secondary structure states, α-helix and β-strand, were used and all other secondary structure states were considered as coil. Furthermore, a strand was defined as a consecutive run of three or more residues assigned as strand, and a helix, four or more residues consecutively assigned as helix. This simpler scheme was used to clarify secondary structure element assignments allowing a clearer picture of secondary structure variability between homologous domains.

2.2.2 CORA: Multiple Structural Alignment

The CORA algorithm (Orengo, 1999) uses dynamic programming at two levels. The use of double dynamic programming was initially developed for the pairwise comparison of protein structures and is based on the comparison of intra-molecular C_β vectors between two structures (SSAP, Taylor & Orengo (1989)). Firstly, putative structurally equivalent residues are determined by comparing features such as secondary structure state, solvent accessibility and torsion angles. Structurally equivalent residues are then compared by constructing a two-dimensional matrix with scores based on internal C_β vectors. The first level of dynamic programming, the residue level dynamic programming, is then used to pick the highest scoring alignment path (section 1.5.2.1) between potentially equivalent residues. If the highest scoring path is above a certain threshold it is added to a summary matrix (Figure 2.1).

After all the residue level comparisons have been made, the second level of dynamic programming is then used to find the highest scoring alignment through the summary matrix. From this alignment CORA then calculates the average structural properties and encodes them into a consensus template. This consensus encodes information such as internal vectors, residue accessibility and torsion angles and records the variability of these properties across different relatives in the alignment. A multiple structural alignment is generated by successively aligning proteins to the evolving consensus template. Proteins are aligned one by one in order of decreasing structural similarity (measured by SSAP score). After each protein is aligned, the consensus template is recalculated to take account of any additional structural features of the newly aligned protein.

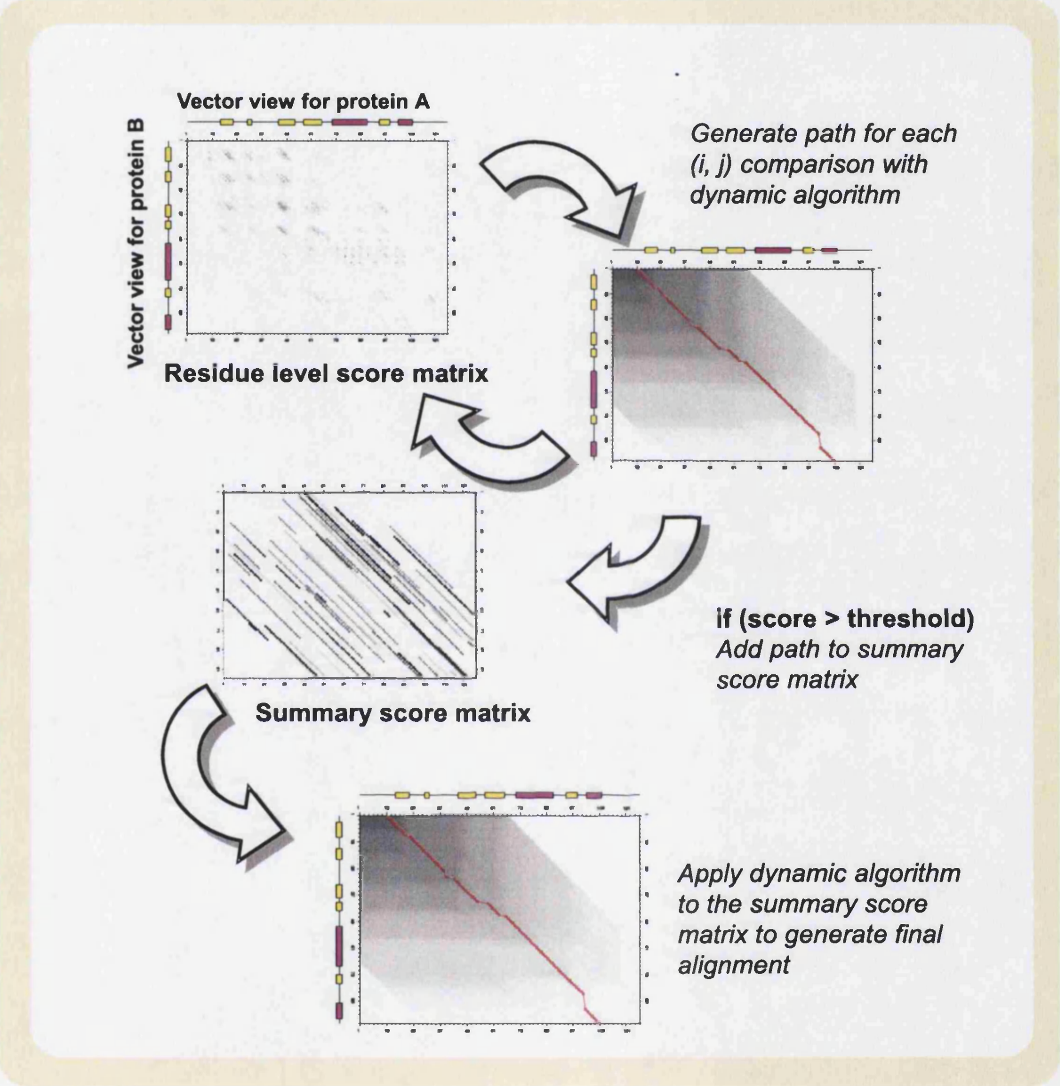

Vector view for protein A
Vector view for protein B
Residue level score matrix
Generate path for each (i, j) comparison with dynamic algorithm
if (score > threshold)
Add path to summary score matrix
Summary score matrix
Apply dynamic algorithm to the summary score matrix to generate final alignment

Figure 2.1: Flowchart describing the double dynamic programming algorithm for the comparison of two structures. CORA extends this pairwise method by constructing a consensus template encoding average structural properties and information on variability after each structure is aligned. The consensus structural information from this template is then used to align the next structure.

2.2.3 2DSEC: Two-Dimensional Secondary Structure Summary of Three-Dimensional Structure.

2DSEC produces a schematic representation of protein domains within a superfamily as aligned by CORA. 2DSEC uses this structural alignment to create a summary of the secondary structures present in each structure, and show which ones are equivalent in each structure (consensus) and which ones are present only in one or a few superfamily members (embellishments). This summary is shown in a cartoon representation, written in Postscript (Adobe systems Inc., 1985) (Figure 2.2) . The three-dimensional structure of each aligned domain is represented in a two-dimensional horizontal plot of its constituent secondary structure elements, α-helices and β-strands. Each domain sequence is labelled by its corresponding PDB code and CATH domain identifier, whilst a consensus description of the secondary structures common to the superfamily members is also shown (Figure 2.4). The diagrams are in the form of a series of linked symbols, where a circle denotes an α-helix and a triangle a β-strand. The size of the circle or triangle is determined by the size of the respective α-helix or β-strand. Core secondary structure elements are represented as light pink circles for α-helices and yellow triangles for β-strands. α-helical embellishments are coloured dark pink and β-strands are brown (Figure 2.2).

The example used throughout this chapter is the oligomerisation domain of the NADP binding oxidoreductases (3.30.360.10). This superfamily is an $\alpha\beta$-2 layer sandwich architecture. It contains five S35Reps: dihydrodipicolinate reductase (1dih02), glucose-fructose oxidoreductase (1ofgA2), biliverdin reductase (1gcuA2) and glucose 6-phosphate dehydrogenases (1dpgA2 and 1qkiA2). The quaternary structure of all five proteins comprises an oligomer. The monomer comprises the NADP binding domain and the oligomerisation domain (members of this superfamily).

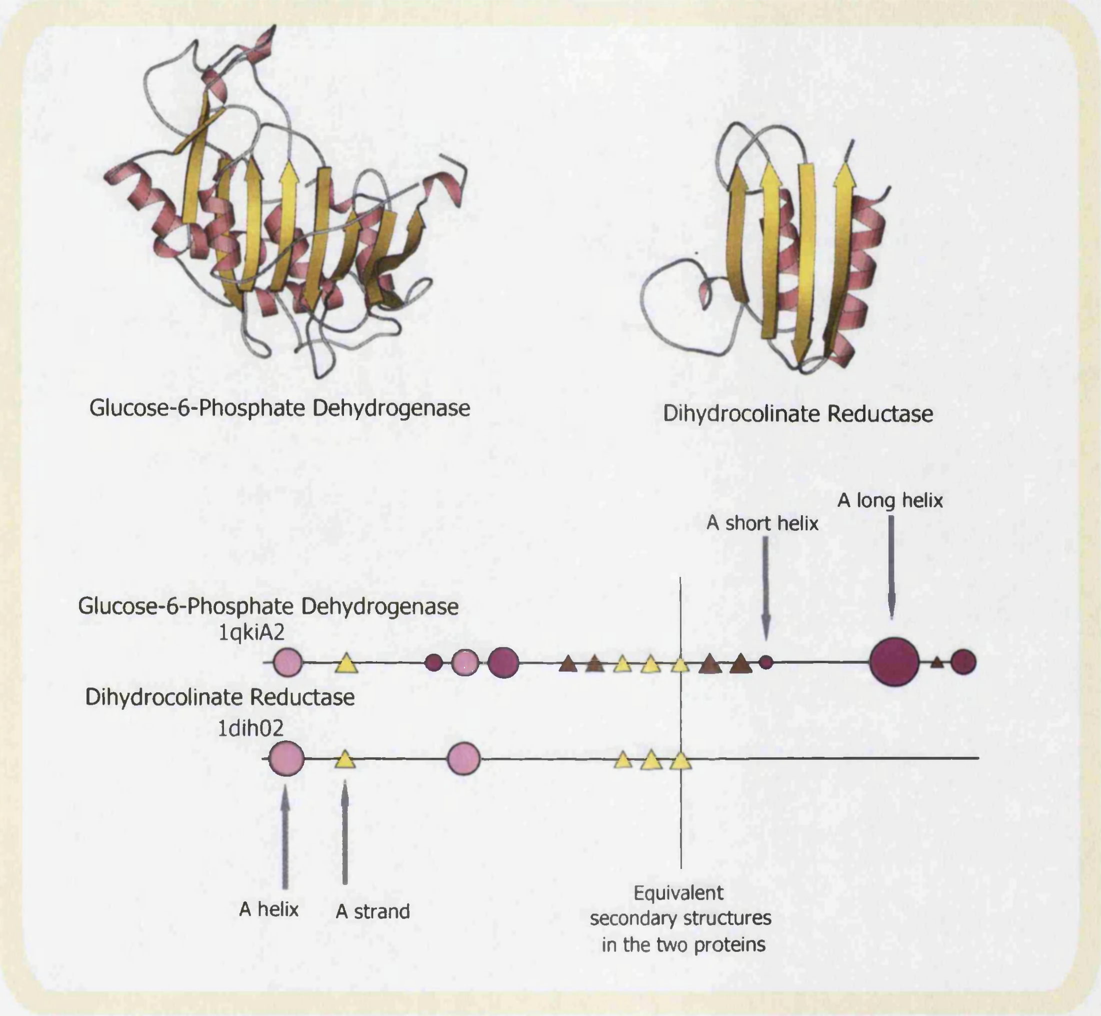

Glucose-6-Phosphate Dehydrogenase
Dihydrocolinate Reductase
A long helix
A short helix
Glucose-6-Phosphate Dehydrogenase
1qkiA2
Dihydrocolinate Reductase
1dih02
A helix
A strand
Equivalent
secondary structures
in the two proteins

Figure 2.2: A 2DSEC cartoon created from a CORA structural alignment of two homologous structures, glucose-6-phosphate dehydrogenase and dihydrocolinate reductase. The α-helices are shown as pink circles, the larger the circle the longer the α-helix. Light pink α-helices are consensus α-helices, that is, they are present in 75% or more of the structures aligned. Dark pink α-helices are embellishments. Yellow triangles are beta strands and as for the helices, the larger the triangle the longer the strand. Consensus β-strands are shown in dark yellow and embellishments are in brown. Equivalent secondary structures in each domain are aligned horizontally.

2.2.3.1 The 2DSEC Algorithm

The secondary structure information within the CORA structural alignment is summarised into segments along the alignment. Regions of extended coil in any of the domains are not represented. The program seeks to bring out the similarities and differences between the secondary structure elements only. Excluding information about length and orientation of loops enables the program to consider the variability of the domains in the alignment purely on the basis of the presence or absence of the secondary structures.

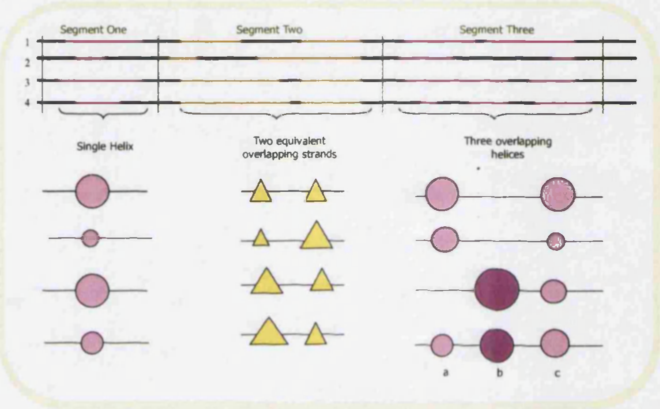


Figure 2.3: The 2DSEC algorithm divides the multiple alignments into segments containing at least one secondary structure in any of the aligned domains and a clear break before the next secondary structure in all aligned domains. Equivalence is calculated in any segments containing more than one secondary structure in any of the domains by assessing which secondary structures overlap most.

Firstly, the multiple alignment is divided up into segments. A segment is created if a secondary structure is present in *one or more* of the domains in the alignment and a clear break in all secondary structures can be seen in all aligned domains between secondary structures. For example, in Figure 2.3 three segments can be seen. Segment one contains only one secondary structure in each of the aligned domains but segments 2 and 3 contain a number of overlapping secondary structures in each protein. Again in this example, a decision is applied to segments 2 and 3 to determine the greatest overlap, so that equivalent secondary structures can be determined. In cases such as this, the most overlapping secondary structure is aligned with the large secondary structure and the secondary structure which overlaps less is considered as an additional element. For segment 2 the algorithm detects that each domain contains two β-strands in this segment and is able to identify them as equivalent. For segment 3, domain 4 contains three α-helices whereas all others contain two. Domains 1 and 2 overlap with α-helices a and c in domain 4 whereas domain 3 overlaps best with α-helices b and c (Figure 2.3). Occasionally, the CORA structural alignment will break up a secondary structure in one domain to align it to two secondary structures in another, forcing a gap in the alignment in the secondary structure. In these cases, 2DSEC treats the original secondary structure as a single unit and aligns it to the secondary structure which has the greatest overlap between the two. 2DSEC is designed meaningfully, to reduce the CORA structural alignment into information concerning the absence or presence of secondary structures in the aligned domains and how they relate to one other. The program calculates a summary, containing the number and type of secondary structures found at each position in the alignment.

Consensus secondary structures are identified and shown in a final representation below the domain structures. A secondary structure is considered a consensus secondary structure if it is present in 75% or more of the domain structures. Such consensus elements are identified in a different colour; consensus β-strands are yellow and consensus α-helices are light pink. Additional β-strands are represented in brown and additional α-helices in dark pink.

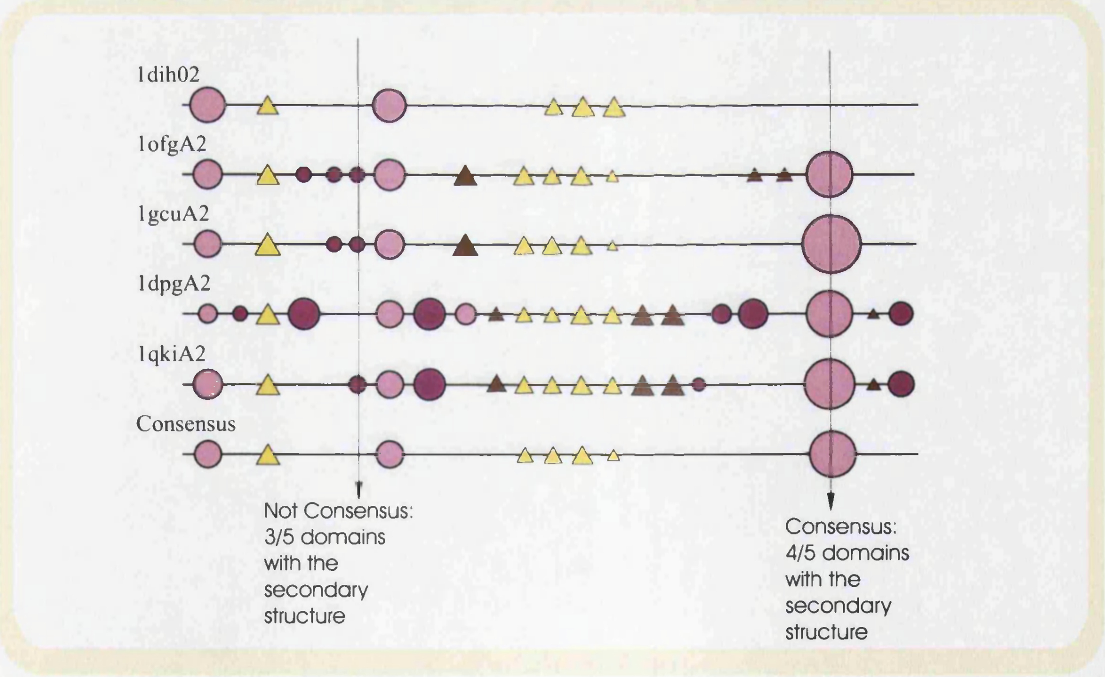


Figure 2.4: 2DSEC diagram of the oligomerisation domain of the NADP binding oxidoreductase superfamily (3.30.360.10). A secondary structure is considered consensus if it is in 75% or more of the aligned domains. In this example of 5 aligned domains the secondary structure must be in 4 or more to be considered as a consensus secondary structure.

2.2.3.2 Mapping the Embellishments Onto the Tertiary Structure

The 2DSEC program also outputs a Postscript file producing a graphical representation for Molscript (Kraulis, 1991), colouring residues in the consensus secondary structures. If 75% of residues at that alignment position are in the α-helical or the β-strand conformation the alignment position is flagged as consensus. This information can be used to label the temperature factor field in the PDB file so that Molscript representations can be coloured according to whether the residue position is in consensus secondary structure or not. Figure 2.5 shows a Molscript diagram of glucose-6-phosphate dehydrogenase. Regions of consensus secondary structure are coloured in red and the embellished secondary structures and loop regions (SVRs) in blue. It can be seen that there is an extension of two antiparallel β-strands onto the right side of the β-sheet and two extra α-helices. It is possible to see regions of loop which are coloured red. This indicates that across the superfamily in more than 75% of structures the strand is longer, forming a larger consensus region. However in this structure the longer β-strand is not present.

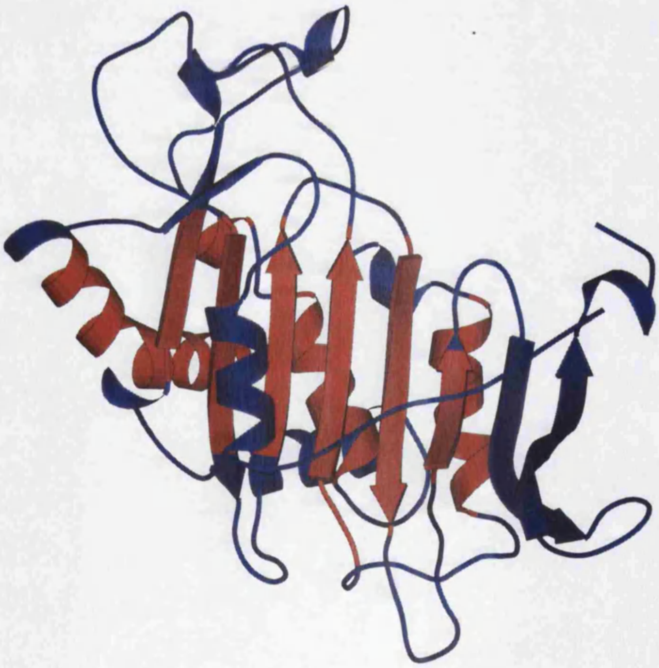
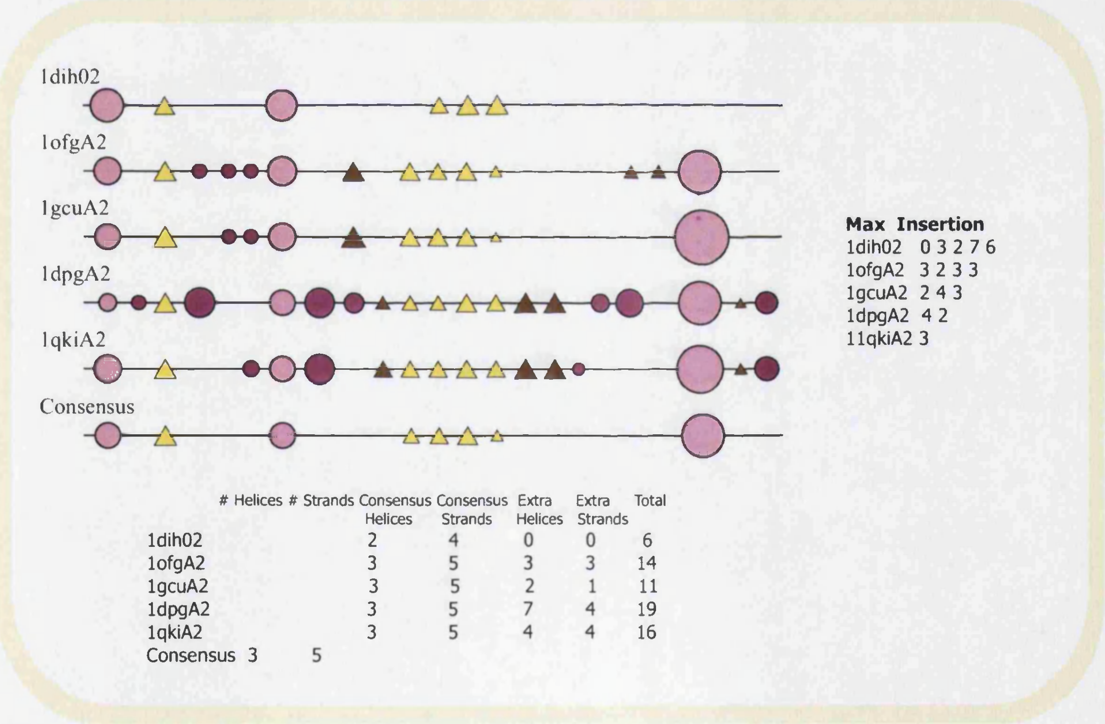


Figure 2.5: The oligomerisation domain of glucose-6-phosphate dehydrogenase coloured to show the residues in the alignment which are part of a consensus secondary structure (coloured red). Other regions such as the embellished secondary structures and the loop regions are coloured in blue.

In addition to the Postscript cartoon, 2DSEC produces a summary file containing information on the secondary structures for each domain in the alignment:

- The total number of consensus helices and total number of consensus strands across the superfamily.
- For each domain, the number of helices and strands present in the consensus identified for the superfamily.
- The number of extra helices and strands in each domain.
- The Total number of secondary structures in each domain.
- The length of the largest insertion present.

This information is used automatically and manually for the analysis of CATH homologous superfamilies in Chapters 3 and 4. Figure 2.6 gives an example of the measurements which are recorded for each superfamily.



Max Insertion

1dih02	0 3 2 7 6
1ofgA2	3 2 3 3
1gcuA2	2 4 3
1dpgA2	4 2
11qkiA2	3

	# Helices	# Strands	Consensus Helices	Consensus Strands	Extra Helices	Extra Strands	Total
1dih02			2	4	0	0	6
1ofgA2			3	5	3	3	14
1gcuA2			3	5	2	1	11
1dpgA2			3	5	7	4	19
1qkiA2			3	5	4	4	16
Consensus	3	5					

Figure 2.6: 2DSEC calculates a summary of secondary structure information for each member in the structural alignment. The information is shown here using the oligomerisation domain of the NADP oxidoreductase superfamily (3.30.360.10). The number of consensus and embellished secondary structures in each relative are also calculated (shown here in the table below the 2DSEC cartoon.)

2DSEC also describes a profile for the superfamily. The information recorded is shown in Figure 2.6. For each domain, the number of consensus strands present is recorded. For example, dihydrocolinate reductase (1dih02) has 2 of the 4 consensus helices and 4 of the 5 consensus strands. The total number of extra helices and strands in each protein are also recorded. Dihydrocolinate reductase has no extra helices and no extra strands. In addition, the largest number of inserted secondary structures in one region, pairwise maximum insertion, is recorded and shown here as a half matrix. In this superfamily, glucose 6-phosphate dehydrogenase (1qkiA2) has the largest insertion of seven SSEs when measured with dihydrocolinate reductase (1dih02).

2.3 Summary

This chapter describes a new algorithm which enables the secondary structure of domains within a given superfamily to be analysed and visualised. 2DSEC reads a CORA multiple structural alignment and summarises the secondary structures aligned. This enables a secondary structure profile of the superfamily to be formed, identifying those secondary structures which are common in the superfamily and those which embellish the core secondary structures. By identifying embellishments it is possible to consider mechanisms of domain evolution and also how these embellishments may affect the function of the protein.

The example of the oligomerisation domain in the NADP binding oxidoreductases used throughout this chapter is illustrated in a final summary in Figure 2.7. The 2DSEC diagram shows that embellishments are present in four main places along the peptide chain. Strands embellish either edge of the β-sheet and α-helices mostly pack against existing α-helices on one side of the β-sheet (Figure 2.7). The embellishments largely promote the interactions between the domains in the multidomain structure. This is discussed in more detail in Chapter 3.

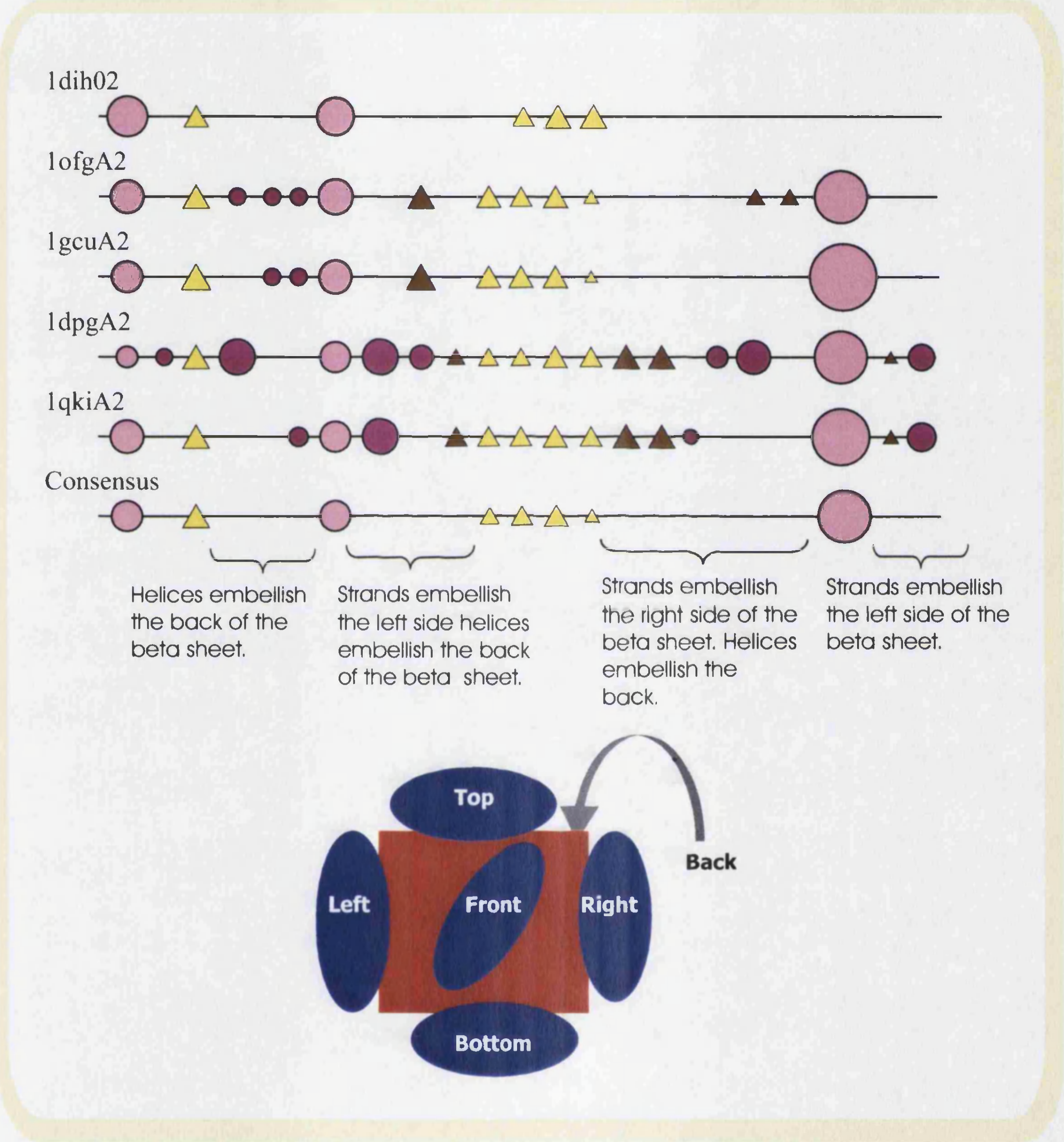

1dih02
1ofgA2
1gcuA2
1dpgA2
1qkiA2
Consensus
Helices embellish the back of the beta sheet.
Strands embellish the left side helices embellish the back of the beta sheet.
Strands embellish the right side of the beta sheet. Helices embellish the back.
Strands embellish the left side of the beta sheet.
Top
Left
Front
Right
Bottom
Back

Figure 2.7: A 2DSEC example. The oligomerisation domain of the NADP binding oxidoreductase superfamily. The embellishments are viewed in two-dimensions using the 2DSEC output and can be mapped onto the three-dimensional structure using a simple description. Embellishments can be seen in four places along the peptide chain. In this example, four regions of embellishments are contributing to the left and right of the central β-sheet and α-helices to the back and the front.

Chapter 3

Structural Evolution in Protein Superfamilies

3.1 Introduction

3.1.1 Background

At present, there are many more known protein sequences (1,497,800 in GenBank) than there are known proteins structures (17,248 structures in the Protein Databank (PDB)). This is, in part, due to the ongoing success of whole genome sequencing projects and the time consuming nature of experimental structure determination methods such as X-ray crystallography and NMR. The elucidation of over 132 published and completed genomes over the last eight years has lead to a vast amount of experimentally uncharacterised sequence data derived from a variety of eukaryotic, bacterial and archaeal sources. Locked into these genome sequences is a wealth of structural and functional information that holds the key to the determination of biochemical processes in which proteins participate.

A challenge of the 'post-genomic' era is the high-throughput assignment of structural and functional information to genes and gene products. Such annotation of sequence data is reliant on the elucidation of those rules that associate sequence and structure and in turn function from structure. Annotation can be achieved in part by the detection of significant sequence similarity between a sequence with unknown structure and a protein of known structure. In cases where significant sequence similarity is found the three-dimensional structure may be constructed by comparative modelling using the backbone conformation of the known structure. Many such bioinformatics-based methods (Tramontano *et al.*, 2001) are now being used to predict protein structure based on current knowledge of experimentally determined structures. However, at present, these methods are unreliable, particularly for more distant relatives as homologous proteins can vary significantly in structure in some superfamilies. Nevertheless, it is not feasible to solve the structures of all proteins by experimental methods as they are far too numerous and so *in silico* prediction methods need to be improved to act as a feasible alternative to experimental methods Chapter 5 on Comparative Modelling). In order to do this a greater understanding of protein evolution is required.

Proteins evolve due to mutations in the DNA encoding them. These mutations may vary from very small changes such as single base insertions and deletions through to much larger changes such as gene duplications (which may be responsible for additional protein domains)(Heringa & Taylor, 1997). As is described by the central dogma of molecular biology, DNA to RNA to protein, changes in the DNA sequence may have effects on the amino acid sequence of the protein it encodes. As changes within protein sequences have accumulated slowly and gradually over time, protein families have sometimes evolved to gain different functions. Close evolutionary relatives may be found through sequence analysis alone – if their sequence identity is greater than or equal to 35% two proteins may be assumed to adopt a similar structure (Chothia & Lesk, 1986). In some relatives sequences may have changed to the extent where they are no longer recognisable as relatives by sequence based methods alone. However, related proteins can have similar or identical structures without having high sequence identity. As the library of these structures becomes more complete, the ability to infer a protein structure by assigning sequences to a given structural family will also increase. Classification of proteins into related groups helps us to see evolutionary relationships between them, and also enables the characterisation of the structural variability between relatives in a homologous superfamily. Databases of protein structures such as CATH (Pearl *et al.*, 2001) and SCOP (Lo Conte *et al.*, 2000) have made this possible by the hierachical categorisation of protein structure.

Prediction methods may be improved by analysing information stored in protein family databases such as CATH in order to understand the structural mechanisms by which proteins evolve and identify which families are the most variable and how structural variability affects function.

3.1.2 Identifying Evolutionary Relationships in the CATH Database

In the CATH database (Version 2.4) (Pearl *et al.*, 2001), a significant evolutionary relationship, i.e. homology, is defined if at least two of the following three criteria are satisfied:

- High sequence similarity (>35% sequence identity, or significant E-Value using PSI-BLAST).

- High structural similarity (SSAP score over 70) or significant E-Value from CATHEDRAL (sections 1.5.5.2 and 1.5.6.1).
- Evidence of functional similarity.

If two of these criteria are met then proteins are clustered into the same homologous group known as a superfamily. Since CATH is a hierarchical classification database, the superfamilies themselves are further clustered into fold groups, or topologies (T-level), that share a similar spatial and sequential arrangement of secondary structures. Proteins that share the same topology but belong to different homologous superfamilies within CATH are given the term analogues. Proteins that do not have similar structures, that is they are not in the same homologous superfamily or fold group, are termed non-relatives. This database provides a valuable source of verified homologous domains, enabling the analysis of sequence, structure and functional relationships.

3.1.3 Convergent and Divergent Evolution

Analogues are generally thought to share a similar folding arrangement by convergent evolution, (Chothia, 1992; Orengo *et al.*, 1994) where proteins arrive at the same fold through an independent evolutionary pathway. Homologues evolve by divergent evolution, where the proteins share a common ancestor. In some cases the relationship is so distant that the only evidence is the structural similarity. The distinction between analogues and homologues is often difficult to identify due to a lack of evolutionary evidence, in terms of sequence or functional similarity. Negligible sequence and functional similarity does not necessarily mean that they are unrelated, only that no evidence of the relationship is currently available. Often a protein sequence or structure will be found that has evolutionary relationships to more than one superfamily and this provides a 'missing link'. This allows superfamilies to be merged, resulting in analogous relationships being redefined as homologous relationships.

3.1.4 Sequence and Structural Variability in Homologous Domains

Generally, it is considered that domains with greater than 35% sequence identity will share a similar structure (Chothia & Lesk, 1986) and will provide a good template for homology modelling of relatives of unknown structure. Therefore, CATH domains clustered into the same superfamily are further sub-clustered into 35% sequence families. In some superfamilies domains remain structurally well conserved at less than 35% sequence identity. However in others, relatives below this threshold can exhibit an extensive amount of structural change including the additions and subtractions of secondary structures around the core structural elements. In some cases these secondary structural changes or embellishments may be constrained by functional requirements and in other cases, it may seem that the differences are more tolerated, where these embellishments do not affect the stability of the structure or the ease of folding.

The number of structures in each architecture in CATH is uneven (Figure 3.1), the architectures 3-layer ($\alpha\beta\alpha$) sandwich (3.40), 2-layer sandwich (3.30), β sandwich (2.60), β barrel (2.40) and orthogonal bundle (1.10) comprise 69% of the total number of non-identical representatives in the database. This may be because these architectures are more regular enabling optimal packing of hydrophobic residues in the core. Alternatively, structures adopting these architectures may crystallise more easily. However, it may also be that these architectures contain many enzymes which have been more extensively studied by structural biologists.

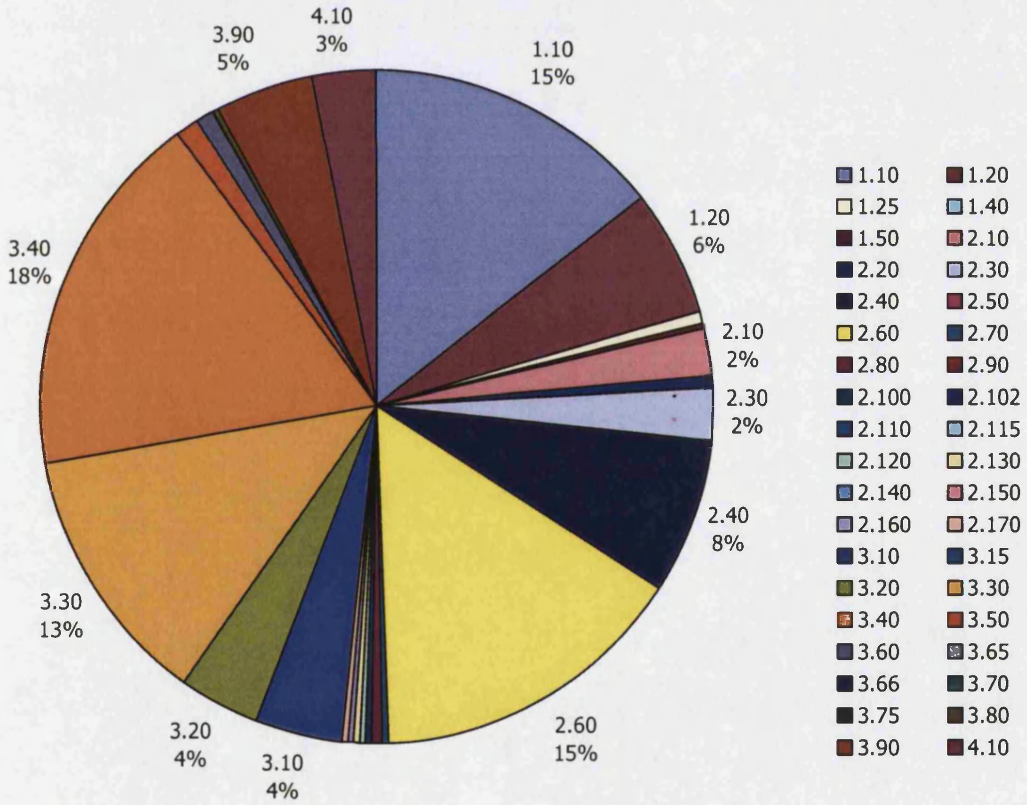


Figure 3.1: The number of non-identical domains in each architecture in CATH shows an uneven distribution. Some architectures are more highly populated: 3-layer ($\alpha\beta\alpha$) sandwich (3.40), 2-layer sandwich (3.30), β sandwich (2.60), β barrel (2.40) and orthogonal bundle (1.10). These five architectures constitute just under 70% of the database.

Structural evolution through insertions and deletions of secondary structure elements can sometimes give rise to profound changes in the fold or architecture of a protein. For example, the addition of α-helices can change a 2-layer ($\alpha\beta$) sandwich to a 3-layer ($\alpha\beta\alpha$) sandwich. Grishin (2001) discusses fold change in homologous structures in several superfamilies, showing how certain evolution events can change the topology of a protein structure (section 1.3.1). Grishin's study shows that there are some homologues which can be detected by sequence analysis that fold into different structures producing contradictions in the protein classification schemes. By contrast, measuring structural similarity between proteins can identify common structural motifs present in a number of different folds and homologous superfamilies. Harrison *et al.* (2002) identify a number of these motifs naming the folds that possess them 'gregarious' folds.

3.1.4.1 Previous Analysis of Sequence/Structure Relationships in Protein Families

The relationship between structure and sequence can also be studied in terms of the number, length and structural location of insertions and deletions tolerated within a set of homologous structures. Since an insertion in one sequence of an aligned pair implies a deletion in the remaining sequence, it is referred to as an indel. Pascarella & Argos (1992) studied the indels in protein structural families, each consisting of non-redundant and multiple tertiary structural superpositions. They discovered that indels prefer to be between 1-5 residues in length with very few examples (1-2%) of indels greater than 10 residues. Flores *et al.* (1993) carried out a similar study on a set of homologous pairs of proteins ranging from 0 to 100% sequence identity and similarly reported that indels prefer to be 1-6 residues in length. A continuation of the study by Pascarella & Argos (1992) using a larger dataset is carried out in this chapter (section 3.3.5).

The relationship between sequence and structure has been the subject of a number of studies. Two models have been proposed to explain how the tertiary structure of a protein is encoded in its linear sequence of amino acids. Firstly, the local model, postulates that fold specificity is coded by just a few critical residues (10–20% of the sequence) (Chothia & Lesk, 1986). The second model, the global model, postulates that the fold is formed by interactions involving the entire sequence (Lattman & Rose, 1993); and more recently Wood & Pearson (1999) conclude that, on average, most sequence changes cause detectable structural changes and that the amount of structural change per sequence change is relatively constant within a protein family. In addition Wood & Pearson (1999) conclude that the rate of change in sequence identity produces a different rate of change in structure for different structural families.

Evidence supporting the local model is extensive. The logarithmic relationship between sequence change and structural change was first reported by Chothia & Lesk (1986) and again in the study by Flores *et al.* (1993). A number of studies involving individual families have also identified conserved amino acids considered important for protein folding (Ptitsyn, 1999, 1998). Ptitsyn (1998) analysed seven subfamilies of cytochromes c. He found four completely conserved positions in all cytochrome c subfamilies which form a network of conserved contacts connecting the N– and C–terminal helices. The importance of the contacts between the interfaces of these helices has been confirmed by their existence in molten globule-like folding intermediates. These residues have no apparent functional role further suggesting that they are of importance in protein folding. A similar study was also carried out on the globin family (Ptitsyn & Ting, 1999) where a cluster of conserved residues with no functional importance was identified. These residues are located between helices which are known to fold in the early stages of folding and remain relatively stable in the equilibrium molten globule state.

Protein superfamilies evolve at different rates and by different mechanisms and it is useful to know how structural properties vary as sequence identity varies. Many different criteria can be used to characterise protein structures, for example, the difference in secondary structure content or the change in residue solvent accessibility. How do the structural properties change with evolution of a particular family? Flores *et al.* (1993) measure the conservation of secondary structure in 90 pairs of homologous pairs having sequence identities ranging from 5 to 100%. The percentage of residues in the same secondary structures decreased linearly as percentage sequence identity decreased. In another study conducted by Russell & Barton (1994) it was found that secondary structure identities in distant homologous proteins can fall as low as 41%, equivalent to what might be expected by chance. Secondary structure variation in terms of conservation and substitution was also examined by Mizuguchi & Blundell (2000). Secondary structures were classified into categories according to the length of the secondary structure and its solvent exposure. From this, a secondary structure element (SSE) substitution table was calculated. Substitutions from SSEs to coil, or deletions of SSEs were calculated and it was found that length was the biggest factor in determining the probability of deletion, although, short and medium buried strands are much less likely to be substituted by coil than accessible ones. This substitution table is useful for the comparison of known secondary structures and sequences with predicted secondary structures.

The above examples illustrate the ways in which structural and sequence variation can be measured and how the identification of structural or sequence commonalities between homologous family members can enlighten the mechanisms of evolution. It is important to elucidate why structures in some folds remain conserved at low sequence identities whilst other folds are more tolerant to structural variation and secondary structure embellishments, as this can help in predicting the structures of new sequence members. At the very least, methods for the identification of core, or structurally invariable, regions present in all family members can lead to measures of confidence in core regions of the structure when modelling new members. Furthermore, understanding more about the rate of evolution, type of evolutionary mechanisms and the effect these mechanisms have on individual folds is important for the classification of new members into the homologous superfamily.

3.1.5 Aims of the Chapter

This chapter reports the analysis of highly populated families in the CATH domain database in order to characterise the types of structural changes occurring during evolution. This is important for understanding constraints on protein evolution and the mechanisms by which proteins evolve structurally. In this chapter, structural variability is assessed in terms of the insertions and deletions tolerated throughout CATH superfamilies, both at the sequence and the secondary structure level. CATH domain superfamilies are characterised in terms of the secondary structures they have in common and those secondary structures which are embellishments to the core of the fold. This characterisation is useful when adding new structural members to the superfamily and also when modelling the structures of sequence relatives.

Analysis of the secondary structure variation in relatives between superfamily members is then performed using the 2DSEC program described in Chapter 2 which automatically identifies those secondary structures which are core to the fold and those which are embellishments based upon the use of a multiple structure alignment produced by CORA (Orengo, 1999). The 2DSEC program calculates a measurement of variability within each superfamily which is used for identifying those superfamilies which are most embellished and those superfamilies which are most conserved. In addition to this, 2DSEC is used to analyse how these structural embellishments are inserted into the peptide chain. Are many SSEs inserted into one place or are a few inserted into many places and how are they arranged in the three-dimensional structure? Are they localised in one region on the structure or in a number of regions? Additionally, sequence/structure variation is also investigated for these superfamilies to help elucidate the extent to which the structure changes as sequence identity falls. Studies measuring the relationship between sequence (measured by percentage sequence identity) and structure (measured by SSAP score (Taylor & Orengo, 1989) described in section 1.5.5.2) and the relationship between the number of indel residues (using a program called IndelCalc described in section 3.2.3) and sequence identity were carried out to help elucidate possible mechanisms for structural evolution.

3.2 Methods

3.2.1 Selecting Datasets

Each CATH (V2.4) superfamily is sub-classified according to different levels of sequence identity. In a given superfamily, structural comparisons were made between representative domains from 35% sequence clusters (S35Reps), as structures are considered to be similar at greater than or equal to 35% sequence identity. In section 3.3.6 where structures are compared, only superfamilies with more than three S35Reps were considered. The total number of superfamilies with three or more S35Reps was 235 containing 1403 S35Reps altogether. For sections 3.3.3 and 3.3.5 relationships at a higher sequence identity are measured. For this the superfamily was divided into clusters at 95% sequence identity (N95Reps) to remove redundancy. A total of 3311 N95Reps were included in the study. In both datasets, the domain with the best resolution was selected. If only NMR structures were available in a cluster, no structure was selected as the secondary structure elements tend to be not as well defined.

3.2.2 Structural Change versus Sequence Change

The structural similarity between two proteins is measured by the program SSAP (section 1.5.5.2) as a value between 0–100, with identical proteins returning a value of 100. SSAP also outputs the percentage sequence identity of the pair, calculated from the structural alignment.

3.2.3 Measuring Indels Between Protein Pairs

IndelCalc

IndelCalc was developed to measure the length and secondary structure type of inserted residues between domain pairs. Domain pairs were aligned using the pairwise SSAP structural alignment program. Each S35Rep was aligned to the N95Reps in the superfamily and the position and length of each indel position is recorded. The positions of the indel residues are then cross-referenced to check that no indel is recorded twice. Figure 3.2 illustrates the overlap of indels between an S35Rep and two N95Reps. The second indel is in both N95Reps but is counted only once.

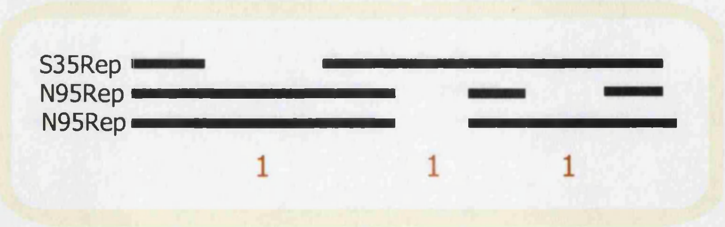


Figure 3.2: Calculating the length of insertions using IndelCalc. Each S35Rep is aligned with all N95Reps, any overlapping alignments are counted only once. The red numbers below show how many times each indel is counted.

3.2.4 Secondary Structure

3.2.4.1 Identifying Superfamilies with Domain Embellishments

2DSEC

Although homologous superfamilies share core structural similarity, many of these superfamilies exhibit extensive structural embellishments which enlarge the domain unit. Superfamilies with secondary structure embellishments are identified using 2DSEC (Chapter 2).

The percentage of secondary structure elements (SSE) present in the largest but not present in the smallest structure is calculated to represent the variability in each superfamily (Equation 3.1).

$$\text{Percentage Variability} = \frac{\text{Total SSE in the Smallest}}{\text{Total SSE in Largest}} \times 100 \qquad (3.1)$$

Analysis of secondary structure insertions

The structures were inspected manually and the length (number of SSEs) and types of embellishments were described. This analysis provides information on how these secondary structures embellish the tertiary structure. Are the secondary structure insertions distributed throughout the structure or do they aggregate in one structural location to produce a larger motif or additional lobes on the tertiary structure? If they are inserted in many regions in the polypeptide chain, are they nevertheless contiguous in 3D or located in different positions throughout the structure?

Description of embellishment falls into the following categories:

- How many regions in the peptide chain contain insertions?
- Are these insertions α-helices or β-strands?

- How many regions in the three-dimensional structure are embellished with insertions?
- If the insertions are β-strands. Do they:-
    1. Form an extension to a central β-sheet or form a separate β-hairpins or β-sheet?
    2. Form an extra lobe or do the embellishments enlarge the structure uniformly?
- If the insertions are α-helices. Do they:-
    1. Form a separate lobe on the structure?
    2. Pack uniformly around the structure to enlarge the domain unit uniformly?

This categorisation was made by comparing all of the proteins in a family by eye using Rasmol, and as such is a qualitative measure of variation.

3.2.5 Tertiary Structure

3.2.5.1 Identifying Structural Variability using SSAP

For each superfamily pairwise SSAP scores were calculated. From this, structural variability of the superfamily was assessed by calculating the mean and the standard deviation (σ). An indication of the extent of structural variability of the superfamily can be gained from this information. If a superfamily has a low mean SSAP score (<60) and a large standard deviation the family is uniformly variable, indicating a plastic fold. By contrast, a superfamily with a high mean SSAP score (>80) with a low standard deviation indicates a superfamily which, at present, is structurally conserved. It is also important to consider the number of relatives in the superfamily when identifying structurally conserved superfamilies.

3.2.5.2 A CATH Resource: Structural Sub-groups

Due to the extensive structural embellishments in some homologous superfamilies, Structural Sub-Groups (SSGs) are used (Sillitoe, 2002). These divide the superfamilies into structurally similar, sequence dissimilar clusters. Redundancy in the superfamily was first removed by selecting the highest resolution structure from each 35% sequence cluster (S35Rep). A pairwise matrix of SSAP scores is read in and, starting with the highest pairwise SSAP score, clusters are built up based on structural similarity. Structures join a cluster if the SSAP score is above a given threshold (SSAP $>$ 80) to all other members of that cluster. Multiple linkage clustering was chosen over single linkage clustering so that internally consistent clusters were produced, that is, all members of the cluster exhibit the same core structural features (Figure 3.3). In superfamilies exhibiting significant structural variation more accurate structural alignments are sometimes obtained by using the sub-groups. In these superfamilies the SSGs were used and the consensus secondary structures were cross referenced between SSGs to calculate the percentage variability. In addition, the number of clusters each superfamily produces is a good indication of the structural variability present.

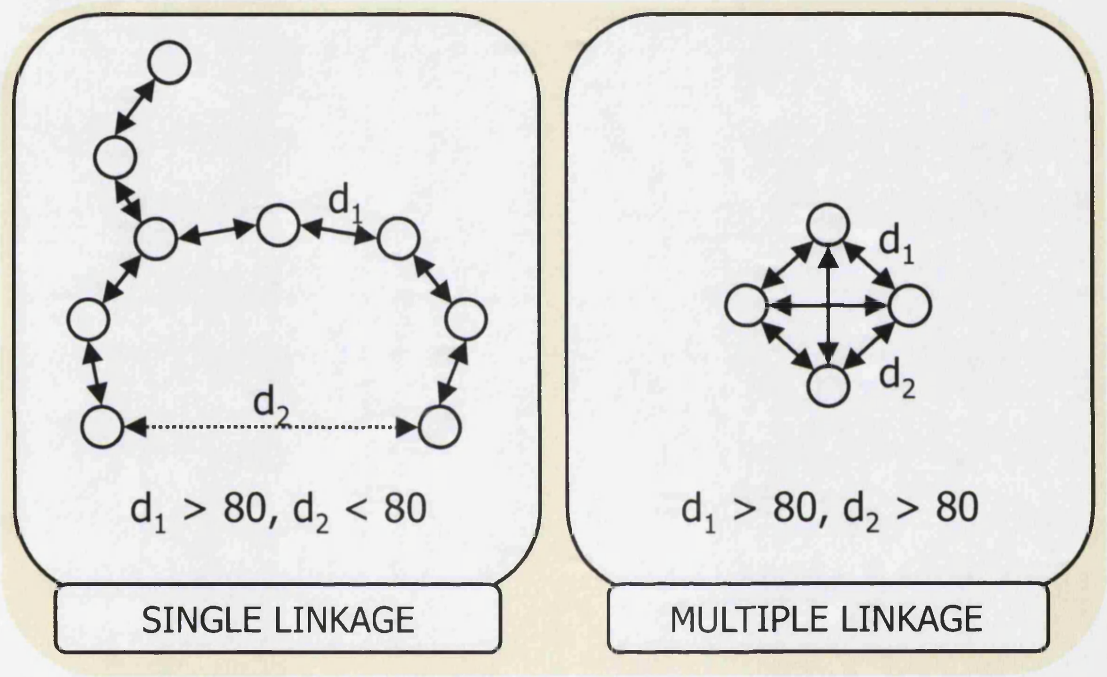


Figure 3.3: Single and multiple linkage clustering. Single-linkage clustering only requires one comparison to meet the clustering criteria (e.g. SSAP score $d_1 > 80$) for a structure to be included in a cluster. This allows structures to be chained together and can result in clusters containing very diverse structures (e.g. SSAP score $d_2 < 80$). Multiple-linkage will only allow a structure to join a cluster if the clustering criteria is met with all members of the cluster (e.g. SSAP score for $d_1, d_2, ..., d_n > 80$).

3.3 Results and Discussion

3.3.1 Distribution of Sequence Identities between Homologues in the CATH Database

Because structural biologists have largely focused on solving the structures of putative novel folds and mutants, sequence identities between non-identical representatives of the same superfamily are mainly between 5–25% and 95–100% with few structures in the 30–95% range. Figure 3.4 shows the sequence identities between N95Reps in the same homologous superfamily in the CATH database. The majority have between 0–30% sequence identity leaving a range of sequence identities from 30–95% with fewer examples.

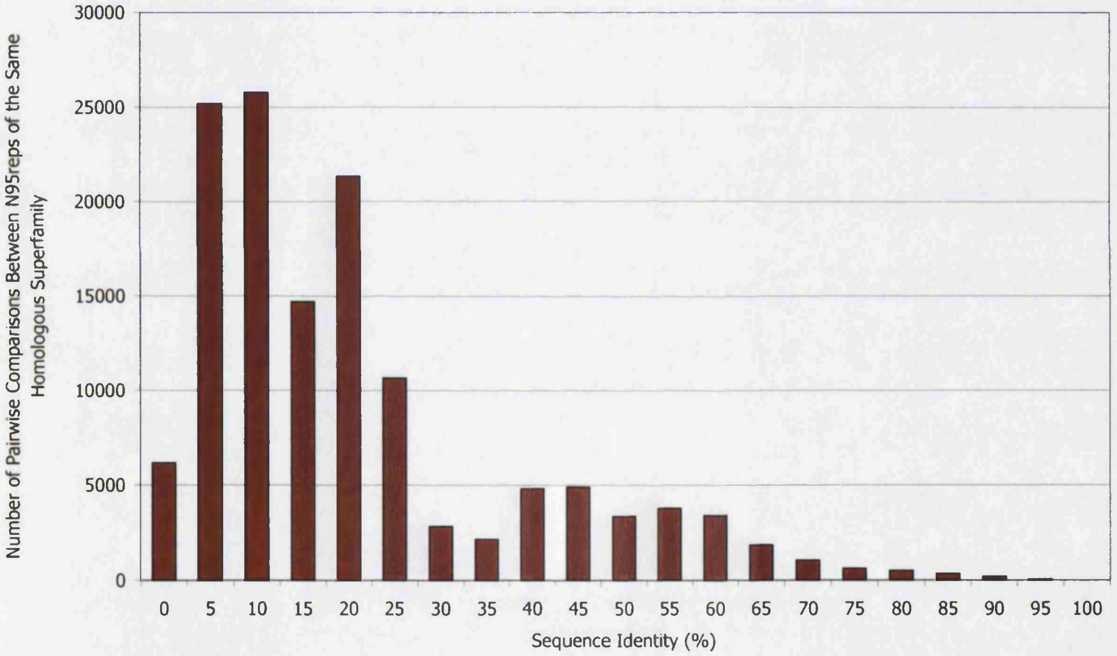


Figure 3.4: Sequence identities between non-identical representatives (N95Reps) in each homologous superfamily shows that most sequences have between 5–30% sequence identity with a lack of structures sharing 30–95% sequence identity.

3.3.2 Identification of Structurally Conserved and Variable Superfamilies by Analysis of Pairwise Structural (SSAP) Similarity Scores.

Global structural variability between pairs of S35Reps in each superfamily was measured using the SSAP structural alignment program, outputting a score ranging from zero for unrelated proteins to one hundred for identical proteins. The average and associated standard deviation of each SSAP score was calculated for each superfamily. Figure 3.5 plots the values calculated for each superfamily. Mean SSAP scores do not generally fall below 65 and only the most diverse families lie between 65 and 75. Superfamilies with a very high mean SSAP score (>85) could be considered to be conserved although many of these superfamilies contain five or fewer S35Reps. It is not yet known if these superfamilies will be found to be more structurally divergent as they become more populated.

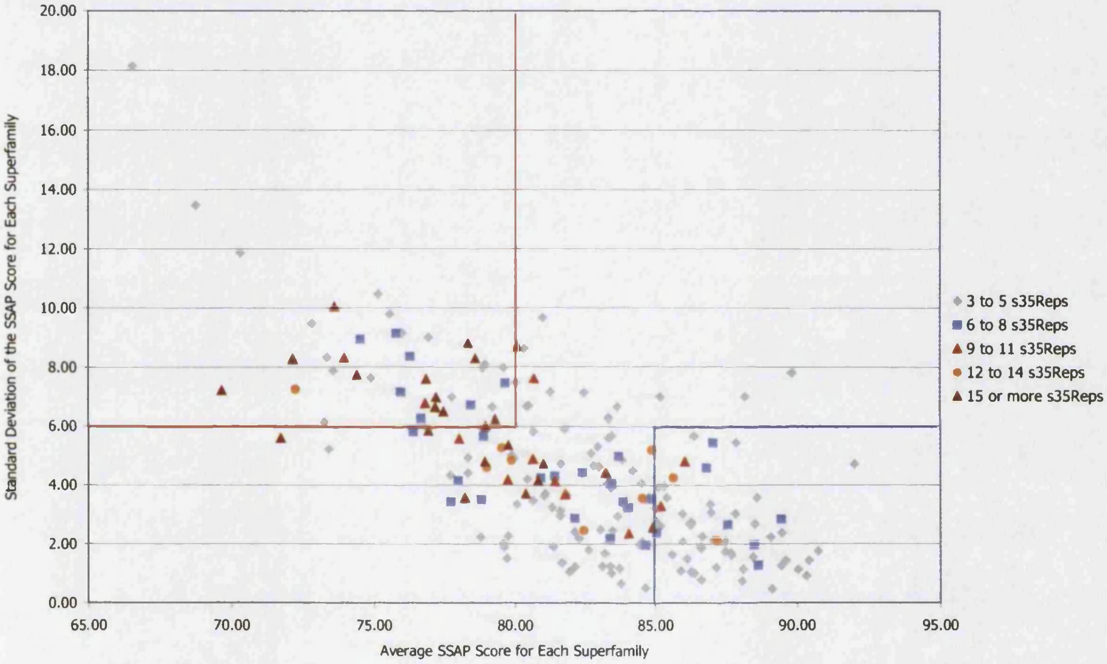


Figure 3.5: The average SSAP score and standard deviation calculated for each superfamily in the dataset. The mean SSAP score ranges from 70 for diverse to 95 for conserved superfamilies. Superfamilies are coloured according to the number of S35Rreps. Two boxes have been marked on the graph. The blue box indicates those particularly conserved superfamilies with high mean SSAP scores and low standard deviation and the red box indicates those superfamilies which are particularly variable with low mean SSAP scores and a higher associated standard deviation.

3.3.3 The Evolutionary Relationship Between Sequence and Structure

The relationship between sequence and structural evolution has been examined in a number of studies. As discussed in section 3.1.4.1, two types of sequence/structure evolution have been proposed: The global model, which suggests that every sequence mutation creates some structural modification (Lattman & Rose, 1993; Wood & Pearson, 1999)and by contrast, the local which model implies a small percentage of the residues are responsible for large structural changes (Chothia & Lesk, 1986; Flores *et al.*, 1993; Ptitsyn, 1999, 1998).

To examine the relationship between structural similarity and sequence identity pairwise SSAP structural comparisons were carried out for all non-identical representatives (N95reps) in each superfamily in CATH. Plots were generated of pairwise SSAP scores versus sequence identity, calculated from the structural alignment. Figure 3.6 shows that for four of the most populated superfamilies, there was gradual decrease in structural similarity with decreasing sequence identity, above 20–30%. Whilst below 25% sequence identity, sequence change had a much greater effect on the structure. The same behaviour was found for 62 superfamilies containing sufficient numbers of S35Reps to study this range of sequence identity. This behaviour agrees with the early observations of Chothia & Lesk (1986) but for a much larger dataset of structural superfamilies.

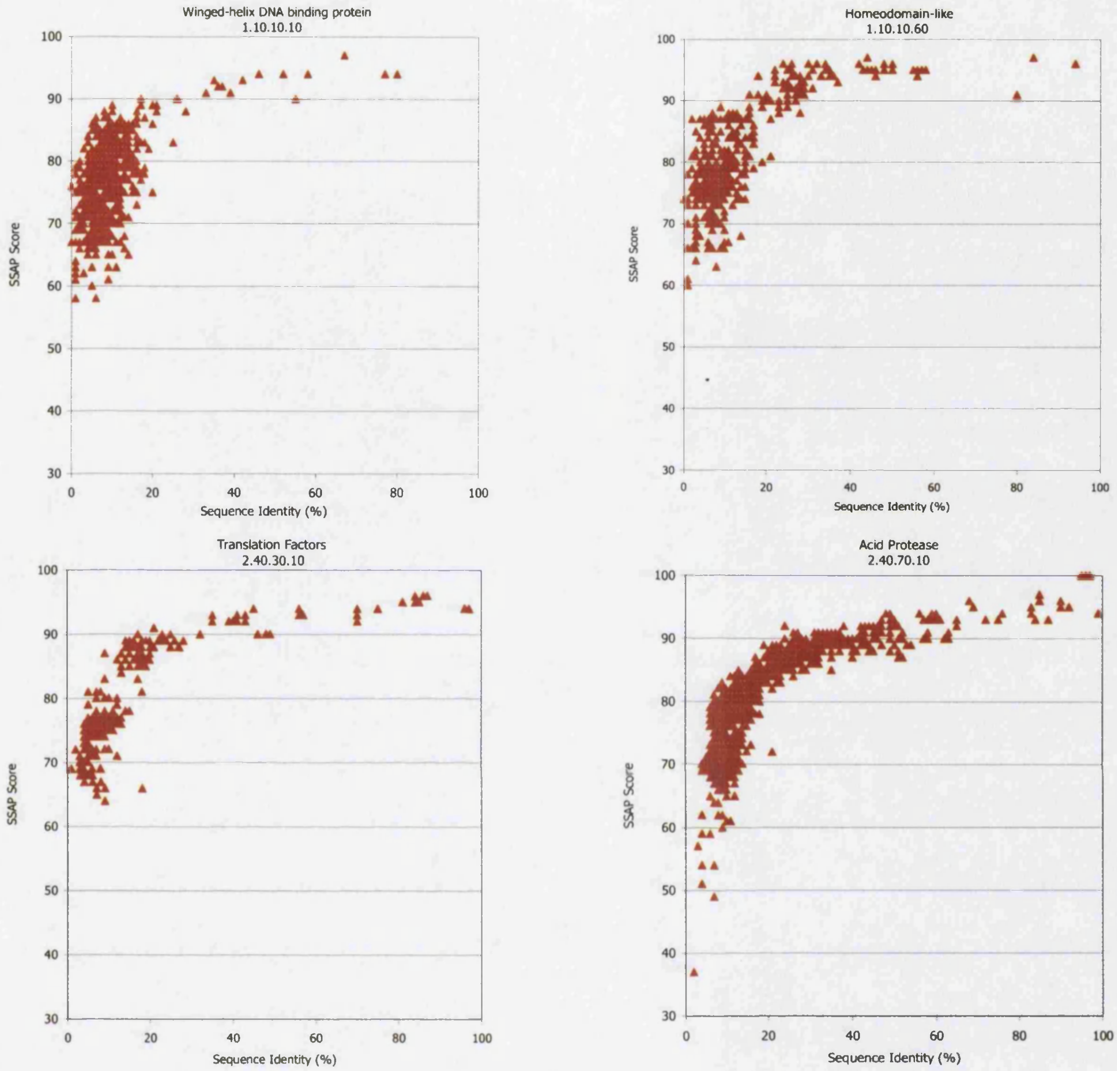


Figure 3.6: The relationship between change in sequence and structure is shown here in four well populated superfamilies. Above ~30% sequence identity there is a gradual change in structure with decreasing sequence identity whereas below ~30% sequence identity the change in sequence has a much greater effect on the structure.

The linear portion of the SSAP/sequence identity plot, >~30% sequence identity, corresponds to the global mode of structural variation proposed by Wood & Pearson (1999). In this region there appears to be a linear dependence of structural similarity with sequence identity. The rate of decrease of structural similarity with decreasing sequence identity was referred to by Wood and Pearson as structural mutation sensitivity (SMS). Their analysis considered 36 structural families. Here, only those superfamilies with at least 10 S35Reps and with a correlation coefficient (R^2) greater than 0.5 for this region, have been considered with at least 10 pairs having sequence identities in the range 35–100% sequence identity. This ensures that the superfamily has been sufficiently sampled to allow general conclusions to be made about the structural 'plasticity' of the superfamily.

There are 62 superfamilies with more than 10 pairwise comparisons showing >35% sequence identity. Selecting those linear relationships with a correlation coefficient (R^2) of greater than 0.5 left 29 superfamilies listed in Table 3.1.

CATH	**Architecture**	**Gradient (>35% sequence identity) SMS**	**Correlation Coefficient**	**Number of pairwise comparisons >35% sequence identity**	**Lowest pairwise SSAP score in superfamily**
1.20.85.10	Up-down Bundle	0.19	0.64	12	61
3.40.50.10	3-layer ($\alpha\beta\alpha$) sandwich	0.15	0.91	11	65
3.30.540.10	2-Layer Sandwich	0.14	0.72	10	74
2.60.120.20	Sandwich	0.138	0.68	112	36
3.10.100.10	Roll	0.13	0.63	51	66
3.30.500.10	2-Layer Sandwich	0.122	0.635	122	80
2.60.120.180	Sandwich	0.12	0.818	36	81
3.20.30.70	Barrel	0.18	0.67	17	66
1.20.1050.10	Up-down Bundle	0.12	0.60	44	76
3.40.420.10	3-layer ($\alpha\beta\alpha$) sandwich	0.12	0.78	24	87
3.40.710.10	3-layer ($\alpha\beta\alpha$) sandwich	0.11	0.67	33	57
1.10.520.10	Orthogonal Bundle	0.11	0.81	25	80
2.60.120.60	Sandwich	0.11	0.74	301	45
2.40.50.110	Barrel	0.11	0.74	20	72
3.40.190.10	3-layer ($\alpha\beta\alpha$) sandwich	0.11	0.55	115	45
2.60.40.420	Sandwich	0.11	0.54	77	44
3.50.50.60	3-Layer($\beta\beta\alpha$) sandwich	0.11	0.53	36	43
3.90.180.10	Complex	0.10	0.86	29	84
3.40.50.80	3-layer ($\alpha\beta\alpha$) sandwich	0.10	0.74	12	80
2.40.70.10	Barrel	0.10	0.67	167	37
3.30.390.30	2-Layer Sandwich	0.09	0.76	15	82
2.60.40.1180	Sandwich	0.09	0.69	22	77
1.10.530.10	Orthogonal Bundle	0.09	0.62	63	56
1.10.420.10	Orthogonal Bundle	0.09	0.70	26	79
3.40.192.10	3-layer ($\alpha\beta\alpha$) sandwich	0.08	0.85	14	84
3.90.70.10	Complex	0.08	0.55	95	73
3.40.50.1460	3-layer ($\alpha\beta\alpha$) sandwich	0.07	0.56	10	86
3.40.640.10	3-layer ($\alpha\beta\alpha$) sandwich	0.06	0.62	24	68

Table 3.1: The sequence/structure correlation at sequence identity above 35%. The structural mutation sensitivity and correlation coefficient (R^2) is recorded for all superfamilies with more than 10 homologues with more than 35% sequence identity and a linear correlation coefficient of greater than 0.5. Also recorded is the SSAP score of the most structurally diverged pair.

Structural mutation sensitivity ranges from 0.19 to 0.06 in this dataset of superfamilies with an average measurement of 0.11. Figure 3.7 illustrates this range for three superfamilies, one with a high SMS, one with an average SMS and one with a low SMS. Even for the most plastic superfamilies the structural changes are not profound. That is, a large change in sequence identity (35–95%) corresponds to a small degree of structural change

(<10%). This tends to confirm that performing comparative modelling for relatives with more than 35% sequence identity is likely to be reliable whilst below 35% the structural divergence is greatly increased and therefore introduces inaccuracy in the models. The range in SMS value could also be useful for predicting the quality of models. Parent structures for more plastic superfamilies, with higher SMS values, need to be selected at the highest possible sequence identity. Also recorded in Table 3.1 is the lowest pairwise SSAP score for each superfamily. In some superfamilies this score can be as low as 45 (2.60.120.60, 2.40.190.10, 3.50.50.60) however, this does not correlate with the associated structural mutation sensitivities.

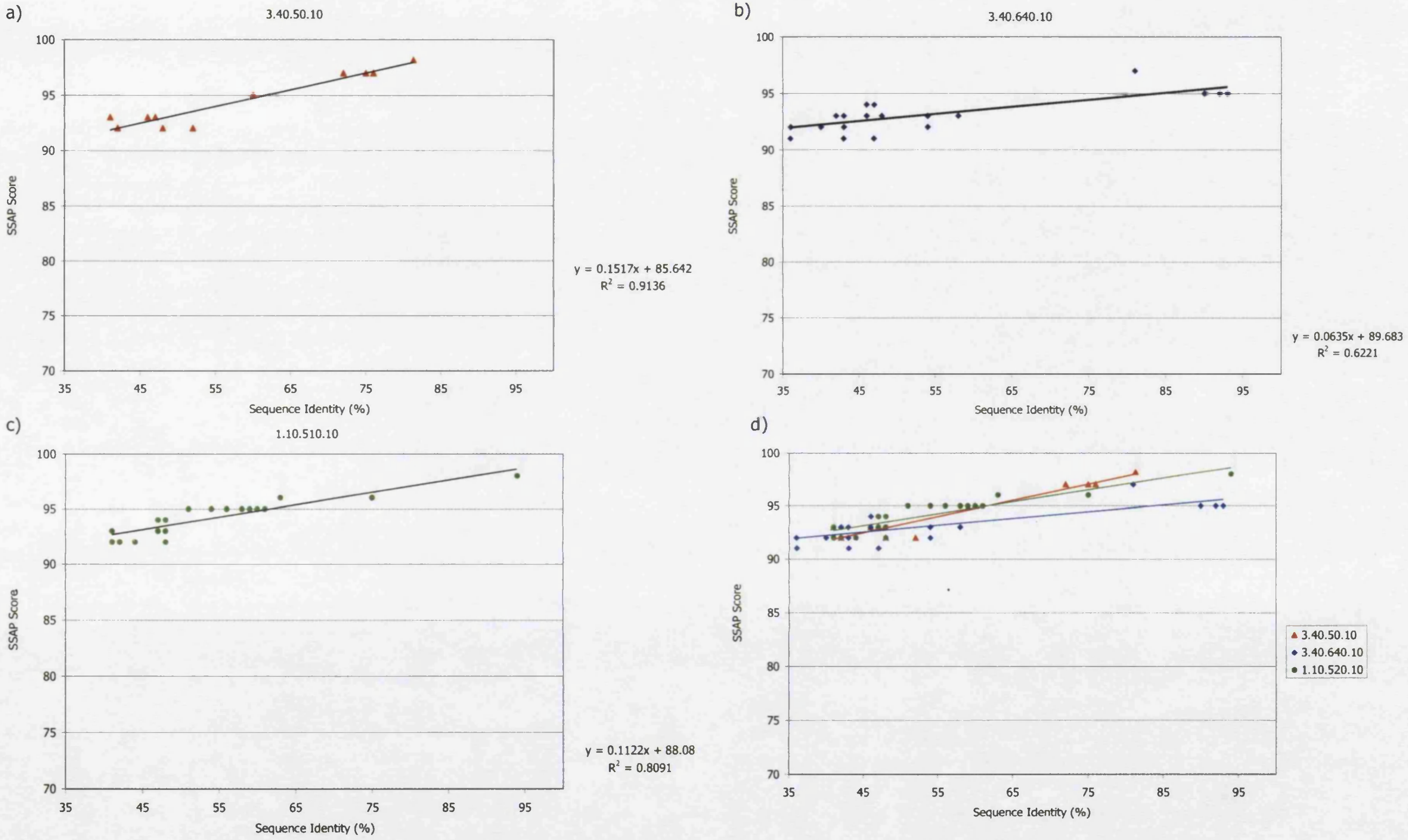


Figure 3.7: Showing the relationship between three superfamilies in the range of 35 to 95% sequence identity. Plot a shows the superfamily 3.40.50.10 with the highest strutural mutation sensitivity of 0.15, b shows the superfamily 3.40.640.10 with the lowest structural mutation sensitivity 0.063 and c shows the superfamily 1.10.510.10 with a middle range structural mutation sensitivity of 0.11.

The data were further divided into structural classes (Mainly α, Mainly β and $\alpha\beta$). Table 3.2 illustrates the number of entries for each class above the mean SMS value, below or equal to the mean SMS value and also the percentage of each structural class present in the dataset above the mean SMS value. The Mainly $\alpha$and the $\alpha\beta$ classes have a higher percentage of superfamilies with an above average SMS value suggesting that for the dataset of structural superfamilies analysed, these classes appear to be more plastic. This was in contrast to Wood and Pearson who reported no correlation between structural class and SMS. However, more structural data would need to be included into the study for a reliable conclusion to be drawn.

	Mainly α	Mainly β	$\alpha\beta$
SMS above average of 0.11	2	2	6
SMS below or equal to average	3	5	10
Number of superfamilies analysed	5	7	16
Percent of class above average	66%	40%	60%

Table 3.2: The structural mutation sensitivity correlated with structural class. The number of homologous superfamilies with an SMS below the average, above the average and the percentage of superfamilies above the average is shown.

3.3.4 Correlation of Structural Variation With Functional Variation

Todd *et al.* (2001) found that function is typically conserved down to 35–40% sequence identity but below this threshold functional conservation falls sharply. Figure 3.8 illustrates the correlation between sequence, structural and functional similarity in four superfamilies (Todd, 2001). It is important to consider the distinction between orthologues and paralogues when analysing these plots. Orthologues are genes in different species that have evolved from a common ancestral gene by speciation and paralogues are genes related by duplication within a genome. Orthologues typically retain the same function in the course of evolution, whereas paralogues often evolve new functions, even if these are related to the original one, possibly allowing greater deviation between paralogous structures since paralogues are not constrained to perform the same function and are often retained in the genome because they perform other useful functions. However it is clear from Figure 3.8 that the relationship between sequence/structure/function is not a simple one. In Figure 3.8 the TIM barrel glycosyl hydrolases, the thioredoxins and the di-iron carboxylate proteins all reveal most functional change occuring at below 30% sequence identity. However, in the α/β hydrolase superfamily structurally variable domains with negligable sequence identity have the same function. Variations functional and structural change in superfamilies could possibly be due to other domains are responsible for the main function of the protein or perhaps the structural variation does not have any bearing on the active site. In addition, mutation of a few residues in the active site can dramatically alter the function. The effect that structural variation has on the function of the domain is studied more closely in three of these functionally annotated superfamilies in Chapter 4.

A continuation of this study of structural plasticity in protein superfamilies would be to consider the functions of the proteins and also whether they are orthologous or paralagous. However, such an analysis was outside the scope of this chapter, as presently, relatives in the CATH database do not always have sufficient functional annotation and there is currently no distinction between orthologues and paralogues.

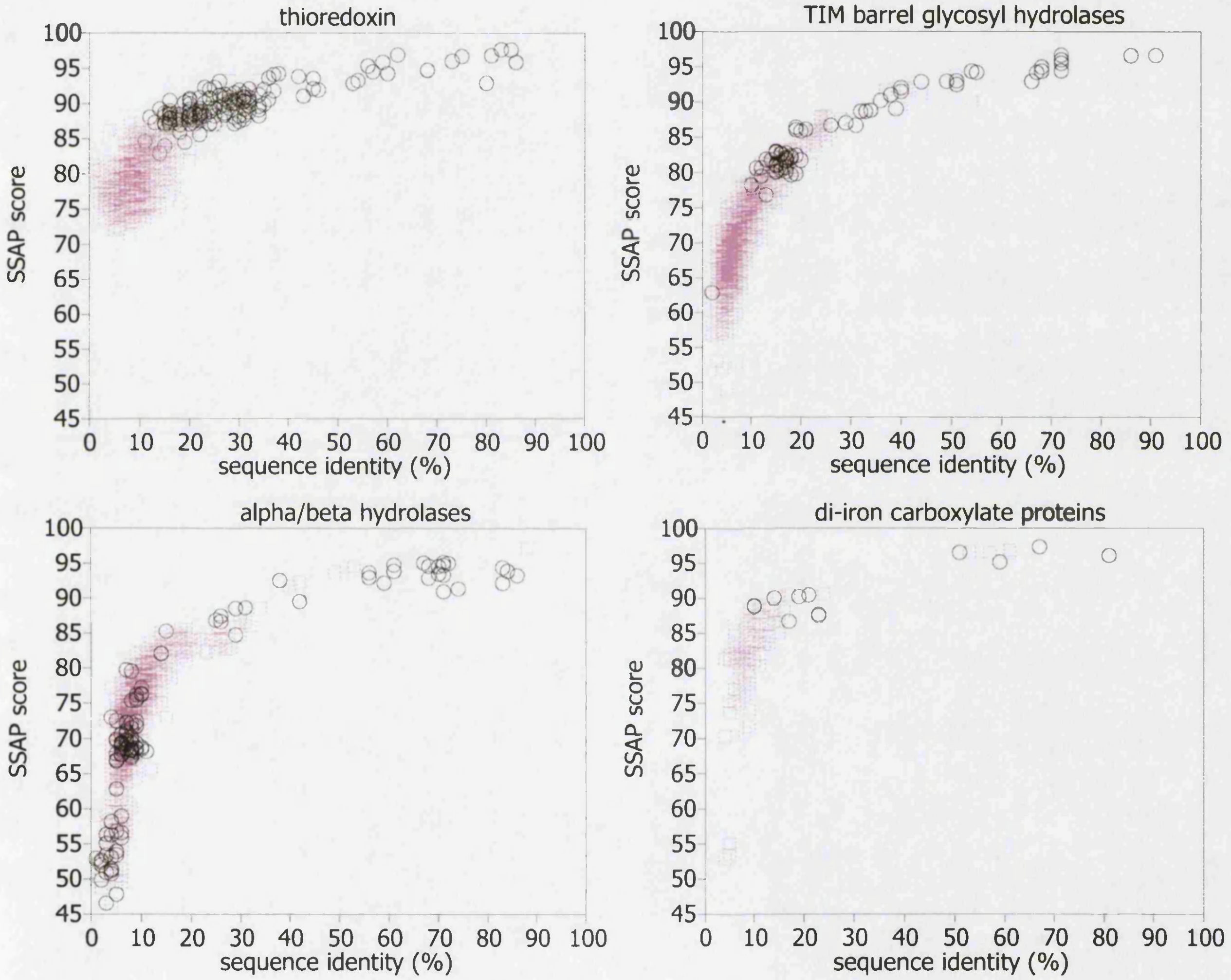


Figure 3.8: The graphs above plot the pairwise sequence identity versus SSAP score for non-identical homologues in domain pairs, with points coloured to distinguish pairs having identical (black circles) and different (pink squares) functions.

3.3.5 Extent of Insertions and Deletions Between Superfamily Members

Evolution at the sequence level occurs by two mechanisms. Residues can be mutated from one type to another, known as substitution, or residues can be inserted or deleted from the protein sequence forming regions known as indels. Indels may consist of a single residue or a whole string of residues, which can sometimes correspond to the addition or subtraction of a whole secondary structure unit. In this section the average length of insertions and deletions between an aligned protein pair is considered and compared to the study undertaken by Pascarella & Argos (1992).

Using IndelCalc (section 3.2.3) the relationship between sequence identity and number of indels in each pair was measured. In this study, pairwise analyses of relatives from 235 homologous superfamilies (45 500 pairwise comparisons) was carried out.

3.3.5.1 Average Indel Lengths

Pascarella & Argos (1992) reported the analysis of the insertions and deletions present in 32 structural families (755 comparisons of non-redundant proteins). They concluded that indels tend to be short, between two residues (at >40% sequence identity) and five residues (at <10% sequence identity). The increase in length was found to be exponential with decreasing sequence identity. Calculating the average indel length and the percentage of indels less than ten residues at 0–20%, 20–40% and 40–80%, indicated the relatively narrow distribution of indel lengths (Table 3.3). The percentage of indels below ten residues remains high at all three sequence identity ranges. The average indel length decreases from 4.6 for a sequence identity from 0–20% to 2.3 for a sequence identity from 40–80% although the standard deviations show that the spread of indel length in the 0–20% range is wide.

Sequence Identity Range (%)	**Mean (s.d.)**	**Percentage Indels Less than 10 residues**
0–20	4.6(5.9)	93%
20–40	3.0(3.4)	94%
40–80	2.3(2.1)	99%

Table 3.3: Results from Pascarella & Argos (1992). Calculating the average indel length and the percentage of indels less than ten residues at 0–20%, 20–40% and 40–80%, indicated the relatively narrow distribution of indel lengths.

The indel analysis was repeated for a larger dataset, here, of 235 superfamilies (45 500 comparisons of non-identical representatives) and also revealed an exponential decay between average indel length and sequence identity (Figure 3.9). Additionally, at 0–10% sequence identity average indel lengths are between six and twelve residues long. This suggests that the more extensive sampling of distant structural homologues in the CATH superfamilies reveals that larger indels are possible in very remote homologues.

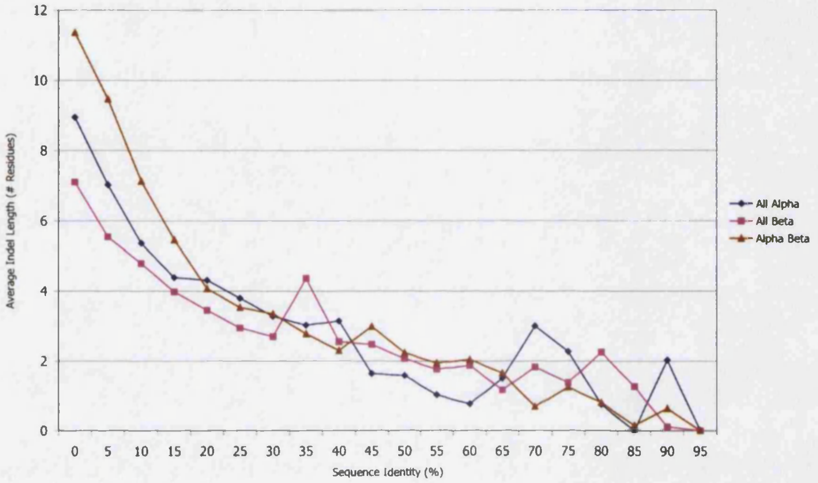


Figure 3.9: The average indel length plotted with sequence identity bins of 5% shows a decrease in indel length as sequence identity increases. At 0–10% sequence identity average indel lengths are between 6 and 12 residues long and at sequence identities of greater than 60% indel lengths are less than 2 residues.

The results for each sequence identity bracket 0–20%, 20–40%, 40–95% are shown in Table 3.4. As for the results obtained by Pascarella & Argos (1992), they show that the majority of indels are less than ten residues and that the average indel length ranges from six in the 0–20% sequence identity range, to two in the 40–95% sequence identity range which is just one residue higher than in the previous study despite the fact that this study was carried out with more than 60 times the number of data in the original dataset. This information could be used to modify gap penalties in sequence alignments.

In Table 3.4 the data are divided into separate classes showing that the $\alpha\beta$ class is more tolerant to indels than the other two classes at lower sequence identities. This can also be seen in Figure 3.9. Table 3.4 also shows a significant decrease (to 70%) in the number of indels below ten residues in the 0–20% sequence identity range.

	α		β		$\alpha\beta$		All	
Sequence Identity (%)	*Average Indel Length*	*% < 10 Residues*	*Average Indel Length*	*% < 10 Residues*	*Average Indel Length*	*% < 10 Residues*	*Average Indel Length*	*% < 10 Residues*
0–20	5.8(5.4)	89%	5.0(2.6)	97%	8.5(5.5)	70%	5.9(4.06)	89%
20–40	4.0(4.6)	95%	3.3(2.3)	99%	3.7(3.0)	96%	3.52(2.91)	98%
40–95	1.7(1.9)	97%	2.0(3.15)	98%	2.3(5.5)	97%	2.07(4.3)	97%

Table 3.4: The average length of indels and percentage of indels less than 10 residues for all of the data in this study and also divided into the α, β and $\alpha\beta$ classes. The sequence identity bins chosen are those used by Pascarella & Argos (1992).

However since some highly populated architectures dominate the different classes (for example $\alpha\beta$ sandwiches and β sandwiches in the $\alpha\beta$ and mainly β classes respectively) the data were also divided into separate architectures to investigate whether any of the architectures are more tolerant to larger indels than others. The results can be seen in Figure 3.10 which shows the average indel length and the percentage of indels with less than 10 residues for all architectures with more than 10 pairwise comparisons in each sequence identity bracket 0–20%, 20–40%, 40–95%. Each sequence identity cluster can be interpreted differently. Pairwise comparisons in the 0–20% sequence identity range show the most variation in average indel length and percentage of indels more than 10 residues. For sequence identities of greater than 20% there is no influence of architecture on the tolerance to indels. Below 20% sequence identity, 3-layer($\beta\beta\alpha$) sandwich (3.50), 3-layer($\alpha\beta\alpha$) sandwich (3.40) and up-down bundle (1.20) architectures display a much higher indel length average in the 0–20% sequence identity range than the other architectures. This may suggest that these architectures are much more tolerant to structural change than the others. In the 20–40% sequence identity range average indel lengths range from aproximately 2 in the 3-layer($\beta\beta\alpha$) sandwiches to approximately 6 in the up-down bundles.

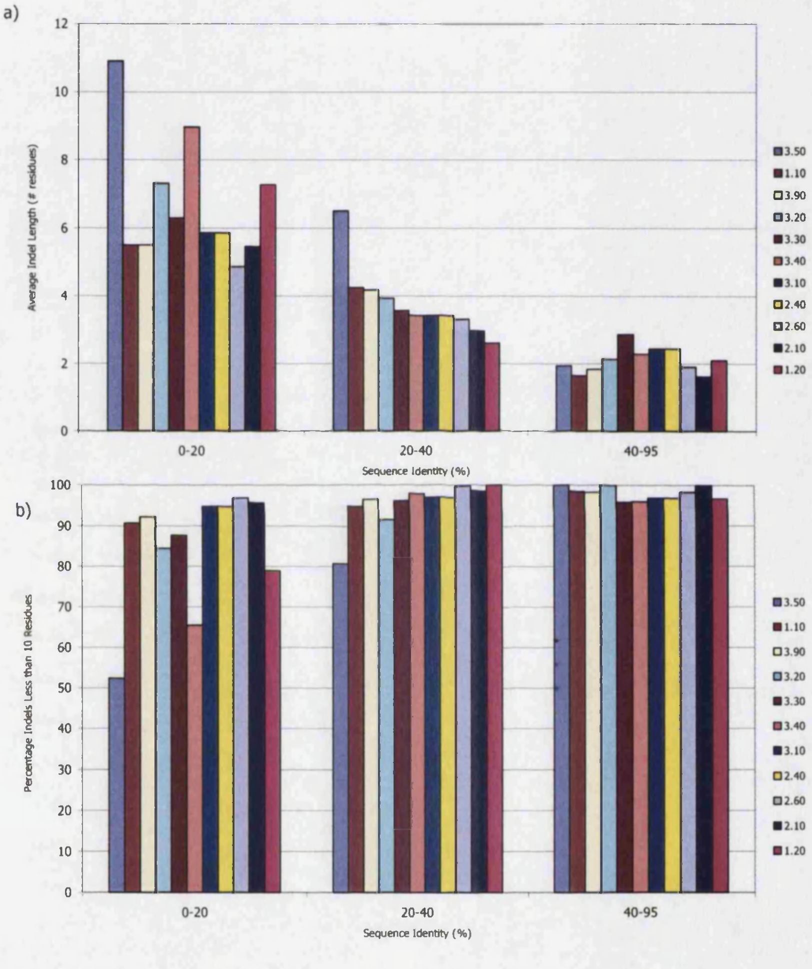


Figure 3.10: Average indel length and percentage of indels less than 10 residues in the eleven most populated architectures in CATH shown in three sequence identity clusters, 0–20%, 20–40% and 40–95%. The 0–20% sequence identity bin identifies the 3-layer($\beta\beta\alpha$) sandwich, 3-layer($\alpha\beta\alpha$) sandwich and up-down bundle architecture as being able to tolerate larger indels. The 20–40% range identifies the 3-layer($\beta\beta\alpha$) sandwich architecture as tolerating larger indels than the others.

3.3.5.2 Secondary Structure Composition of Indels

The IndelCalc program also calculates the type of secondary structure (helix, strand or coil) of each indel residue. Figure 3.11 shows the percentage of each secondary structure type for different sequence identity bins. As sequence identity increases, the number of indel residues decreases. As a result, the histogram bars are based on fewer data at higher sequence identities. For all classes, most indel residues are coil, showing that the majority of insertions and deletions occur in the loops between the secondary structures. However, in all three classes some of the residues adopt secondary structure conformations. In some cases these indels extend the secondary structure elements already present and in other cases whole secondary structure elements are inserted. This is reviewed in section 3.3.6. Figure 3.11 shows that in the $\alpha\beta$ class α-helices are more frequently inserted and deleted and feature more highly than β-strands. This may be because an α-helix is stabilised by hydrogen bonds between residues within the α-helix whereas a β-strand does not exist in isolation but occurs in hydrogen bonded pairs. So, when an insertion occurs, it is easier for the inserted peptide to fold into an α-helix.

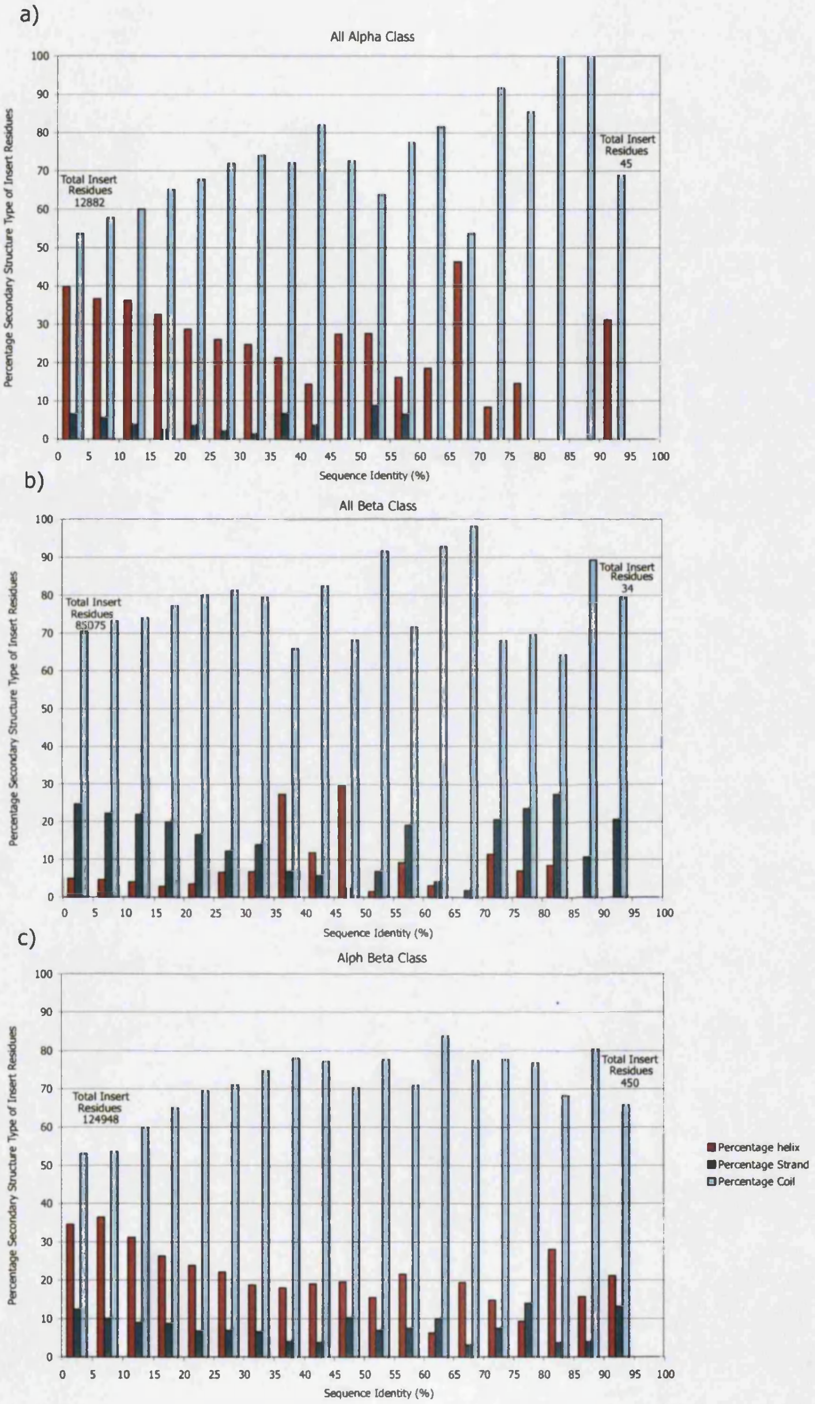


Figure 3.11: The histograms show the types of secondary structures formed by indel residues for different sequence identity bins. The data is divided into the mainly-α class, the mainly-β class and the $\alpha\beta$ class. The majority of indel residues are in coil structures. However, those indel residues that do adopt a secondary structure state are more likely to be α-helical. As the sequence identity increases there is a decrease in the total number of indel residues.

3.3.6 Secondary Structural Embellishments

Figure 3.11 demonstrates that at high sequence identities (>35%) most (>70%) of the indel residues present between homologous structures are coil, although at lower sequence identities (<15%) there is also a significant percentage (~30%) of inserted residues in β-strand and α-helix conformation. It is possible that these secondary structure residues represent the insertions and deletions of whole secondary structure elements.

Extensive secondary structure insertions may be tolerated because they confer additional functional properties or modify an existing active site. They may also alter or facilitate additional protein – protein interactions by altering surface geometry. To explore whether secondary structure embellishments are tolerated as neutral changes having no impact on the function or stability of the protein or whether they are tolerated because of beneficial changes in the functions of the relatives are obtained, a dataset of superfamilies was selected containing relatives which had been significantly embellished by secondary structure insertions. In order to examine secondary structure indels present in some superfamilies a set of highly embellished superfamilies was identified using 2DSEC (see section 2.2.3). A dataset of S35Reps for each superfamily were used in this study. Only superfamilies with more than three S35Reps were included in the dataset which provided a total number of 235 superfamilies. 71% of these superfamilies have only three to five S35Reps.

3.3.6.1 Percentage Variability of Secondary Structures within a Superfamily.

Superfamilies particularly susceptible to domain enlargement by secondary structure embellishment were identified using percentage variability of secondary structures (as described in section 3.2.4.1). Figure 3.12 shows the distribution of variability in the number of secondary structures, for structures ranging from 0% variability (i.e. no additional secondary structures) to 80% variability, where the largest relative has almost double the number of secondary structures than the smallest. Figure 3.12 shows that at present, most of the superfamilies in the study have three to five S35Reps. Therefore, for those superfamilies which appear to be conserved, this may simply be a consequence of not having sampled the superfamily widely enough.

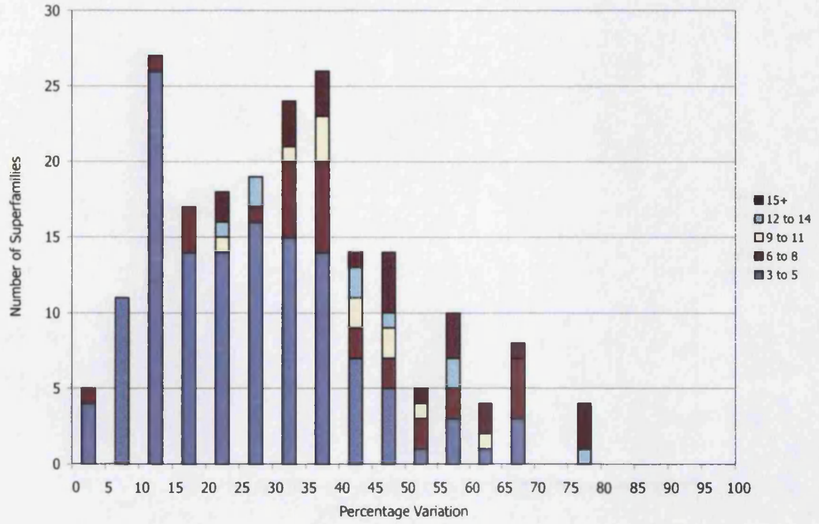


Figure 3.12: Percentage variability in number of secondary structure elements in members of the same superfamily. The histogram shows that some superfamilies show no secondary structure variation between superfamily members whereas others can almost double in size from the smallest superfamily member to the largest. The histogram also shows that most (71%) of the superfamilies have only 3–5 S35Reps. It is not known whether the conserved, less populated superfamilies will become more variable as more relatives are structurally determined.

Percentage variability is dependent on two factors which need to be considered when identifying the most embellished superfamilies. Figure 3.13 shows the relationship between percentage variability and number of superfamily members. The more populated sequence families tend to have a higher percentage variability, perhaps suggesting a more complete evolutionary picture of the superfamily.

Another bias is caused by the size of the protein; that is the number of SSEs it contains. If the largest member of a superfamily has a total of four secondary structures, removing only one secondary structure from the consensus results in a score of 25% variablility. As the domains in the superfamily get larger, more secondary structures must be removed to get the same percentage variability. Therefore, the dataset was reduced to those superfamilies with an average of at least five secondary structures to exclude any which scored a high percentage variability simply by losing a single secondary structure element. Figure 3.14 shows the dependence of the percentage variablility on the number of secondary structures. No superfamilies with 20 secondary structures or more show a percentage variability above 50%.

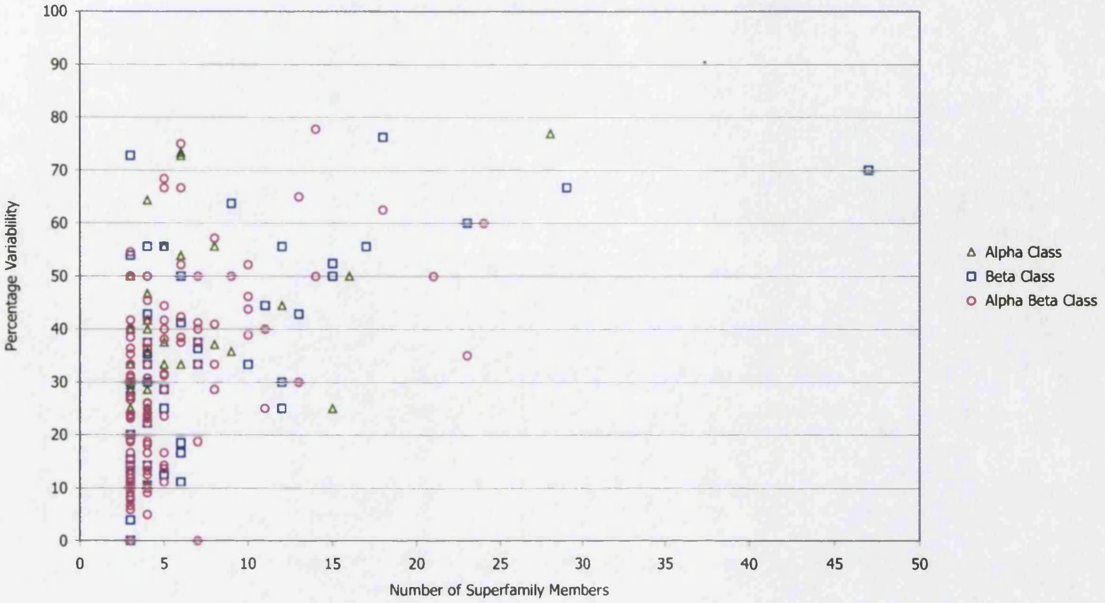


Figure 3.13: The percentage variability in each superfamily increases with the number of relatives. However it can be seen that some highly populated superfamilies are particularly conserved.

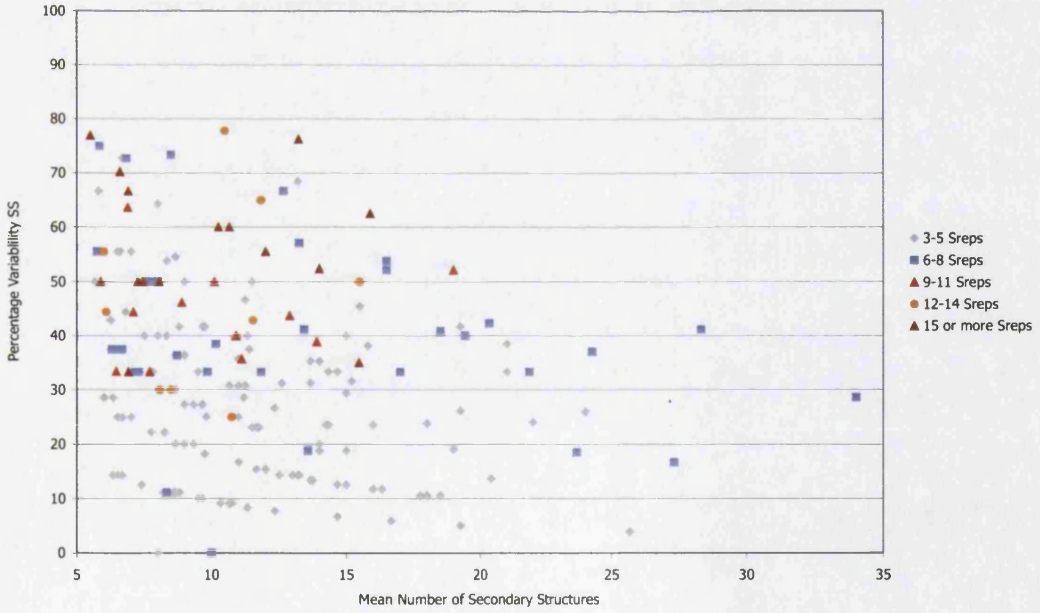


Figure 3.14: The percentage variability in each superfamily versus average numbers of secondary structures.

3.3.6.2 Selecting Variable and Conserved Superfamilies

Taking into consideration the dependence on number of S35Reps and the size of the domains, percentage variability was used to select a subset of particularly conserved and particularly embellished superfamilies in order to analyse and to characterise ways in which these families have evolved. Percentage variation gives an indication of the proportion of the secondary structure elements which are embellishments. But if variable superfamilies were selected on percentage variability alone some of the larger domains with considerable variability would be missed. The most embellished superfamilies were selected on a sliding scale derived empirically from Figure 3.14. The higher the average number of secondary structures the lower the percentage variability cut off (Table 3.5). This sliding scale identified 39 variable superfamilies shown on Figure 3.15. These variable superfamilies are examined further in section 3.3.8. In selecting conserved superfamilies, it is important to consider only those superfamilies which have been sufficiently sampled. The three selected superfamilies all contained more than ten diverse relatives (i.e. <35% identity, >10 S35Reps).

Average Number of SSEs	**% Variability Cutoff**
6 to 9	60
10 to 14	50
15 to 19	40
20+	30

Table 3.5: The sliding scale used to select the most embellished superfamilies. Any superfamilies with an average number of secondary structures between 6 and 9 are only selected if they have more than 60% variability. Superfamilies with more secondary structures are selected at the lower percentage variabilities listed in the table.

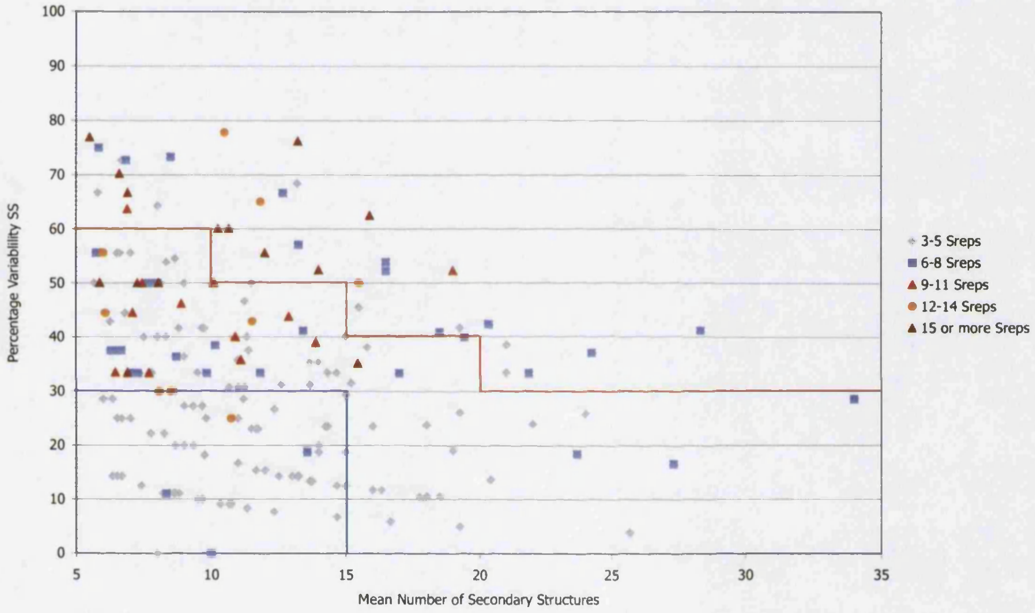


Figure 3.15: Selection of the most embellished and the most conserved superfamilies in CATH. The red line shows the sliding scale above which the more embellished superfamilies are plotted. The blue line shows the selection of the more conserved members of the superfamily. In this region of the graph there are many structures with only 3–5 S35Reps. It is not known whether these superfamilies will become more embellished as new members are added. The most conserved superfamilies selected for analysis are those which are well populated (have more than 12 S35Reps as shown by the yellow circle and brown triangle) and show less than 30% variablility.

3.3.7 Particularly Conserved Superfamilies

A number of superfamilies show considerable conservation in secondary structure (less than 30% variablility) (Figure 3.13) but many of these have only three to five representatives (Figure 3.14). These superfamilies may appear conserved with the data available today, but may become more diverse as more members are added. Therefore three superfamilies with more than ten S35Reps, which implies considerable sequence diversity, and 30% or less variation in secondary structures were selected as particularly conserved superfamilies.

CATH	Superfamily	Architecture	Variation (%)	Number of S35 Families	Average Number of SS	Average SSAP Score (s.d.)
2.30.29.30	Pleckstrin Homology and Phospho -tyrosine Binding	Roll	30	12	9	84.54(3.56)
3.10.100.10	C-Type Lectin-Like	Roll	30	13	8	84.8(5.21)
3.30.200.20	Kinase	2-Layer Sandwich	25	11	6	86.06(4.80)

Table 3.6: Well populated superfamilies showing high conservation in secondary structures.

All three of these superfamilies show significant sequence diversity, clustering into 11 or more 35% sequence families, but significant structural similarity, with an average SSAP score no lower than 84.

3.3.7.1 Structural Conservation Due to Functional Constraints?

Why do these three families show structural conservation at low sequence identity? Is the secondary structure conservation in these families important for the function of the domain? This subsection addresses these two important questions.

Pleckstrin Homology and Phosphotyrosine Binding Domains

This structural superfamily comprises two distinct functional families, a set of pleckstrin homology (PH) and phosphotyrosine binding (PTB) domains. The structure consists of two nearly orthogonal anti-parallel β-sheets capped with an amphipathic α-helix which interacts with the hydrophobic core of the β-sandwich (Figure 3.16). The phosphotyrosine binding domain is slightly larger, containing an extra N terminal α-helix and a long insert containing an α-helix and a long loop. Although these functional families are similar in structure, they contain very little sequence similarity. PH domains are found in a large number of proteins, cellular signalling, cytoskeletal organisation, regulation of intracellular membrane transport and modification of membrane phospholipids. PH domains interact directly with the cell membrane by binding phosphoinositides with a range of binding specificities. PTB domains are found in cellular signalling proteins such as SHC. SHC contains three domains of which the PTB domain is responsible for binding to a phosphorylated receptor that activates the protein allowing it to bind the next enzyme in the signalling cascade. A similarity in these two functional families is that they both associate with membrane phospholipids and are both located in the cytoplasm.

The domains in this superfamily are small stable structures associated with many cellular functions but exhibiting a common, useful function as switch proteins in specific protein – protein interactions in signalling cascades. In many cases, especially in cellular signalling, this domain is part of a much larger assembly of interacting proteins suggesting that structural conservation may be conserved to preserve these specific protein – protein interactions. Additionally, Figure 3.16 shows that the architecture of the structure is more irregular than a sandwich architecture, and unlike a sandwich architecture it is difficult to suggest where extra secondary structure embellishments might be located and still allow optimal packing of hydrophobic residues. It may be the case that addition of extra secondary structure elements would disrupt the fold.

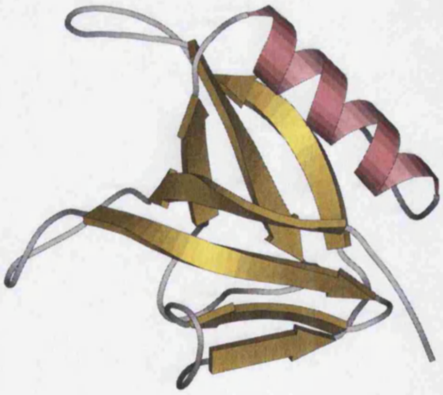


Figure 3.16: The pleckstrin homology domain from dynamin (2dynA0).

C-Type Lectin-Like Domains

The C-type lectin-like domain (CTLD) superfamily contains carbohydrate recognition domains (CRDs) from the lectins but have been recruited for other functions such as NK cell receptors (MHC ligands), phospholipase receptors, type II antifreeze proteins and coagulation factor binding proteins (Drickamer, 1999). The structure can be divided into two parts: one region contains the elements of regular secondary structure and the other region consists of an extended loop region and is the carbohydrate recognition site in the CRDs (Poget *et al.*, 1999) (Figure 3.17). A member of this superfamily, the sea raven antifreeze protein comprises an ice-binding site of residues which correspond to the calcium binding site of the lectins (Gronwald *et al.*, 1998). The CRDs bind ligands in distinct ways, mediated by the variable loop regions of the structure. It could be that the loops which mediate the function are located on a stable structural framework produced by the core of the domain structure. However, like the pleckstrin homology and phosphotyrosine binding superfamily, the roll architecture is more irregular than some architectures like the β-sandwiches. It is therefore also difficult to see where additional secondary structures could be located in order to maintain optimal residue packing.

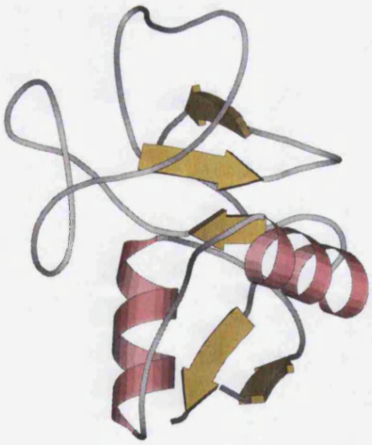


Figure 3.17: C-Type mannose-binding protein (1msbA0)

N-Terminal Protein Kinase Domains

Protein kinases are involved in every aspect of signal transduction in eukaryotic cells, from primary transmembrane signalling to control of transcription and cellular metabolism. Specificity is regulated by unique phosphorylation events and binding interactions. Protein kinases have two domains. The N-terminal domain, the conserved domain, consists of a single β-sheet and one α-helix whilst the C-terminal domain is composed almost entirely of α-helices. The ATP binding site is situated between the two domains and regulation of these protein kinases occurs through phosphorylation of a loop close to the active site (Figure 3.18). As with the PH domain superfamily, this is a small domain, involved in specific protein – protein interactions in which a large region of the surface is mediating the interactions. Therefore, to maintain these specific interactions there may be restrictions on insertions or embellishments which might change the geometry of the surface.

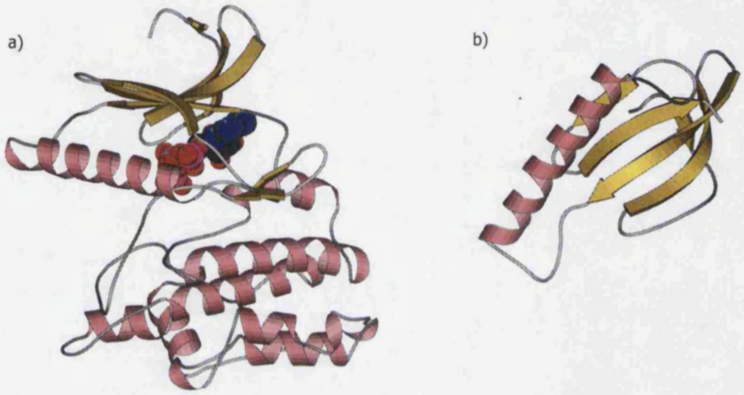


Figure 3.18: Rabbit muscle phosphorylase kinase (1phk). The whole protein is shown in figure (a) and the structurally conserved domain is shown in (b).

3.3.7.2 Conclusions About Conserved Superfamilies

All three superfamilies contain small domains. One of the reasons for their structural conservation could be that a large proportion of their structure is involved in the function as the loops and protein surfaces are involved in ligand binding and also in protein – protein interactions. Both the kinases and the PH, PTB domains are involved in cellular signalling and are involved in similar functions, phosphorylating and dephosphorylating signalling proteins in enzyme cascades. Protein interactions between cellular signalling proteins are complex and it may not be advantageous for these domains to grow in size or change in structure. Another reason for their conservation may be the irregularity of the architecture. Insertions into the structure may disrupt the packing and therefore would be less tolerated.

Three superfamilies have been identified as conserved in terms of secondary structure content. Considering also the pairwise structure similarity (SSAP) (Figure 3.5), all superfamilies fall in the conserved area of the graph (with >= 85 average SSAP score and <6 associated standard deviation).

3.3.8 Particularly Variable Superfamilies

Superfamilies with three or more S35Reps with a great deal of secondary structure embellishment were also identified using the percentage variability measurement. Once identified, the position of these embellishments on the peptide chain was examined using 2DSEC and their positions on the three-dimensional structure were characterised using the three-dimensional protein structure viewer, Rasmol (Sayle & Milner-White, 1995).

CATH	**Superfamily**	**Architecture**	**S35Reps**	**Percentage Variability**	**SSAP Score Average (s.d.)**	**SSGs Singletons**	**Average SSEs (s.d.)**	**Largest Total SSEs**	**Smallest Total SSEs**	**Largest Number Embellished SSEs**	**Largest Single SS Insertion**
2.60.40.10	Immunoglobulin-like	Immunoglobulins	48	70	81.01(4.72)	7 (9)	7(1.82)	10	3	H0 E4	H0 E3 1bec02
1.10.10.10	DNA binding domain	Orthogonal Bundle	28	76.92	78.94(5.99)	8 (12)	6 (2.08)	13	3	H4 E7	H1 E4 1repC1
3.40.30.10	Thioredoxin-like	3-Layer($\alpha\beta\alpha$) Sandwich	13	41.67	78.91 (4.80)	5 (11)	9 (1.71)	12	7	H3 E3	H1 E2 1qq2A0
1.25.40.10	Serine threonine protein phosphatase	Horseshoe	6	73.33	86.81(4.59)	2 (2)	9(3.89)	15	4	H9 E0	H6 E0 1qqeA0
2.40.50.100	Oxidoreductases	Barrel	3	72.73	82.45(7.11)	1 (3)	7(4.04)	11	7	H4 E5	H1 E7 1htp00
1.20.85.10	Membrane-spanning α helix pairs	Up-down bundle	6	72.73	75.79(9.12)	3 (3)	7(3.31)	11	3	H7 E2	H3 E0 1jgiA0
3.90.10.10	Cytochrome C3	Complex	5	66.67	81.07(3.71)	1 (2)	6(2.68)	9	6	H6 E2	H6 E2 19hcA1
2.60.40.30	Fibronectin type III	Sandwich	29	66.67	83.23(4.41)	4 (6)	7(1.05)	9	7	H1 E2	H1 E2 1hft02
1.10.275.10	Fumarase/aspartase	Orthogonal Bundle	4	64.29	73.55(7.85)	1 (3)	8(4.08)	14	5	H4 E6	H2 E2 1b8fA1
2.10.90.10	Cystein-knot cytokines	Ribbon	9	63.64	78.00(5.58)	2 (4)	7(1.90)	11	4	H3 E5	H0 E3 1bet00
3.30.420.10	Nucleotidyl transferase	2 layer sandwich	14	77.78	72.23(7.24)	3 (8)	11(3.71)	18	4	H6 E3	H2 E2 1noyA
2.40.70.10	Acid proteases	Barrel	18	76.19	78.59(8.28)	2 (2)	13(5.21)	21	5	H4 E14	H0 E5 1qdmA2
3.30.360.10	Dihydrodipicolinate Reductase, domain 2	2-Layer Sandwich	5	68.42	81.03(3.62)	2 (3)	13(4.97)	19	6	H6 E4	H2 E2 1dpgA2
3.80.20.10	Nuclear Protein/RNA binding	Horseshoe	6	66.67	82.38(4.44)	1 (3)	3(5.85)	21	7	H10 E5	H4 E3 1fvqA0
3.30.470.20	ATP dependent carboxylate-amine/thiol ligase	2-Layer Sandwich	7	60.00	77.46(6.47)	2 (8)	12(5.01)	20	8	H9 E6	H6 E5 1bncA2
2.60.40.420	Cupredoxin	Sandwich	19	57.00	79.76(5.35)	4 (5)	10(2.47)	16	7	H2 E5	H0 E4 1aozA2
3.40.190.10	Periplasmic binding protein-like II	3-Layer($\alpha\beta\alpha$) Sandwich	24	60	73.60(10.03)	4 (9)	11(2.3)	15	6	H7 E3	H4 E2 1anf01

continued on next page

CATH	Superfamily	Architecture	S35Reps	Percentage Variability	SSAP Score Average (s.d.)	SSGs Singletons	Average SSEs (s.d.)	Largest Total SSEs	Smallest Total SSEs	Largest Number Embellished SSEs	Largest Single SS Insertion
2.60.120.20	Virus Coat Protein	Sandwich	17	55.56	72.15(8.26)	4 (13)	12(2.98)	18	8	H4 E7	H1 E4 2bbvA0
2.60.120.60	Lectin	Sandwich	15	52.38	76.85(7.58)	3 (8)	14(3.25)	21	10	H1 E10	H1 E5 1a8d01
3.40.630.30	N-Acetyltransferase	3-Layer($\alpha\beta\alpha$) Sandwich	9	50	81.77(3.68)	2 (2)	10(2.32)	14	7	H3 E3	H2 E2 1nmtA2
3.40.47.10	Peroxisomal Thiolase, subunit A, domain 1	3-Layer($\alpha\beta\alpha$) Sandwich	4	50	75.57(9.78)	1 (10)	12 (3.42)	16	8	H6 E3	H4 E2 1bq6A1
3.60.20.10	Glutamine Phosphoribosylpyrophosphate, subunit 1, domain 1	4-Layer Sandwich	18	62.5	80.09(8.67)	2 (2)	16(3.20)	24	9	H8 E6	H4 E2 1ecbA1
1.10.620.20	Di-iron carboxylate proteins	Orthogonal Bundle	6	53.85	76.27(8.35)	1 (4)	17(4.97)	26	12	H13 E2	H6 E0 1mtyD0
3.40.50.1240	Phosphoglycerate mutase	3-Layer($\alpha\beta\alpha$) Sandwich	6	52.17	77.71(3.43)	3 (3)	17(4.28)	23	11	H11 E3	H8 E2 1dkqA0
3.40.690.10	Aspartyl tRNA Synthetase, subunit A, domain 2	3-Layer($\alpha\beta\alpha$) Sandwich	10	52.17	76.81(6.75)	3 (5)	19(3.89)	23	11	H10 E4	H2 E4 1atiA0
3.40.50.970	DHS-like NAD/FAD-binding domain	3-Layer($\alpha\beta\alpha$) Sandwich	14	50	79.86(4.84)	3 (5)	16(3.11)	24	12	H10 E3	H5 E2 1b0pA6
3.40.50.610	Adenine nucleotide alpha hydrolases	3-Layer($\alpha\beta\alpha$) Sandwich	4	45.45	77.69(4.33)	1 (3)	16(4.51)	22	12	H11 E1	H6 E0 1ct9A2
3.20.20.30	FMN dependent fluorescent proteins	Barrel	4	41.67	83.28(4.05)	1 (2)	19(4.11)	24	14	H5 E2	H4 E1 1ezwA0
3.40.710.10	DD-peptidase/β-lactamase	3-Layer($\alpha\beta\alpha$) Sandwich	8	40.91	78.79(3.51)	3 (4)	19(3.07)	22	13	H2 E6	H2 E4 2bltA0
3.40.50.300	P-loop containing nucleotide triphosphate hydrolases	3-Layer($\alpha\beta\alpha$) Sandwich	52	40	69.66(7.20)	13 (28)	15(3.08)	20	12	H10 E5	H6 E1 1gajA0
3.40.50.950	$\alpha\beta$ hydrolase	3-Layer($\alpha\beta\alpha$) Sandwich	33	40	74.39(7.71)	6 (14)	19(3.50)	25	15	H14 E3	H7 E0 1hlgA0

continued on next page

CATH	Superfamily	Architecture	S35Reps	Percentage Variability	SSAP Score Average (s.d.)	SSGs Singletons	Average SSEs (s.d.)	Largest Total SSEs	Smallest Total SSEs	Largest Number Embellished SSEs	Largest Single SS Insertion
3.40.510.10	Class I aminoacyl-tRNA synthetases (RS), catalytic domain	3-Layer($\alpha\beta\alpha$) Sandwich	6	42.31	74.51(8.93)	2 (4)	20(4.32)	26	15	H9 E8	H2 E4 1ile03
2.130.10.20	Trp-Asp repeat (WD-repeat)	7 Propellor	6	41.18	78.42(6.71)	2 (3)	28(5.01)	26	16	H5 E10	H3 E5 1gotB0
3.20.20.90	FMN-dependent oxidoreductase	Barrel	24	38.46	78.21(3.57)	7 (9)	21(4.28)	26	16	H7 E4	H3 E3 1gox00
1.10.630.10	Cytochrome P450	Orthogonal	8	37.04	82.10(2.88)	1 (3)	24(3.45)	27	17	H5 E2	H1 E1 1cpt00
3.20.20.150	Divalent-metal-dependent TIM barrel enzymes	Barrel	3	33.33	81.60(3.13)	1 (2)	21(4.36)	24	16	H7 E2	H5 E0 1xib00
3.20.20.80	TIM barrel glycosyl hydrolase	Barrel	33	33.33	71.71(5.58)	8 (18)	22(3.72)	27	18	H11 E2	H4 E0 1eswA0

Table 3.7: The most structurally embellished superfamilies in CATH are listed. The table lists the CATH numbers, homologous superfamily names, architectures, number of S35Reps, percentage variation, average SSAP score and associated standard deviation, number of Structural Sub-Groups (SSGs) and singletons (not clustered into an SSG), average number of secondary structures and associated standard deviation, number of secondary structures in the largest S35Rep and also in the smallest and finally, the largest single insertion in the superfamily and the CATH code of the member which contains that insertion.

3.3.8.1 Characterisation of Embellishments in the Variable Superfamilies

β-strand embellishments frequently occur as additions or extensions to existing β-sheets or form β-hairpins, whereas helices can exist as an entity on their own, as all hydrogen bonds are satisfied. Furthermore, α-helices are not as constrained in their orientations and pack with a greater variation in angles (Chothia *et al.*, 1977). Because of this, it is often harder to align these secondary structure elements in remote homologues and therefore more difficult to assess equivalent α-helices, often making embellishments in helical families harder to characterise.

Two architectures were found to be featured more frequently than others in the table of most embellished superfamilies; mainly β 2-layer sandwiches and 3-layer ($\alpha\beta\alpha$) sandwiches. These architectures are extremely highly populated in the database. This may be because they contain proteins which have been studied more thoroughly. However, it may also suggest that these structures are more tolerant to structural embellishments and therefore, nature may have used these regular architectures to recruit novel functions more frequently, which explains their popularity. These two architectures together with the 2-layer $\alpha\beta$ sandwiches were examined separately for similarities and differences in the types and positions of the secondary structure embellishments. The positions of the insertions in every superfamily in each architecture are illustrated by examining a single superfamily in more detail.

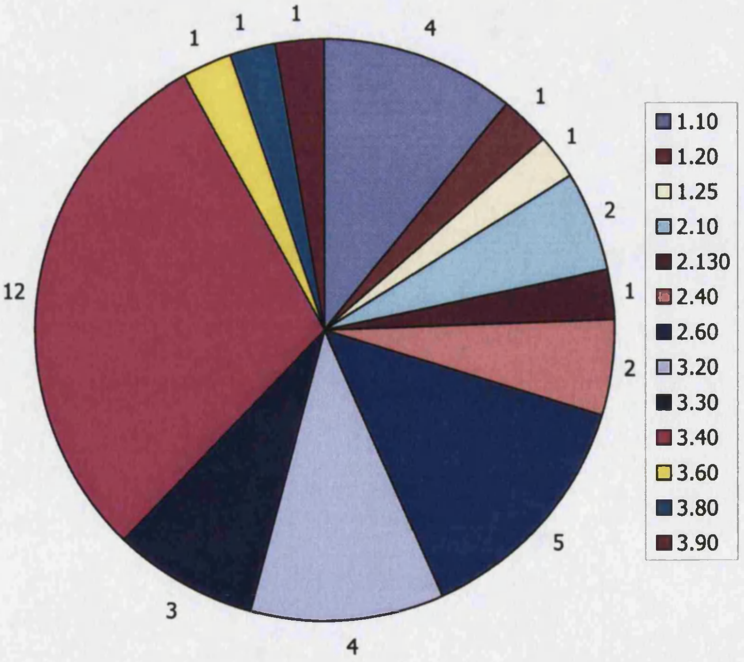


Figure 3.19: The number of embellished superfamilies for each architecture.

Mainly β 2-Layer Sandwich Architecture

CATH	Maximum Insert	Num Strands and Helices in Maximum Insert	Total SS	Total Extra Strands and Helices	Description of Embellishments
2.60.40.10	3 1bec02	3E 0H	10	4E 0H	Insertions in 2 places. Largest insertion localised at the edges of both β-sheets.
2.60.40.30	3 1hft02	2E 1H	9	2E 1H	Insertion is in a single region. Localised at the edge of one β-sheet.
2.60.40.420	4 1aozA2	4E 0H	13	5E 2H	Insertions in 4 areas, one at beginning, two in the middle and one at the end of the chain. Most insertions are localised on the edges of the β-sheet.
2.60.120.60	6 1a8d01	5E 1H	19	6E 3H	Insertions in 4 areas, one at beginning, two in the middle and one at the end of the chain. Insertion at the beginning of 1a8d01 extends the edges of both β-sheets. Middle insertions embellish the other end of the β-sheet.
2.60.120.20	4 2bbvA0	4E 1H	18	7E 4H	Insertions throughout the structure. They extend both ends of each β-sheet in the sandwich.

Table 3.8: Description of secondary structure embellishments in the mainly β 2-layer sandwiches. The table shows the highly embellished superfamilies identified by 2DSEC from this architecture. It records the number of α-helices and β-strands in the largest continuous insertion, the total number of secondary structures in that particular representative, the number of embellished secondary structures in total in that representative. Finally, a brief description of the insertions throughout the peptide chain and how they are orientated in the three-dimensional structure.

Almost all of the insertions found in these particularly embellished β-sandwiches contribute to the ends of the two β-sheets (Table 3.10). Often insertions occur in a number of regions within the peptide chain but congregate at the edges of the β-sheets.

Examples of these embellishments are shown by the galectin-type CRD domain superfamily, a family of lectins (CATH 2.60.120.60). The superfamily members share sequence similarities in the carbohydrate recognition residues. The crystallisation of galectin-7 complexed with galactose (Leonidas *et al.*, 1998) revealed the carbohydrate binding site to be in the loops above and below the sandwich, binding the carbohydrate ligand with great specificity. Binding interactions are provided by the loops on both ends of the sheets. The examples in Figure 3.21 show the types of embellishments in this superfamily. 1bkzA0 is the least embellished member of this superfamily showing two anti-parallel β-sheets, each with five β-strands. Embellishments of two types can be seen. 1a8d01 shows the more typical type of embellishment with additions to either side of the β-sandwich. The first set of inserted secondary structures (Figure 3.20) occurs to one side of the β-sandwich and the single inserted β-strand in the centre of the peptide contributes to the other side of the β-sandwich. All relatives have active sites in the same location throughout the superfamily. There are significant changes in the binding pocket shaped by the extensive β-strand embellishments occurring at the edges of the β-sheets (Figure 3.21). K-carrageenans (sulfated $\alpha\beta$-glactands, PDB code 1dyp) has a tunnel-shaped active site thought to be responsible for the degradation of polysaccharides (Michel *et al.*, 2001).

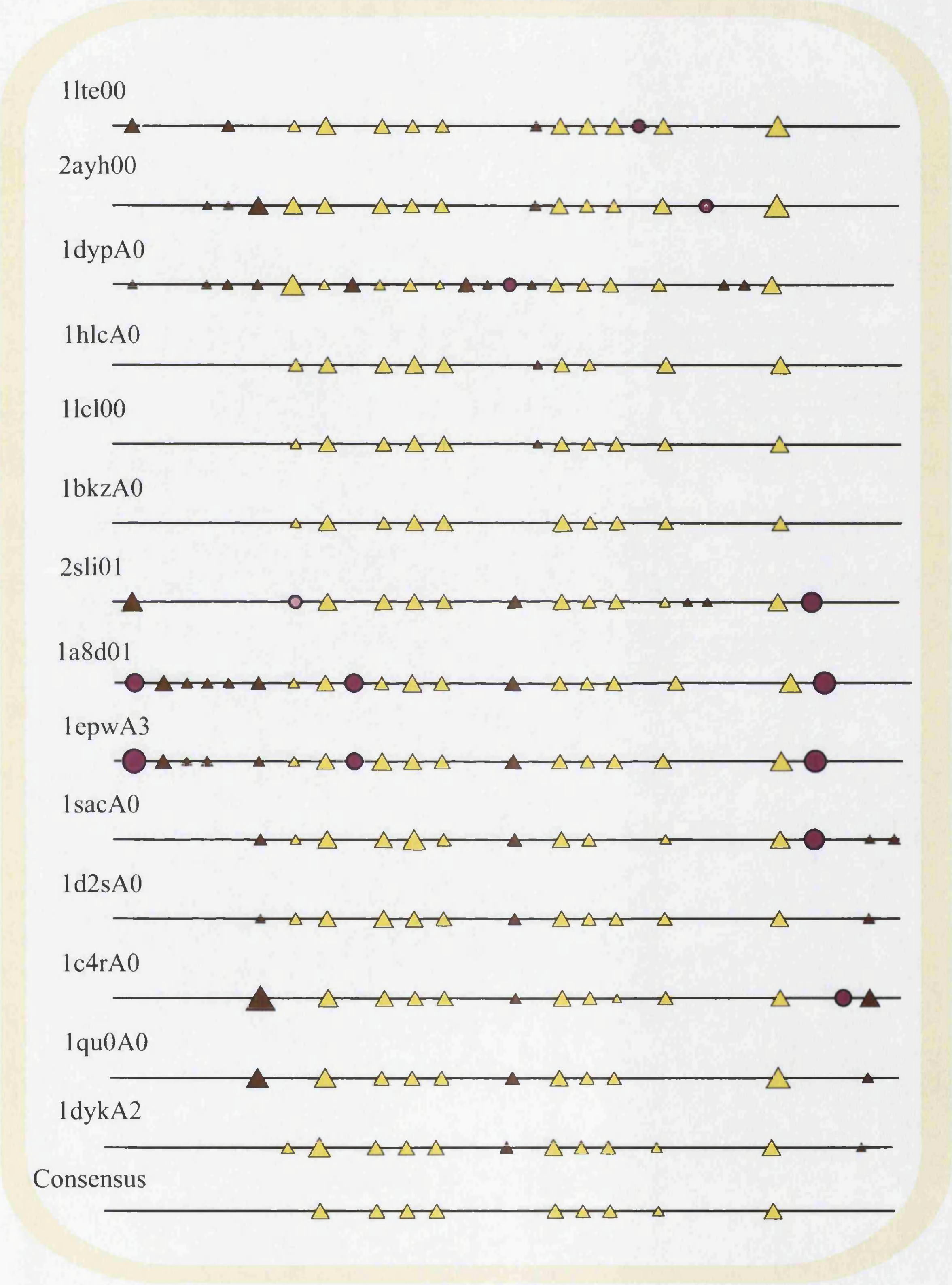


Figure 3.20: 2DSEC diagram showing four areas of embellishment in the CATH superfamily 2.60.120.60. The N-terminal embellishment occurs on the left side of the β-sandwich as it is orientated in Figure 3.8(1a8d01). The other embellishments are located on the top, bottom and right side of the β-sandwich.

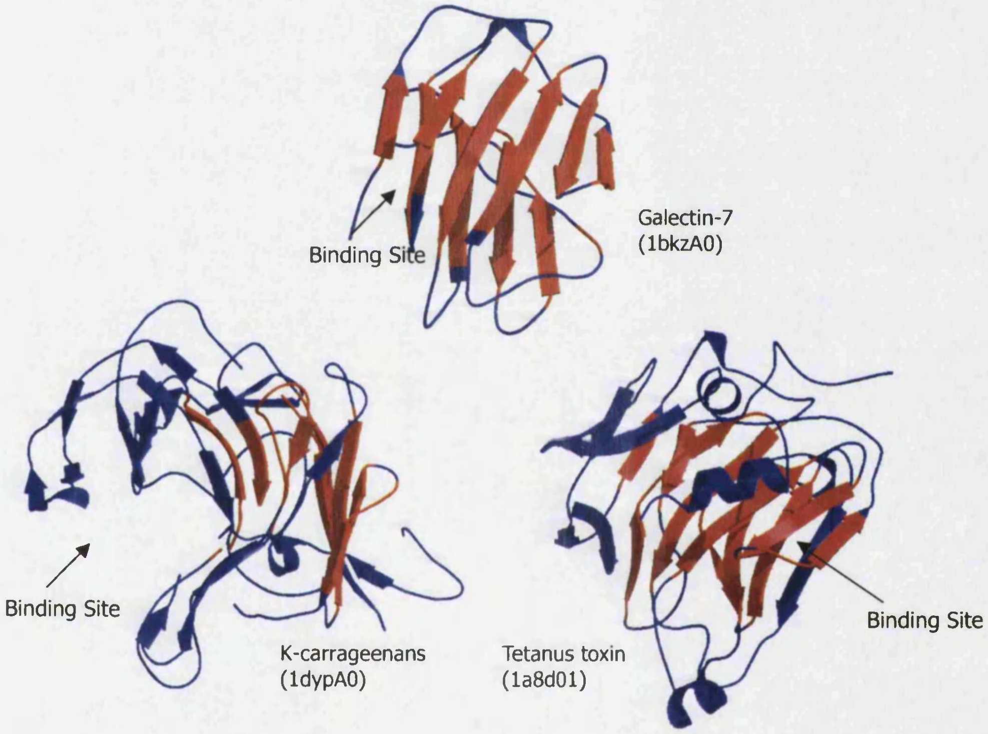


Figure 3.21: Three domains from the galectin-type carbohydrate recognition domain superfamily. The domains are coloured according to their structural conservation. Residues in the same secondary structure throughout the superfamily are coloured in red and residues without secondary structure or with an additional secondary structure not present in all of the other members are coloured in blue. Domains 1bkzA0 and 1a8d01 are in the same orientation so that the embellishments can be seen. 1dypA0 in a different orientation shows how these embellishments can modify the geometry of the binding site. The binding site remains in the same place in all members of this superfamily.

$\alpha\beta$ 2-Layer Sandwich

In all three superfamilies (Table 3.10), the β-strand insertions extend the central β-sheet and the inserted α-helices pack onto the existing α-helices on one side of the β-sheet.

CATH	**Max Insert**	**Num Strands and Helices in Maximum Insert**	**Total SS**	**Total Extra Strands and Helices**	**Description of Embellishments**
3.30.360.10	4 1dpgA2	2E 2H	19	4E 4H	Insertions in 4 places in the chain. Additional strands occur on either side of the β-sheet. Additional helices congregate on one side of the β-sheet.
3.30.470.20	11 1bncA2	5E 6H	20	7E 10H	Insertions on beginning and end (1bncA2 beginning, 1bxrA2 end) of the chain embellish the edges of the β-sheet. Additional helices.
3.30.420.10	4 1noyA2	2E 2H	14	3E 1H	Insertions in 3 areas, two in the middle and one at the end of the chain. Middle insertion extends β-sheet.

Table 3.9: Description of secondary structure embellishments in the $\alpha\beta$ 2-Layer Sandwiches. The Table shows the highly embellished superfamilies identified by 2DSEC from this architecture. It records the number of α-helices and β-strands in the largest continuous insertion, the total number of secondary structures in that particular representative, the number of embellished secondary structures in total in that representative. Finally, a brief description of the total number of insertions throughout the peptide chain and how they orientate themselves in the three-dimensional structure.

Examples of embellishments in this architecture are shown by the oligomerisation domain of the NADP dependent oxidoreductase proteins which use NADP as a cofactor (3.30.360.10). Each member in this superfamily contains two domains, an NADP binding domain and the embellished domain which is involved in the tetramerisation of the biological unit. The oligomerisation domain contains a mixed β-sheet with α-helices on one side. In all members of this superfamily, interactions between the NADP binding domain and the oligomerisation domain are mediated by the α-helices. The β-sheets of the oligomerisation domain then typically interact in two ways to form the tetramer. The oligomerisation domains in two monomers interact edge to edge across the β-sheet, forming an extended β-sheet. Each pair of monomers then interact to form the tetramer by forming an open faced interaction across the β-sheets (Figure 3.22). In each member of this superfamily, the size and shape of the oligomerisation domain is important for the contacts and the orientation of the NADP binding domain (Rowland *et al.*, 1994; Scapin *et al.*, 1995).

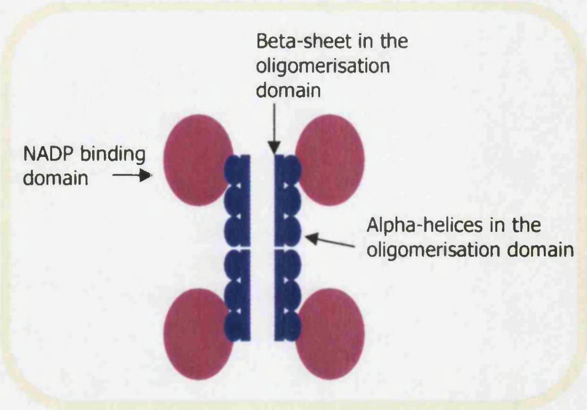

Beta-sheet in the oligomerisation domain
NADP binding domain
Alpha-helices in the oligomerisation domain

Figure 3.22: Typical domain orientation in members of the NADP oxidoreductase family is a tetramer. This is promoted by the oligomerisation domain shown in blue in which the β-sheets interact edge to edge and across the face.

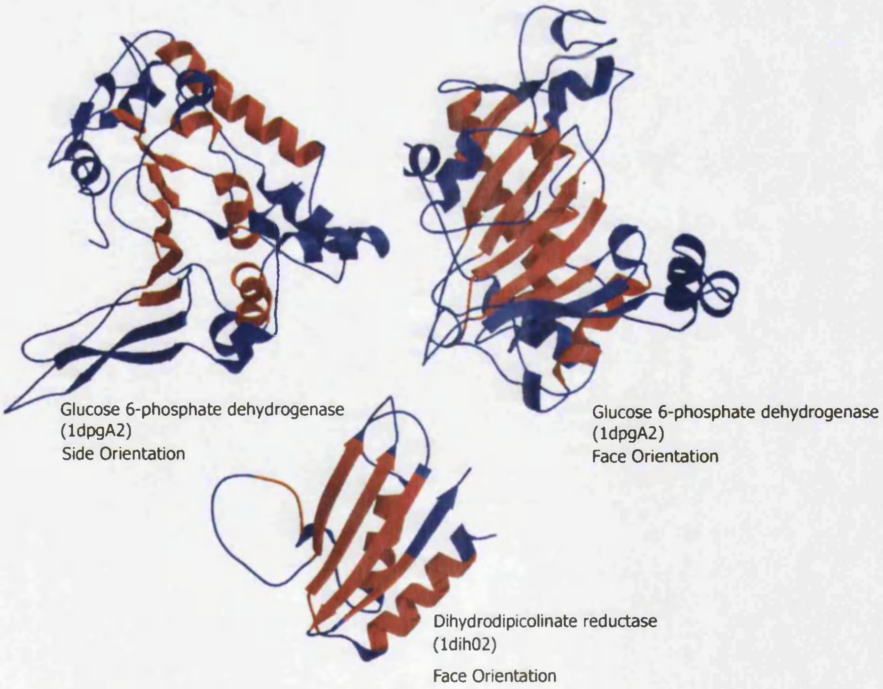

Glucose 6-phosphate dehydrogenase (1dpgA2) Side Orientation
Glucose 6-phosphate dehydrogenase (1dpgA2) Face Orientation
Dihydrodipicolinate reductase (1dih02) Face Orientation

Figure 3.23: Two domains from the oligomerisation domain in the NADP oxidoreductase superfamily. The helices are used to interact with the NADP binding domain and the strands form interfaces in the tetramer. The domains are coloured to show embellished secondary structures and loops in blue and consensus secondary structures in red. Glucose 6-phosphate dehydrogenase (1dpgA2) is shown in two orientations and non-embellished dihydrodipicolinate reductase (1dih01) is shown β-sheet facing.

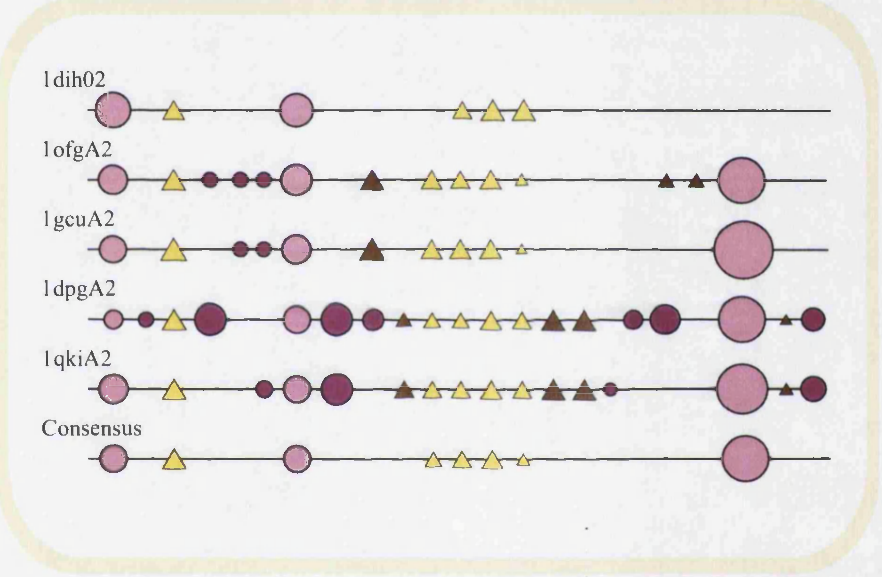


Figure 3.24: 2DSEC diagram shows the embellishments present in the oligomerisation domain in the NADP oxidoreductase superfamily. There are four places with embellishments. The β-strands extend both sides of the β-sheet and extra helices pack mainly against the consensus helices at the back of the β-sheet leaving the face free for tetramerisation.

3.3.8.2 Alpha Beta 3-Layer ($\alpha\beta\alpha$) Sandwich

In this architecture extra β-strands often contribute to the central β-sheet. As for the other embellished superfamilies, insertions can occur in a number of places in the peptide chain but are often co-located in the three-dimensional structure. In addition, a number of superfamilies in this architecture have embellishments which form an extra lobe on the structure which could be considered as a separate α-helical domain. However, these structures are all classified in CATH and SCOP and the literature as a single domain structure.

CATH	Max Insert	Num Strands and Helices in Maximum Insert	Total SS	Total Extra Strands and Helices	Description of Embellishments
3.40.190.10	7 1anf01	2S 5H	14	3S 6H	Insertions mostly at the C-terminal end of the structure do not contribute to central sheet. The embellishment forms an extra lobe.
3.40.630.30	4 1nmtA2	2S 2H	14	3S 3H	Insertions are at the C- and N-termini but they embellish the same side of the sheet.
3.40.47.10	6 1by6A2	2S 4H	16	2S 6H	Insertions form an extra lobe of helices.
3.40.50.1240	10 1dkqA0	2S 8H	23	3S 11H	Insertions form an extra lobe of helices.
3.40.690.10	7 1atiA1	5S 2H	23	9S 6H	Insertions are in 3 main areas, one at the N-terminal and two in the middle. One side of the sheet is embellished by the two central insertions.
3.40.50.970	7 1bopA6	2S 5H	24	3S 10H	Insertions are at the C- and N-termini. Most insertions form a separate lobe. The β-strands in largest insertion in 1bopA6 contribute to the central sheet (only case).
3.40.50.610	6 1ct9A2	0S 6H	22	4S 5H	Insertions are throughout the structure. The β-strands contribute to the central parallel β-sheet the α-helices pack against the consensus α-helices.
3.40.710.10	6 1bltA0	3S 2H	22	4S 5H	Insertions are in 5 main areas. The largest insertion (1bltA0) adds strands to the main sheet. Other insertions contribute to one side of the β-sheet.
3.40.50.300	5 1gajA0	0S 5H	20	5S 10H	Extra strands form a new β-sheet. Extra α-helices added to opposite side of core β-sheet.
3.40.50.950	7 1hlgA0	0S 7H	25	3S 14H	Insertions are throughout the structure. Inserted strands are added to one side of the β sandwich. The inserted α-helices form an extra lobe.
3.40.510.10	6 1ile03	4S 2H	26	8S 8H	Insertions are throughout. Insertions are co-located in 3D forming a large lobe.
3.40.30.10	3 1qq2A0	1E 2H	12	3S 3H	Insertions in two main places. Embellish one side of the β-sheet and the loops.

Table 3.10: Description of secondary structure embellishments in the 3-Layer ($\alpha\beta\alpha$) sandwiches. The table shows the highly embellished superfamilies identified by 2DSEC from this architecture. It records the number of α-helices and β-strands in the largest continuous insertion, the total number of secondary structures in that particular representative, the number of embellished secondary structures in total in that representative. Finally, a brief description of the insertions throughout the peptide chain and how they are orientated in the three-dimensional structure.

An example of embellishments in this $\alpha\beta$ 3-layer architecture is given by the $\alpha\beta$ hydrolases superfamily (3.40.50.950). The structure is described as an 8 stranded mostly parallel $\alpha\beta$ structure. The sheet is twisted so that it forms a half barrel and the active site is situated on the loop of the β-strands at the top of the β-sheet (Figure 3.25). It has been shown that the embellished secondary structures form a lobe in this superfamily which forms a lid to the active site in some members with hydrophobic substrates (Ollis *et al.*, 1992). Therefore, as with some of the other superfamilies analysed in this section, some of the secondary structure embellishments are modulating or facilitating the functions of the proteins. Whilst other secondary structure embellishments which are remote from the active sites may be tolerated as they do not significantly alter the overall shape or architecture of the protein.

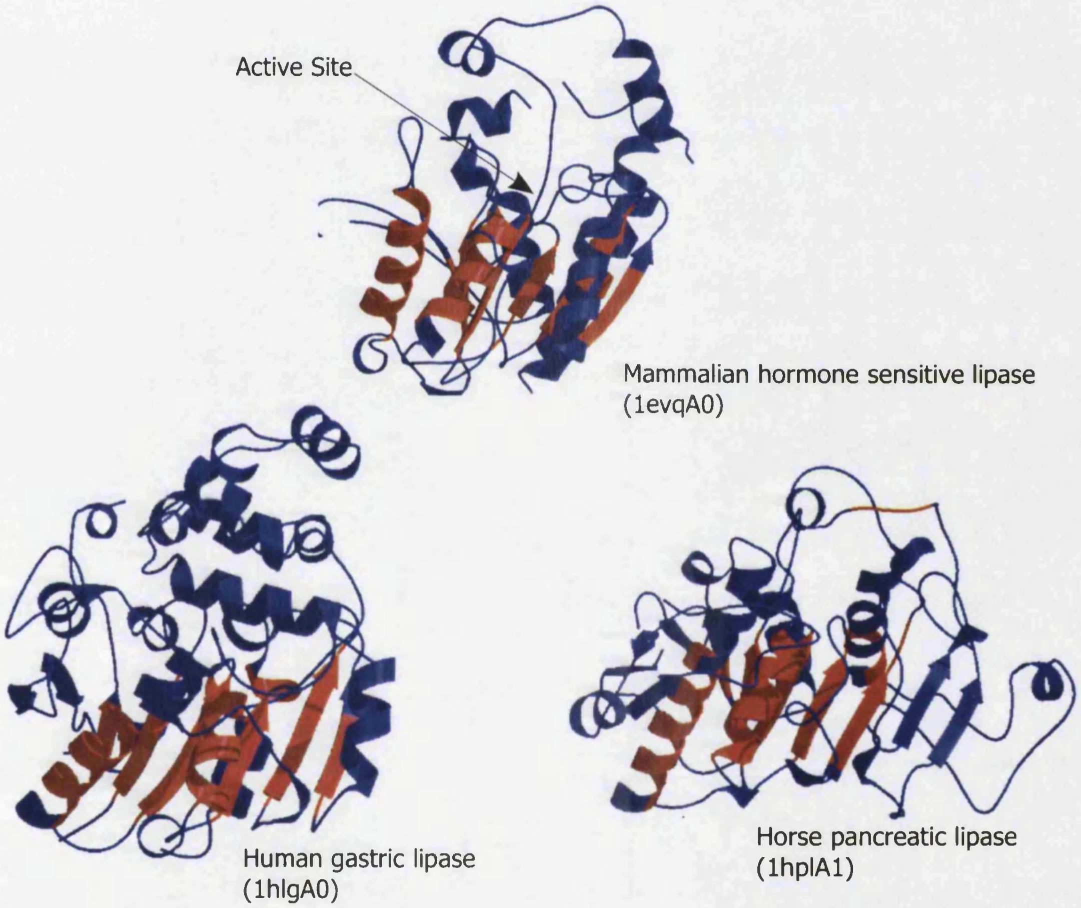


Figure 3.25: Three domains from the $\alpha\beta$ hydrolase superfamily (3.40.50.950) showing the extent of structural embellishments in the superfamily. In mammalian hormone-sensitive lipase (1evq) and human gastric lipase (1hlg) the embellished helices form an extra lobe on the structure. This lobe could be described as a separate α-helical domain, however the structure is classified as a single domain chain by CATH, SCOP and the literature. Halophilic malate dehydrogenase (1hlp) does not have this lobe, but has β-strands that embellish the β-sheet.

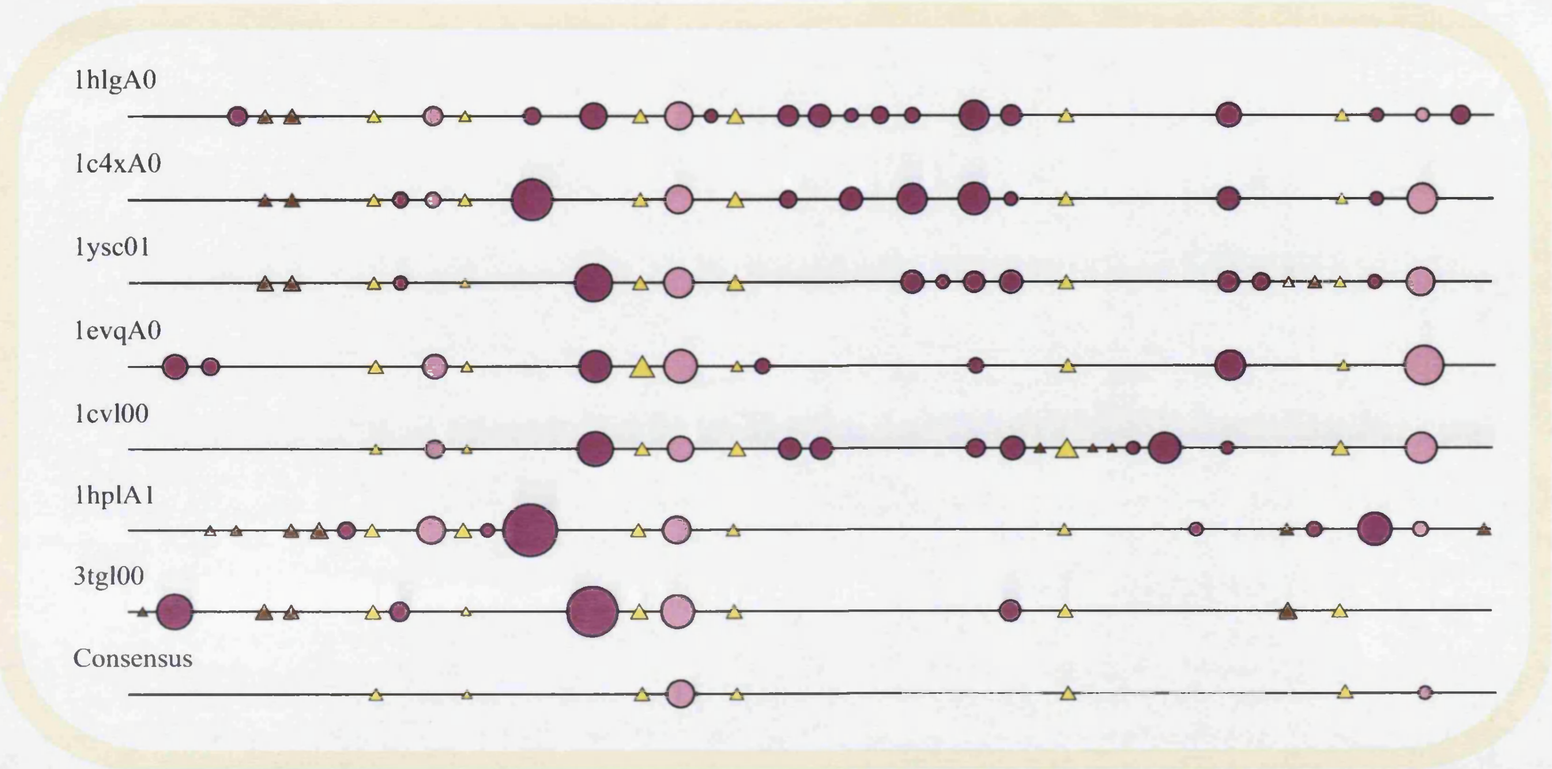


Figure 3.26: The 2DSEC diagram illustrates the positions of the embellishments along the peptide chain in selected members of the $\alpha\beta$ hydrolase superfamily. There are four main areas for embellishments in this superfamily. Extra β-strands are added onto the central β-sheet. The extra helices pack against one another, mainly at the back and the top (as it is orientated in figure 3.25). The third embellished region, situated between the fourth and fifth consensus strands, contains a string of α-helices in four superfamily members. These helices form an extra lobe on the structure together with the α-helices inserted into other areas of the peptide chain.

3.3.8.3 General Conclusions on Secondary Structure Insertions

From this study, a number of conclusions can be drawn:

- The most represented architectures are the sandwich architectures, the β sandwiches and the 3-layer ($\alpha\beta\alpha$) sandwiches.
- α-helical embellishments are harder to characterise as they satisfy their own hydrogen bonds and can pack against the consensus structure in a number of orientations.
- β-strands often form additions to a main β-sheet, otherwise forming β-hairpins.
- The most common type of embellishment in the current dataset comprises extra α-helices.
- In looking at consecutive inserts it is apparent that most of the insertions in the sequence correspond to a small number of secondary structure elements, but often when the three-dimensional structure is examined these small insertions have come together to form a large embellishment which sometimes modifies the geometry of the active site or creates additional interfaces for interactions with other proteins. So secondary structures which are next to each other in the embellishment as seen in the three-dimensional structure may not be adjacent to each other in the sequence. There are very few large continuous insertions, suggesting that evolution of structure though insertion of small structural motifs by gene shuffling and fusion is not a common mechanism.
- The 3-layer ($\alpha\beta\alpha$) sandwiches (3.40s) almost always have a β-sheet extension as part of their embellishment. This may be indicative of the types of additions these architectures can tolerate. A β-sheet extension would not significantly alter the packing of α-helical and β-strand layers. Similarly, with the α-bundles, extra helices could pack alongside the helices which are the major composition of proteins in the architecture.

It would be of interest to assess the structural impact of these embellishments on the active site of the protein. Do many residues from the inserted segments provide contacts to the substrate in the active site? Currently it is difficult to perform a systematic analysis of this because information on protein – ligand interactions is not available for all members in each superfamily. Presently, work is being undertaken to annotate CATH superfamilies with GO annotations (Ashburner *et al.*, 2000) and with ligand binding data from the MSD (Boutselakis *et al.*, 2003) database which would make a functional and structural study more possible. However, a more detailed study of three superfamilies is carried out in Chapter 4.

3.4 Conclusions

Examining the correlation between sequence and structure revealed a bi-phasic relationship, in agreement with Chothia & Lesk (1986). It was found that above 35% sequence identity the structural change was gradual and linear whereas below 35% sequence identity there is a greater structural divergence between homologous pairs.

The extent of residue insertions and deletions in relatives in the different superfamilies was then examined. It was found that as sequence identity increased, the average indel length decreased ranging from six residues for sequence identities between 0–10% to four residues between 20–40% sequence identity and two residues between 40–95% sequence identity. Dividing the data into the 3 classes showed that structures in the mainly β class were slightly less tolerant to indels, having lower average indel lengths and that the structures in the $\alpha\beta$ class were more tolerant. The sandwiches 3.40 ($\alpha\beta\alpha$) and 3.50 ($\beta\beta\alpha$) show most toleration to indels. Inserted β-strands are frequently located at the edges of the β-sheet which does not significantly disrupt the packing of the layers. When the secondary structure content of these indels was determined (Figure 3.11) it was found that most of the indel residues were in the coil form but α-helices are favoured over β-strands in the $\alpha\beta$ class. It was also observed that most of the domains in the structurally embellished superfamilies belong to multi domain structures. Manual inspection of some of these structures suggests that the embellishments in these superfamilies could be modulating the domain – domain interactions. This is studied in more detail in Chapter 4 when three superfamilies are selected for further study.

Structural analysis of the secondary structure indels revealed three conserved, but well populated, superfamilies and functional reasons for the structural conservation were considered. This analysis revealed structurally conserved superfamilies tend to be small in size, containing averages of nine (Pleckstrin Homology and Phospho-tyrosine Binding), eight (C-Type Lectin-Like) and six (Kinases) secondary structures (Figure 3.14) and all three are involved in protein – protein interactions in enzyme cascades.

Having identified the most structurally variable superfamilies in the CATH database it is evident that superfamilies adopting an $\alpha\beta$ sandwich architecture are more tolerant to structural change than other superfamilies, for example, by extending the β-sheets without significantly affecting the packing of the β-sheet and α-helix layers in the structure. The high number of these superfamilies could be due to a biased dataset since over 30% of N95Reps in the database are $\alpha\beta$ sandwich architectures. However this bias could also be due to the fact that Nature has re-used these structures more extensively because they can be structurally modified more easily, leading to new functions and protein – protein interactions.

It was also discovered during the course of the investigation that large structural embellishments are usually created through the addition of a few secondary structures in the loops connecting core secondary structures which then localise in one part of the structure. This is in contrast to insertions of large structural motifs.

This chapter provides a method for the identification of structurally embellished superfamilies. It would be interesting to see what effects these embellishments have on the function of the proteins, when more data are available on specific protein – ligand interactions. In order to set up a protocol to study the functional implications of these structural embellishments, three superfamilies are looked at in more detail in Chapter 4.

Chapter 4

Superfamilies with Domain Enlargement

4.1 Introduction

4.1.1 Background

The increase in structural data in recent years has revealed considerable structural diversity in homologous superfamilies. Indeed, Holm & Sander (1997) introduced the concept of a minimal structural and functional core of related proteins. Large scale diversions are permissible provided that key residues stay intact. Nature has probably embellished more ancestral folds in the evolution of protein function. The increase in complexity of organisms through evolution, from single cell to multi cellular and the increase in complexity of biochemical pathways and functions has lead to an increase in domain complexity through embellishment. However, the reverse scenario does occur. The TIM barrel glycosyl hydrolases present a loss of two α-helices in endo-β-N-acetylglocasiminidase to accommodate its substrate (Van Roey *et al.*, 1994). Likewise, the FAD/NAD(P)(H)-dependent disulphide oxidoreductase superfamily shows a probable truncation in the evolution of flavocytochrome *c*:sulphide dehydrogenase to accommodate the cytochrome *c* subunit (Van Driessche *et al.*, 1996).

Due to the increase in the number of experimentally determined structures it has been possible to carry out meaningful analyses into their structural and functional evolution. In a collective analysis of 31 functionally diverse enzyme superfamilies Todd *et al.* (2001) observed a number of mechanisms the superfamilies have evolved in order to vary the function of a protein. Considering structural change in the active site revealed that the same active site framework may be used to catalyse a host of diverse activities or conversely a different catalytic apparatus may exist in related proteins with very similar functions.

On a larger scale, considerable structural variation is present in some superfamilies. This may be loss or gain of different domain partners or, instead extensive domain enlargement. The variation in domain size involves variation in the loop lengths or embellishments in the structural cores such as β-sheet extensions. Out of the 31 superfamilies analysed, 11 exhibit a two-fold increase in domain size between homologous members. These changes often play a role in substrate specificity or subunit oligomerisation. Domain organisation and subunit assembly plays a large part in the modification of function in homologous superfamilies. In 27 of the 31 superfamilies the domain organisation varies. Additional domains may play a role in regulation, oligomerisation, cofactor dependency, sub-cellular targeting or substrate specificity. Additionally, in 23 of the superfamilies considered, subunit assembly varies between members.

4.1.2 Aims

This chapter provides an analysis of the functional implications of the embellishments found in three superfamilies identified in Chapter 3. The cupredoxin, the ATP-grasp and the thioredoxin-like superfamilies are described, both by the embellished domain common to all superfamily members as well as the whole biological unit or quaternary structure of each member. The modifications in the active site, domain interactions and subunit assembly that might result from these evolutionary embellishments and any functional consequences of these changes are discussed. An overview is shown in Figure 4.1.

4.2 Methods

Three superfamilies were selected from the embellished superfamilies identified in Chapter 3 using 2DSEC (section 2.2.3). These superfamilies were chosen from a dataset of 31 superfamilies which had previously been extensively functionally characterised (Todd, 2001). The cupredoxin, the ATP-dependent carboxylate-amine/thiol ligase (ATP-grasp) and the thioredoxin-like superfamilies exhibit a range of functions within each superfamily, and also vary in domain partners and subunit aggregation.

Superfamily members were clustered into families in which relatives possessed a sequence identity of 35% or greater with at least one other family member and the representative with the highest resolution was taken from each cluster (S35Rep). Some protein structures comprise domain repeats and these repeats may be clustered into the same or different sequence families. Where they have been clustered separately, they are treated separately when describing the embellishments to the superfamily and described together when considering the whole protein.

4.2.1 Description of the Embellished Domain

The embellished domain in each member of the dataset is identified by its six letter CATH code: the PDB identifier followed by the chain to which it belongs and the domain number. Structural variation of the S35Reps was measured using SSAP (section 1.5.5.2) and plotted versus percentage sequence identity. SSAP gives a global score for structural similarity based on the orientations of the secondary structures, the length and the orientations of the loops and the secondary structural embellishments. SSAP was used to identify the most representative structure and the mean and standard deviation were recorded.

Secondary structural variability was measured by 2DSEC. The variations in total number of secondary structures in each domain in the dataset are summarised on a histogram and the positions and types of secondary structure are illustrated on the 2DSEC cartoon.

4.2.1.1 The OC Clustering Program

In two of the three superfamilies, members are clustered by SSAP score using a simple distance measurement tree drawing program, OC (Barton, 2002) to give a visual representation of the pairwise structural similarities between superfamily members.

4.2.2 The Biological Unit of Each Protein in the Superfamily

4.2.2.1 Description of the Domain Partners

The full biological unit of each protein in the dataset in terms of domain partners in the monomer and subunits in the quaternary structure were determined from the literature. Protein quaternary structure was viewed in Rasmol (Sayle & Milner-White, 1995) using the PQS server described below (Henrick & Thornton, 1998; Ponstingl *et al.*, 2000) to create the PDB files with all subunits. Several questions were addressed. Do the proteins have different domain partners? If they have the same domain partners, are they orientated in the same way in each relative? How do variations in domain partners and/or subunit assembly affect the active site? Are the domain/subunit interactions mediated by the secondary structural embellishments?

Protein Quaternary Structure (PQS)

The protein quaternary structure (PQS) (Henrick & Thornton, 1998; Ponstingl *et al.*, 2000) database provides the co-ordinates for the likely quaternary states of structures found in the protein data bank that have been solved by X-ray crystallography. As outlined in the documentation for the PQS server, the crystallographic co-ordinates obtained for a given protein are not independent of the crystallographic symmetry (space group and unit cell), and therefore may not represent the complete molecule that is under study, or may include several copies of the molecule. The method underlying PQS aims to recognise multiple copies and/or generate protein co-ordinates that describe the biological assembly of a particular protein from symmetry. Biologically relevant protein-protein interaction sites are distinguished from those considered to be a result of crystal packing by measuring the size of the solvent accessible surface area buried in the interface, solvation energies of folding, salt bridges and disulphide bonds formed at the interface. The PQS database web server (http://pqs.ebi.ac.uk) was used to visualise the quaternary structures described by the literature.

4.2.2.2 Functions of Relatives in the Superfamilies

The individual functions (where known) for each protein in the superfamily were collected from the detailed functional analysis (Todd, 2001) and the literature. The particular roles of the embellished domains that together contribute to the overall function of the protein were described. Particular reference was made to the positions of the active sites and any differing size in the substrates bound.

4.2.3 Calculating the Proximity of Embellishments to Other Domains or Subunits

The proximity of the embellishments on each domain to the other domains and subunits in the quaternary structure was measured using KdTree (Jonathan Barker, personal communication) a resource from the European Bioinformatics Institute. KdTree measures all atoms in a given sub-set of residues. In this chapter, KdTree measures the distances between residues in the embellished secondary structures (calculated by 2DSEC) and all other domains and subunits in the protein. These residue ranges are encoded by a Perl wrapper (Hugh Shanahan, personal communication). The atoms in the embellishments were measured at two cut-off distances. Those atoms less than 5Å were considered to be in direct contact with another domain or subunit via hydrogen bonds or close electrostatic interactions and those atoms between 5 and 10Å were considered to have a long range electrostatic influence on the domain and subunit interactions.

The embellishments were selected using the method described in section 2.2.3.2. If a residue in the CORA multiple structural alignment is an α-helix or a β-strand and it is present in less than 75% of the aligned structures, it is counted as an embellishment. Strings of four or more embellished residues were selected for measurement.

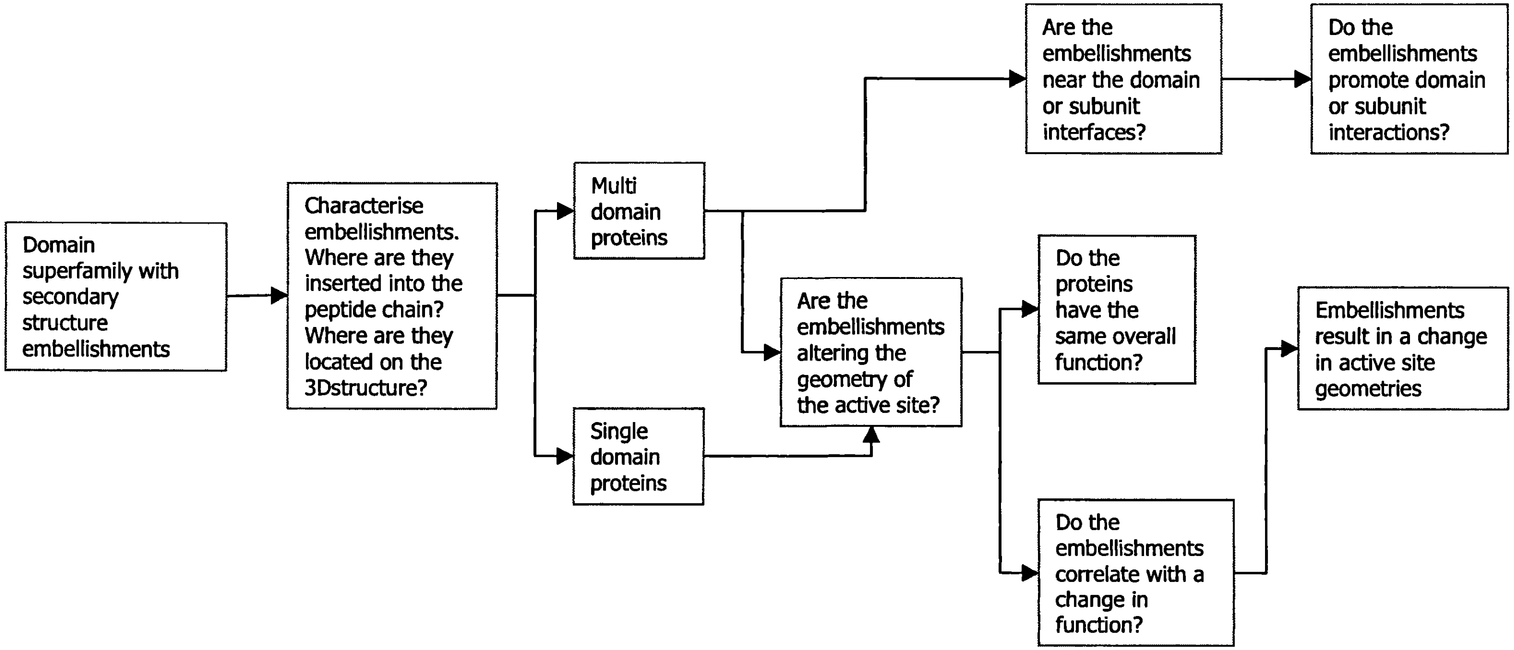

Domain superfamily with secondary structure embellishments
Characterise embellishments. Where are they inserted into the peptide chain? Where are they located on the 3Dstructure?
Multi domain proteins
Single domain proteins
Are the embellishments near the domain or subunit interfaces?
Do the embellishments promote domain or subunit interactions?
Are the embellishments altering the geometry of the active site?
Do the proteins have the same overall function?
Do the embellishments correlate with a change in function?
Embellishments result in a change in active site geometries

Figure 4.1: A flow diagram showing the steps in the analysis of impacts of structural embellishments on protein functions in three protein superfamilies. Firstly, the embellishments of the selected CATH superfamily were characterised. Then, for each member of the superfamily, the whole biological unit including associated domains and subunits were examined. The embellishments were then characterised to examine any impact on the active site or the interactions between the domains.

4.3 Cupredoxin

4.3.1 Overview

Cupredoxin is a stable mainly β sandwich domain (Ryden & Hunt, 1993) which has been adopted by a number of proteins exhibiting different functions. Some members of this superfamily consist of a non-enzymatic single domain and others have evolved to form multidomain enzymes, made up of repeats of the cupredoxin fold (Figure 4.2). The cupredoxin superfamily of domains in CATH contains 19 sequence families with >35% sequence identity. The S35Reps can be clustered into three functional groups: small blue electron transfer agents, oxidases and the heme-copper respiratory oxidases.

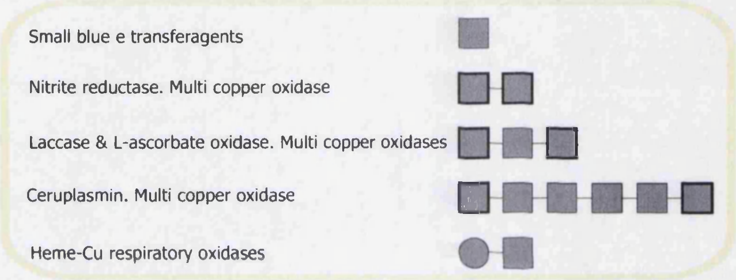

Small blue e transferagents
Nitrite reductase. Multi copper oxidase
Laccase & L-ascorbate oxidase. Multi copper oxidases
Ceruplasmin. Multi copper oxidase
Heme-Cu respiratory oxidases

Figure 4.2: The domain assembly in the cupredoxin superfamily. The small blue electron transfer proteins are single domain. The multi copper oxidases comprise multiple cupredoxin domains and the heme-Cu respiratory oxidases comprise a single cupredoxin domain and a non cupredoxin domain. Domains represented with a bold square are involved in the catalytic activity of the protein.

The cupredoxin domain typically binds copper and several types of copper sites are known to exist within the fold, such that homologous domains differ in copper content (Murphy *et al.*, 1997a). All small blue electron transfer agents are single domain and have one copper binding site, the multi-copper oxidases contain multiple cupredoxin domain repeats and have more than one copper binding site. The heme-Cu respiratory oxidases have a single cupredoxin domain and a membrane associated domain, and bind copper in a binuclear binding site. The single domain electron transfer agents are most likely to closely represent the common ancestor and the multi copper oxidases evolved later (Ryden & Hunt, 1993). Table 4.1 provides functional and structural information for individual S35Reps in this study.

Type	Name	PDB	Copper sites[1]	Function	# Subunits & Domains
Electron Transfer Protein	Rusticyanin	1rcy00	Type I	Electron transfer. (Botuyan *et al.*, 1996)	Monomeric. Single domain
	Stellacyanin	1jer00	Type I	Electron transfer. Low redox potential. (Hart *et al.*, 1996)	Monomeric. Single domain
	Azurin	1jzeA0	Type I	Electron transfer in respiratory chain. (Hammann *et al.*, 1996)	Monomeric. Single domain
	Pseudoazurin	1paz00	Type I	Electron transfer. (Petratos *et al.*, 1988)	Monomeric. Single domain
	Plastocanin	1plc00	Type I	Electron transfer in photosynthesis. Has an additional capacity to bind to the thylakoyd membrane by a highly charged negative spot on the membrane. Unusually high redox potential. (Xue *et al.*, 1998)	Monomeric. Single domain
	Cucumber basic protein	2cbp00	Type I	Electron Transport. (Guss *et al.*, 1996)	Monomeric. Single domain
Multi Cu Oxidoreductases	Copper containing nitrite reductase	1nif01	Type I & II	Nitric Oxide Reduced to Nitrite. (Murphy *et al.*, 1997b)	Homo-trimer, each monomer with two domains
		1nif02	Type II		
	Laccase (polyphenol oxidase)	1a65A1	Type II& III	It is found in the development of the large hydrophobic molecule lignin and metabolism in fungi and plants. (Ducros *et al.*, 2001)	Monomeric, three domains
		1a65A2	None		
		1a65A3	Type I, II & III		
	L-ascorbate oxidase	1aozA1	Type II &III	May be involved in redox system involving ascorbic acid. (Messerschmidt *et al.*, 1992)	Dimeric, three domains in each
		1aozA2	None		
		1aozA3	Type I, II & III		
	Ceruplasmin (ferroxidase)	1kcw01	Type II &III (some residues from dom 6)	Blue copper glycoprotein found in plasma. Four possible functions are ferroxidase activity, amine oxidase activity, copper transport and homeostasis, and superoxide dismutase activity. (Murphy *et al.*, 1997a)	Monomeric, six domains
		1kcw02	Type I (also on dom 4 & 6)		

continued on next page

Table 4.1: *continued*

Type	Name	PDB	Copper sites	Function	# Subunits & Domains
Heme Cu-respiratory oxidases	Cytochrome C oxidase polypeptide II	2cuaA0	Haem Cu Subunit I	Subunit I and II form the functional core of the enzyme complex. Electrons originating in cytochrome c are transferred via haem and Cu(A) to the catalytic binuclear centre in another non-homologous subunit. (Williams *et al.*, 1999)	Membrane bound hetero-oligomer
	Nitrous oxide reductase	1qniA2	CuA	Elimination of nitrous oxide from biosphere by denitrifying bacteria. (Brown *et al.*, 2000)	Homodimeric. One domain in each monomer
	Cytochrome C oxidase	1occB2	CuA	Component of the respiratory chain. Electrons originating in cytochrome c are transferred via the CuA centre (sub 2) and haem of subunit 1 to the bimetallic centre formed in another non-homologous subunit. (Tsukihara *et al.*, 1996)	13 different polypeptide subunits

Table 4.1: The S35Reps from the cupredoxin superfamily. The table lists the name, the PDB code, chain identifier and domain number, the number and type of copper sites, the function and the quaternary structure of the 19 S35Reps selected for structural and functional analysis.

[1] Type I Cu site almost invariably comprises two His residues, and one Cys and one Met. Several domains supply ligands for Type II and binuclear Type III inter-domain copper sites. Adjacent Type II and III sites form a trinuclear site when both are present. The three copper ions are co-ordinated by eight His residues, located within four His-X-His motifs and contributed by two domains.

4.3.2 Structural Description

The cupredoxins are a superfamily from the mainly β class, with a sandwich architecture and immunoglobulin-like topology. In the literature the fold is described as an eight stranded β greek key barrel. The fold begins with the strand on the right hand side of the front sheet (Figure 4.3). The first and third strands of the cupredoxin barrel are parallel in contrast to the other greek key barrels such as superoxide dismutase and the immunoglobulins which only contain anti-parallel β-sheets.

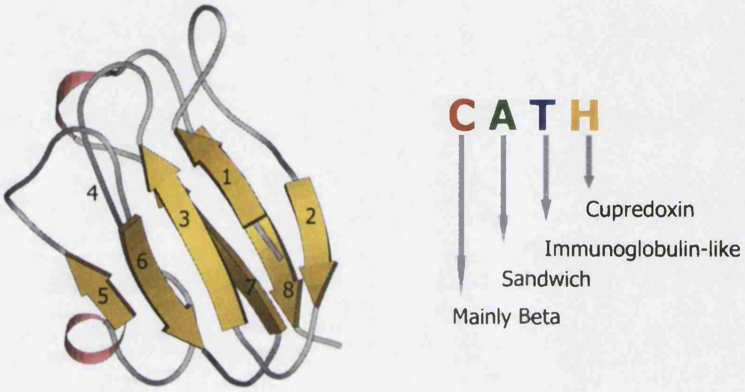


Figure 4.3: Plastocyanin (1plc). The representative structure for the cupredoxin superfamily. This structure represents the eight beta strands described in the literature as the cupredoxin fold. Strands 1 and 3 are parallel.

4.3.3 Structural Variation

The average SSAP score for this superfamily is 79.76 and the associated standard deviation is 5.35 based on pairwise comparisons of the S35Reps. The lowest score is 68.98 and there are seven pairs which are below 70. The most representative structure is plastocyanin (1plc) having a SSAP score above 80 with 17 other representatives. The smallest domain in this superfamily contains seven (plastocyanin, 1plc00) secondary structures, less than half the number in the largest domain with 16 (L-ascorbate oxidase, chain A, domain 3, 1a65A3) (Figure 4.4).

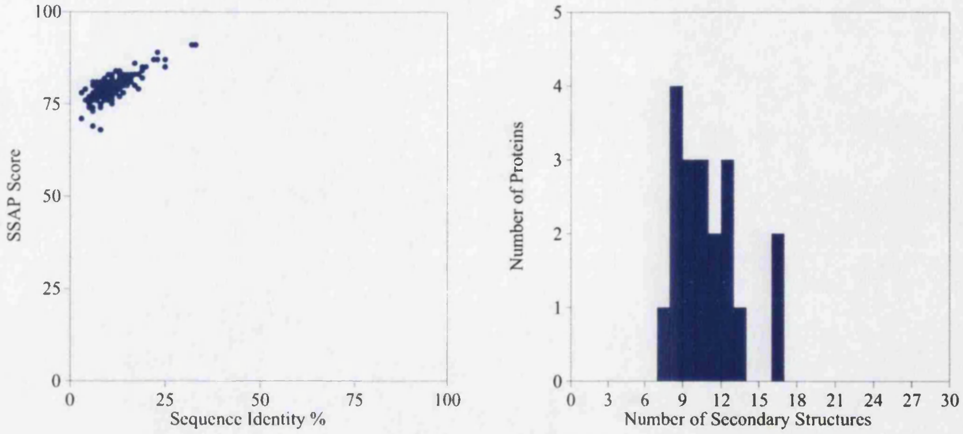


Figure 4.4: The relationship between change in sequence and structure is shown here for the S35Reps in the cupredoxin superfamily. The average SSAP score is 79.76 with associated standard deviation of 5.35. The smallest domain in this superfamily contains 7 (1plc00) secondary structures, which is less than half the number in the largest domain with 16 (1a65A3).

The structural evolution of this superfamily can be categorised into mechanisms including, domain enlargement, domain duplication, segment elongation and subunit aggregation (Ryden & Hunt, 1993). Examples of these four mechanisms are illustrated in Table 4.2. The duplication of the cupredoxin domain in the oxidoreductases produces a variation in the size of the structure from 100 residues in the single domain electron transfer proteins to 1040 residues (6 cupredoxin domains) in a single chain in the oxidoreductases. The number of copper ions bound by the proteins varies from one in the singular domain, small blue electron transfer proteins to eight in the multi domain oxidoreductases (Ryden & Hunt, 1993). Some of the domains in the multi copper oxidoreductases have lost their copper binding sites completely.

Domain Mechanism	Protein
Enlargement	L-ascorbate Oxidase
Duplication	ceruplasmin
Recruitment	cytochrome C
Subunit Aggregation	laccase

Table 4.2: Types of domain modification in the cupredoxin superfamily and examples of proteins in which these mechanisms are observed. More information on these examples is in section 4.3.6.

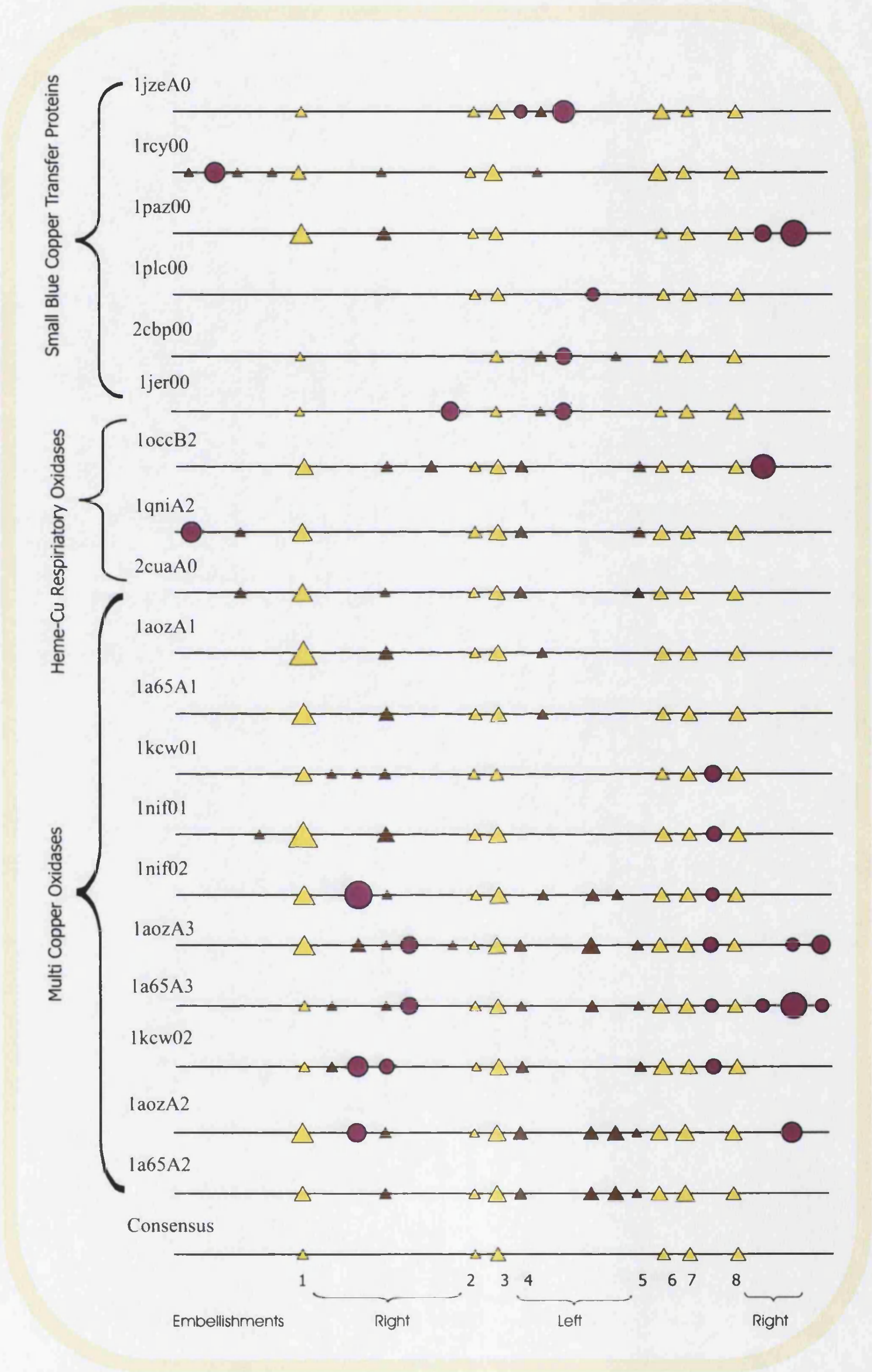


Figure 4.5: 2DSEC cartoon for the S35Reps in the cupredoxin superfamily. This is based on a CORA structural alignment (section 2.2.2). The first six domains are the small blue electron transfer proteins, represented below are the three heme-Cu respiratory oxidases and finally the domains of the oxidoreductases. Multiple domains of the same protein are shown when they have clustered into different 35% sequence families. The consensus defined by the 2DSEC plot is numbered according to the literature numbering of the β-strands (Ryden & Hunt, 1993). There are three main areas of embellishments marked by the horizontal curly brackets at the bottom of the cartoon. The first and the third embellishments both appear on the right side of the β-sandwich as it is positioned in Figure 4.3 and the second appears on the left.

Figure 4.5 was used to illustrate the CORA alignment of each CATH S35Rep taken from the cupredoxin superfamily. The first six domains in the alignment are the small blue electron transfer proteins. Represented below these are the three heme-Cu respiratory oxidases and finally the domains of the multi copper oxidases. Multiple domains of the same protein are shown when they have clustered into different 35% sequence families, such as the three domains of ascorbate oxidase (1aoz) which share less then 35% sequence similarity. Two domains in ceruplasmin are represented in the 2DSEC diagram and the other four are clustered into the same S35 family with these representatives. The consensus β-strands defined by the 2DSEC plot are numbered according to the literature. However there are some descrepancies between the consensus structure cited in the literature and that were identified using CORA and 2DSEC. The fourth and fifth β-strands, which are located as the furthermost left hand strands on the back sheet (Figure 4.3) are sometimes absent and often in a different orientation. As a result, they have not been shown as consensus on the 2DSEC plot. β-strand 5 is present only in some members of the superfamily (Ryden & Hunt, 1993).

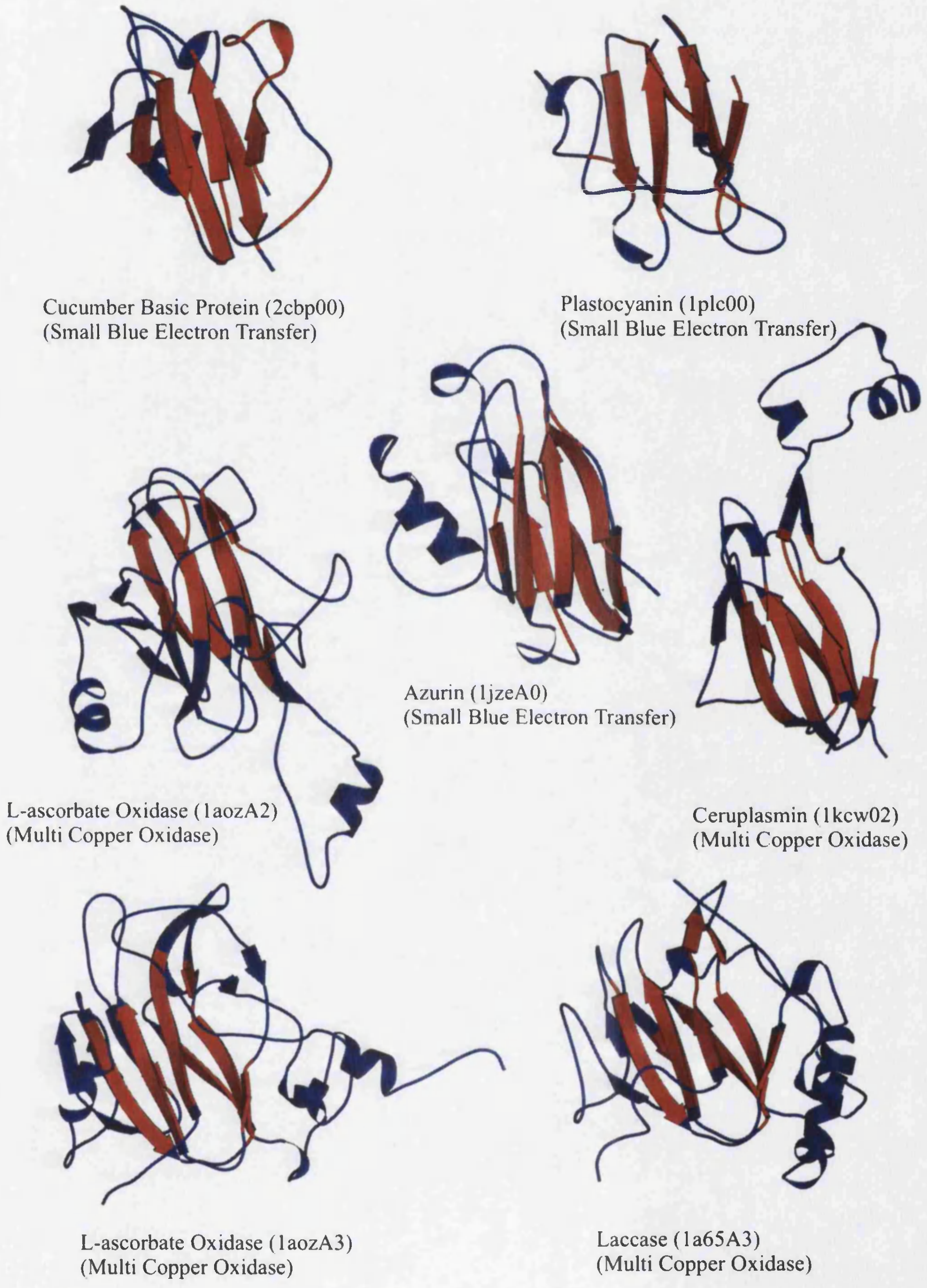

Cucumber Basic Protein (2cbp00)
(Small Blue Electron Transfer)
Plastocyanin (1plc00)
(Small Blue Electron Transfer)
Azurin (1jzeA0)
(Small Blue Electron Transfer)
L-ascorbate Oxidase (1aozA2)
(Multi Copper Oxidase)
Ceruplasmin (1kcw02)
(Multi Copper Oxidase)
L-ascorbate Oxidase (1aozA3)
(Multi Copper Oxidase)
Laccase (1a65A3)
(Multi Copper Oxidase)

Figure 4.6: Showing the extent and types of secondary structure embellishments present in the cupredoxin superfamily. Domains are coloured according to the consensus in the CORA alignment. If more than 75% of aligned residues form part of a β-strand or α-helix, they are coloured red. The cucumber basic protein and plastocyanin show typical cupredoxin structure. Azurin shows additional α-helices at the back of the structure as it is orientated here. L-ascorbate oxidase domain 2 and ceruplasmin domain 2 show embellishments which elongate the structure of the domain vertically. L-ascorbate oxidase domain 3 and laccase domain 3 show embellishments to either side of the sandwich, elongating the structure horizontally.

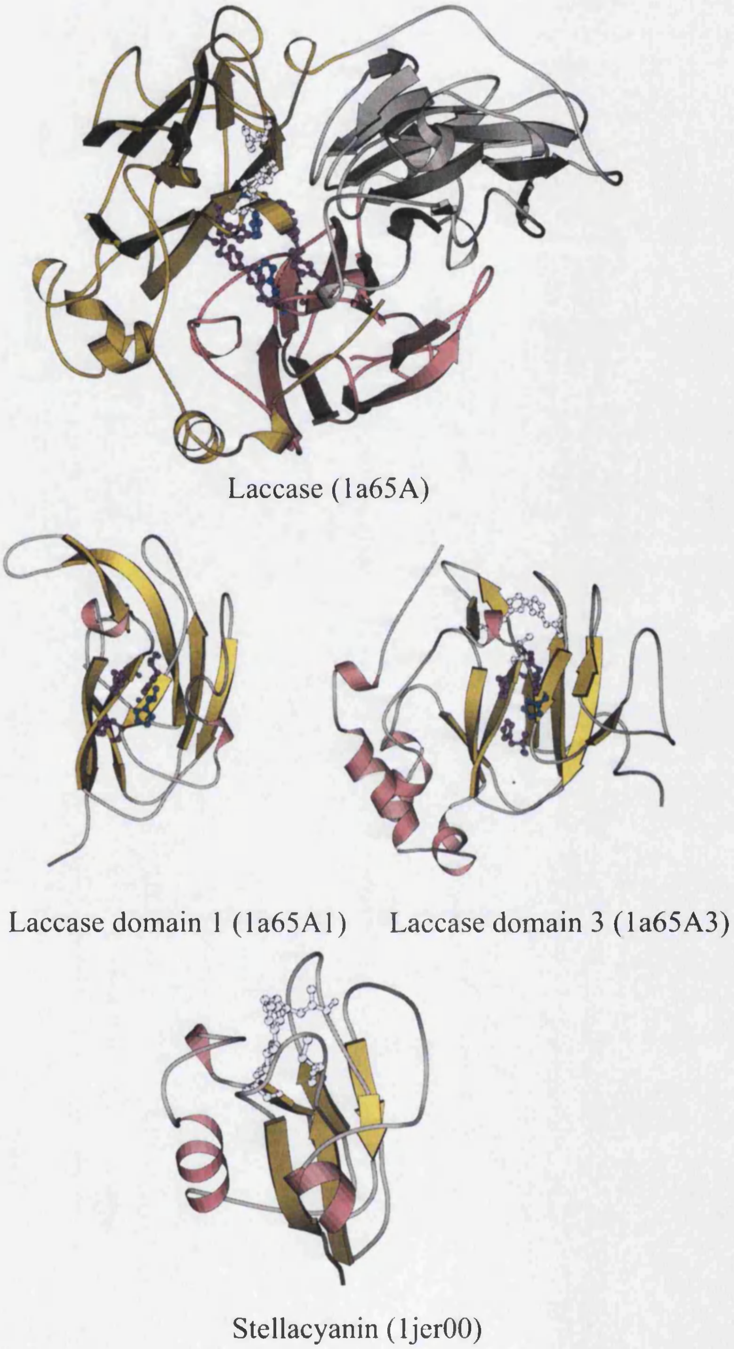
The majority of structural variation in this superfamily is in the extension of the β-sheets which form the β-sandwich. In some cases extra strands have been added onto the end of the sheets (shown by the cucumber basic protein, azurin, ceruplasmin domain 2 and L-ascorbate oxidase domain 3 on Figure 4.6) and in other, rarer cases the embellishment form hairpins and extra helices away from the central β-sheets (shown by ceruplasmin domain 2 and L-ascorbate oxidase domain 2 in Figure 4.6). The 2DSEC plot in Figure 4.5 shows three places in which embellishments are inserted in this superfamily. Firstly between strand 1 and strand 2 in the consensus structure (1aozA3 and 1kcw02). This embellishment is located on the right side of the structure as it is orientated in Figure 4.6. The additional secondary structures present on the N-terminal of 1rcy00, 1qniA2 and 1nif02 also help to form the same embellishment in the three-dimensional structure. The second is between strand 4 and strand 6 and includes strand 5 which creates an embellishment on the three-dimensional structure on the left side of the domain as it is orientated in Figure 4.6. The final insertion is at the end of the peptide chain. These secondary structures are packed adjacent to the insertion between consensus strands 1 and 2, on the right of the structure as it is orientated in Figure 4.6.

4.3.4 Functional Descriptions

Individual functions of each protein are described in Table 4.1. However, general descriptions can be made for each functional class:

Small Blue Electron Transfer Proteins

The blue-copper electron transfer agents are all single domain and bind Type I copper (Figure 4.7). The Type I Cu site which exists in many cupredoxins almost invariably comprises two His residues, and one Cys and one Met. Type I copper sites are found exclusively in the loops at the top of the cupredoxin domain as it is orientated in Figure 4.7. Small blue proteins are found in bacteria and plants and are involved in the transfer of a single electron from a donor to an acceptor molecule. Plastocyanin, in addition, has the ability to bind to the membrane by using a highly charged negative patch on its surface.



Laccase (1a65A)

Laccase domain 1 (1a65A1) Laccase domain 3 (1a65A3)

Stellacyanin (1jer00)

Figure 4.7: Copper binding sites in the cupredoxin superfamily can be found in three places. The Type I and binuclear copper binding sites are situated in the loops of the structure, represented here by the Type I binding site of stellacyanin. Types II and III binding sites are situated together in the interface between two domains, represented here by domains 1 and 3 in laccase. Some cupredoxin domains have lost their copper binding sites altogether shown here by domain 2 of laccase represented in grey. Type I binding site is shown in white, Type II in turquoise and Type III in purple.

Heme-Cu respiratory oxidases

The cupredoxin domains of the heme-Cu respiratory oxidases are part of large multi-subunit complexes. The cupredoxin domain in cytochrome *c* oxidase transfers electrons from the donor (cytochrome *c*) to the catalytic oxidase centre by way of a binuclear CuA site situated in the same place as the Type I binding site. The catalytic centre is located on a separate polypeptide chain where O_2 is reduced to water, and protons are then pumped across the membrane. The cupredoxin domain provides the reducing substrate binding site (cytochrome *c*). Nitrous oxide reductase contains an identical binuclear site to cytochrome *c* oxidase and probably plays an analogous electron transfer role.

Multi Copper Oxidoreductases

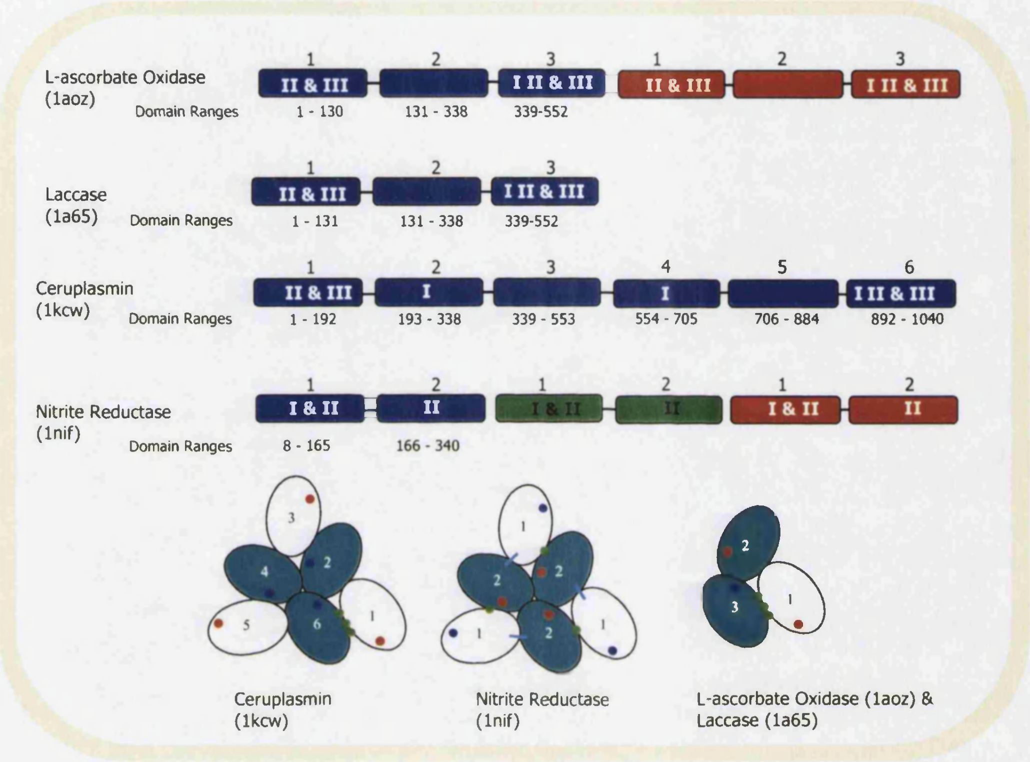


Figure 4.8: Showing the number of domains and the position of the copper binding sites for the four multi copper oxidase proteins in the dataset. L-ascorbate oxidase, ceruplasmin and nitrite reductase all have six domains but laccase functions with three. The domain cartoon is coloured according to chain indicating the number of subunits in the quaternary structure. Also listed are the types of binding sites present on each domain. Below this the arrangement of the domains in two-dimensional space with the copper binding sites are shown. Green dots show the Type II and III copper binding sites, red dots show the inactive Type I copper binding sites (for orientation) and blue dots show the active Type I copper binding sites.

As the multi-copper oxidases evolved from the small blue proteins they acquired three new Cu binding sites close to the existing Type I copper binding site. Types II and III sites form a trinuclear site comprising eight His residues, located within four His-X-His motifs in the interface between two domains represented in Figure 4.7 by laccase. This inter-domain Type II and III binding sites can be found in three members (see Table 4.1) ceruplasmin, laccase and L-ascorbate oxidase. The multi-copper oxidases oxidise their substrate from an electron source *via* a four electron reduction of O_2 to water. Laccase acts on a wide variety of inorganic compounds, L-ascorbate oxidase has a much narrower specificity oxidising L-ascorbate and ceruplasmin uses iron as its reducing substrate. Nitrite reductase, the fourth enzyme of this subset, may be classified with this group of proteins, although it does not function as an oxidase, and instead catalyses the reduction of nitrite to nitric oxide and water, using pseudoazurin as the electron source. All four proteins have a Type I site which functions as the primary electron acceptor, receiving electrons from the reducing substrate and the Type II/III site reduces O_2.

Laccase oxidase is constructed exclusively from three different cupredoxin domains. Only domain 3 binds Type I copper (Messerschmidt *et al.*, 1992). The trinuclear copper site (Type II and III) is formed at the interface between domains 1 and 3. L-ascorbate oxidase also contains three domains in the monomer but unlike laccase it dimerises. Like laccase, domain 3 has Type I Cu site and domains 1 and 3 have Type II and III (Ducros *et al.*, 2001). Ceruloplasmin is a monomer built up by six cupredoxin domains binding six Cu atoms in three mononuclear sites in domains 2, 4 and 6 and one trinuclear interface between domains 1 and 6. Domains 2, 1 and 6 in ceruloplasmin corresponds to 2, 1 and 3 respectively in L-ascorbate oxidase and laccase except for the Type I binding sites on domain 2 of ceruplasmin not present in L-ascorbate oxidase or laccase. Based on an evolutionary study it has been predicted that ceruplasmin has evolved by repeating two cupredoxin domains three times (Messerschmidt & Huber, 1990). Consistent with this prediction, domains 1, 3 and 5 are classified into the same 35% sequence cluster and domains 2, 4 and 6 are classified together in CATH. Nitrite reductase (1nif) is a trimer of two cupredoxin domains, suggested to be similar to its distant relative ceruplasmin (Godden *et al.*, 1991). However, copper binding is quite different. Only Type I and II sites exist. This information is shown schematically in Figure 4.8.

4.3.5 Structure/Function

Amongst the enzymes, the cupredoxin domains have a mixture of enzymatic and non-enzymatic roles. In the heme-Cu respiratory oxidases the cupredoxin domain shares the same function as the electron transfer agents and the enzymatic activity is carried out by another subunit. In the multi copper oxidases there is a mixture of copper binding and non copper binding cupredoxin domains. However, generally, as the domain becomes more embellished, the number of copper binding sites present on the whole biological unit also increases, although the copper binding sites are situated in the conserved core of the protein where there is little change to the structure and not on the embellishments.

Clustering the domains using the SSAP score (section 4.2.1.1) indicates four structural subclusters (Figure 4.9). Group I containing the multi copper binding oxidoreductases, group II containing the non-enzymes, group III and group IV containing the non copper binding and the CuA binding sites.

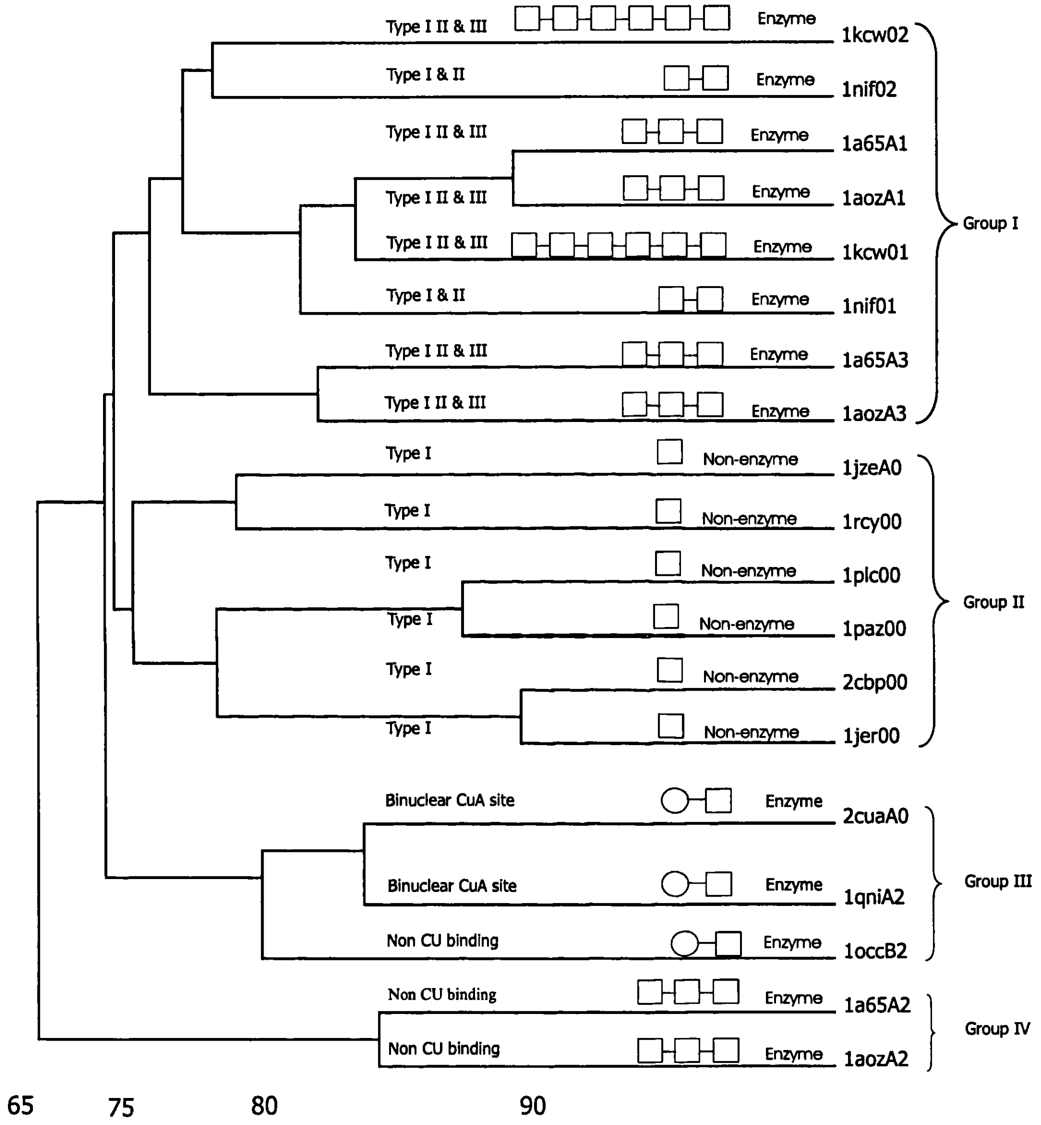


Figure 4.9: Domains of the cupredoxin superfamily clustered by SSAP score using OC. Information about the number and type of domains, whether the protein is an enzyme or a non-enzyme and what types of Cu binding are listed also. Four groupings can be seen: Group I, the Cu-binding domains of the multi copper oxidoreductases; Group II, the small blue electron transfer proteins; Group III, CuA binuclear and non Cu binding; Group IV, non Cu binding domains of the multi copper oxidases. The numbers indicate the approximate SSAP scores between domains linked at that position in the tree.

4.3.5.1 Group I: The Multi Copper Oxidoreductases

This cluster contains all the domains associated with Type II and III copper binding sites in multi copper oxidases. The copper binding sites are not located on the additional embellishments but are on the consensus β-strands.

4.3.5.2 Group II: The Small Blue Electron Transfer Proteins

These small blue electron transfer proteins have remained relatively small but exhibit some structural embellishments throughout the cluster, although there is high structural similarity in all metal binding sites. Each has distinct chemical properties due to surrounding amino acids (Walter *et al.*, 1996). The embellishments do not play a role in modification of the metal site but may play a role in the interactions with other proteins in electron transport chains. They have remained monomeric as they are responsible for connecting protein complexes in the electron transport chains via diffusion. For example, plastocyanin (1plc) connects cytochrome b6f and the PSI complex in the plant thylakoid membrane by its diffusion (Romero *et al.*, 1998). Another possible reason for remaining monomeric and single domain is because they do not need to form clefts with other proteins in order to form active sites. Their sole role is to receive and pass on electrons to other members of the electron chain in the complex. These domains are most likely to represent the ancestor of this superfamily (Ryden & Hunt, 1993). The chains in these proteins vary in length from 97 residues (cucumber basic protein) to 129 (azurin). The enlargements are mainly due to added length in the loops between the strands. The largest variations occur between β-strands 4 and 5 in azurin (1jzeA0) where the residues form a flap which contains a helix (Ryden & Hunt, 1993). From the 2DSEC diagram it is possible to see that both the cucumber basic protein (2cpb) and stellacyanin (1jer) also contain this flap with a helix (Figure 4.5).

4.3.5.3 Group III: CuA Binuclear

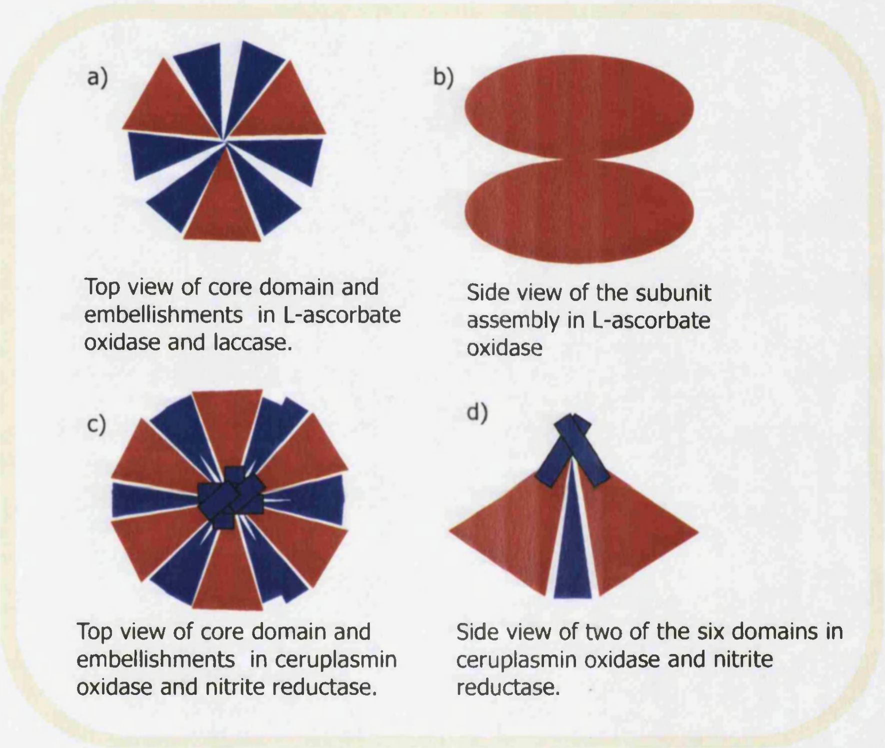
Other groupings within the tree show that all single cupredoxin domains associated with another membrane-associated domain cluster together. The embellishments present in the three examples (1occB2, 1qniA2 and 2cuaA0) are different (Figure 4.5). However they have high general structural similarity as they have been clustered by SSAP score (Figure 4.9). The embellishments could be interacting with the membrane associated domain. However no structural data are available for the membrane domain in any of these examples.

4.3.5.4 Group IV: Non Cu Binding

There are only two domains in this cluster (Laccase domain 2 and L-ascorbate oxidase domain 2). Both domains are part of a three domain subunit of a multi copper oxidoreductase and neither functions as a copper binding domain but are both extensively embellished.

4.3.6 Embellishments in the Multi Copper Oxidases

The most interesting relationship in this superfamily is the evolution from the small blue electron transfer proteins to the multi copper oxidoreductases. The transition includes extensive embellishments, domain duplications and subunit aggregation and from this, evolution has created an additional copper binding site the Type II and Type III binding sites which are situated in the cleft between domains. The analysis in this chapter shows no direct correlation between the number of copper binding sites and domain enlargement as the copper binding sites are all situated in the consensus structure. However, there is an indirect correlation. As number of domains rises the amount of embellishment increases also. By identifying the embellished secondary structures in this superfamily it is possible to see that the embellishments form the interactions between the domains stabilising the multi domain structures (Figure 4.10).



Top view of core domain and embellishments in L-ascorbate oxidase and laccase.

Side view of the subunit assembly in L-ascorbate oxidase

Top view of core domain and embellishments in ceruplasmin oxidase and nitrite reductase.

Side view of two of the six domains in ceruplasmin oxidase and nitrite reductase.

Figure 4.10: The arrangement of the domains and embellishments in the cupredoxin multi copper oxidases. The red regions show core secondary structures and the blue regions are the embellishments. The domain arrangement in L-ascorbate oxidase and laccase is shown in (a). Embellishments on either side of the domains interact to form the full unit. Diagram (b) shows a side view on how these three domain monomers interact to form the homodimer L-ascorbate oxidase. The domain organisation of ceruplasmin (six domains in one single chain) and nitrite oxidase (three subunits, each with two domains) is shown in (c) Like L-ascorbate oxidase and laccase, there are embellishments between the domains but in addition to this there are also embellishments above the domain which interact with each other, shown in (d).

Laccase and L-ascorbate oxidase have the same domain organisation. The only difference is that L-ascorbate oxidase forms a homodimer with subunit placed on top of the other (Figure 4.14). The metal binding site residues for the Type II and III binding sites are situated on the consensus strands (Figure 4.7) on strands 4 and 6 in domains 1 and 3. The embellishment to the left as it is called in Figure 4.5 does not have any direct role in the binding of the additional copper ions. When looking at the formation of all three domains in this monomer it is apparent that there are additional strands promoting domain-domain interactions. The interaction between domain 1 and domain 3 is stabilised by the α-helices stretching round the back of domain 3 and to the right of domain 1. This is shown in Figure 4.11 by laccase and a schematic of the domain organisation is shown in Figure 4.10. The regions of core secondary structure are shown in red and the structurally variable regions (SVR) are coloured blue as calculated by 2DSEC (section 2.2.3.2).

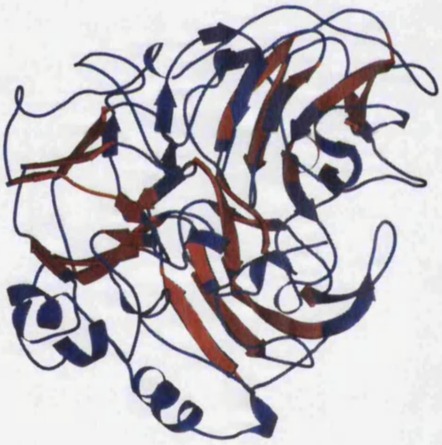


Figure 4.11: The biological unit of laccase (1a65). The domains are coloured so that residue positions in the CORA structural alignment with the same α-helix or β-sheet conformation in 75% of domains appear red. The blue regions represent those secondary structures which are only present in less than 75% of domains or where the structure is coil. This shows those secondary structures which are embellishments to the core fold and their position in the quaternary structure. Most of these embellishments appear to interact to form domain contacts.

The major embellishment in ceruplasmin is present in all domains in this superfamily. It is between consensus strands 1 and 2. Together these six embellishments assemble together in the structure stabilising the domain organisation (Figure 4.12). The oxidoreductase site in this protein is between the right side of domain 1 and the right hand side/back of 6. During evolution, the three most recently acquired sites were again lost in the two new double domains after triplication occurred while the Type I site retained its copper. Nitrite reductase shows the same domain organisation. The arrangement of all six domains is shown in Figure 4.13 in orthogonal representations. Each domain has an embellishment which lies above the β-sandwich and packs against similar embellishments from other domains, promoting the domain interactions (Figure 4.10).

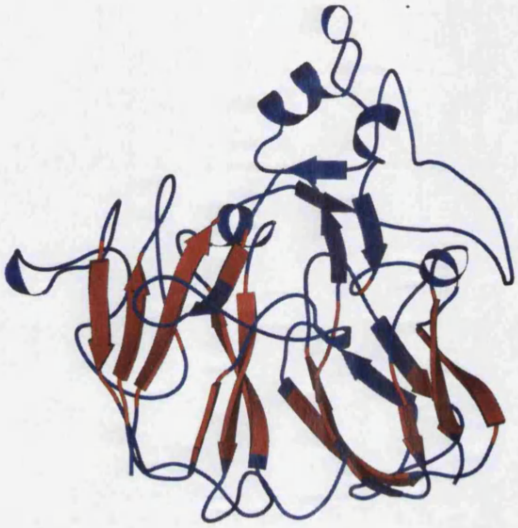


Figure 4.12: Two of the six domains from ceruplasmin (1kcw domains 1 and 2) show how the main embellishment stabilises the domain interaction from above.

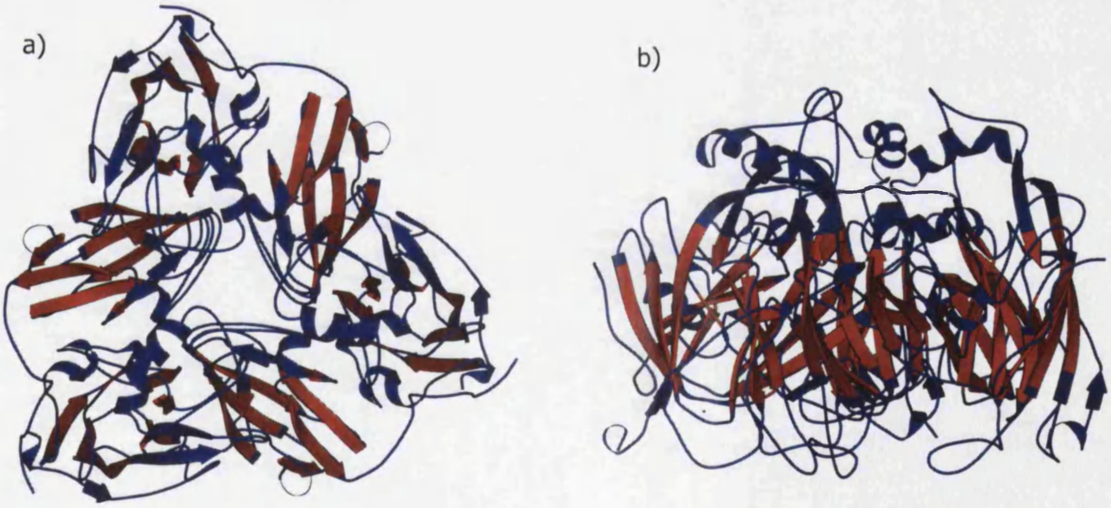


Figure 4.13: The biological unit of nitrite reductase (1nif) shown in two orthogonal representations. Figure (a) shows the interactions between the six domains. The embellishments (shown in blue) are located between each domain. Figure (b) shows a similar embellishment to that in ceruplasmin (Figure 4.12) interacting above the core domain structures.

4.3.7 Measuring the Embellishment Domain/Subunit Interactions

The proximity of the embellishments to other domains or subunits were measured using KdTree (section 4.2.3). Figure 4.14 shows the quaternary structure of L-ascorbate oxidase (1aoz) and the position of all residues, in the embellishments in domain 3 at a distance of 5Å, between 5 and 10Å and greater than 10Å to other domains and subunits.

For all multi-copper oxidase domains in this dataset the distances of the embellishments to other domains or subunits were measured. The data are shown in Figure 4.15. For seven of the ten domains (L-ascorbate oxidase domains 1 and 3, laccase domains 2 and 3, ceruplasmin domain 2 and nitrite reductase domains 1 and 2) more than 50% of the residues in the secondary structure embellishments are 5Å or less from another domain or subunit and in laccase domain 1 and ceruplasmin domain 1 more than a quarter of the residues in the secondary structure embellishments are greater than 10Å.

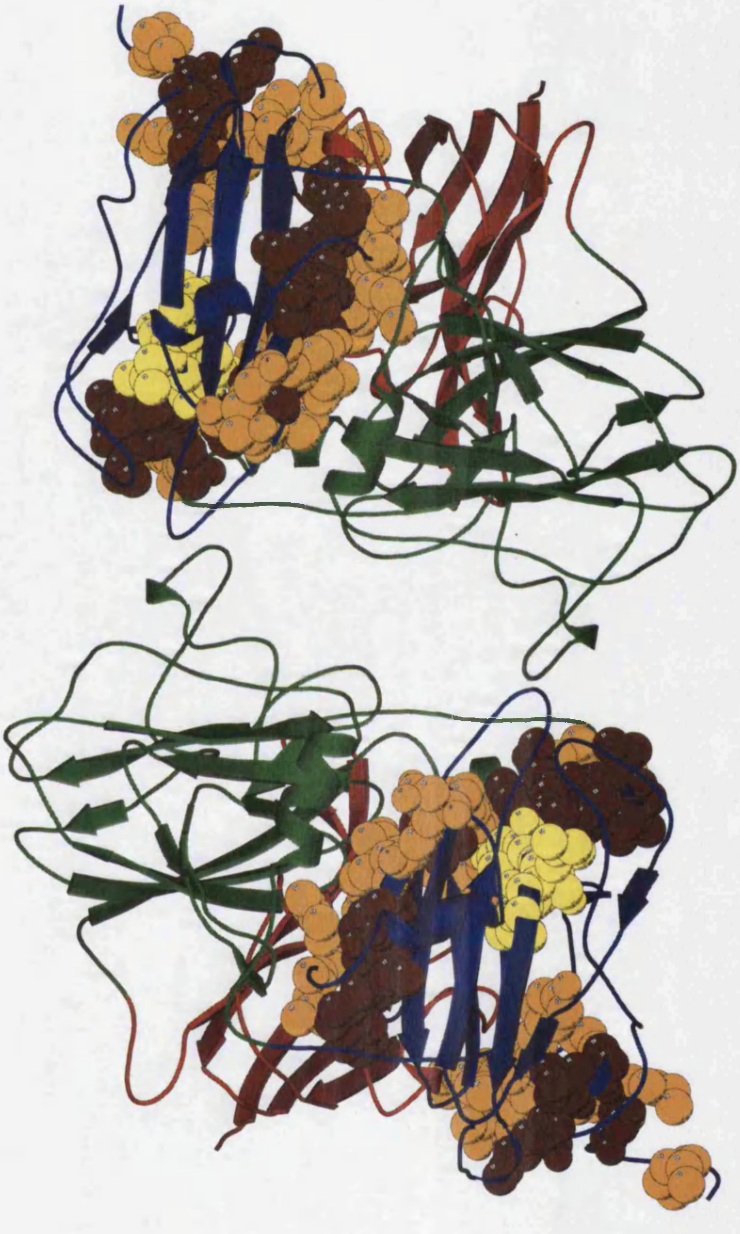


Figure 4.14: The quaternary structure of L-ascorbate oxidase and the interactions between the embellishments on domain 3 and the rest of the protein. Domain 1 is coloured red, domain 2 is coloured green and domain 3 is blue. The interactions between the embellished secondary structures and the other domains are shown in brown, orange and yellow. Residues shown in orange are less than 5Å from another domain or subunit. Residues between 5 and 10Å are shown in brown and residues greater than 10Å, having no contact with another domain or subunit are shown in yellow.

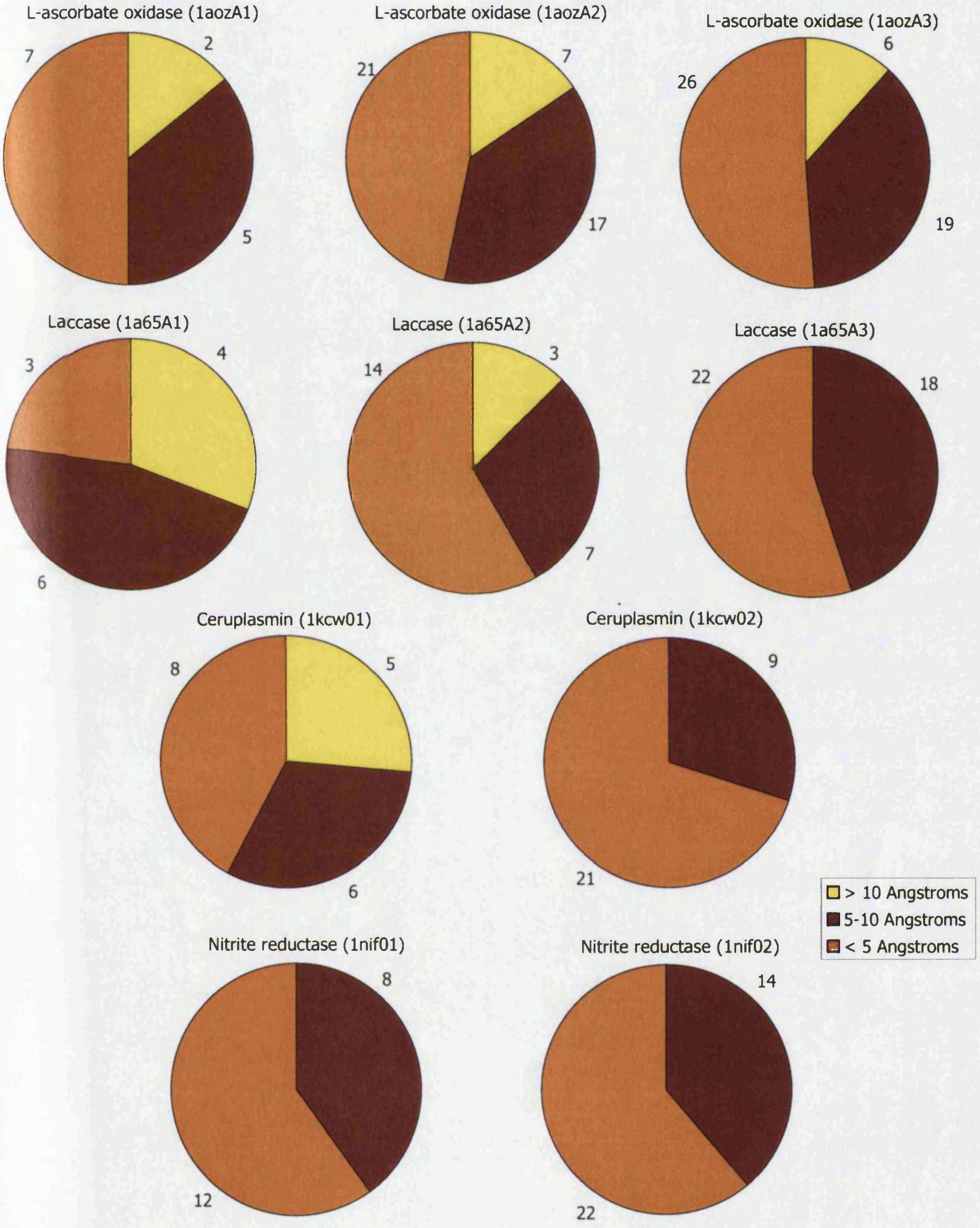


Figure 4.15: The interactions between residues in secondary structure embellishments and other domains and subunits in the cupredoxin multi copper oxidases. The proportion of residues with a distance of 5Å or less from another domain or subunit are represented in the orange segment, the brown segment represents residues between 5 and 10Å and the yellow segment represents residues at a distance of greater than 10Å.

4.3.7.1 Embellishments in the Loops

Another type of variation shown within the multi copper oxidases is the variation in substrate specificity via the loops surrounding the Type I copper binding site. The Type I Cu binding site is at the 'north end' of the structure in the loops. In 1a65A the Type I copper binding site is in a groove, 6Å from the surface. The sides of the groove consist of three loops, including the loop containing the first embellished helix between consensus strands 1 and 2 in 1aozA2 (Figure 4.5), the loop including the first embellished strand and helix before consensus strand 2 in 1aozA3 and the loop region between consensus strands 2 and 3 in 1aozA3 (Figure 4.5) (Ducros *et al.*, 2001). These loops could also be primary substrate binding areas. The fact that they are highly variable in both sequence and size of the groove could account for varying substrate specificity. There are extended loop regions in L-ascorbate oxidase (1aoz) which has a narrow specificity, oxidising L-ascorbate. The loops are implicated in substrate specificity and are completely absent in laccase, which acts on a wide variety of aromatic and inorganic compounds.

4.3.7.2 Cupredoxin Conclusion

Members of this superfamily have enzymatic and non-enzymatic functions. The more simple non-enzymatic proteins are single domain, have fewer secondary structural embellishments and have been described in the literature as the ancestors of the superfamily. As the domains have evolved they have formed multi-domain complexes consisting of multiple copies of the cupredoxin domain unit. This has enabled new copper sites to form in the crevices between the domains. Evolution has also embellished the domains in these multi-domain superfamily members with extra α-helices and β-strands. The embellishments tend to be to the right and left of the structure as they are orientated in Figure 4.6. A possible reason for these embellishments could be to stabilise the structure in its multi-domain state so that the domains are orientated correctly for the function of the types II and II copper binding sites in the crevice between two of the domains (Figures 4.7 and 4.8).

It was found that, although the number of embellishments increases as the number of copper binding sites and the number of domains in the quaternary structure increases, the new copper binding sites are not situated on the embellishments. Instead, the embellishments mediate the domain interactions allowing the new copper binding sites to form in the crevices between the domains.

4.4 ATP-dependent carboxylate-amine/thiol ligase

4.4.1 Introduction

Members of the ATP-dependent carboxylate-amine/thiol ligase (ATP-grasp) superfamily have all been found to be multidomain enzymes. Members of this superfamily typically catalyse ATP-dependent ligation of a substrate carboxylate to an amine or thiol group of a second substrate (Todd, 2001). Nearly all members of this superfamily share three common domains. ATP is bound in the cleft between two of the domains which are referred to as the small and the large ATP binding domains. The third domain, the biotin carboxylase, N-terminal domain-like domain is referred to as the B domain in the literature. The large ATP binding domain, with seven S35Reps in CATH was identified as a structurally embellished superfamily in Chapter 3. The functions, together with domain and subunit conformations of members of this superfamily are summarised in Table 4.3.

Name	PDB	Carboxylate substrate	Amine/thiol substrate	Function	Subunits & Domains
D-alanine D-alanine ligase	1iow02	D-alanine	D-alanine	Cell wall formation. (Fan *et al.*, 1995)	Dimeric. Three common domains.
Synapsin Ia	1auvA3	Biochemical function unknown	-	Neuronal phosphoprotein that coats synaptic vesicles, binds to the cytoskeleton, and is believed to function in the regulation of neurotransmitter release.(Esser *et al.*, 1998)	Seven domains, three of them common.
Glutathione synthetase	1gsh02	γ-glutamylcysteine	glycine	One of two enzymes involved in the production of glutathione. (Yamaguchi *et al.*, 1993)	Homotetramer. Three common domains.
Glycinamide ribonucleotide synthetase	1gsoA3	glycine	5-phosphoribosyl -amine	Catalyses the second step of the purine biosynthesis pathway. (Wang *et al.*, 1998b)	Monomer. Three common domains. The embellishment in this protein is described as a separate domain by Wang *et al.* (1998b)
Phosphoribosylamino-imidazole carboxylase	1b6rA2	HCO_3^-	5-phosphoribosyl-5-aminoimidazole	ATPase activity that is dependent on the presence of aminoimidazole ribonucleotide. (Thoden *et al.*, 1999)	Homodimer. Three common domains.
Acetyl-CoA carboxylase, biotin carboxylase subunit	1bncA2	HCO_3^-	biotin-enzyme	Component of the acetyl CoA carboxylate complex. Biotin carboxylase catalyses the carboxylation of the carrier protein. (Waldrop *et al.*, 1994)	A hetero-oligomer. Biotin carboxylase has the three common domains and exists as a dimer in the acetyl-CoA carboxylase complex.
Succinyl-CoA synthetase, β chain	2sucB1	succinate	coenzyme A	Carries out the substrate level phosphorylation of GDP or ADP in the citric acid cycle. (Fraser *et al.*, 1999)	Four subunits $\alpha 2$ $\beta 2$. The beta subunit comprises ATP-grasp domains and a third domain unrelated to the domains in this superfamily.

Table 4.3: The S35Reps from the ATP-Grasp superfamily. The table lists the name, the PDB code, chain identifier and domain number, the function and the quaternary structure of the seven S35Reps selected for structural and functional analysis. See also Figure 4.20.

4.4.2 Structural Description

The large ATP binding domain is classed in CATH as an $\alpha\beta$ 2-layer sandwich. The literature describes the structure as a five stranded β-sheet flanked by an α-helix with the sequential order of the β-strands being 32145 (Figure 4.16).

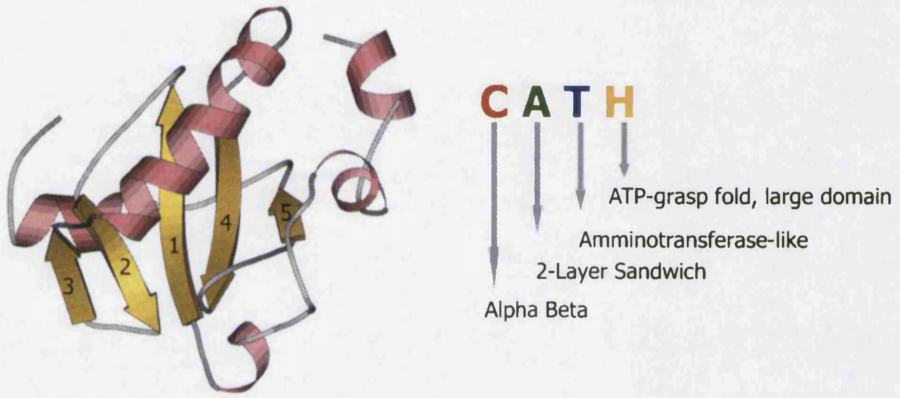


Figure 4.16: The ATP-grasp large domain fold represented by glutathione synthetase (1gsh02).

4.4.3 Structural Variation

4.4.3.1 The Large ATP-grasp Domain

The structure of the large domain described in section 4.4.2 varies in size from five strands to 11 in some members. These domains have an average SSAP score of 77.46 with an associated standard deviation of 6.47 and the most distant pair score of 68.00 between glycanamide ribonucleotide synthetase (1gsoA3) and succinyl-CoA synthetase (2scuB1). The total number of secondary structures varies from eight in 1auvA3 to 20 in 1bnc02 which is a 2.5 fold increase (Figure 4.17). The best representative is phosphoribosylaminoimidazole carboxylase (1b6rA2) with the most (five) pairwise scores above 80.

The 2DSEC diagram for the large domain (Figure 4.18) describes the embellishments present in members of this superfamily. Additional α-helices are found appended to the N-termini of several members of the superfamily. There is also a small embellishment between consensus strand 2 and consensus α-helix 1. The extent of the embellishment varies from one or two extra α-helices in D-alanine D-alanine ligase (1iow02) and glutathione synthetase (1gsh02) and an extra β-strand and α-helix in succinyl-CoA synthetase (2scuB1) to more extensive embellishments in glycinamide ribonucleotide synthetase (1gsoA3), biotin carboxylase (1bncA2) and phosphoribosylaminoimidazole carboxylase (1b6rA2) (Figures 4.18 and 4.19).

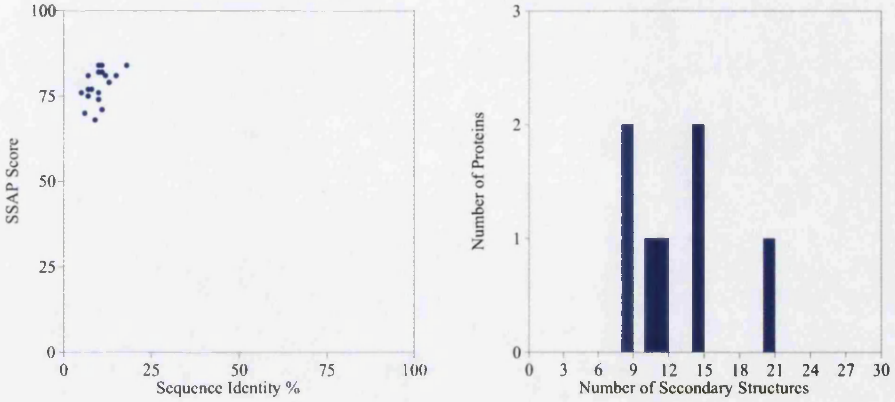


Figure 4.17: Pairwise SSAP score plotted against sequence identity. The large domains have an average SSAP score of 77.46 and the most distant pair score 68.00 between 1gsoA3 and 2scuB1. The distribution of total number of secondary structures in each member of the superfamily ranges from eight in synapsin Ia (1auvA3) to 20 in biotin carboxylase (1bnc02) which is a two and a half fold increase. The best representative is phosphoribosylaminoimidazole carboxylase (1b6rA2) with five pairwise scores above 80.

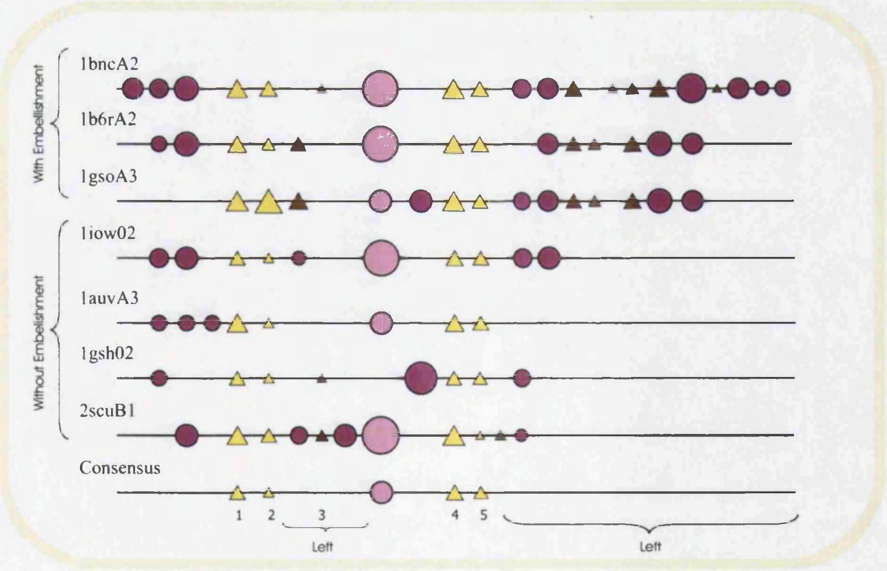
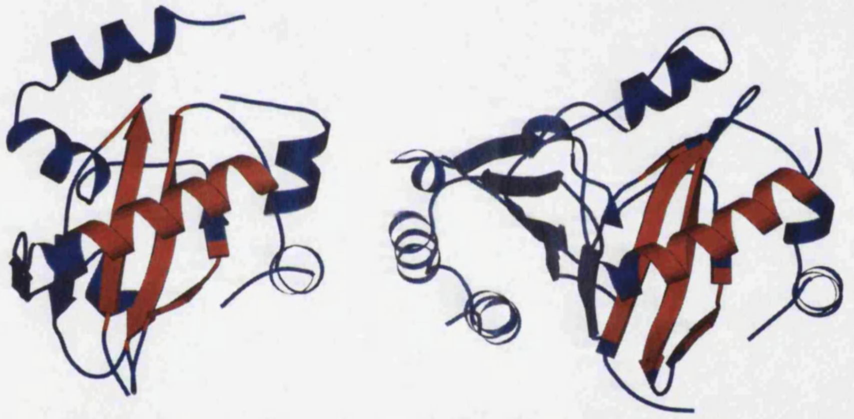


Figure 4.18: A 2DSEC plot of the ATP-grasp large domain. The domains are arranged with those containing the large C-terminal embellishment shown first. Both embellished regions are located together on the left hand side of the three-dimensional structure as it is orientated in Figure 4.19.



D-alanine-D-alanine Ligase (1iow02)

Phosphoribosylamino-imidazole carboxylase (1bncA2)

Figure 4.19: D-alanine-D-alanine ligase (1iow02) without the extension to the β-sheet and phosphoribosylaminoimidazole carboxylase (1b6rA2) with the β-sheet embellishment.

4.4.3.2 ATP-grasp Protein Unit

The small domain comprises three or four β-strands and two or three α-helices and remains structurally conserved with an average SSAP score of 85. The ATP-grasp structures have three domains in common with the exception of succinyl-CoA synthetase (2scu) in which the third domain is a different fold. However among those with the same three domains, synapsin Ia comprises a much larger domain complex with five extra domains.

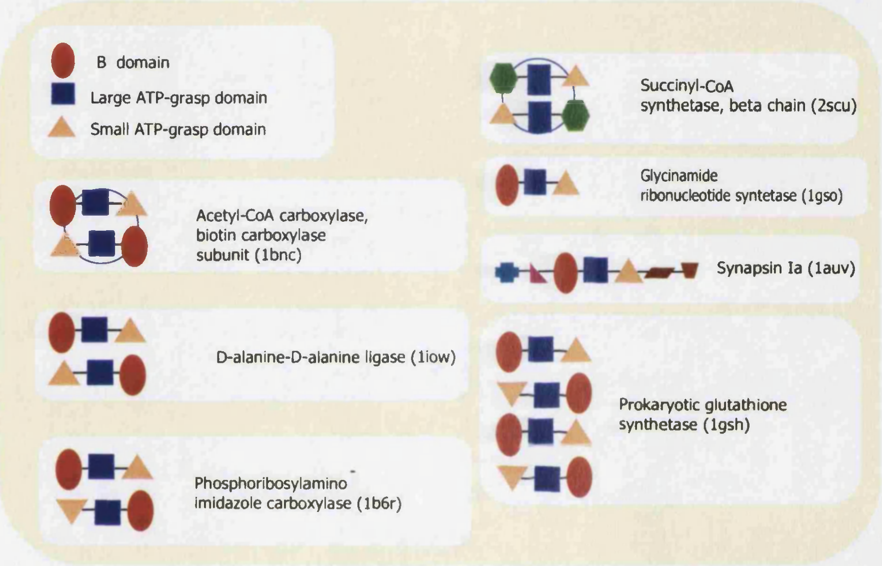

B domain
Large ATP-grasp domain
Small ATP-grasp domain
Acetyl-CoA carboxylase, biotin carboxylase subunit (1bnc)
D-alanine-D-alanine ligase (1iow)
Phosphoribosylamino imidazole carboxylase (1b6r)
Succinyl-CoA synthetase, beta chain (2scu)
Glycinamide ribonucleotide syntetase (1gso)
Synapsin Ia (1auv)
Prokaryotic glutathione synthetase (1gsh)

Figure 4.20: Domain and subunit interactions of members of the ATP-grasp superfamily. Glycinamide ribonucleotide synthetase and synapsin Ia are both monomeric. Each monomer containing the three consensus domains, the small and large ATP binding domain and the B domain, however synapsin Ia comprises five additional domains. The three common domains of biotin carboxylase dimerise as part of the larger acetyl-CoA carboxylase complex (shown as a blue ring). Phosphoribosylamino-imidazole carboxylase is a dimer, each monomer contains the common domains and prokaryotic glutathione synthetase is a tetramer. The ATP-binding domains of succinyl-CoA synthetase dimerise to form the β_2 unit of an $\alpha_2\beta_2$ hetero-oligomer. The two α domains are represented by a blue ring.

4.4.4 Functional Descriptions

Members of the ATP-dependent carboxylase-amine/thiol ligase protein superfamily bind Mg^{2+}ATP in the cleft between the small and large domains. This ATP binding function is conserved throughout the superfamily. However, the overall function of the whole protein varies due to the substrate bound, although the mechanism of the enzyme reaction is conserved throughout. All mechanisms involve the ATP-dependent ligation of a substrate carboxylate to an amine or thiol group of a second substrate, forming C-N or C-S bonds.

Collectively, the enzymes act on a vast array of donor and acceptor substrates (Table 4.3). In all reactions ATP is converted to ADP and inorganic phosphate.

4.4.4.1 The ATP Binding Site

Mg^{2+}ATP is bound in between two anti-parallel β-sheets, one from the small and one from the large domains (Artymiuk *et al.*, 1996). The ATP binding site residues were collated from the literature (where known). The ATP binding site is conserved throughout the family, binding with two Lys residues a Gln, and a Leu. Synapsin Ia is the only member which binds Ca^{2+} instead of Mg^{2+}. In this case studies have shown that ATP does not bind unless Ca^{2+} is present (Esser *et al.*, 1998). A conserved Lys has been identified as the critical residue for catalysis, and is thought to serve as a hydrogen bond donor (Wang *et al.*, 1998b).

4.4.4.2 The Substrate Binding Site

The substrate binding residues are located on the large domain (Figure 4.22). From the literature and the functional summary (Todd, 2001) very few residues involved in the binding of the substrate are described. Information on D-alanine D-alanine ligase (1iow) was collected from the literature (Fan *et al.*, 1994). D-alanine is orientated in the active site by the residues Tyr 216 and Ser 281 and reaction intermediates are stabilised by residues in the small and B domains.

4.4.5 Structure/Function

4.4.5.1 The Whole Chain

The structural similarity between the three common domains were measured using the SSAP structural alignment program and then the scores were clustered using the OC tree program (section 4.2.1.1). Clustering the members of this superfamily by whole chain has revealed two groups. Group one contains those structures with the C-terminal embellishment described in section 4.4.3 and group two contains those structures without. This suggests that the embellishment is the greatest structural change between members of this superfamily. The third domain of succinyl-CoA synthetase (2scu) is different, so this protein is considered separately.

Group I

The proteins clustered together in group one are phosphoribosylaminoimidazole carboxylase (1br6A), biotin carboxylase (1bncA) and glycinamide ribonucleotide synthetase

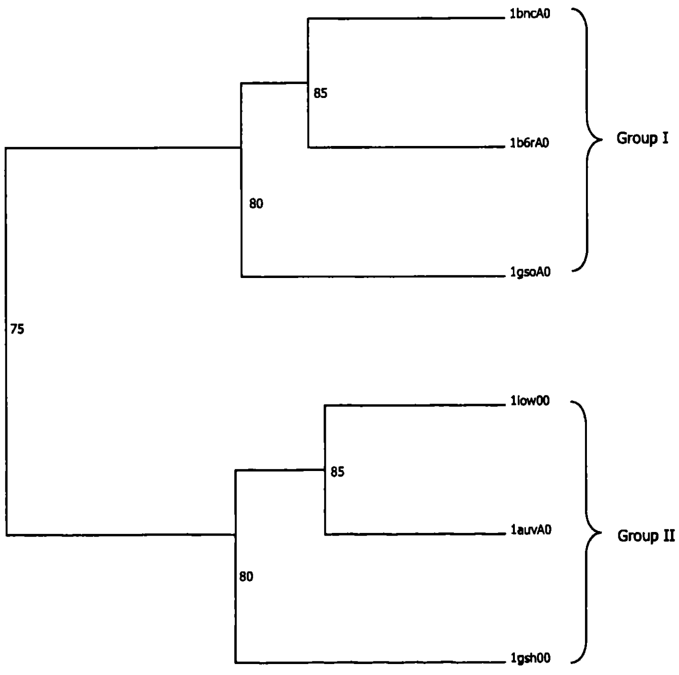
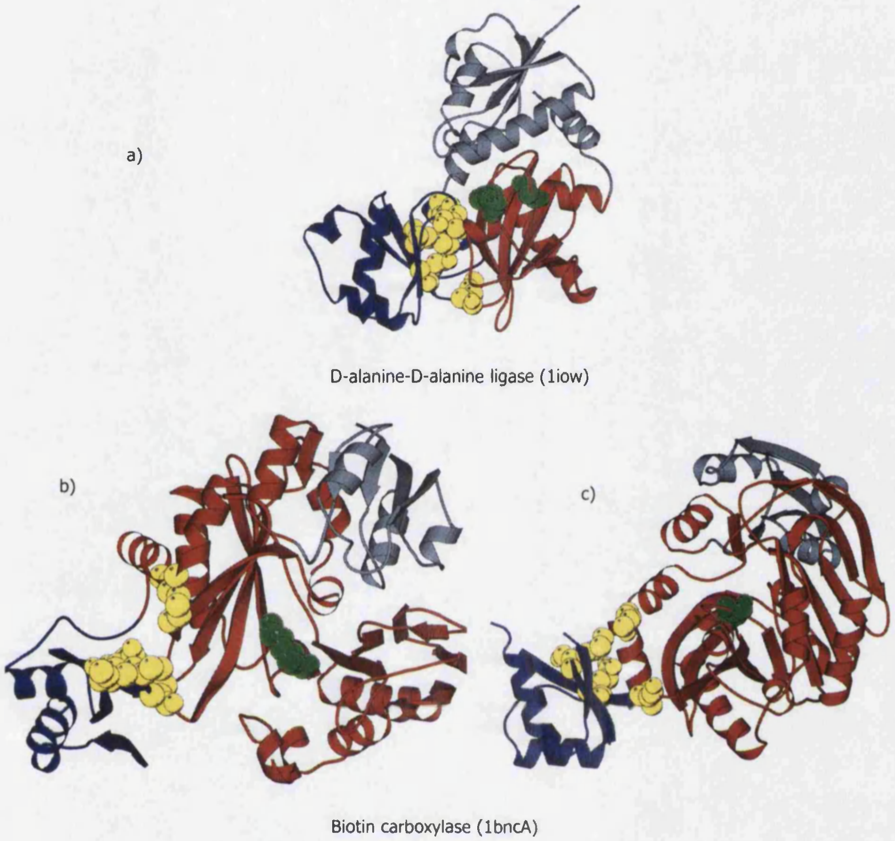


Figure 4.21: Members of the ATP-grasp superfamily clustered by the SSAP scores of the three common domains.

(1gsoA). All three have an extensive C-terminal embellishment. These secondary structures come together to form an extension to the β-sheet. This embellishment is so extensive that it has been described as an additional fourth domain in 1gso (Wang *et al.*, 1998b). Of these three members, two are carboxylases (1b6rA and 1bncA) and the other is a synthetase (1gso). The substrates, biotin for biotin carboxylase, phosphoribosylamine for glycinamide ribonucleotide synthetase and 5-aminoimidazole ribonucleotide for phosphoribosylaminoimidazole carboxylase are all small in size in comparison with the substrates in group two. In this group, the three domains are arranged so that the active site is enclosed in a box forming a narrow active site. This is represented by biotin carboxylase (1bncA), Figure 4.22.

Group II

The second group includes synapsin Ia (1auv), D-alanine-D-alanine ligase (1iow) and glutathione synthetase (1gsh) also known as L-glutamyl-L-cystiene:glycine ligase. The active site in these members is more accessible allowing for larger substrates to access the active site, peptidoglycans in D-alanine-D-alanine ligase and glutathione in glutathione synthetase. The absence of the embellishment leaves an L-shaped arrangement with the small and large domain forming the backbone and the B domain forming the base. The ATP cleft and the active site are more exposed compared to group I. This is represented by D-alanine-D-alanine ligase (1iow), Figure 4.22.



D-alanine-D-alanine ligase (1iow)

Biotin carboxylase (1bncA)

Figure 4.22: A figure showing the three domains of the ATP-grasp family. In red the large domain, in blue the small domain and in grey the B domain. The Molscript labelled as a. shows D-alanine D-alanine Ligase (1iow) in the L conformation. Molscripts b and c show biotin carboxylase. Molscript b illustrates the box conformation formed by the embellishment present in some of the members of this family and Molscript c shows biotin carboxylase in the same orientation as D-alanine D-alanine ligase. Residues shown in yellow are involved in ATP binding residues and the green residues represent those involved in substrate binding.

4.4.5.2 Succinyl-CoA synthetase (2scu)

Succinyl-CoA synthetase (2scu) is a $(\alpha\beta)2$-tetramer comprising two α domains and two β domains. There are two copies of each subunit, with each $\alpha\beta$ dimer coming together to form an asymmetric unit. The $\alpha\beta$ dimers are similar in structure, with each having a nucleotide binding motif. Conserved residues between D-alanine D-alanine ligase and Succinyl-CoA synthetase β subunit suggest ATP binding to be in the same place, in the cleft between the two ATP-grasp domains (Fraser *et al.*, 1999). Coenzyme A is bound to the nucleotide binding motif in the α subunit. The initial transfer of the phosphate group from ATP bound in the ATP-grasp cleft in the β subunit, to the carboxylate group of succinate is done *via* an intermediate His group located on the α chain (Fraser *et al.*, 1999).

4.4.6 The Role of the Embellishment in Subunit Aggregation

Both biotin carboxylase and phosphoribosylaminoimidazole carboxylase in Group I dimerise to form the biological unit. In biotin carboxylase (1bnc, Group I) the C-terminal embellishment is directly involved in the dimerisation interface (Artymiuk *et al.*, 1996) (Figure 4.23). Phosphoribosylaminoimidazole carboxylase (1b6r, Group I) is also a homodimer. The embellishment is also part of the dimerisation interface (Figure 4.24), however, the interface between the subunits are formed by interactions between the embellished secondary structures of the large domain to the consensus secondary structures of the large domain in the other subunit. Contacts are through the α-helices and not β-sheet to β-sheet as in biotin carboxylase. Glutathione synthetase (1gsh), in which the C-terminal embellishment is absent, forms a tetramer containing two interfaces, neither of which involve the large domain. Here, in contrast to biotin carboxylase and phosphoribosylaminoimidazole carboxylase, the interfaces are formed by the interaction between the small and B domains (Yamaguchi *et al.*, 1993). Additionally, the contacts between the α and β subunits of succinyl-CoA synthetase are between the loops at the top of the β-sheet (as the domain is orientated in Figure 4.19) of the large and B domain for the interactions between α and β subunits and the β subunits interact back to back through the α-helices of the small and B domains. Figure 4.25 shows that only a small proportion of the residues in the secondary structure embellishments are involved in the interactions with the other subunit.

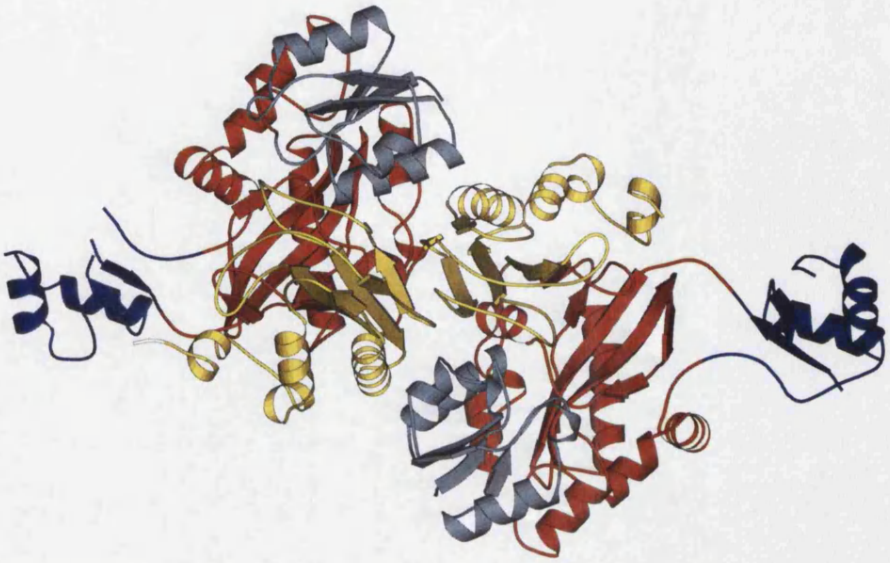


Figure 4.23: The biological unit of biotin carboxylase (1bnc, Group I) visualised using PQS. This diagram shows how the embellishment promotes the oligomerisation of the protein into its biological form. The small domain is shown in blue, the B domain in grey, the large domain consensus region in red and the embellishment in yellow.

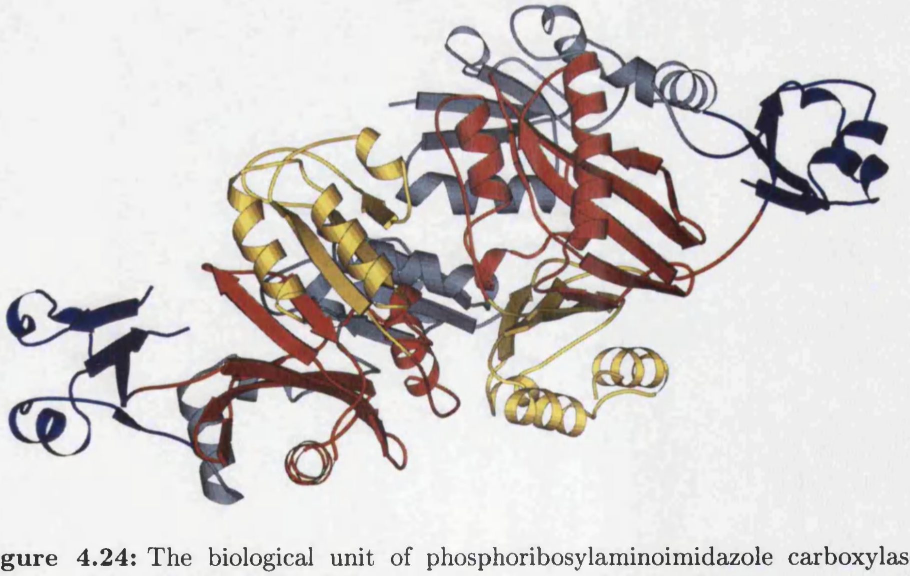


Figure 4.24: The biological unit of phosphoribosylaminoimidazole carboxylase (1b6r). The embellishment to the large domain, shown in yellow, interacts with the consensus structure of the large domain, shown in red. The small domain is shown in blue and the B domain is shown in grey.

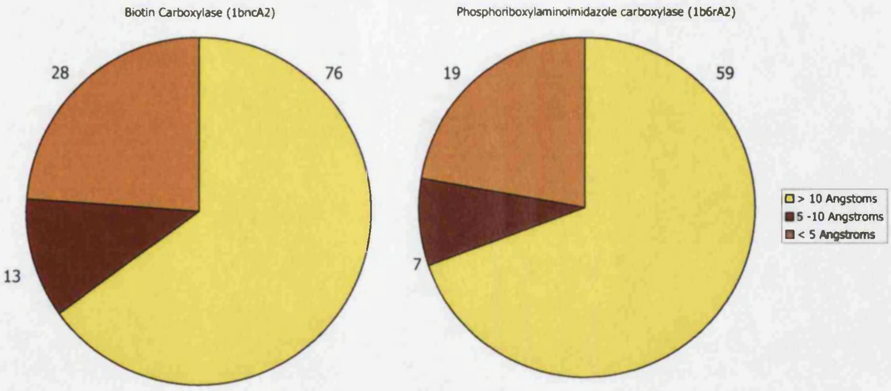


Figure 4.25: C-terminal embellishment interactions in biotin carboxylase and phosphoribosylaminoimidazole carboxylase in oligomerisation. In both cases, most of the residues in the embellished secondary structures are not involved in the interactions between subunits. Residues less than 5Å contact distance are shown in orange and correspond to those residues which are in direct contact with the other subunit. The proportion of residues between 5 - 10 Å are shown in brown and the proportion of residues >10Å from the other subunit are shown in yellow. In both cases just under three quarters of the residues are not involved in the subunit oligomerisation.

4.4.6.1 Conformational changes

Work by Thoden *et al.* (2000) has shown that there is a major conformational change in biotin carboxylase (1bnc) upon the addition of ATP. A rotation of approximately 45^o of one domain occurs relative to the other domains, thereby closing off the active site pocket.

4.4.7 ATP-Grasp Conclusions

Members of the ATP-grasp superfamily usually comprise three domains designated as the small, large and B domains. The small domain is highly conserved throughout the superfamily and the large domain varies in size with some of the domains in this superfamily possessing a C-terminal embellishment, which mainly comprises an extension to the main β-sheet. There are two binding sites within this superfamily, the ATP and the substrate binding sites. The ATP binding site is within a cleft formed by the small and large domains and this is very well conserved throughout the superfamily. Clustering of structural similarities revealed that members of the superfamily sharing three common domains can be divided into two groups, corresponding to the presence or absence of the large domain embellishment. Those proteins with the embellishment have a closed active site and tend to be carboxylases. Those proteins lacking the embellishment have an exposed active site and tend to be ligases.

4.5 Thioredoxin

4.5.1 Introduction

The thioredoxin superfamily contains a mixture of both enzymes and non-enzymes. The non-enzymes form single and multi domain structures, whilst the enzymes are multi domain. There are 13 thioredoxin-like S35Reps in CATH version 2.4 (Table 4.4). The proteins in the thioredoxin superfamily can be divided into four functional/structural clusters: single domain non-enzymes, multiple domain non-enzymes, multidomain enzymes consisting of thioredoxin-like domains only and multidomain enzymes consisting of thioredoxin-like and α-helical dimerisation domains.

Name	PDB	Enzyme/Non Enzyme	Function	# Subunits & Domains
Glutaredoxin	1kte00	Non-Enzyme	A protein reductant; essential for glutathione-dependent reduction of ribonucleotides by ribonucleotide reductase. Has the CXXC motif essential for redox activity. Reduction mechanism is analogous to that of thioredoxins. They preferentially reduce GSH-containing mixed disulphides due to the GSH binding site on the protein. (Katti *et al.*, 1995)	Monomeric. Single domain.
Thioredoxin	2trxA0	Non-Enzyme	A protein reductant; regulation of enzyme activity by disulphide activity. Has the CXXC motif essential for redox activity. The oxidised form contains a disulphide bridge formed by the Cys residue in the highly conserved CGPC motif and this is reduced to a dithiol by NADPH and the flavoprotein, thioredoxin reductase. (Katti *et al.*, 1990)	Monomeric. Single domain.
Peroxidase hORF6	1prxA1	Enzyme	Regulation of the intracellular concentration of H_20_2. Does not require selenium but a single Cys residue which is structurally equivalent to the less accessible C-terminal Cys residue in the CXXC motif in thioredoxin. (Choi *et al.*, 1998)	Homodimer. Each monomer has two domains, one thioredoxin-like and one alpha helical domain.
Thioredoxin peroxidase	1qq2A0	Enzyme	Unknown	Homodimer. Each monomer has one thioredoxin-like domain.
Protein disulphide oxidoreductase	1a8l01 1a8l02	Enzyme	Probably the Archean counterpart of protein disulphide isomerase which catalyses protein disulphide formation and breakage during protein folding. Has the CXXC motif essential for redox activity. (Ren *et al.*, 1998)	Monomer. Two thioredoxin-like domains.
Glutathione peroxidase	1gp1A0	Enzyme	Protection of cellular components from oxidative damage. Catalyses the reduction of hydrogen peroxide and a variety of organic hydro-peroxides (ROOH) to water or the corresponding alcohol (ROH) using GSH as the reducing substrate. The active site contains a selenocysteine residue which probably shuttles between a selenolate anion (RSe-)and a selenic acid (Rse-OH) in the catalytic cycle. (Ren *et al.*, 1997)	Homotetramer. One thioredoxin-like domain in each monomer.
Thiol-disulphide interchange protein (DbsA)	1fvkA0	Enzyme	Periplasmic protein oxidant. Has the CXXC motif essential for redox activity. (Guddat *et al.*, 1998)	Dimer. Two domains in each monomer, one thioredoxin-like and one α-helical.

continued on next page

Table 4.4: *continued*

Name	PDB	Enzyme/Non Enzyme	Function	# Subunits & Domains
Glutathione S-transferase	1gseA1	Enzyme	Detoxification of compounds. They catalyse the S-conjugation between the thiol group of GSH and an electrophilic moiety in the hydrophobic and toxic substrate. (Sinning *et al.*, 1993)	Homodimer. Two domains in each monomer, one thioredoxin-like and one α-helical.
Calsequestrin	1a8y01 1a8y02 1a8y03	Non-enzyme	Calcium storage in muscle. (Wang *et al.*, 1998a)	Polymer. Three thioredoxin-like domains form the monomer.
Phosducin	2trcP1	Non-enzyme	Regulation of phototransduction (dark/light adaptation). (Gaudet *et al.*, 1999)	Hetero-oligomer. Monomer comprises a thioredoxin-like domain and an α-helical domain. Complexed with β and α subunits of transducin.

Table 4.4: The S35Reps from the Thioredoxin superfamily. Listed in the table are the names, CATH numbers, PDB codes and chain and domain identifiers, whether the protein has an enzymatic or a non-enzymatic function, a more detailed function of the protein and a description of the biological unit.

4.5.2 Structural Description

Domains of this superfamily are three-layer $\alpha\beta\alpha$ sandwiches. All members have a minimum of four mixed (parallel and anti-parallel) β-strands in order 4312 with respect to the sequential assignment of the β-strands. The third strand is anti-parallel to the rest. The β-sheet is typically flanked by three α-helices (Martin, 1995).

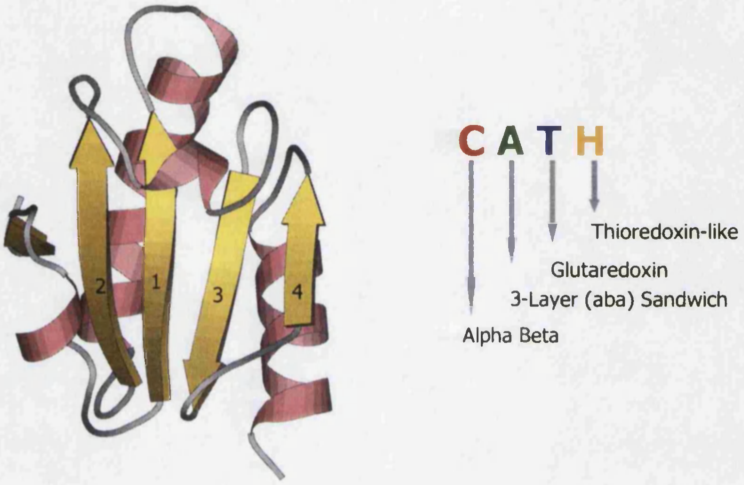


Figure 4.26: The most representative structure of the thioredoxin-like superfamily, thioredoxin (2trx). The four consensus β-strands are numbered sequentially. The third strand is antiparallel to the rest.

4.5.3 Structural Variation

General structural variability between members of the superfamily was measured using the SSAP structural alignment program. The average score is 78.91 with an associated standard deviation of 4.80 and the lowest pairwise SSAP is 67.36 between 1gseA1 and 1prxA1. The most representative structure is 2trxA0 with ten pairwise scores of above 80. The smallest has seven secondary structures, four helices and three strands (1a8l01) and the largest has twelve with five helices and seven strands (1qq2A0).

The thioredoxin-like fold has been embellished at two main points in the structure (Figure 4.29). These embellishments occur on the right of the central β-sheet as it is orientated in Figure 4.28 and between β-strands 3 and 5 which occur at the top of the structure. Embellishments are present at the N-terminus and after the second consensus strand. Variations occur in the number of β-strands forming the β-sheet and in the number of α-helices. In addition to the variation in number of α-helices, visual inspection indicates that there is also a great deal of variation in their orientation.

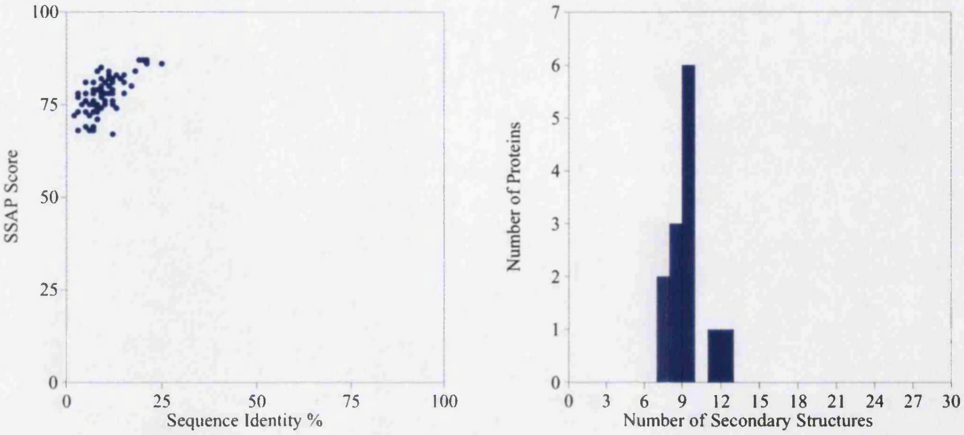
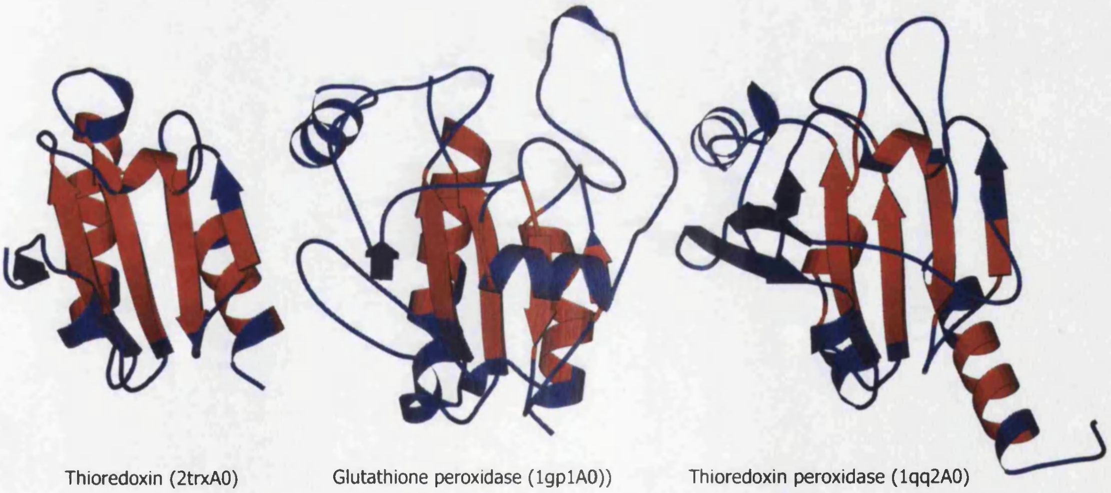


Figure 4.27: The relationship between change in sequence and structure is shown here for the S35Reps in the thioredoxin-like superfamily. The average SSAP score is 78.91 with associated standard deviation of 4.80. The smallest domain in this superfamily contains seven (1a8l01) secondary structures which is about half the size of the largest with 12 (1qq2A0).



Thioredoxin (2trxA0) Glutathione peroxidase (1gp1A0)) Thioredoxin peroxidase (1qq2A0)

Figure 4.28: Three domains belonging to the thioredoxin-like superfamily showing the extent of the embellishments. Thioredoxin is the representative fold for this superfamily having only one extra strand on the left of the β-sheet as it is orientated. Glutathione peroxidase and thioredoxin peroxidase show embellishments to the left and the top of the structure as it is orientated. The position of the active site in members of this superfamily is in the loops at the top of the β-sheet.

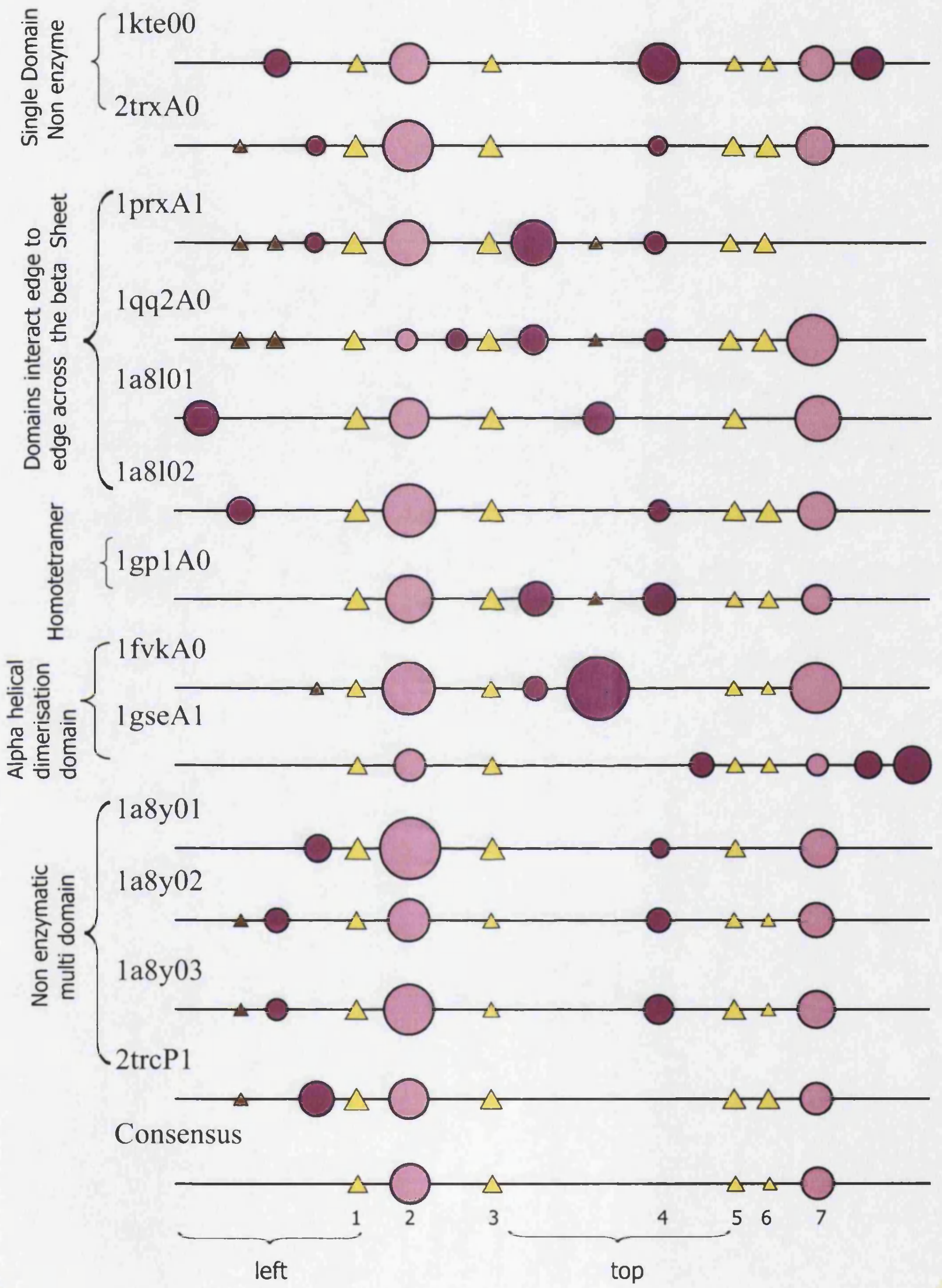


Figure 4.29: 2DSEC plot for the S35Reps of the thioredoxin-like fold superfamily. Consensus strands and helices have been numbered consecutively and the two main areas where embellishments are present are marked as the left embellishment as it is located on the left side of the structure as it is orientated in Figure 4.28 and the top embellishment as it is located at the top of the domain. This is also the region of the active site.

4.5.4 Functional Descriptions

The active site in all members is in the same place on the thioredoxin-like structure (Martin, 1995; Todd, 2001), at one end of the core β-sheet (Figure 4.32). The environment of the active site depends greatly on the orientation of the domains and subunits creating the full biological unit. In some proteins the single domain exists on its own, whilst in others the structural domain is repeated so that two thioredoxin-like domains exist on the same chain. Additionally, in some members the thioredoxin-like domains oligomerise to make a dimer, homotetramer or a polymer. With some of the members the thioredoxin domain is fused with an α-helical domain which functions as a dimerisation domain. A summary of these domain interactions is shown in Figure 4.30. Individual functions for members of this superfamily are listed in Table 4.4.

The redox active members of this superfamily involve sulphur redox chemistry. All must stabilise a cysteine thiolate (or selenolate) ion during catalysis. In some members this cysteine thiol group is intra-molecular whilst in others it binds as an external substrate. Many members contain a conserved glutathione (GSH)-binding site, where GSH is a tripeptide formed by Glu, Cys and Gly. The redox active site is found as a CXXC motif present in thioredoxin and glutaredoxin. In protein disulphide oxidoreductase and DbsA the solvent accessible N-terminal Cys of the CXXC motif forms a mixed disulphide bridge with the substrate. Glutathione peroxidase requires a selenium atom for its redox activity whereas other peroxiredoxin enzymes use a single Cys residue. Calsequestrin and phosducin are not redox active.

4.5.5 Structure/Function

The orientation of the domains upon oligomerisation is extremely variable in members of this family, making the structural environment of each active site very different, even though it remains in the same place in each thioredoxin-like domain (Figure 4.30).

In many cases the embellishments to the thioredoxin-like fold appear to promote the interactions between the domains and subunits. In the previous examples in this chapter, clustering of the members of the superfamily was carried out by SSAP score using the OC program (sections 4.3.5 & 4.4.5). However, in this example, clustering this way did not elucidate any structural/functional relationships due to extensive changes in the orientations of the secondary structures and changes in the loops. Therefore, in order to examine the relationship between structure and function in this superfamily, the proteins were categorised into five groups according to their quaternary structures: single domain non enzymes, members in which the thioredoxin-like domain interface is across the β-sheet either formed by two subunits or two domains on the same chain, homotetramers,

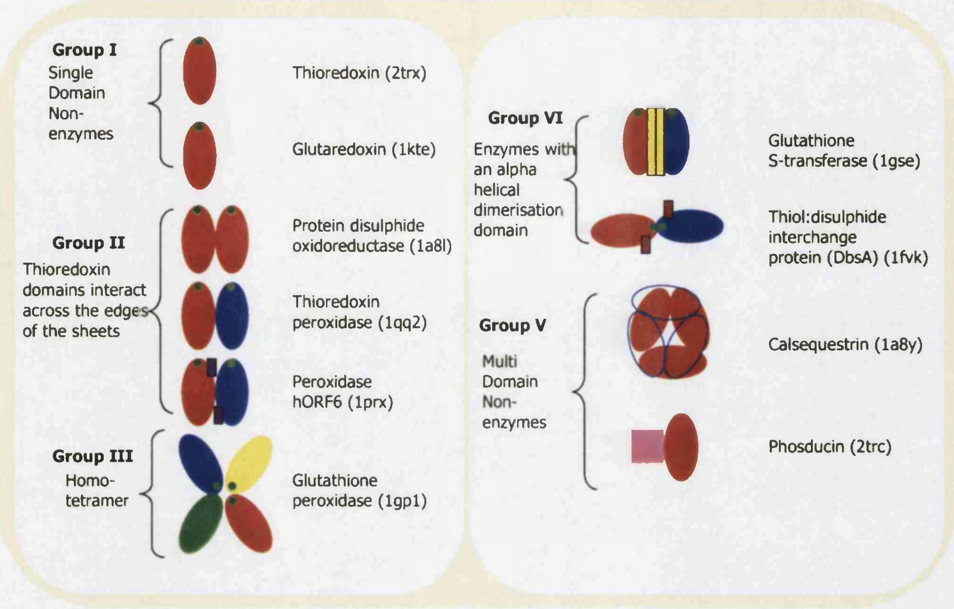

Group I
Single Domain Non-enzymes
Thioredoxin (2trx)
Glutaredoxin (1kte)
Group II
Thioredoxin domains interact across the edges of the sheets
Protein disulphide oxidoreductase (1a8l)
Thioredoxin peroxidase (1qq2)
Peroxidase hORF6 (1prx)
Group III
Homo-tetramer
Glutathione peroxidase (1gp1)
Group VI
Enzymes with an alpha helical dimerisation domain
Glutathione S-transferase (1gse)
Thiol:disulphide interchange protein (DbsA) (1fvk)
Group V
Multi Domain Non-enzymes
Calsequestrin (1a8y)
Phosducin (2trc)

Figure 4.30: The quaternary structures of the proteins in the thioredoxin-like superfamily are varied. In this figure, thioredoxin domains are shown as ovals. Red and blue indicate separate chains and a green dot indicates the position of the active site. Thioredoxin and glutaredoxin are single domain non-enzymes. In protein disulphide oxidoreductase, thioredoxin peroxidase and peroxidase hORF6 two thioredoxin-like units interact edge to edge to form a long β-sheet. In protein disulphide oxidoreductase the domains are fused whilst in thioredoxin peroxidase and peroxidase hORF6 they are located on separate chains (Figure 4.32). All three are stabilised by interactions between the last consensus α-helix. However, in peroxidase hORF6 this last α-helix is part of a separate dimerisation domain shown here in purple. Protein disulphide oxidoreductase has a single chain whilst thioredoxin peroxidase and peroxidase hORF form the same orientation from different subunits. Glutathione peroxidase is a homotetramer comprising four thioredoxin-like subunits, the orientation of the domains is completely different from those which dimerise across the central β-sheet. The thioredoxin-like domains of glutathione S-transferase and thiol:disulphide interchange protein dimerise with the aid of a separate α-helical domain, minimising the contact between the thioredoxin-like domain monomers. Calsequestrin and phosducin are both non-enzymatic, multi domain complexes. In calsequestrin, each chain comprises three domains which stack onto each other. The packing of two subunits are shown here in red (underneath) and blue (on top).

including an α-helical dimerisation domain and non-enzyme multidomain. The few embellishments on the two single domain non-enzyme domains thioredoxin and glutaredoxin do not interfere with the active site and since very little is known about their protein interactions they are not considered in this section.

4.5.5.1 Group II: Peroxidase hORF6, Thioredoxin Peroxidase and Protein Disulphide Oxidoreductase. Proteins in which the edges of the β-sheets of the two thioredoxin domains interact.

Peroxidase hORF6 (1prx) and thioredoxin peroxidase (1qq2) are both homodimers and their dimerisation interfaces are located in the same place. The α-helical and β-strand embellishments do not promote the domain-domain or subunit-subunit interactions (Figure 4.33). As illustrated by the pie chart in Figure 4.34 and by the cartoon in Figure 4.37. There are no embellished residues less than 5Å from the subunit interface. The dimer interface in both these proteins is stabilised by a long consensus α-helix. In thioredoxin peroxidase it is the second consensus helix but in peroxidase hORF6 this helix belongs to a separate four-stranded $\alpha\beta$-2-layer sandwich domain. In this case, additional embellishments forming a small β-sheet and an extra α-helix to the C-terminal end of the peroxidase hORF6 have caused the formation of a new domain. The 2DSEC diagram (Figure 4.29) of the thioredoxin-like fold shows the dimerisation helix present at the C-terminal end of thioredoxin peroxidase (1qq2A0) but no equivalent secondary structures in peroxidase hORF6 as they have been assigned to another domain. As described before, the active site is situated in the loops of the thioredoxin fold. In peroxidase hORF6 the active site is a narrow pocket (Choi *et al.*, 1998) created by some of the residues in domain 2, the dimerisation domain. Residues in the entrance of the active site create a narrow and positively charged environment suggesting that dimerisation is important for activation. In thioredoxin peroxidase, this active site is more open (Hirotsu *et al.*, 1999) principally because it lacks the β-sheet which is added by the separate dimerisation domain in peroxidase hORF6 (Figure 4.32).

Protein disulphide oxidoreductase (1a8l) consists of two thioredoxin domains which are the result of an ancient domain repeat (Ren *et al.*, 1998). The interface between the two domains is so extensive that Ren *et al.* (1998) describes it as a single domain protein. There are two CXXC motifs situated in the loops of the thioredoxin-like fold. However, the N-terminal domain is made up of CQYC and the C-terminal of CPYC. The axis between the domains is mediated between domain 1, the consensus α-helix 3 and consensus α-helix 1 in domain 2 and consensus β-strands 4 in domain 1 and consensus β-strand 2 in domain 2. The second embellished helix in domain 2 interacts with a loop in domain 1 which can be seen in Figure 4.34.

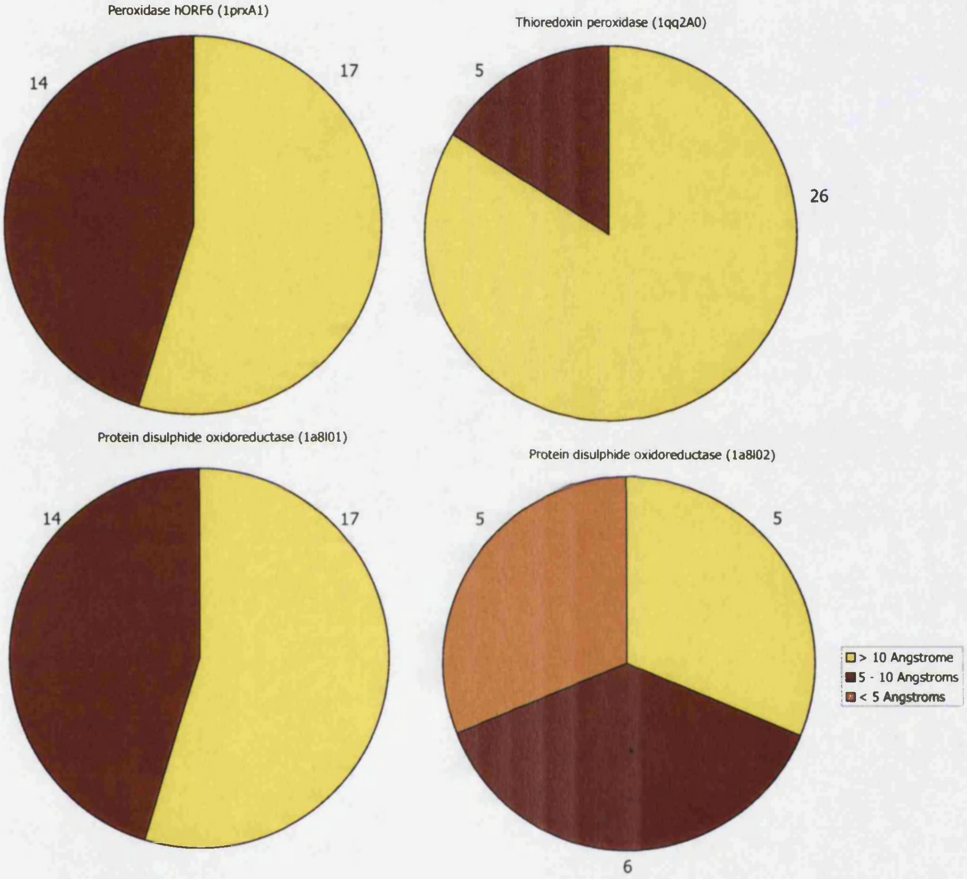


Figure 4.31: Interactions between the embellished secondary structures and other domains and subunits in the thioredoxin-like superfamily. The charts show that no residues in peroxidase hORF6 (1prxA1), thioredoxin peroxidase (1qq2A0) and protein disulphide oxidoreductase (1a8l01) have any embellished residues within 5 Å from the other domains and subunits. However, the chart for protein disulphide oxidoreductase (1a8l02) illustrates a short range interaction and this involves the second embellished α-helix and the N-terminal thioredoxin domain.

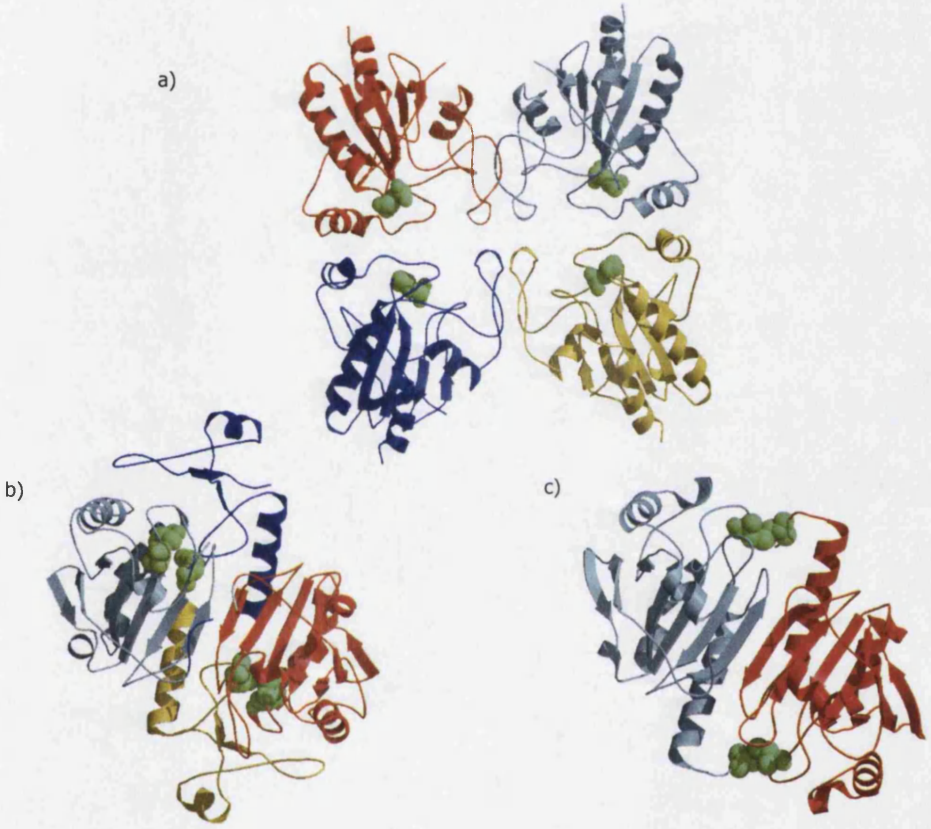
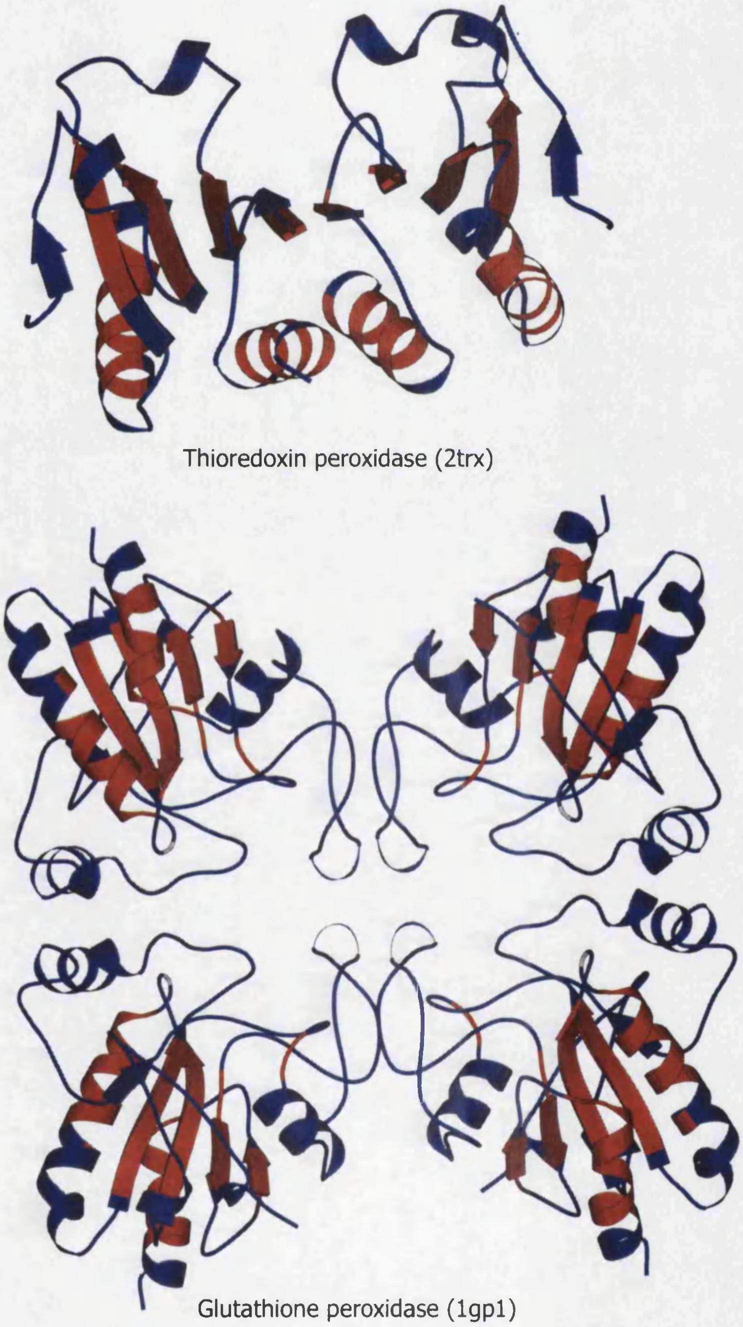


Figure 4.32: The biological units of three members of the thioredoxin superfamily. Glutathione peroxidase (a) is a tetramer. The domain interactions are mainly from the first embellished helix and the loops between the first and consensus β-strand and α-helix. Peroxidase hORF6 and thioredoxin peroxidase form dimers. In peroxidase hORF6 (b) the last consensus helix forms another domain with four strands and another helix. The two thioredoxin domains in peroxidase hORF6 are shown in grey and red and the two dimerisation domains are shown in dark blue and yellow. These two dimerisation domains also enclose the active site shown in green. In thioredoxin peroxidase (c) the dimerisation interface is stabilised by an α-helix which is equivalent in structural orientation to the long α-helix in the dimerisation domain of peroxidase hORF6.



Thioredoxin peroxidase (2trx)

Glutathione peroxidase (1gp1)

Figure 4.33: Thioredoxin peroxidase and gluathione peroxidase coloured according to core and structurally variable regions.

4.5.5.2 Group III: Glutathione Peroxidase. The homotetramer

The biological unit of glutathione peroxidase (1gpl) is a tetramer (Figure 4.32). The tetramerisation interfaces are completely different from the dimerisation interfaces described previously. The interfaces between the four subunits are formed mainly from the first embellished helix and the loops between the first consensus β-strand and α-helix, between the second consensus strand and the first embellished helix and the second embellished helix and the third consensus strand (Figure 4.33). Upon tetramerisation, a pocket is formed on the surface of the protein. The catalytic, modified cysteine (Cso45) is situated near the interface of the two subunits. Four active sites are present, two in the pockets on one side and two in the pockets on the other. These pockets are formed by the interacting domains which means that the quaternary structure is very important for the activity of the protein. In this example it is apparent that the embellished α-helices promote the domain interfaces together with loops (Figure 4.37).

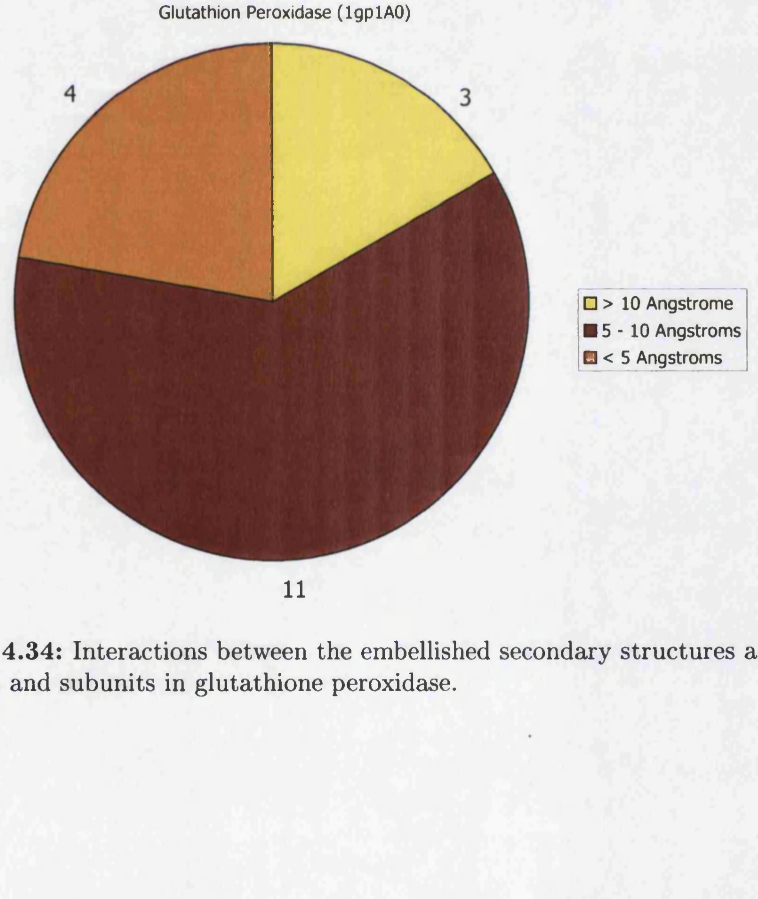


Figure 4.34: Interactions between the embellished secondary structures and other domains and subunits in glutathione peroxidase.

4.5.5.3 Group IV: Human Alpha Glutathione Transferase and Thiol-disulphide interchange protein (DbsA). Containing an α-helical dimerisation domain.

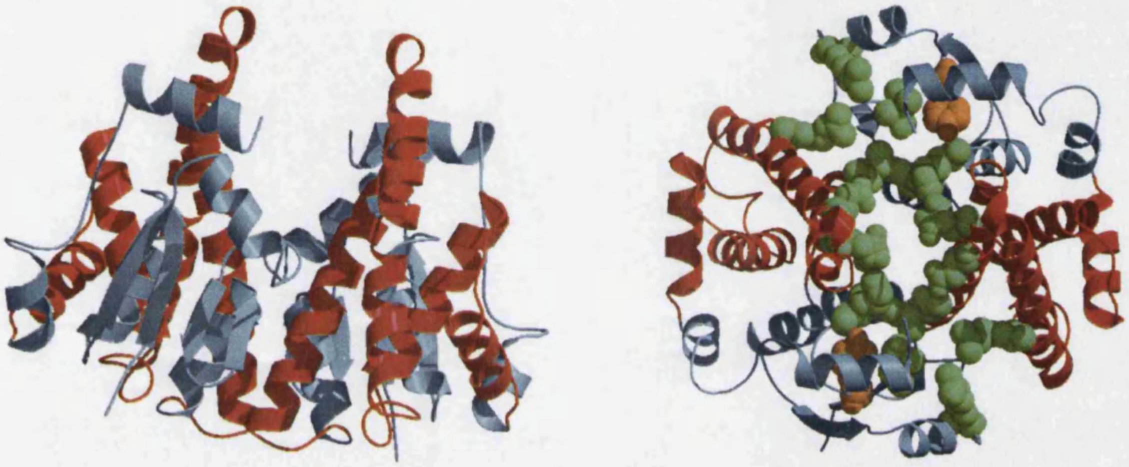


Figure 4.35: The domain interactions involved in human alpha glutathione transferase. The structure forms a large V-shape crevice about 40 Å long. Binding sites for glutathione and substrate are created, mainly from the loops of the thioredoxin-like domain (shown in grey).

Human alpha glutathione transferase forms a homodimer with the monomeric unit consisting of two domains; the thioredoxin-like domain and an alpha helical up-down bundle domain inserted after the last consensus helix and before the last two embellished helices. In the three-dimensional structure it is packed against consensus helices 1 and 3. The two final embellished helices pack against the thioredoxin-like structure. The quaternary structure is made up by extensive sidechain interactions between the thioredoxin domain of one monomer and the up-down α-bundle of the other (Sinning *et al.*, 1993) (Figure 4.37). The structure forms a large V-shape crevice about 40 Å long. Binding sites for glutathione and substrate are created, mainly from the loops of the thioredoxin-like domain (shown in grey). This active site can be blocked by the movement of the first embellished α-helix (located on the left embellishment as described in Figure 4.29) and the last (located in the top embellishment as described in Figure 4.29) embellished α-helices which play a part in modifying the active site, activating and deactivating the protein. In this protein, the interfaces between thioredoxin-like domains forming the quaternary structure are mediated by the consensus helices.

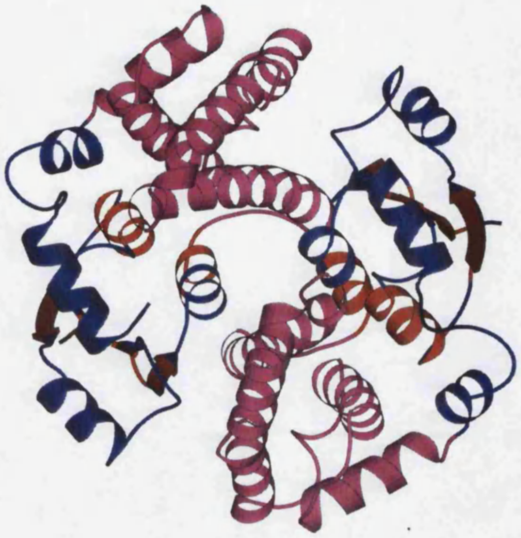


Figure 4.36: The two thioredoxin-like domains of glutathione synthetase dimerise with the help of two α-helical domains and the role of these domains is analagous to that played by the secondary structure elements in the cupredoxin family.

Thiol:disulphide interchange protein DsbA (1fvk), another dimeric protein, also contains an α-helical domain as well as the thioredoxin-like domain. Interaction between the alpha helical domain and the thioredoxin-like domain are mediated by consensus helix 1 (Figure 4.29) in the thioredoxin-like domain and the two embellished helices in the embellishment located at the top of the thioredoxin β-sheet. It is debatable as to whether these two embellished helices are part of the α-helical domain or the thioredoxin-like domain, or if in fact, the α-helical domain can be described as a large embellishment to the thioredoxin domain. As with glutathione transferase, this α-helical domain mediates the dimer interaction. However, these α-helical domains are inserted into the thioredoxin-like domain at two different points. The α-helical domain of glutathione transferase is inserted between the last consensus α-helix and the two C-terminal α-helices and in DsbA, the α-helical domain is inserted after the second consensus β-strand (Guddat *et al.*, 1998) (Figure 4.29). Both dimerisation domains could be a result of a build up of α-helices, promoting and stabilising the formation of the dimerisation of the proteins. This suggests a different mechanism for the formation of multi domain proteins from the domain recruitment mechanism suggested for the evolution of most proteins.

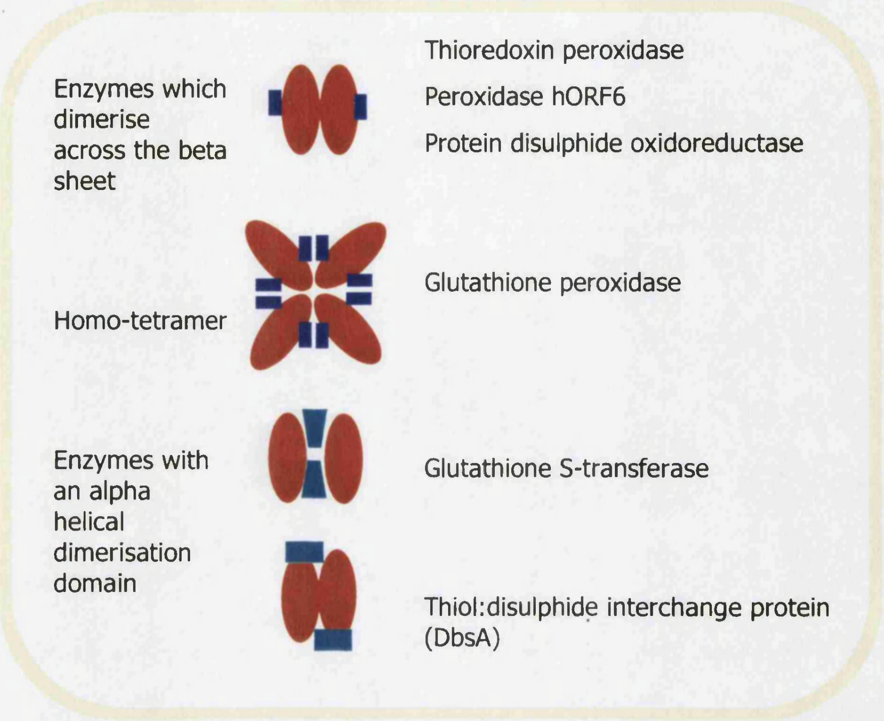

Enzymes which dimerise across the beta sheet
Thioredoxin peroxidase
Peroxidase hORF6
Protein disulphide oxidoreductase
Homo-tetramer
Glutathione peroxidase
Enzymes with an alpha helical dimerisation domain
Glutathione S-transferase
Thiol:disulphide interchange protein (DbsA)

Figure 4.37: Domain orientations and positions of the embellishments in the thioredoxin superfamily. The orientations of the domains and the positions of the embellishments are shown for 6 members of the thioredoxin family. In thioredoxin peroxidase, peroxidase hORF6 and protein disulphide oxidoreductase the embellishments (shown in blue) are located on the other side of the β-sheet to the domain interface, but for glutathione peroxidase, the embellishments interact in the homotetramer. For both glutathione transferase and thiol:disulphide interchange protein the additional α-helical domain (shown in light blue) plays an analagous role to the embellishments in glutathione peroxidase and the multi copper oxidases in the cupredoxin family by promoting the domain or subunit interactions.

4.5.5.4 Group V: Calsequestrin and Phosducin. Non-enzyme multi domain proteins

Calsequestrin, a polymer, is the major storage protein for muscle and is used as a Ca^{2+} buffer. The monomer comprises three negatively charged thioredoxin-like domains. In the quaternary structure, the three domains form a disk-like structure and these are packed on top of each other. Calcium is not bound in a distinct binding site like the EF hand. Instead, pairs of acidic residues bind Ca^{2+} through the net charge density.

In the quaternary structure, each monomer makes two extensive dimerisation contacts, a back-to-back contact and a front-to-front contact (Figure 4.38). Both sets of contacts form electronegative pockets which are the suggested binding sites for Ca^{2+}. By correlating the information from the literature (Wang *et al.*, 1998a) with the 2DSEC diagram it is possible to see that the front-to-front pocket is mediated by residues in the first consensus helix of domain 1 and stabilised by the second embellished helix in domain 3. The highly extended N-terminal region of one monomer becomes inserted into a hydrophobic cleft of the other (Figure 4.38). The back-to-back interface is stabilised through intermolecular interactions between α-helix 4 of domain 2 and α-helix 3 of domain 1. Strong interactions are formed by three intermolecular salt bridges between Glu 215 and Lys 86, Glu 216 and Lys 24, Glu 169 and Lys 85. These residues are located on α-helix 1 and 3 on domain 1 and α-helix 4 of domain 4. The strand and helix embellishment at the N-terminus of domains 2 and 3 form interfaces with domains 1 and 2 respectively. Domain 1 does not have this interface and does not have the embellishment (Figure 4.29).

Phosducin consists of two domains, the thioredoxin-like domain and an α-helical domain which form a complex with the β-subunit of transducin. The two domains are described as completely independent (Gaudet *et al.*, 1999) and no active site residues are recorded for the thioredoxin-like domain. Gaudet *et al.* (1999) suggests that the orientation of the thioredoxin-like and the α-helical domain triggers the activation of the protein but no other information is available.

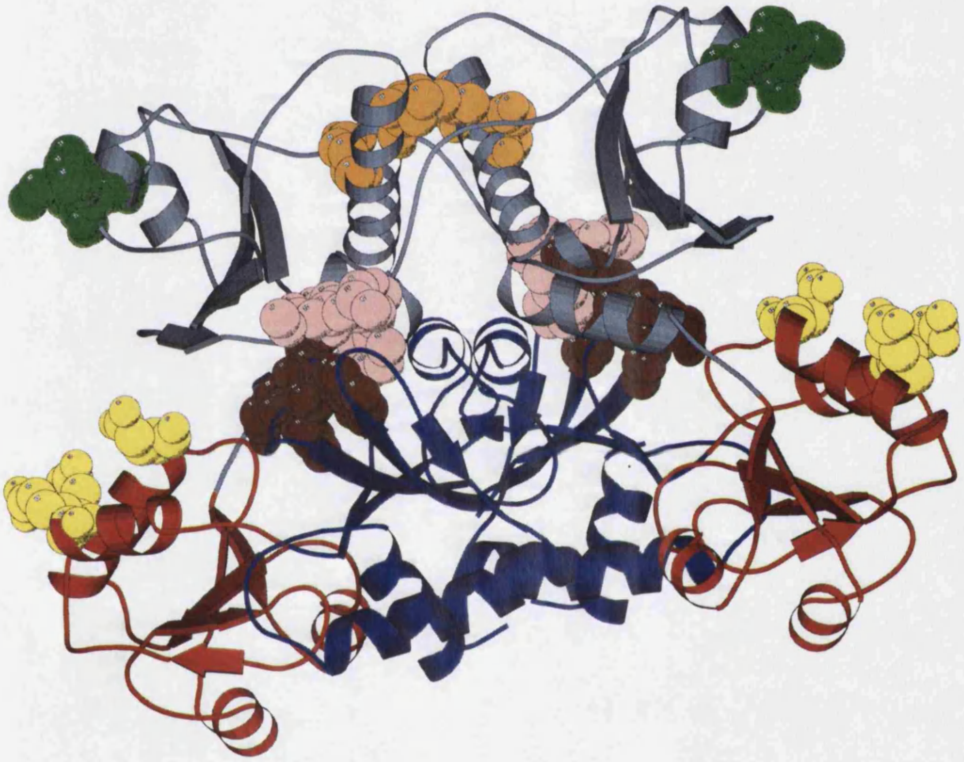


Figure 4.38: Calsequestrin is shown here interacting front-to-front. Domain 1 is shown in light blue, domain 2 in red and domain 3 in dark blue. The front-to-front interaction is stabilised by residues Lys 49 and Glu 55 shown here in orange, on consensus helix 1 in domain 1 and assisted by N-terminal residues and embellished helix 2 of domain 3 by salt bridges and hydrophobic interactions. The back-to-back domain interactions are mediated by salt bridges made from residues shown here in yellow and green. These interactions are present on the two embellished helices on domain 1 and the second consensus helix in domain 2.

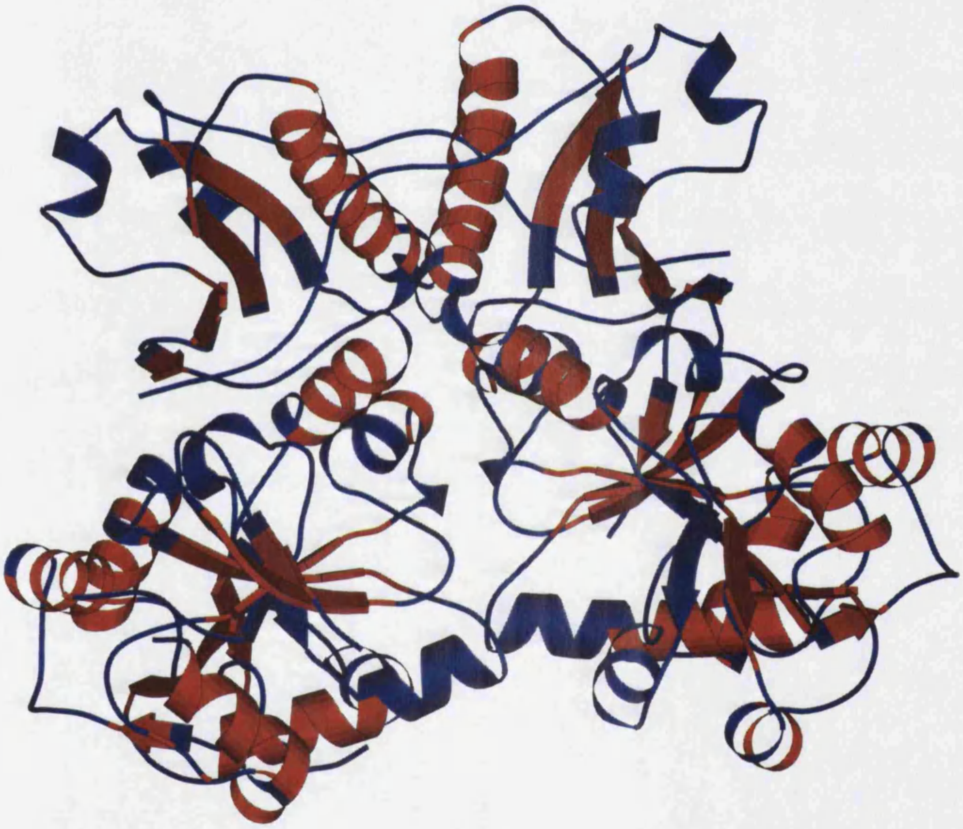


Figure 4.39: The biological unit of calsequestrin with consensus secondary structures coloured in red and structurally variable regions coloured in blue. The diagram shows that the embellished secondary structures play a role in promoting the domain interactions and subunit oligomerisation.

4.5.6 Thioredoxin Conclusions

This superfamily shows a number of different oligomerisation states in which the embellishments promote the domain interactions in only some cases. In group II, the three examples in which two thioredoxin domains interact across the edges of the β-sheets, the interactions are promoted by the conserved secondary structures. Extensions to the β-sheet occur on the opposite side of the β-sheet. However, a C-terminal embellishment in peroxidase hORF has caused the definition of another domain, with the boundary defined before the C-terminal consensus α-helix. The β-strands in this separate domain enclose the active site of the protein.

Glutathione peroxidase forms a homotetramer from four thioredoxin domains. In this example, the embellished secondary structures promote the tetramer formation. Similarly, the interactions between the domains of calsequestrin are mediated by the embellished secondary structures.

The dimerisation of two proteins in this superfamily, glutathione peroxidase and thiol-disulphide interchange protein (DbsA), is promoted by an extra α-helical domain. The domains are inserted into different places on the peptide chain. For thiol disulphide interchange protein the α-helical domain is inserted into an embellished region located at the top of the thioredoxin fold (Figure 4.29). The α-helical domain is described in the literature as a separate domain (Guddat *et al.*, 1998) but the domain is not divided in CATH and could represent a large embellishment which promotes the interaction of the two thioredoxin domains. In glutathione synthetase (1gse) the α-helical domain is fused to the C-terminal end of the thioredoxin fold and is classified in CATH as an up-down bundle. It is possible that these two examples present two different mechanisms for domain formation; gradual secondary structure accumulation and the more represented method of domain recruitment.

4.6 Conclusions

In this chapter the functional consequences of the secondary structural embellishments of three superfamilies have been studied. The secondary structural embellishments have been distinguished from the core secondary structural elements using a multiple structural alignment generated by CORA and the 2DSEC analysis suite. Once identified, the three-dimensional structures of these embellishments were examined with respect to the position of the active site and the interactions between the domains forming the biological unit. It was discovered that in all three superfamilies the identified embellishments played a role in the mediation of domain or subunit interactions. In the cupredoxin superfamily, the proteins range from single domain non-enzymes with one copper binding site to multi-cupredoxin-domain enzymes with additional copper binding sites situated in the clefts between the domains. These more complex proteins contain additional secondary structures embellishing the consensus structure on both sides of the β-sheets. These embellishments correspond exactly to the regions where the domains interact to form the biologically active unit, indirectly causing the formation of the new copper binding sites in the clefts between domains. The large C-terminal embellishment present in the ATP-grasp superfamily is also shown to be important in the stabilisation of the two subunits of biotin carboxylase and phosphoribosylaminoimidazole carboxylase. Additionally, the embellishments were also shown to be important in the domain and subunit interactions in the complex and varied domain and subunit assemblies present in the thioredoxin superfamily.

As well as assisting the formation of the quaternary structures of the proteins in these three superfamilies, some of the embellishments have a direct effect on the geometry of the active site. In the ATP-grasp superfamily the large C-terminal embellishment in two members, biotin carboxylase and phosphoribosylaminoimidazole carboxylase, encloses the active site. The substrates in these two proteins are small. In other members such as D-alanine-D-alanine ligase the absence of the embellishment creates a more open active site. Members of this subset are mainly ligases where larger substrates are joined together. The active site must be more open to allow access for these substrates. In the thioredoxin superfamily, modification of active site is observed in thioredoxin peroxidase and peroxidase hORF6. A C-terminal embellishment to peroxidase hORF6 is characterised as a separate domain and is involved in capping the active site. Other modifications to the active site can be seen with glutathione S-transferase, where the active site is blocked by an embellished α-helix.

The formation of the biological unit in proteins of these three superfamilies can also be considered a modification of the active site. The formation of multi-domain cupredoxin proteins caused the formation of a second site for copper binding, transforming members of this superfamily from non-enzymatic electron transfer proteins to enzymes with a variety of functions. Formation of the active site from the quaternary structure is also apparent in the thioredoxin superfamily.

The thioredoxin superfamily provides an interesting example of the difficulty in defining an embellishment and an extra domain. Both glutathione S-transferase and thiol-disulphide interchange protein have α-helical domains which function as dimerisation domains. The alpha helical domain in thiol-disulphide interchange protein is inserted into the thioredoxin domain at the site of the second embellishment, the embellishment which lies at the top of the thioredoxin β-sheet as it is labelled in Figure 4.29. This could also be considered as a large α-helical embellishment functioning to promote the interaction between the two domains and is presently classified in CATH in this way. The α-helical embellishment in glutathione S-transferase is situated between the last consensus α-helix and the two embellished α-helices at the C-terminal end of the domain. This α-helical domain is classified in CATH with seven other S35Reps as a mainly α up-down bundle. The role of these α-helical domains is to promote interactions between the the thioredoxin domains in the same way as other embellishments present in the three superfamilies examined. However, it happens that they form a big enough embellishment to be classed as an extra domain. This suggests a mechanism of a gradual accumulation of inserted secondary structures to form a separate domain rather than the complete domain insertion through DNA shuffling. The other example of this fine line between extra domain and embellishment is between thioredoxin peroxidase and peroxidase hORF6. In both cases the final C-terminal helix is involved in the stabilisation of the dimer.However in peroxidase hORF6 the final consensus α-helix is classified in a separate domain with a small additional β-sheet. The β-sheet in this extra domain has a role in the modification of the active site on the thioredoxin domain.

Chapter 5

Automatic Comparative Modelling for Diverse CATH Superfamilies

5.1 Introduction

5.1.1 Background

In recent years selected organisms from all three kingdoms of life have been studied giving rise to their full genome sequences. Such genome projects have provided us with amino acid sequences of more than a million proteins - the catalysts, inhibitors, messengers, receptors, transporters and building blocks of the living organisms (Collins *et al.*, 1998). In order to make use of this information, the focus is now moving on to the functional analysis of the genome, addressing the functional and physiological role that proteins have within the cell.

The use of sequence databases, such as PRINTS (Attwood, 2002), Prosite (Sigrist *et al.*, 2002) and Interpro (Mulder *et al.*, 2002) can help predict functional information by the identification of key functional residue motifs. However, protein function is also tightly linked to its three-dimensional structure, and residues far apart on the peptide chain can be in close spatial proximity, performing a catalytic role in the final globular state. Since protein structure is more conserved than protein sequence during evolution, identifying similarities in protein three-dimensional structure can provide evidence of more distant evolutionary relationships. Therefore, it is important to study structure as well as sequence.

The aim of the structural genomics initiatives is to sample fold space completely in order to structurally and functionally characterise protein sequences. Experimental techniques for solving the three-dimensional structure, primarily NMR and X-ray crystallography, are often hampered by technical limitations making them time consuming.

The elucidation of genome sequences has provided impetus towards the improvement of these experimental techniques through the development of high throughput analysis.

The protein sequence databases GenPept (1,497,800 sequences), SWISS-PROT (94,000 sequences) and trEMBL (425,000 sequences), contain approximately 50 times more protein sequences than there are known three-dimensional structures. As a consequence, only about 2% of available protein sequence data has an associated three-dimensional structure. As such, the task of solving the structure of every single protein sequence by experimental methods would be slow and laborious. Focus has turned to modelling three-dimensional structures by comparison with known structures. In virtually all cases, sequences of greater than 30% sequence identity adopt a similar structure (Chothia & Lesk, 1986) and in many cases the fold is conserved at much lower sequence identities. This allows the structure of an uncharacterised sequence to be inferred from a homologous protein.

5.1.2 Comparative Modelling

The traditional steps involved in comparative modelling are outlined in Figure 5.1. Firstly, the amino acid sequence of the target structure is aligned to that of the homologous parent structure(s). The core regions of the structure are modelled, followed by an iterative procedure of loop and sidechain modelling. The model is energy minimised and the quality is assessed. These steps can be carried out manually but have also been automated in software such as COMPOSER (Sutcliffe *et al.*, 1987a,b) and Swiss-Model (Schwede *et al.*, 2003) which in turn have been developed and refined.

5.1.2.1 MODELLER

MODELLER (Sali & Blundell, 1993; Marti-Renom *et al.*, 2000) is a further approach for the construction of models. It differs from other methods, such as COMPOSER, in the way that it describes given parent structures and consequently derives the models. This algorithm is freely available and simple to use and therefore is a widely applied comparative modelling package and, as such, was considered to be a suitable method for the work carried out in this chapter. MODELLER arrives at a three-dimensional model of the target sequence by optimisation of a molecular probability density function (pdf). Following this, the molecular pdf for comparative modelling is optimised with the variable target function procedure in Cartesian space that employs methods of conjugate gradients and molecular dynamics with simulated annealing. The input to MODELLER is the sequence to be aligned to a set of template structures, or a previously generated alignment of a target sequence to its template structures. The distance and dihedral angle

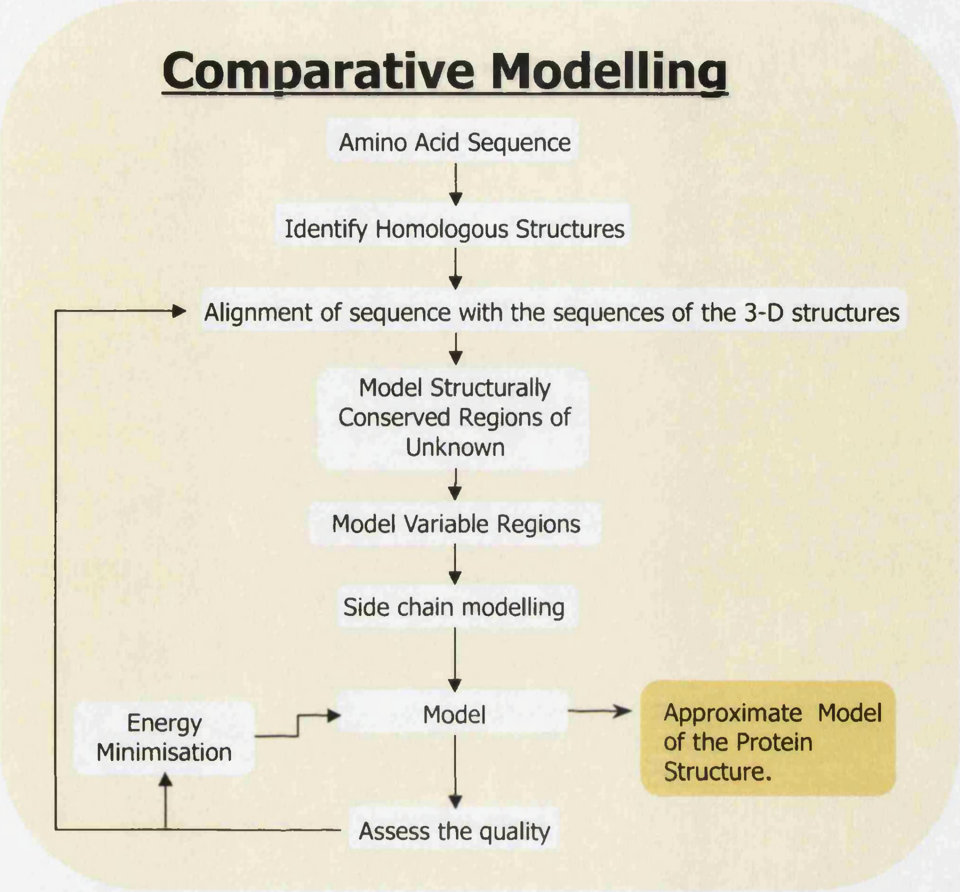

Comparative Modelling
Amino Acid Sequence
Identify Homologous Structures
Alignment of sequence with the sequences of the 3-D structures
Model Structurally Conserved Regions of Unknown
Model Variable Regions
Side chain modelling
Model
Energy Minimisation
Approximate Model of the Protein Structure.
Assess the quality

Figure 5.1: A flow chart showing the basic steps of comparative modelling.

restraints on the target sequence are calculated from the alignment with the template. The model is obtained so that there is minimal violation to the input restraints. The probability density function is a mathematically derived function which fits the trends in a set of observed discrete data values, with minimal error (Figure 5.2).

The form of these restraints was obtained empirically from a statistical analysis of the relationship between many pairs of homologous structures. This analysis used a database of 105 family alignments, including 426 proteins with known three-dimensional structure (Sali & Overington, 1994). The measured features must approximate to a function which is non-negative and integrates to 1 over all possible values. The integration of the curve from two points χ_1 and χ_2 is a function of χ which describes the curve which best fits the data.

A pdf suitable for restraining a certain feature can be described as in Equation 5.1, where, χ is the structural feature and a, b and c are the the elements which affect this structural feature:

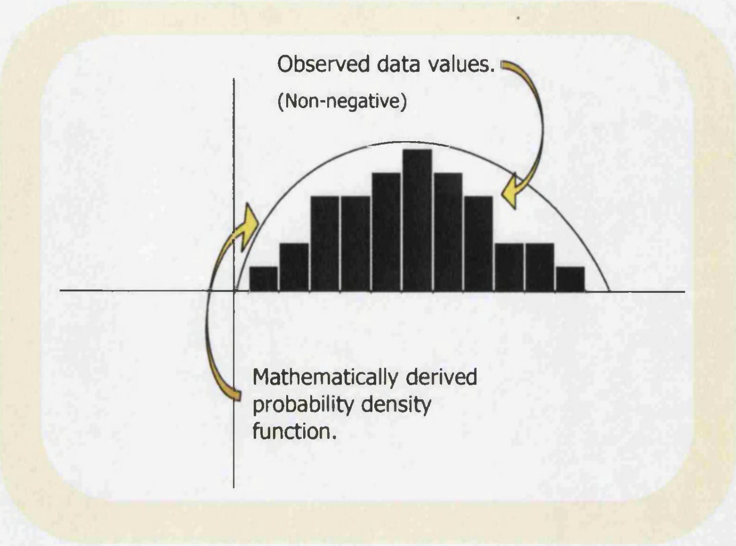


Figure 5.2: A mathematically derived function, which fits the observed data with minimal error, is calculated. The data must approximate to a function which is non-negative and integrates to 1, indicating certainty, over all possible values.

$$p(\chi|a, b....c) \tag{5.1}$$

A conditional pdf gives a probability density function for χ when a,b....c are specified. For example,

$$p(\chi|residuetype, \phi, \psi) \tag{5.2}$$

where χ = sidechain dihedral angle.

The sidechain dihedral angle depends on the residue type and the conformation of the main chain ϕ and ψ angles. Typical pdf restraints include Cα-Cα distances, main chain N-O distances, and main chain and sidechain dihedral angles. The spatial restraints on the target sequence are calculated from its alignment with the template structures. Following this, the spatial restraints and CHARMm energy force field (Momany & Rone, 1992) enforcing proper steriochemistry, are combined into an objective function. Optimising the objective function begins with satisfying sequentially local restraints and slowly introduces longer range restraints until the complete objective function is optimised. Finally simulated annealing implemented by molecular dynamics is used to refine the model.

MODELLER has been used to populate a database of modelled structures called ModBase (Sanchez & Sali, 1999; Pieper *et al.*, 2002). This query-able database of annotated comparative models derived from ModPipe, (Sanchez & Sali, 1998), an automated modelling pipeline which uses the programs PSI-BLAST (Altschul *et al.*, 1997) and MODELLER.

It is important to assess the effectiveness of a given comparative modelling method. This can be carried out by predicting the structure of a sequence for which an experimentally determined structure is already known. The accuracy of the model can then be measured by comparing it with the experimentally determined structure. In more recent years the Critical Assessment of Structure Prediction (CASP) experiments have enabled comparisons to be made between different comparative modelling techniques (section 5.1.5). Results from these experiments have shown that there are still limitations involved in the comparative modelling of protein structures. Two of the most error-prone steps in comparative modelling, after errors to the alignment, involve the modelling of structurally variable regions (SVRs) and the assignment of sidechain orientation.

5.1.2.2 Modelling of the Loops

Building the structurally variable regions remains a significant source of error. So far, three main methods have been used to build the structurally variable regions (SVRs): (1) The use of molecular graphics to manually build SVRs by the modification of backbone torsion angles and performing subsequent energy minimisation. (2) Knowledge-based searches of a database of loop conformation to find the most likely SVR conformation. (3) The use of *ab initio* methods to build SVRs through a conformation search within a given environment guided by an energy function.

The database approach to loop modelling involves finding a loop which fits the two stem regions on the main chain. The search is performed using a database of known loop structures. Loops which fit the stem are selected based on sequence and conformation criteria although this method is limited by the number of known loop conformations Fidelis *et al.* (1994). A variation on the knowledge-based procedure is the prediction of a general conformational class of the loop. This method has been developed by Burke *et al.* (2000) who developed the SLOOP database of loop conformations and connecting secondary structure elements. The loops are classified according to their length, the bounding secondary structures and the conformation of the mainchain. The method selects conformers based on the sequence of the loop and position of the secondary structure elements that anchor the loop region in the model.

An *ab initio* approach to loop modelling is carried out by MODELLER-5 (Fiser *et al.*, 2000). An energy function is calculated using terms from the CHARMm22 force field (Momany & Rone, 1992) and optimisation is based on conjugate gradients and molecular dynamics with simulated annealing (section 5.1.2.1).

5.1.2.3 Modelling of the Sidechains

There are a number of different protocols available for sidechain modelling. The 'Maximum Overlap Protocol' works by inheriting the sidechain torsion angles from the parent to the model where possible. The additional atoms are built from a single standard conformation. The 'Maximum Perturbation Protocol' proposed by Shih *et al.* (1985) finds the best sidechain conformations by rotation about the sidechain torsion angles. The use of rotamer libraries is a common approach for sidechain modelling. Each sidechain is stored in the library in all of the conformations of the torsion angles after steric clashes are eliminated. In the SCRWL method (Dunbrack & Cohen, 1997) the most favoured sidechain positions are set for all sidechains along the peptide backbone. The steric clashes are then resolved by changing a sidechain to a less favourable rotamer. This is carried out until all steric clashes are resolved and the 'minimum steric clash energy' is found.

5.1.3 Measuring Model Accuracy

As already discussed, it is possible to measure the accuracy of a model if the model sequence has a known experimental structure, enabling the comparison of experimentally determined structure and putative model. It is generally thought that for a sequence identity of 75%, with no indels, modelling can achieve an RMSD of 0.6Å, similar to the RMSD between two crystal structures of the same protein (Martin *et al.*, 1997). Between a sequence identity of 75% and 50%, the model is likely to be around 1Å which is comparable to a medium resolution NMR model or a model solved by low resolution X-ray Crystallography. Errors are generally found in the packing of the sidechains, small shifts in the main chain regions in the core, or larger shifts in the loops. Between 30–50% sequence identity the error is frequently increased by sidechain packing, core region and loop distortions. Below 30% the errors in the final model increase rapidly due to errors in the intial alignment between parent structures and target sequence (Baker & Sali, 2001).

There are a number of key measurement techniques freely available for the assessment of the quality of a structure, with reference to its experimentally determined structure.

5.1.3.1 Root Mean Square Deviation

The root mean square deviation (RMSD) is commonly used as a measure of structural similarity between two sets of three-dimensional co-ordinates. Equivalent residues are first identified and then the two structures superposed and rotated so that they overlap in three-dimensional space as closely as possible (Figure 5.3). The distance between each pair of equivalent residues, d_i, is then calculated and squared. The sum of these squared distances between equivalent residues is then taken and divided by the number of pairs, N, to give the mean. The mean is then square rooted (Equation 5.3).

$$RMSD = \sqrt{\frac{\sum_{i=1}^{N} d_i^2}{N}} \tag{5.3}$$

RMSD provides a measurement between two superposed structures and takes into account every pair of superposed atoms. However, a single region of structural variability between the two structures can significantly alter the score, even if overall the structures are similar.

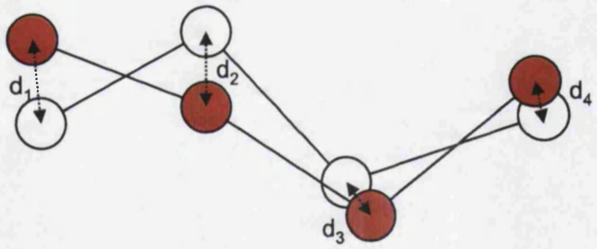


Figure 5.3: Calculating root mean square deviation (RMSD) as a measure of structure similarity following a structural superposition

5.1.3.2 LGScore

The LGScore (Cristobal *et al.*, 2001) method searches for the most significant non-continuous segment of a model (model-fragment) and measures the significance of that region with a P-Value (the probability of the score given by the model-fragment to experimentally determined structure alignment, occurring by chance) (Levitt & Gerstein, 1998). In other words because in many cases only a fraction of the model is similar to the structure, this method enables the detection of the most significant sub-part of the alignment. The method involves two heuristic algorithms.

Top-down Algorithm

1. Start with the whole protein
2. Superpose the model with the experimental structure
3. Calculate and store the P-Value
4. Delete the residues that are furthest apart
5. Repeat with fewer residues
6. Return best P-Values

Bottom-up Algorithm

1. Start with 4 residues
2. Superpose residues 0–4 of model and parent
3. Calculate P-Value
4. Add residues outside segment that are closest in model and experimentally determined structure. Repeat steps 2–3.
5. Return best P-Values

5.1.4 Predicting the Quality of a Comparative Model

5.1.4.1 ProsaII

ProsaII (Sippl, 1993) assesses the quality of a structure by calculating its energy. This calculation is based upon the principals that describe a protein in its native folded state which are derived empirically from a set of three-dimensional structures. The forces involved are extracted and then recombined for a protein of known or unknown structure as a function of the amino acid sequence. The value for the recombination of these forces can be compared with the forces of a model to identify a native fold from a mis-folded protein or an unsuccessful model.

5.1.5 Critical Assessment of Structure Prediction

The assessors of CASP4 (Tramontano *et al.*, 2001) studied whether the models predicted were closer to the experimental structure than the closest parent. It was found that the model was rarely closer to the experimental structure than the parent, that is, algorithms were unable to predict features which did not already exist in the parents. In fact for many predictions, the model structure was much closer (in terms of RMSD) to the parent structure than to the target experimentally determined structure, as shown in Figure 5.4. This was attributed to the non-optimal selection of parents and also errors in the alignments. It was also observed that active site regions within a target sequence are more accurately predicted as they remain much more conserved. Such conserved active site regions give a more accurate sequence alignment between target sequence and parent structures due to the pattern of conserved residues.

Overall, the assessors at the meeting (Tramontano *et al.*, 2001) highlighted a number of areas of improvement for comparative modelling:

- Identifying the best parent structure and the prediction of regions that deviate from the parent structure.
- The prediction of regions which deviate from the parent structure.
- The quality of the alignment when sequence identity is below 50%.
- Improvements to the prediction of relative domain orientations.

Many methods that use multiple parents result in a model which tends to correspond to an 'average' conformation of all the parents used. If a single parent is very similar in sequence identity to the target sequence (>70% sequence identity) it may provide a more accurate model than the 'average' conformation obtained from many parents. In CASP2 Martin *et al.* (1997) a general rule was suggested that using multiple parents does not generally improve the quality of the model over using the closest single parent at higher sequence identities. However, it was also reported that methods using multiple parents perform better than single parents on general fold correctness in CASP4 for the 14 target sequences with between 20–60% sequence identity to the closest parent (Tramontano *et al.*, 2001). Alignment quality also remains a considerable source of error in comparative modelling. Sequence alignments are often manually adjusted so that gaps appear mainly in the loops between the secondary structure elements. It was also notable that RMSDs calculated for the loop regions showed no discriminating results between the groups and the development of method which is able to predict loop regions with consistently good accuracy is still a challenge.

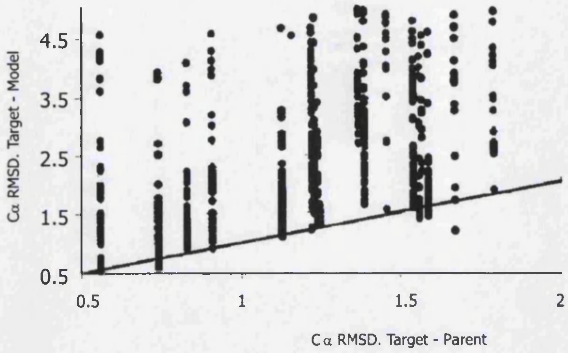


Figure 5.4: The RMSD of model and experimentally determined structure versus experimentally determined structure and closest parent in CASP4 for predictions with RMSDs lower than 5 Å only.

5.1.5.1 The Second Critical Assessment of Fully Automated Structure Prediction

More recently a number of automated comparative modelling procedures have been developed, a number of which have been subsequently assessed in the **C**ritical **A**ssessment of **F**ully **A**utomated **S**tructure **P**rediction (CAFASP2).

The second Critical Assessment of Fully Automated Structure Prediction methods (Fischer *et al.*, 2001) highlighted the current standards of automatic structure prediction servers and their effectiveness compared to the more human intensive methods assessed in the CASP experiments. Results showed that only a handful of manual groups performed better than the fully automatic comparative modelling servers. The automatic methods submitted to CAFASP2 were 3DJigsaw (Bates *et al.*, 1997; Bates & Sternberg, 1999) and the **F**ully **A**utomated **M**odelling **S**erver (FAMS; Ogata & Umeyama (2000)).

3DJigsaw contains many novel ideas to solve some of the major issues presently dogging comparative modelling. The method selects parents using PSI-BLAST with a high sequence identity threshold, ensuring high quality alignments. It selects a single parent if there is considerable similarity between a potential parent structure and the target sequence or if there is so much variation in the selection of multiple parents that there is likely to be no benefit in using the multiple parent method. Multiple parents are chosen when it is thought that the variation in structures may add beneficial information to the final structure. The target sequence is aligned by PSI-BLAST to the chosen parent(s) and, when not being automatically implemented, the alignment is manually altered. If multiple parents are used, the structures are separated into conserved secondary structures and non conserved regions. The conserved regions are modelled on the backbone and the non-conserved regions are modelled by searching a database of fragments. Sidechains are modelled using rotamers as described in section 5.1.2.2. FAMS applies a Smith-Waterman alignment to a structurally aligned template with substitution matrix derived from structure alignments.

In CAFASP2, the models produced by these methods were evaluated on three levels, fold assignment, backbone and sidechain conformations. It was found that for 9 of the 15 comparative modelling targets, FAMS based its model on an incorrect fold assignment whereas 3D-Jigsaw predicted the fold correctly for all. However, FAMS outperformed 3D-Jigsaw on both backbone and sidechain conformation for those folds it predicted correctly.

5.1.6 Aims of the Chapter

The aim of this chapter is to develop an automatic comparative modelling package, to provide structures for genome sequences that have homology to those in CATH structural superfamilies. Hidden Markov Models (HMMs; Karplus *et al.* (1998)) were used to align a target sequence to the structural parents and MODELLER was then used to produce a model. In addition, a method for identifying the structurally variable and structurally conserved regions of a protein has been developed (Mosaic). For each structural region (variable and conserved) the closest parent was chosen by measuring sequence identity with the target, to create a single chimeric parent. The selection of such structurally variable and conserved regions within a particular superfamily can in turn give an indication of the degree of difficulty one may come across when attempting to model a sequence that is related to this superfamily. This information can be used when considering the quality of specific regions within the resultant models for that particular superfamily.

Methods for assessing the quality of the resultant models have been integrated into the pipeline. Two methods, RMSD and LGScore, rely on the target sequence having an experimentally determined structure which is used as a reference against the model. These methods provide a guide to the quality of the modelling method with respect to the specific superfamily. Additionally, ProsaII, which does not use a reference structure is used. This method measures the quality using empirical potentials to calculate the energy of the structure and predicts its accuracy.

5.2 Methods

It is hoped that the automatic comparative modelling pipeline will eventually model the structures of genomic sequences identified as homologous to a CATH superfamily, using the structural representatives in that superfamily as structural parents. In order to test the accuracy of the comparative modelling pipeline outlined in this chapter, sequences with known structure are modelled and the quality of the subsequent model can then be assessed with reference to the known structure. In this study a number of CATH S35Reps were used as parent structures. Related sequences (of known structure) to these parent structures were selected for a range of sequence identities. Models could then be constructed for these target sequences and the effect of different sequence identities to their parent structures could then be assessed. The methods by which the parent structures and target sequences were selected are outlined in more detail below.

5.2.1 GenMod: Modelling Pipeline

GenMod contains several steps for recognising homologues to known structures in CATH, aligning these target sequences to their parent structures and building the comparative models. The modelling pipeline is described in the following sections:

- Identifying Structural Sub-Groups (SSG) in CATH superfamilies to provide reliable structural alignments of parents and build Hidden Markov Models (HMMs) for homologue recognition.
- Using a model library of HMMs (SAMOSA) to recognise homologues for CATH superfamilies and provide target–parent alignments.
- Modelling the target sequences using MODELLER.

5.2.1.1 Identifying Structural Sub-Groups in CATH Superfamilies to Provide Reliable Structural Alignments of Parents and Build Hidden Markov Models (HMMs) for Homologue Recognition.

In well populated CATH superfamilies containing two or more diverse structural relatives (i.e. S35Rep families), domain structures are further clustered into Structural Sub-Groups (SSGs). Representative structures are taken from each CATH S35Rep (<35% sequence identity) and clustered on the basis of their pairwise SSAP scores (Figure 5.5). As a result, these SSGs consist of sequence dissimilar (sequences <35% sequence identity), structurally similar (SSAP scores >80) domains and provide ideal clusters of parent structures for comparative modelling.

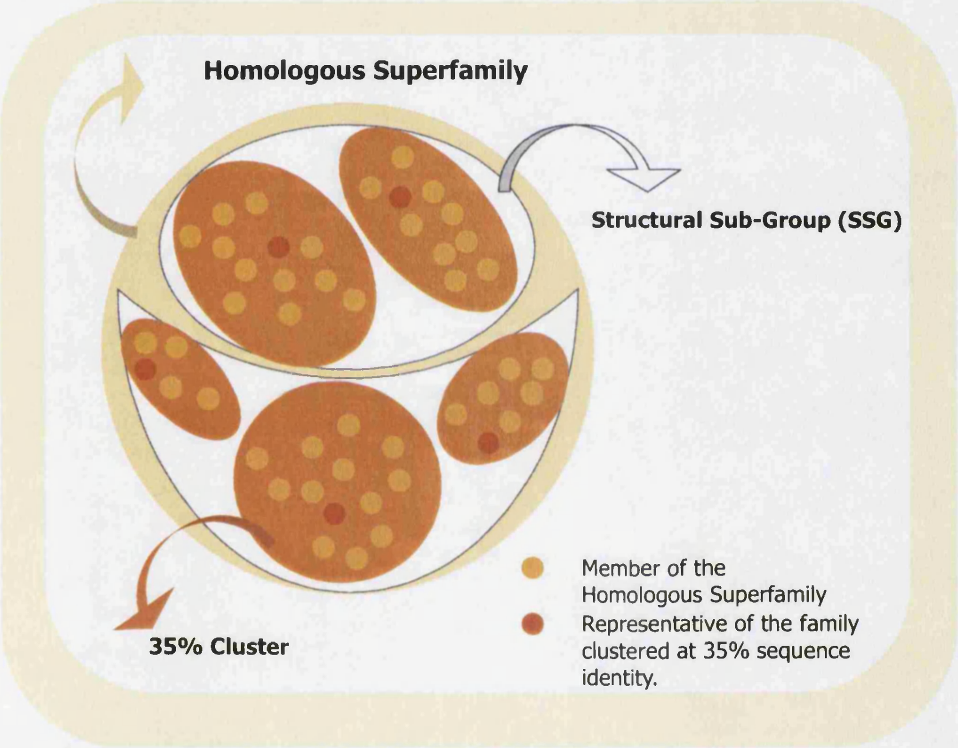

Homologous Superfamily
Structural Sub-Group (SSG)
35% Cluster
● Member of the Homologous Superfamily
● Representative of the family clustered at 35% sequence identity.

Figure 5.5: The homologous superfamily is divided into sequence clusters at 35% sequence identity. Within these clusters, the structures remain very similar. These sequence clusters are further divided into Structural Sub-Groups (SSGs). SSGs provide sub-clusters within the homologous superfamily which exhibit significant structural similarity but also contain sequence diverse relatives.

5.2.1.2 Using a Model Library of HMMs (SAMOSA) to Recognise Homologues for CATH Superfamilies and Provide Target – Parent Alignments.

A method for identifying and aligning homologous sequences to the SSGs is summarised in Figure 5.6. The Hidden Markov Models using SAM–T99 is one of the most sensitive methods for identifying homologous sequences (Park *et al.*, 1998). The SAM–T99 software uses each S35Rep in the SSG as a seed to search the GenBank sequence database (translated GenBank-NRDB, released March 2000). The target99 script in the SAM-T99 software identifies a set of related genomic sequences and generates a multiple sequence alignment.

The CORAXplode protocol (Sillitoe, 2002) was then used to combine these multiple sequence alignments (for each S35Reps in the SSG), guided by a structural alignment of all S35Reps within the given SSG (built using the CORA algorithm (Orengo, 1999)). In summary, the program selects a set of similar structures from a homologous superfamily, and generates a multiple structure alignment of these seed proteins using the CORA algorithm. The SAM-T99 sequence alignments for each seed structure are then combined using the CORAXplode protocol by following the CORA structural alignment. These are known as SAMOSAs (Sequence Augmented Multiple Structure Alignment) models (Figure 5.6).

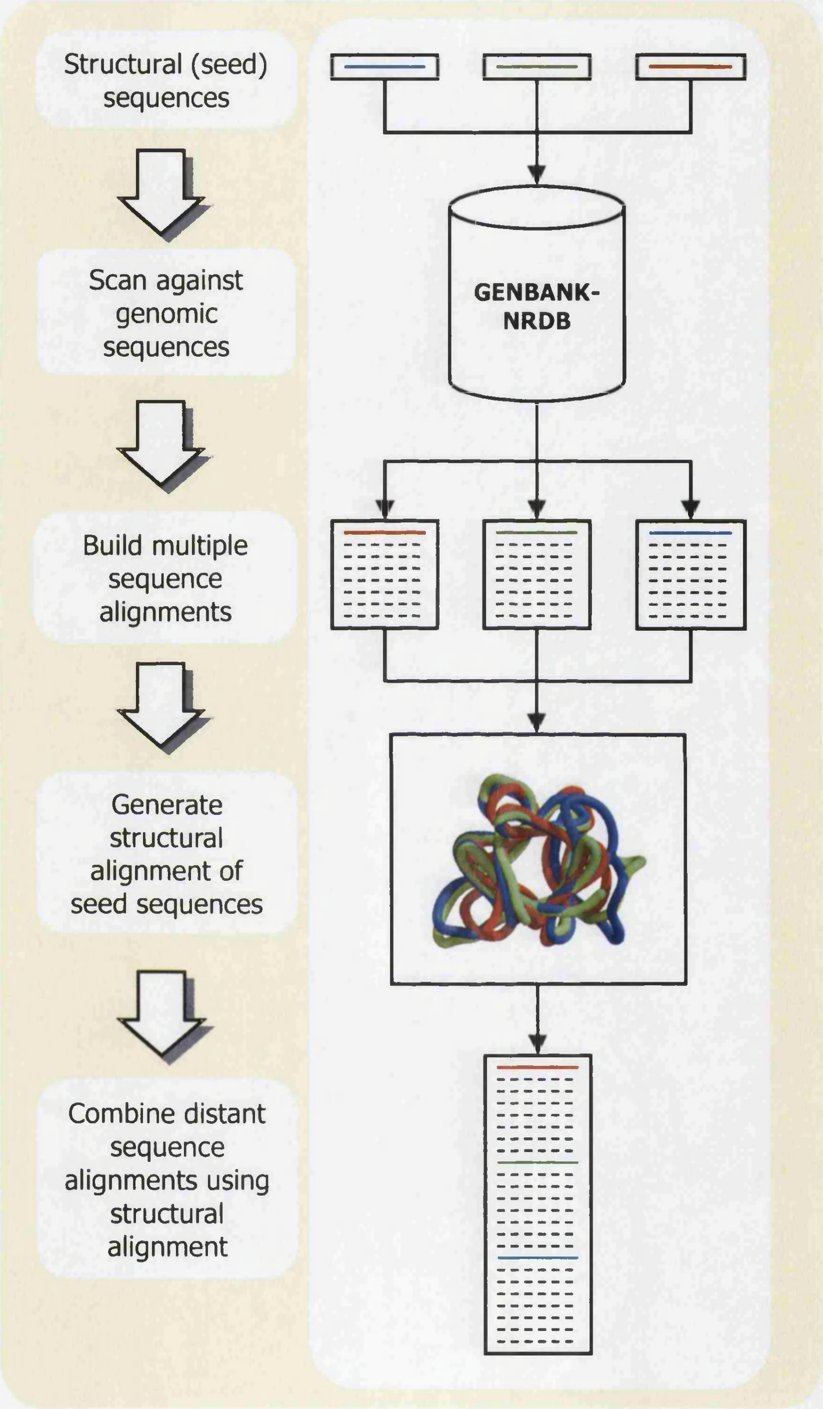

Structural (seed) sequences
Scan against genomic sequences
Build multiple sequence alignments
Generate structural alignment of seed sequences
Combine distant sequence alignments using structural alignment
GENBANK-NRDB

Figure 5.6: Flowchart summarising the protocol for selecting and aligning homologues. Single structures are initially used as seeds to search a sequence database. These distantly related sequence alignments are then combined, using the CORAX-plode protocol, by referring to a multiple structural alignment of the seed structures.

5.2.2 Alternative Methods for Selecting Parent Structures

As previously described the use of multiple parents may lead to the creation of an average structure which can sometimes blur regions of close structural similarity between the target and one parent. In some cases, it may be possible to find an area of local similarity between a parent and target in one region of the structure and another parent and target in a different region of the structure. To investigate this, methods for selecting regions of local similarity between a parent and the target sequence were implemented. These searched for areas of local similarity between parent and target, filtering out data which might detract from the real structure of the target sequence. Four different methods were tested which used either single or multiple parents.

5.2.2.1 Method 1: The Single Closest Parent

Sequence identities between target sequence and each parent were calculated and the highest scoring parent was selected as the closest parent. The structure/sequence alignment obtained from the SAMOSA model (section 5.2.1.2) of the closest parent and target sequence only are input into MODELLER.

5.2.2.2 Method 2: Multiple Parents

The structure/sequence alignment obtained from the SAMOSA model for the target sequence and *all* parents from the SSG (rather than just the closest parent) are input into MODELLER.

5.2.2.3 Method 3: Single Chimeric Parent

A single chimeric parent is created by selecting the closest parent structure for each region. Three programs are used, Findcore, Mosaic and AlignAdjust to select regions of parent structures.

In order to try to combine the sensitivity of using the single parent method (which can be obtained when the parent structure and target sequence are closely related) with the flexibility and extra information provided by multiple parents, a simple method for combining all the best possible regions of the multiple parents was proposed. The multiple alignment (the CORA structural alignment of the S35Reps in the SSG) is divided into those regions of high structural similarity ($\leq$3Å) which correspond to the core regions and the more variable regions (structural similarity >3Å) by the program Findcore. Findcore superposes CORA aligned structures one by one building up a consensus, and measures the distance between the equivalent Cα atoms at each position in the alignment. The program was adapted from pairwise 'Findcore' developed by Martin *et al.* (1997) for the CASP2 assessment.

The sequence similarity of each region in a given parent, to the corresponding sequence in the model sequence, is assessed using an algorithm called Mosaic which uses the BLOSUM62 matrix to score the similarity. For each region the most similar sequence is identified from the available parents to produce the chimeric parent.

AlignAdjust provides a way of selecting which regions of each parent are used by MODELLER to build the structure of the target sequence. The method alters the structure/sequence alignment such that regions of the parents that should be considered are aligned whilst other regions are adjusted by introducing gaps into the alignment (Figure 5.7). The sequence identity of the chimeric parent is used to measure the similarity between the sequence and template, referred to as the chimeric sequence identity.

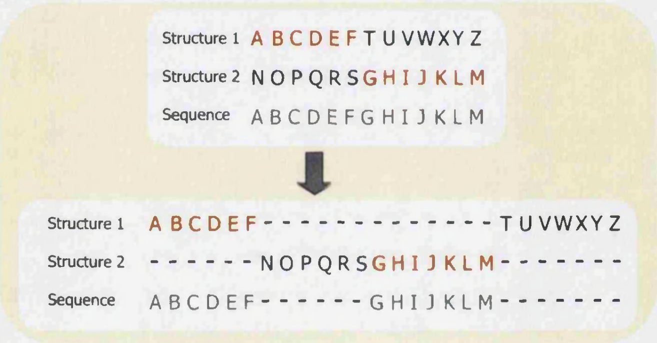

Structure 1 A B C D E F T U V W X Y Z
Structure 2 N O P Q R S G H I J K L M
Sequence A B C D E F G H I J K L M
Structure 1 A B C D E F - - - - - - - - - - - - - T U V W X Y Z
Structure 2 - - - - - - N O P Q R S G H I J K L M - - - - - - -
Sequence A B C D E F - - - - - - G H I J K L M - - - - - - -

Figure 5.7: An illustration of AlignAdjust. The program is instructed which regions of the alignment are most similar (shown here in red) to the target sequence (shown here in green) and adjusts the alignment so that the most similar region is aligned and other sequence regions are not.

5.2.2.4 Method 4: Multiple Parents for Core Regions and Single Chimeric Parent for Variable

The use of a mixture of single parent structures for the variable regions, and multiple parents for the core regions was considered. The model was created by using the best parent selection in the variable regions but all parents for the core regions.

The whole modelling pipeline and the four modelling methods are outlined in Figure 5.8. Any target sequences with chimeric sequence identities above 20% and below 60% were modelled. Although 30% sequence identity is usually suggested as a cut-off for producing a reasonably good model, exploring performance down to 20% sequence identity gave a more thorough test of the 4 methods.

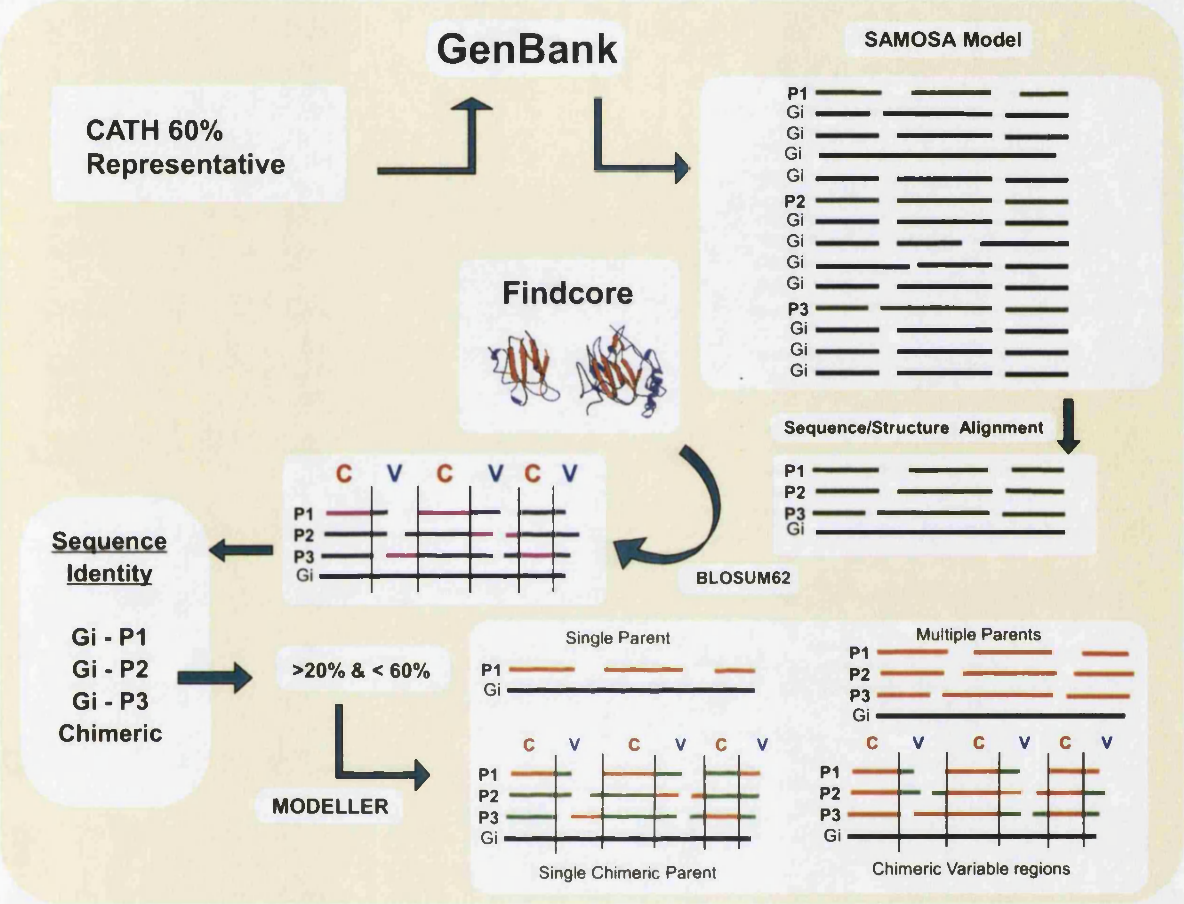


Figure 5.8: The target sequence is identified in the SAMOSA model and extracted, along with the structural parents. The structural parents shown here in green and target sequence in black. The core and variable regions of the parent structures are located using Findcore and mapped to the target sequence. The BLOSUM62 matrix selects which regions are more closely related (shown in pink). Sequence identities are calculated between all parents and target sequence, and also the sequence identity of the the most related segments (chimeric sequence identity). Any target sequence which aligns with a chimeric sequence identity of greater than 20% and less than 60% is modelled. Four different modelling methods are proposed. Each method differs in the template structure selection. Modelled regions of the parent structures are shown in orange and discarded regions shown in light green.

5.2.3 Evaluation of Alignment Quality

In order to assess the performance of the different comparative modelling methods properly, it was necessary to determine the quality of the alignment produced by the modelling pipeline. This was calculated as the percentage of correctly aligned residues, based upon a structural alignment of parent structure(s) and experimentally determined target structure. This involved 3 stages:

1. SSAP comparison is carried out for the experimental structure and its closest relative to determine the correct equivalent residues.

2. The target sequence and its closest relative are extracted from the SAMOSA alignment.

3. The number of aligned positions which are the same in the sequence alignment as in the structure alignment are calculated as a percentage of the total number of positions aligned.

5.2.4 Evaluation of Model Quality

It was also necessary to evaluate the model quality. Three measures were applied; LGScore, RMSD and ProsaII.

5.2.4.1 LGScore

LGScore (Cristobal *et al.*, 2001) (section 5.1.3.2) is used to assess the quality of the models with reference to their experimentally solved structures. The results give the percentage of the structure which gives the best P-Value. The P-Value for closely related proteins is dependent on the size of the target and therefore the P-Value is normalised to a size independent Q-Value. This is a straight line correlation between m, the size of the protein and the minus log of the P-Value. The gradient of this line is 0.0268. This relationship is described by Equation 5.4.

$$-logP = 0.0268m + 0.5115 \tag{5.4}$$

The value for -logP is then brought to the origin of the graph by subtracting 0.5115 and the gradient is made the subject of the formula. The resultant Q-Value is calculated by dividing the gradient by a large number (in this case 1000) to make it negligable, or, multiplying the rest by 1000.

$$\text{Q-Value} = \frac{-logP - 0.5115}{m} \times 1000 \tag{5.5}$$

As such the Q-Value is almost independent on the size of the fragment. A Q-Value of greater than 2 is significant, zero is not at all significant (Elofsson, personal communication).

5.2.4.2 RMSD

As with LGScore, the quality of the models is also assessed by calculating the RMSD (section 5.1.3.1) between model structure and experimentally determined structure. Unlike LGScore, the whole of the structure is measured, regardless of how similar or different the regions are to the experimentally determined structure. This means that one region of great dissimilarity from the experimentally determined structure will reflect in the score, even if the rest of the structure has been modelled well.

5.2.4.3 ProsaII

ProsaII (section 5.1.4.1) predicts the quality of a model from empirical assessment of the forces in native protein structures. This method is different from LGScore and RMSD as it does not use the experimentally determined structure as a reference. It is therefore a useful measurement when modelling sequences with unknown structure.

5.2.5 Selection of the Test Dataset

In order to sample the sequence space evenly and thoroughly within the homologous superfamily, structures were clustered into 60% sequence families by multi-linkage clustering and a representative was taken from each cluster (S60Rep). Models were created for any superfamily with two or more SSGs to ensure that the dataset was largely composed of structurally divergent superfamilies. Figure 5.9 illustrates how the homologous superfamily is divided. Genbank sequences of every structural member of the homologous superfamily, except the members of the SSG which were used as parent structures, were in the test dataset. Any structural relatives with >4Å resolution and those solved by NMR were removed from the dataset.

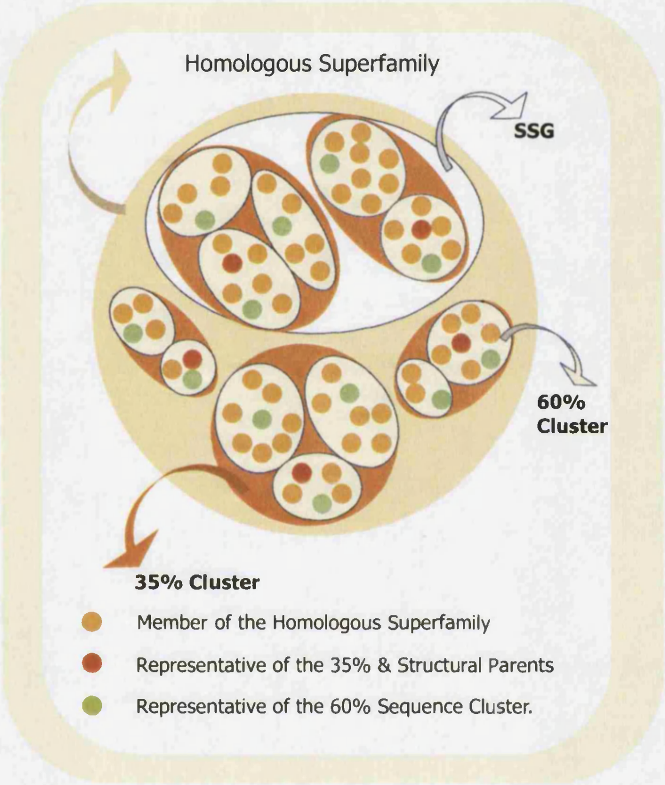

Homologous Superfamily
SSG
60% Cluster
35% Cluster
● Member of the Homologous Superfamily
● Representative of the 35% & Structural Parents
● Representative of the 60% Sequence Cluster.

Figure 5.9: Homologous superfamilies (in dark yellow) are clustered into groups of 35% sequence identity (in orange) and subclustered into groups of 60% sequence identity (in cream). Domains selected to be the 35% representatives are coloured in dark orange. 60% representatives are shown in green. The selected SSG in this homologous superfamily is shown in white and contains two 35% representatives. These two representatives become the structural parents. All other 60% representatives in the SSG and all other SSGs in the superfamily become the target sequences for testing the homology modelling methods. This ensures sequences are chosen to fully sample sequence space around the chosen structural parents.

5.3 Results

5.3.1 Overview of Results

The results in this section are discussed in three main categories:

1. The factors which affect model quality independent of modelling method used.
   - How does structural deviation between the experimentally determined structure and the closest parent affect the model quality?
   - How accurate was the alignment? How much affect does this have on the model quality?
2. Assessment of model quality.
   - How do the three model quality assessment methods perform?
3. Assessment of the four modelling methods.

5.3.2 Selecting Target Sequences to Test the Modelling Methods

This study considers any sequences which align with a chimeric sequence identity between 20% and 60%. It is widely considered that sequences which align with 30% sequence identity or greater, have a similar structure although this varies between superfamilies (section 3.3.3). In order to explore the sequence structure relationship the range was extended to include those which aligned down to 20% sequence identity as it is possible that in some superfamilies relatives with low sequence identities will be modelled more successfully. Superfamilies with only one SSG were removed from the dataset. With the present dataset approximately 2400 S60Reps were identified as possible models and 140 of these target sequences aligned to their template with 20–60% sequence identity.

5.3.3 Measuring the Chimeric Sequence Identity

The chimeric sequence identity is calculated by summing the sequence identities for the best fragments from each of the parent structures, taking the highest sequence identity for each fragment. Figure 5.10 shows that the chimeric sequence identity for most structures is higher than the closest parent. Occasionally, the chimeric sequence identity is worse because the best sequence for each segment has been selected according to a score derived using the BLOSUM62 matrix. This calculates likely exchanges for one amino acid to another and therefore sometimes a string of likely exchanges can score better than a few exact matches.

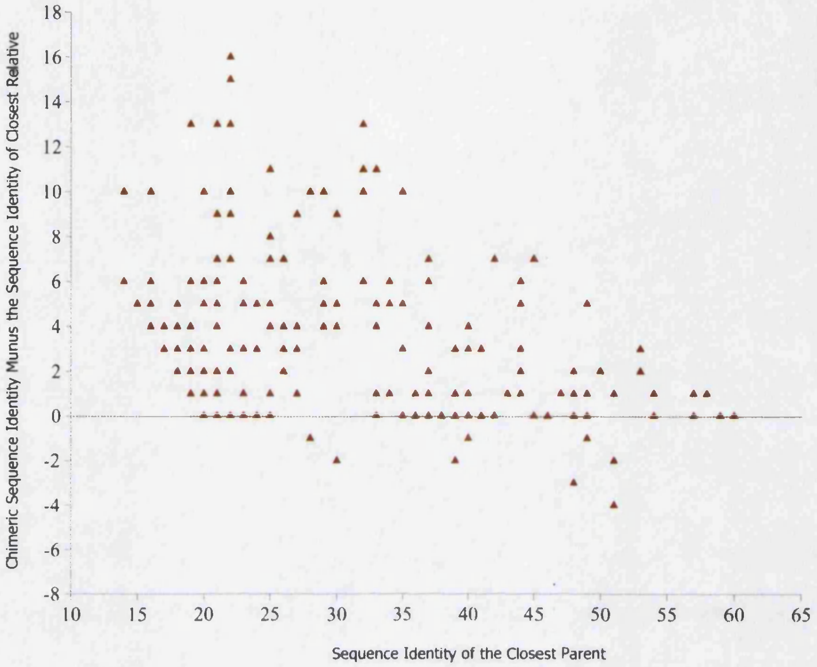


Figure 5.10: Chimeric sequence identity minus sequence identity of closest parent versus sequence identity of the closest parent.

5.3.4 The Effect of Structural Variability Between Experimental Structure and Closest Parent on the Model Quality

As previously described, Tramontano *et al.* (2001) observed that it was rarely the case that a model submitted to CASP4 was closer in structure to the experimental structure compared to its closest parent. This factor must be considered when interpreting the results for comparative modelling assessments.

The two structures (experimentally determined structure of sequence to be modelled and its closest parent) were superposed and aligned using the SSAP structural comparison program and the RMSD was then measured. The results are shown in Figure 5.11. It can be seen that as sequence identity decreases, structural similarity also decreases indicating that at low sequence identities the differences in structures will add to the inaccuracy of models. In other words even given a highly accurate modelling procedure, the final model accuracy will always be dependent upon the sequence similarity between the target and the parent structure. This distribution also shows that there are a number of pairs that show greater structural diversity at a particular sequence identity (such examples are coloured in red) than others at the same sequence identity. A structure pair which shows high structural similarity at a low sequence identity (approximately 24% coloured green) can also be seen. These two structures are DNA binding proteins forming helix-turn-helix homeodomains.

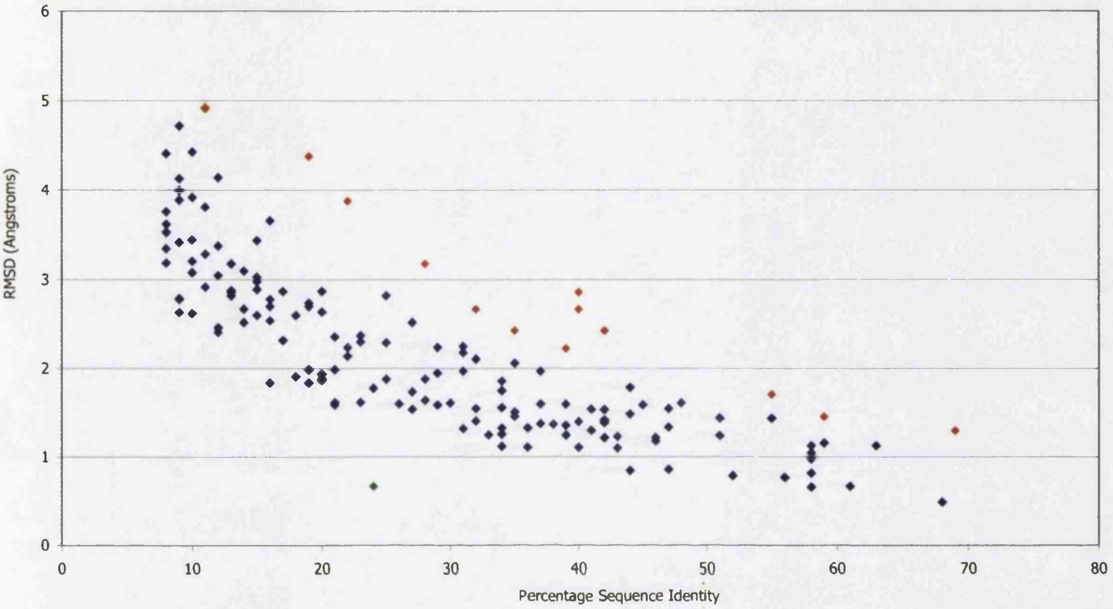


Figure 5.11: Showing the relationship between RMSD of the experimental structures and the closest parents in the dataset. The RMSD is plotted against sequence identity. As the sequence identity between the two structures increases, the RMSD decreases. Some pairs are more structurally variable (coloured in red) and one pair shows considerable structural conservation at a low sequence identity.

5.3.5 The Effect of Alignment Quality on Model Quality

In order to assess GenMod, structural alignments were generated between the experimental structure of the given target sequence and its closest relative using the structural comparison program SSAP. This alignment was considered to be the 'true' alignment. The equivalent residues in the SSAP alignment were then compared with the equivalent positions aligned by the SAMOSA model between the target sequence and parent structure. In this way misaligned residues in the sequence/structure alignment (compared to the 'true' alignment) were identified (section 5.2.3).

The percentage of correctly aligned residues, plotted against the corresponding RMSD between the model and experimentally determined structure is shown in Figure 5.12. A general trend can be seen, whereby as the number of correctly aligned residues increases, the RMSD between model and experimentally determined structure decreases. Between

0–30% correctly aligned residues there is a wide range of RMSDs whilst at higher percentage correctly aligned residues lower RMSDs are generally found. It is apparent that the quality of the alignment has a significant effect on the final quality of the model. However, some RMSDs for models with 90 to 100% correctly aligned residues are still as high as some models which measured a lower percentage alignment.

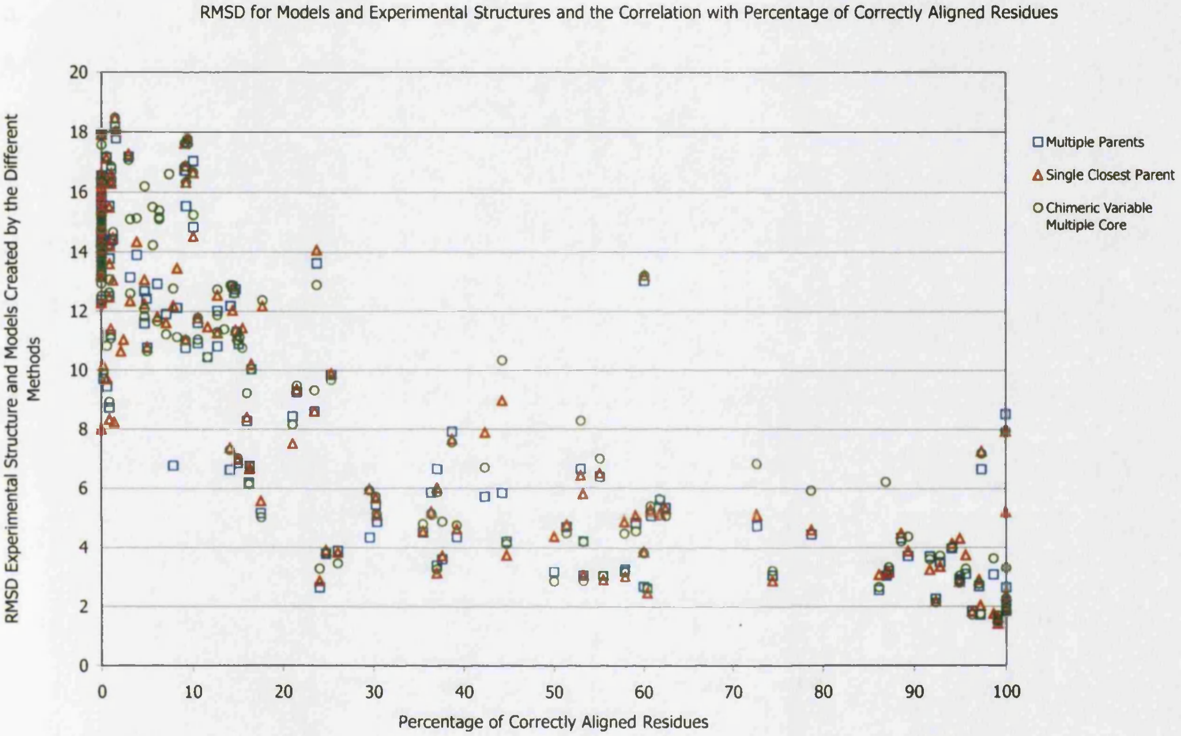


Figure 5.12: The relationship between percentage of correctly aligned residues with the RMSDs of the models and their experimentally determined structures.

For the purposes of measuring the performance of the modelling methods, the dataset was divided into two sub-sets; those with more than 30% correctly aligned residues and those which aligned with less than 30%.

5.3.6 Modelling Sequences Which have Closely Related Target Structures

It may also be possible that for target sequences that are more closely related to parent sequences (for example >50% sequence identity), the process of alignment by a profile based method could create a more inaccurate alignment compared to a simple pairwise alignment between model and closest structure. To analyse this, a Smith-Waterman alignment was carried out between whole chain target sequence and the closest structural relative. Figure 5.13 shows the difference in percentage correctly aligned residues between the SAM–T99 alignment and the Smith-Waterman alignment, using the SSAP alignment as the correct alignment. At low sequence identities both methods perform equally (the difference in percentage correctly aligned residues is around zero) but at higher sequence identities the SAM–T99 alignments improve at a higher rate than the Smith-Waterman alignments. This may be because in GenBank sequences that are multidomain, the SAM–T99 method is able to recognise the local similarity of a single domain, whereas, the Smith-Waterman has less success in recognising the correct domain.

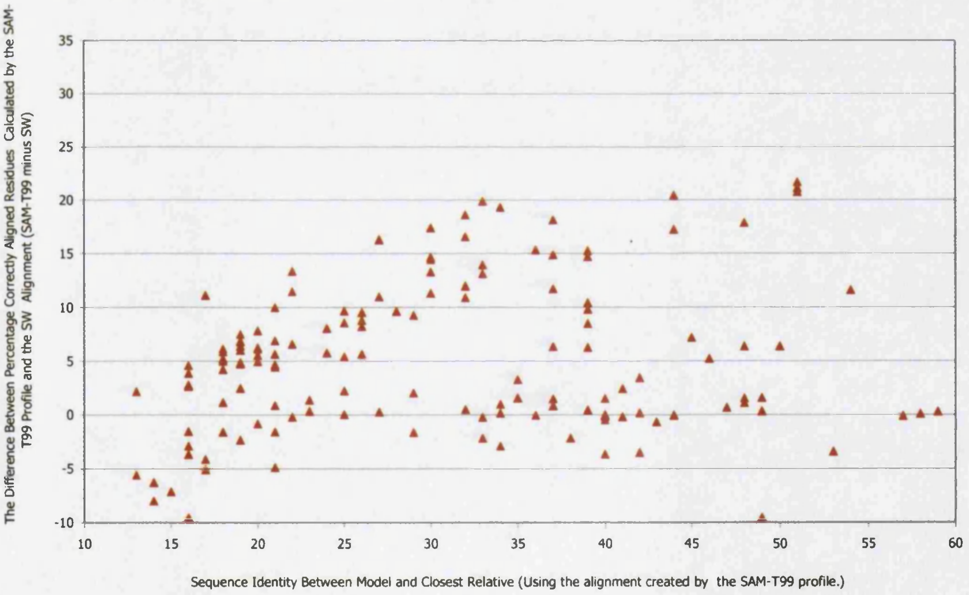


Figure 5.13: Smith–Waterman alignment versus the SAM–T99 alignment for aligning protein sequences shows that for low sequence identities, both methods perform similarly but for higher sequence identities the SAM–T99 method outperforms the Smith Waterman method.

5.3.7 The Effectiveness of different Methods in Assessing Model Quality

5.3.7.1 The Performance of Different Modelling Methods as Assessed by RMSD

RMSD is widely used to assess structural similarity, although it is intolerant to local areas of dissimilarity for a pair of structures which are similar over most regions. Plotting the chimeric sequence identity against RMSD measured between experimentally determined structure and the model, shows that generally, as the chimeric sequence identity increases, the RMSD decreases (Figure 5.14a). Most of these models fall between 20-35% sequence identity. In this percentage bracket, the models display a range of RMSD values from 3–19Å. When only those models which aligned with >30% correctly aligned residues were plotted (Figure. 5.14b) most of the structures in the 20–35% sequence identity region were removed.

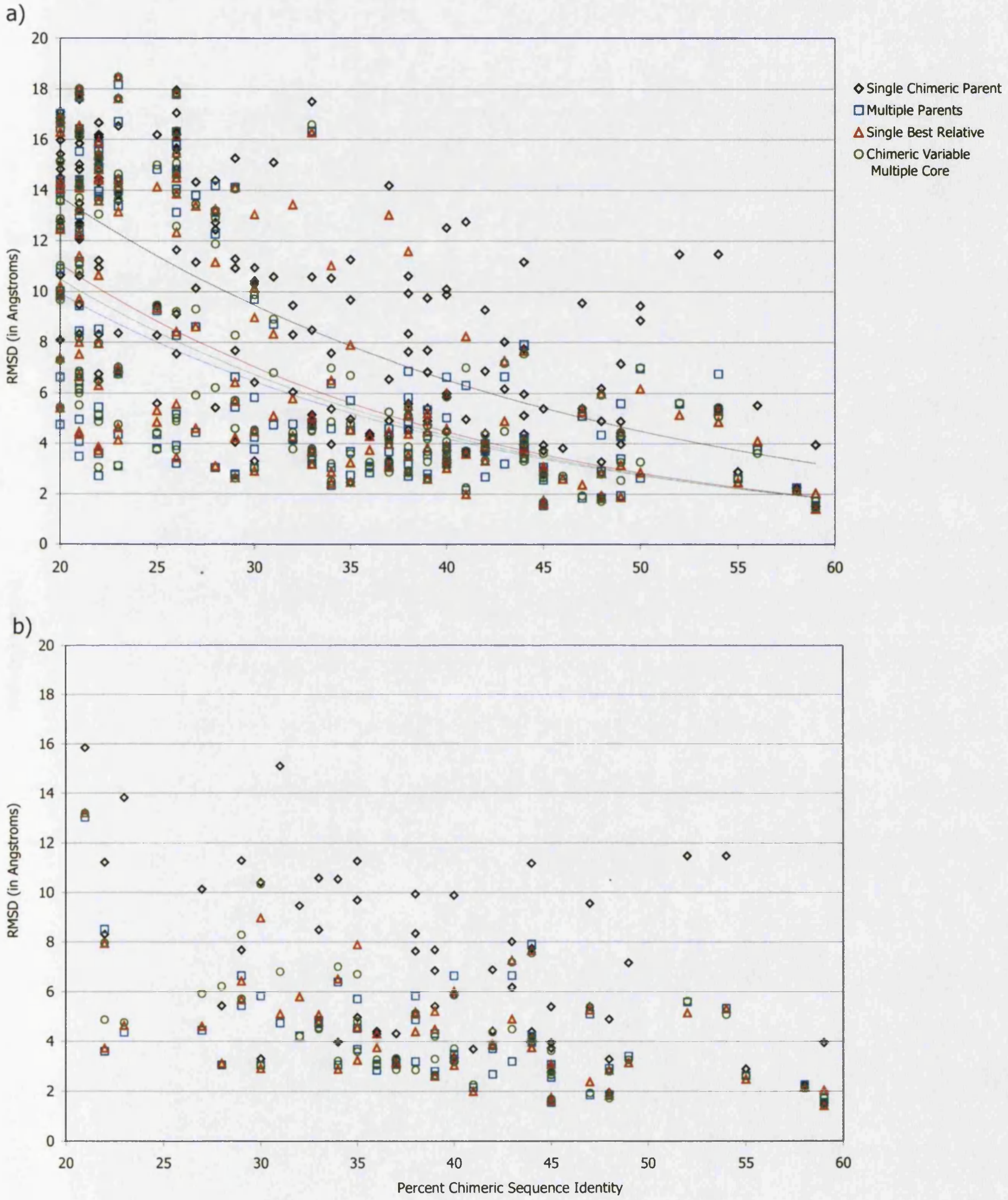


Figure 5.14: Root Mean Square Deviation between model and experimental structure for chimeric sequence identities of 20–60%. Plot b. shows those structures which were measured to have 30% or more of their residues correctly aligned.

Figure 5.15 shows RMSD versus the chimeric sequence identity for models with 30% or more correctly aligned residues, for the Multiple Parent Method, and colours the points according to percentage correctly aligned residue ranges. The line is drawn for those models which have >90% correctly aligned residues only, and shows an increase in model quality as chimeric sequence identity increases. Other percentage correctly aligned residue ranges show a deviation from this trend. The most striking insight gained from this graph is that those models created from alignments with >90% alignment accuracy are not always those with the higher chimeric sequence identities. In fact there are five models with low chimeric sequence identites below 35% which have 70–100 % of their residues correctly aligned.

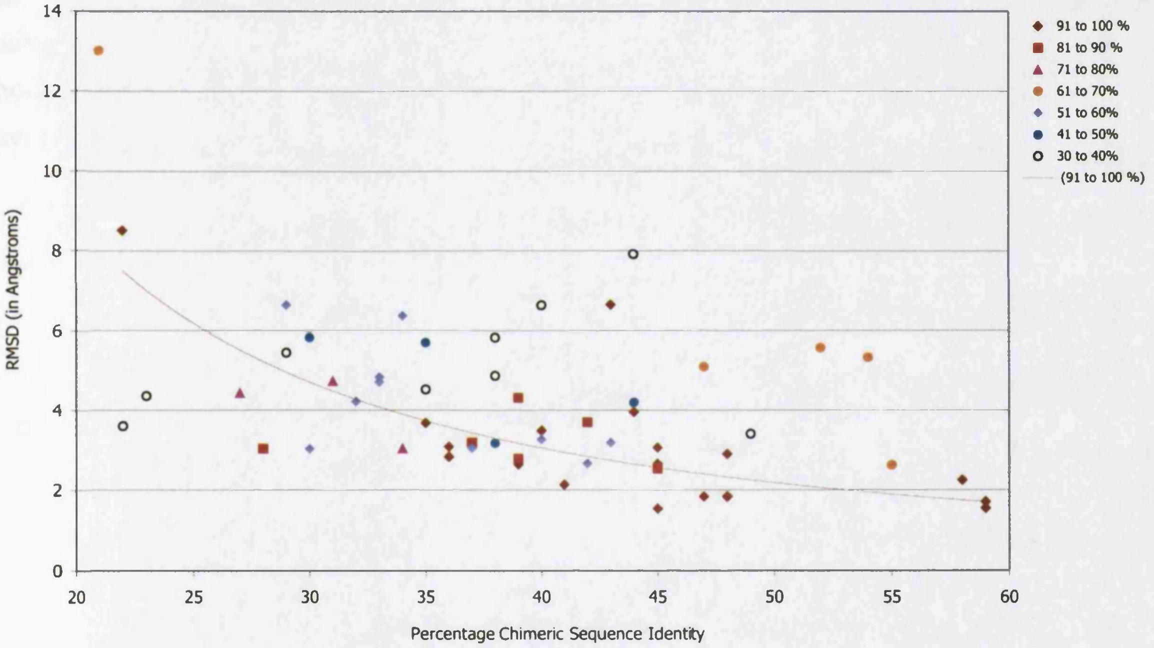


Figure 5.15: The RMSD and percentage chimeric sequence identity are shown for all models with 30% or more correctly aligned residues. Each model is coloured according to the percentage of correctly aligned residues in ranges of 10%. It can be seen that as alignments get more accurate the points get closer to the line of correlation, here plotted for those structures which have been 90–100% correctly aligned.

5.3.7.2 Using LGScore

Like RMSD, LGScore calculates similarity between model and experimentally determined structure but, unlike RMSD, it measures similarity based on the largest discontinuous segment, leaving any areas which differ significantly between model and experimental structure out of the calculation. The similarity is measured with a P-Value, a Q-Value or the number of aligned residues. The Q-Value removes the size dependency from the P-Value measurement. As a general rule Q-Values above 2 indicate a model of significant quality. The better the quality of the model, the higher the Q-Value (section 5.1.3.2).

Figure 5.16a shows the Q-Value score plotted for all data in the study whilst Figure 5.16b shows only those models built from alignments with >30% correctly aligned residues. As expected, the Q-Value increases with increasing sequence identity. Only models built using the Single Chimeric Parent score below the threshold of Q-Value of 2. Again, poor models are removed when models created with less then 30% correctly aligned residues are removed.

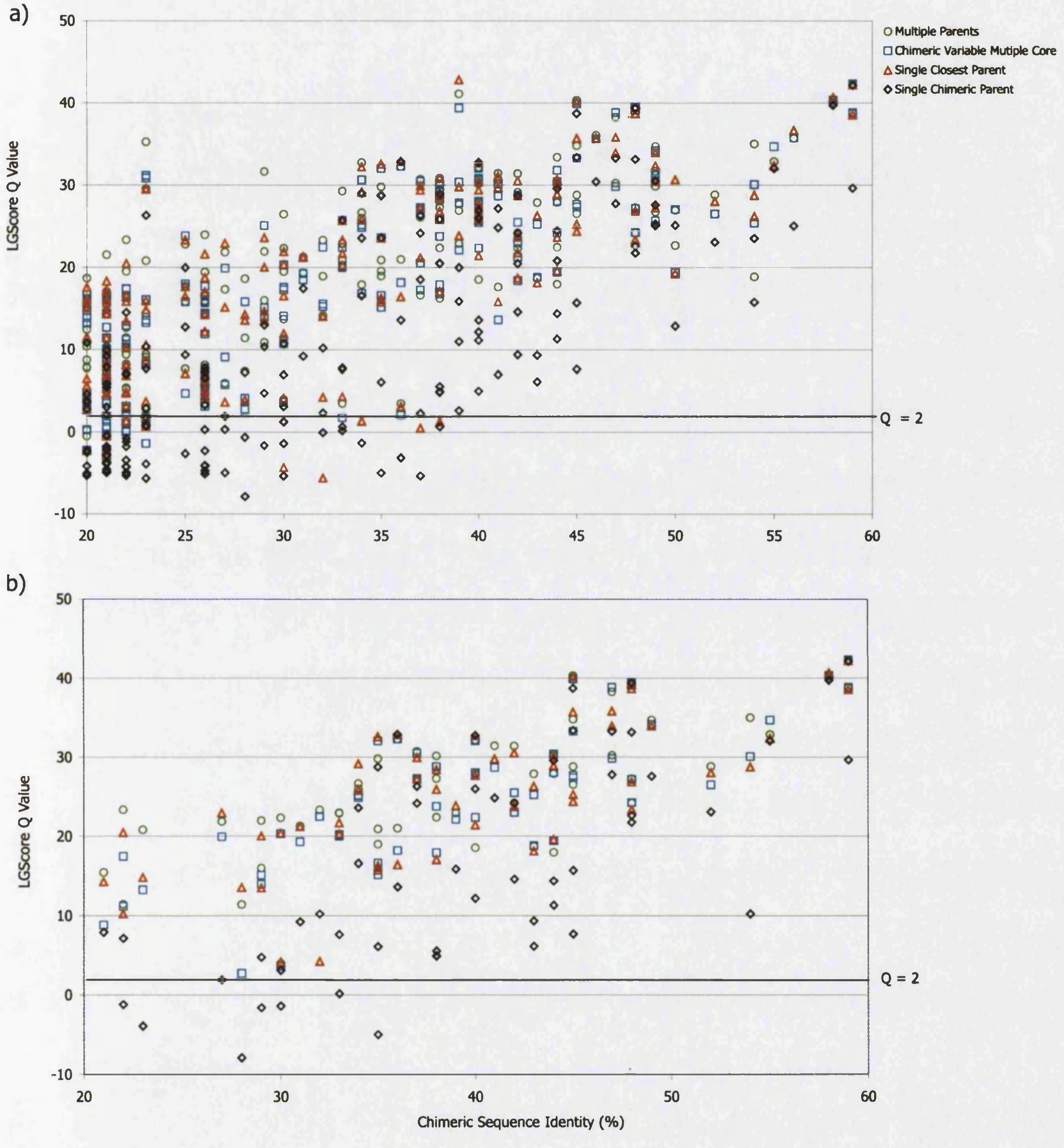


Figure 5.16: LGScore Q-Value plotted versus chimeric sequence identity for model sequence and parents. Plot a. contains all proteins in the dataset. Plot b. contains only those models which aligned to their closest parents with 30% correctly aligned or more.

5.3.7.3 Using ProsaII

An aim of this comparative modelling pipeline (GenMod) is to build models for sequences of unknown structure. However, the reliability of these models must be validated. Unlike LGScore and RMSD, ProsaII does not use any reference to an experimental structure and therefore is useful when the structure has not been experimentally determined. ProsaII measures the quality of models by considering their energy in terms of how this compares to what might be expected for a natively folded protein. Quality can be examined by the overall energy of the model or by the Z-score. Z-scores are size dependent (Figure 5.17). However, it is possible to correct the Z-score so that size dependency is removed.

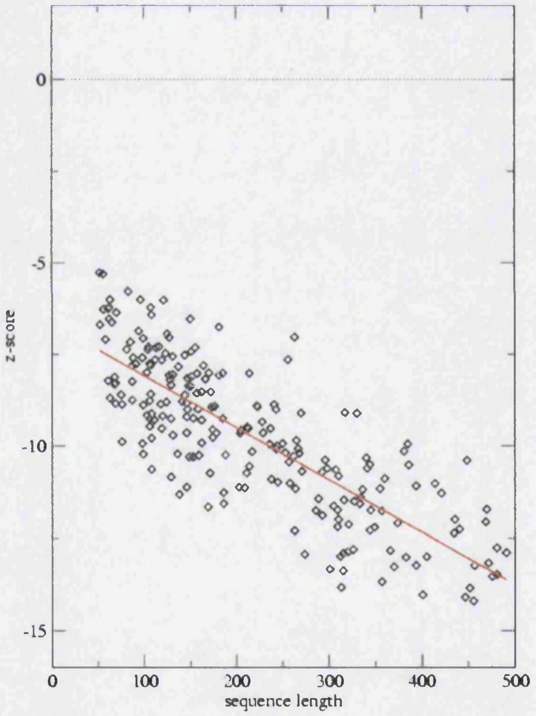


Figure 5.17: Correlation of protein size and ProsaII Z-score carried out on a set of known structures. Figure from the Prosa Manual. Regression equation: $y = -6.67 - 0.0141x$

The length dependency is directly proportional to the Z-Score (Equation 5.6):

$$y = -6.67 - 0.0141x \tag{5.6}$$

The Z-Scores were normalised with a rotation matrix so that all were plotted as if the length of the domain was zero (Equations 5.7 to 5.10).

$$a = arc\ \tan(0.0141) \tag{5.7}$$

Equation 5.7 gives the angle through which to rotate the values about the origin. In order to rotate through the origin, the value 6.67 must then be added to the y value. The point (x, y) is moved such that the straight line goes through the origin:

$$(x, y) \longrightarrow (x, y + 6.67) \tag{5.8}$$

then apply the rotation matrix:

$$\binom{x'}{y'} = \begin{pmatrix} \cos a & -\sin a \\ \sin a & \cos a \end{pmatrix} \binom{x}{y} \tag{5.9}$$

The final normalised Z-Score is then brought back down to the original position:

$$(x', y') \longrightarrow (x', y' - 6.67) \tag{5.10}$$

and the final Z-Score is y'.

Applying this rotation matrix to the data collected in this study normalises the Z-Score so that models can be compared with each other. Figure 5.18 shows that the normalisation adjusts the data to produce better correlation with chimeric sequence identity. All Z-Scores are considered as if the protein was zero residues in length.

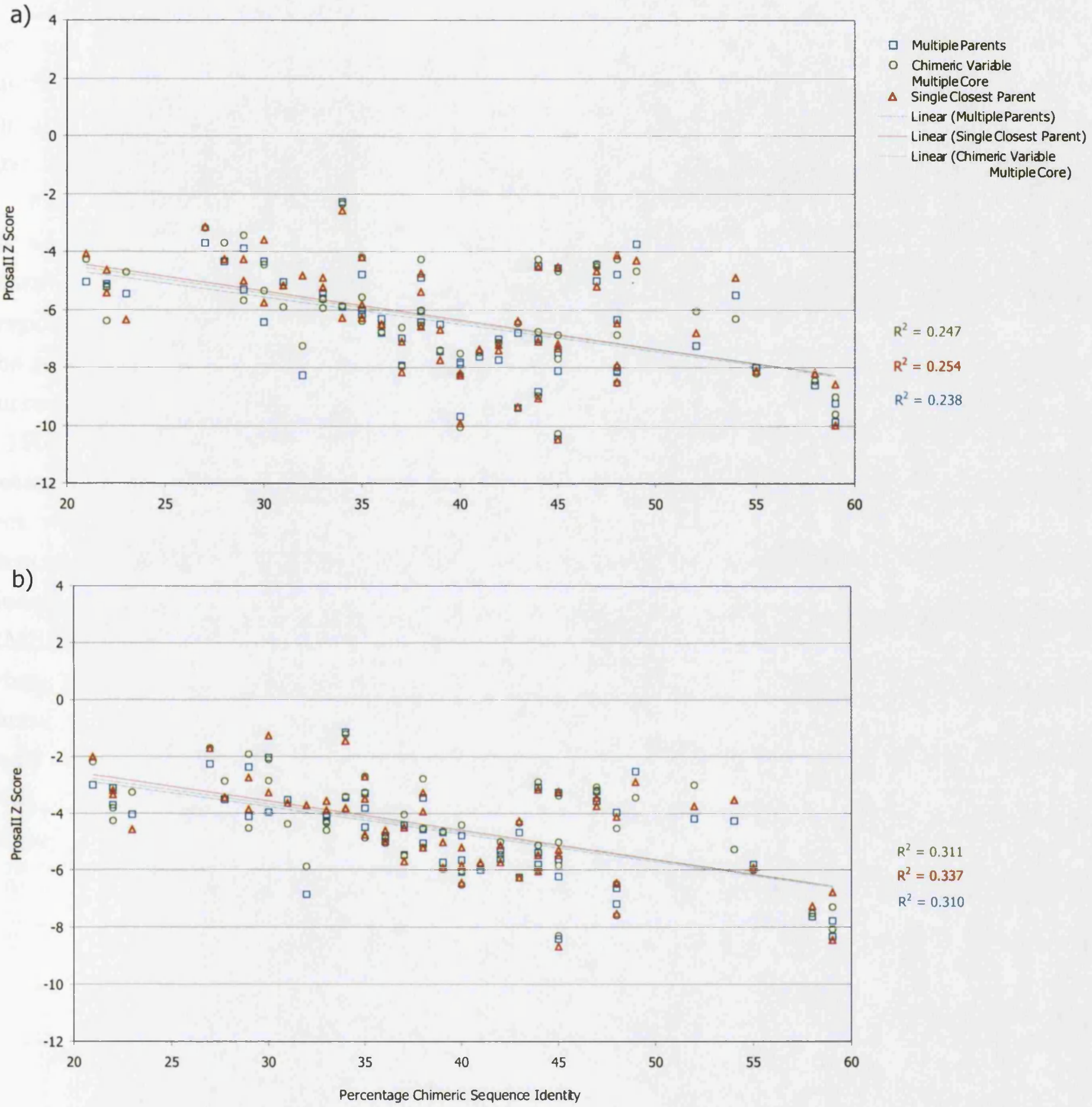


Figure 5.18: Plotting ProsaII with chimeric sequence identity for all models with >30% correctly aligned residues. Plot a. shows unnormalised data and plot b. shows the normalised data. This shows that normalising for size produces a slightly better correlation.

RMSD and LGScore provide an accurate way of measuring the quality of a model because they use the experimentally determined structure as a reference. However, a method for measuring model quality is still required when using GenMod to model sequences with unknown structure. ProsaII can be used for such a measurement, though first its performance must be correlated with RMSD and LGScore.

Figure 5.19a shows the relationship between the ProsaII Z-Score and RMSD for models built with GenMod. It can be seen that in many cases where ProsaII gives a good score (the model is predicted to form an energetically favourable conformation) the corresponding RMSD shows a poorer value indicating an inaccurate model. In other words, the resultant model, although calculated as energetically favourable by ProsaII is not the correct structure. The removal of pairs with <30% correctly aligned residues (Figure 5.19b) removes many of these poor models (as defined by RMSD) and the relationship between ProsaII Z-Score and RMSD becomes stronger. Figure 5.20 shows a similar effect when RMSD is replaced with the LGScore Q-Value. Again in Figure 5.20b those structures with <30% correctly aligned residues are removed producing a stronger relationship between the ProsaII Z-Score and LGScore Q-Value. In summary LGScore and RMSD are able to discriminate between good and bad representatives of the correct fold whereas ProsaII is only able to assess when the final model is energetically favourable and therefore unable to recognise when MODELLER has managed to create an energetically good model from a bad alignment. In general, although ProsaII clearly has limitations with some incorrectly predicted models it is still a valuable tool to measure accuracy for sequences with unknown structure.

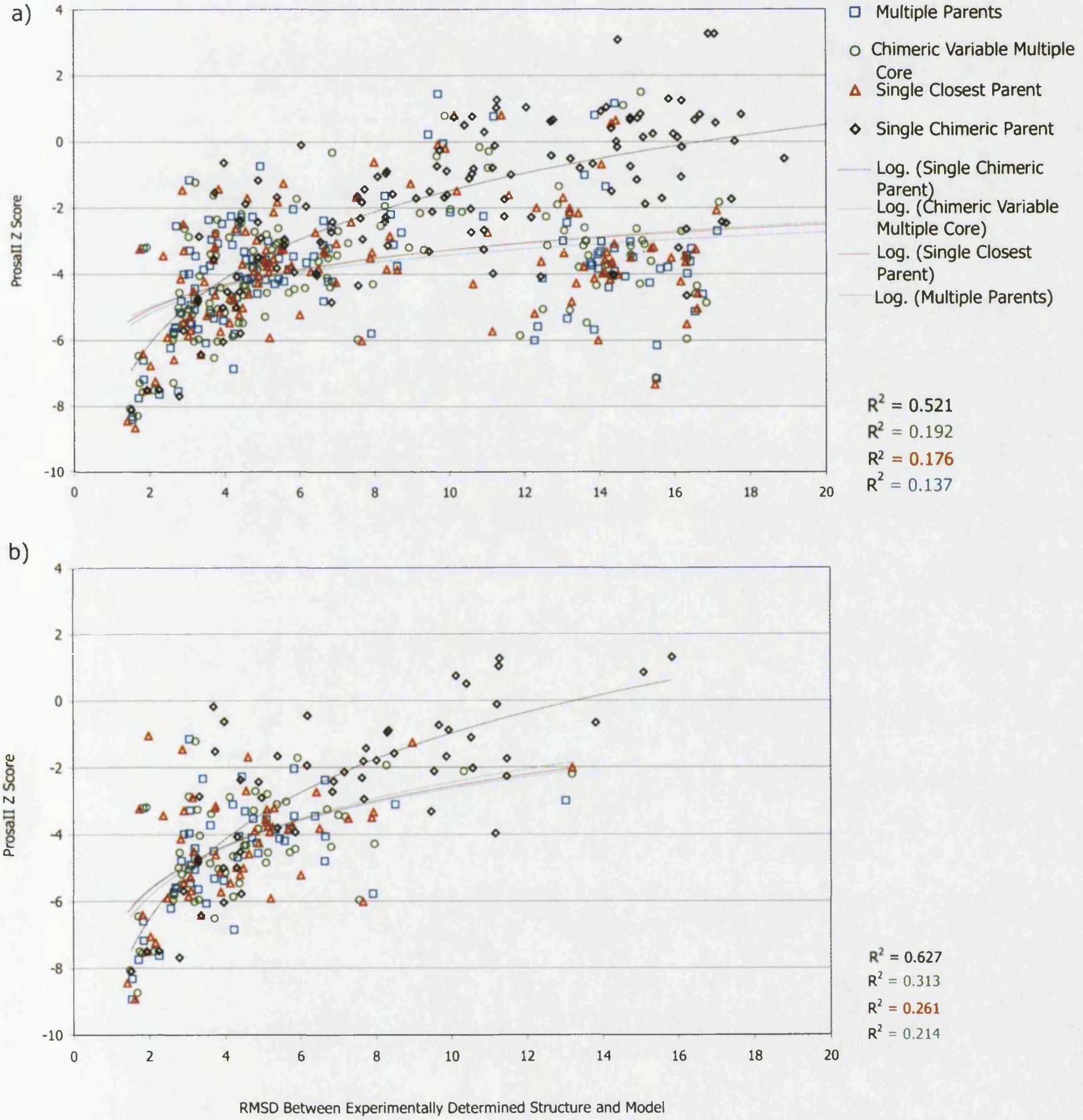


Figure 5.19: The relationship between ProsaII Z-Score and RMSD shows a correlation for correctly aligned structures. However this correlation is not as clear for all structures in the dataset (a).

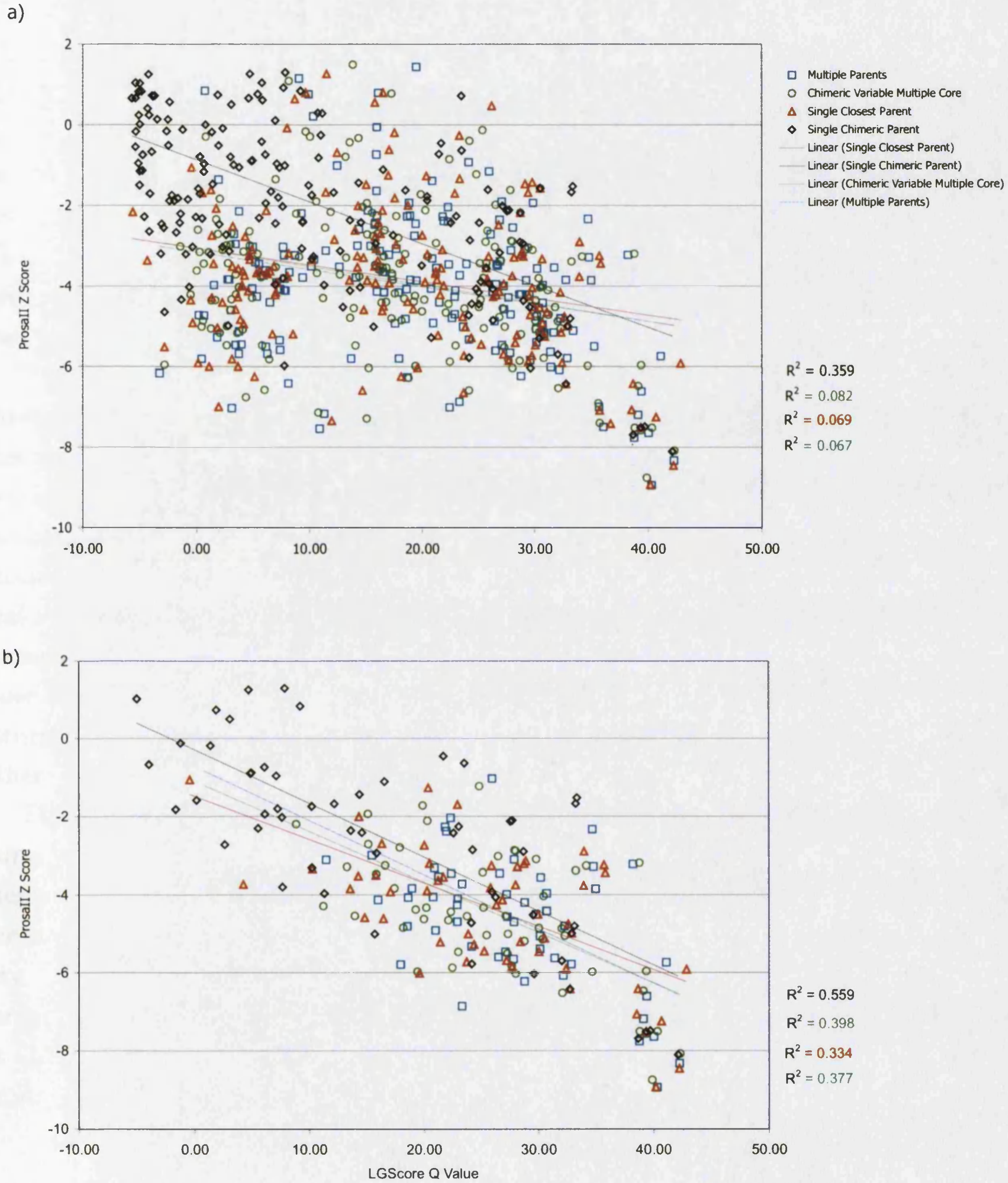


Figure 5.20: The relationship between LGScore with the ProsaII Z-Score reveals a good correlation at higher Q-Values but lower Q-Values exhibit a wide range of Z-Scores (a). When the badly aligned structures are removed from the dataset, the correlation improves (b).

5.3.8 Assessment of the Four Individual Methods for Modelling

The four individual modelling methods were used to predict the structural models of 140 CATH domains for sequences of varying similarity (20–60% chimeric sequence identity). For 56 of these, the alignment to the SAMOSA models resulted in >30% correctly aligned residues. It has been shown that above this cut–off (>30% percentage correctly aligned residues) LGScore, RMSD and ProsaII all give a reliable measurement of model quality. These methods were therefore used to study the individual model quality produced by each of the four methods (Single Closest Parent, Multiple Parents, Single Chimeric Parent and Chimeric Variable Multiple Core).

Figure 5.21 provides a summary of the performance of each method measured by the three assessment measurements. The number of times each of the four types of modelling methods produced the best model was calculated, according to the three measurements. Scoring with LGScore suggests that the Multiple Parent method outperforms all other methods at every sequence identity (Figure 5.21a). The LGScore measures the highest scoring discontinuous fragment, leaving any areas which are particularly dissimilar to the real structure, out of the measurement. RMSD is a global measurement and takes into account the whole model. Measuring with RMSD assigns the Chimeric Variable Multiple Core method as the best method more frequently (Figure 5.21b). This suggests that the structurally variable regions in these models have been modelled more successfully than other methods assessed.

Figure 5.22 divides the RMSD scores measured from model and experimentally determined structure into four separate histograms associated with different chimeric sequence identity ranges, 20–30%, 31–40%, 41–50% and 51–60%. It shows that the majority of the targets with >30% correctly aligned residues possess 31–40% chimeric sequence identity. Figure 5.23 plots the RMSD of the most successful modelling method, for each target sequence. It shows that structures modelled using multiple parents were the best at chimeric sequence identites below 50% whilst above 50% chimeric sequence identity, models created using the single closest parent also feature.

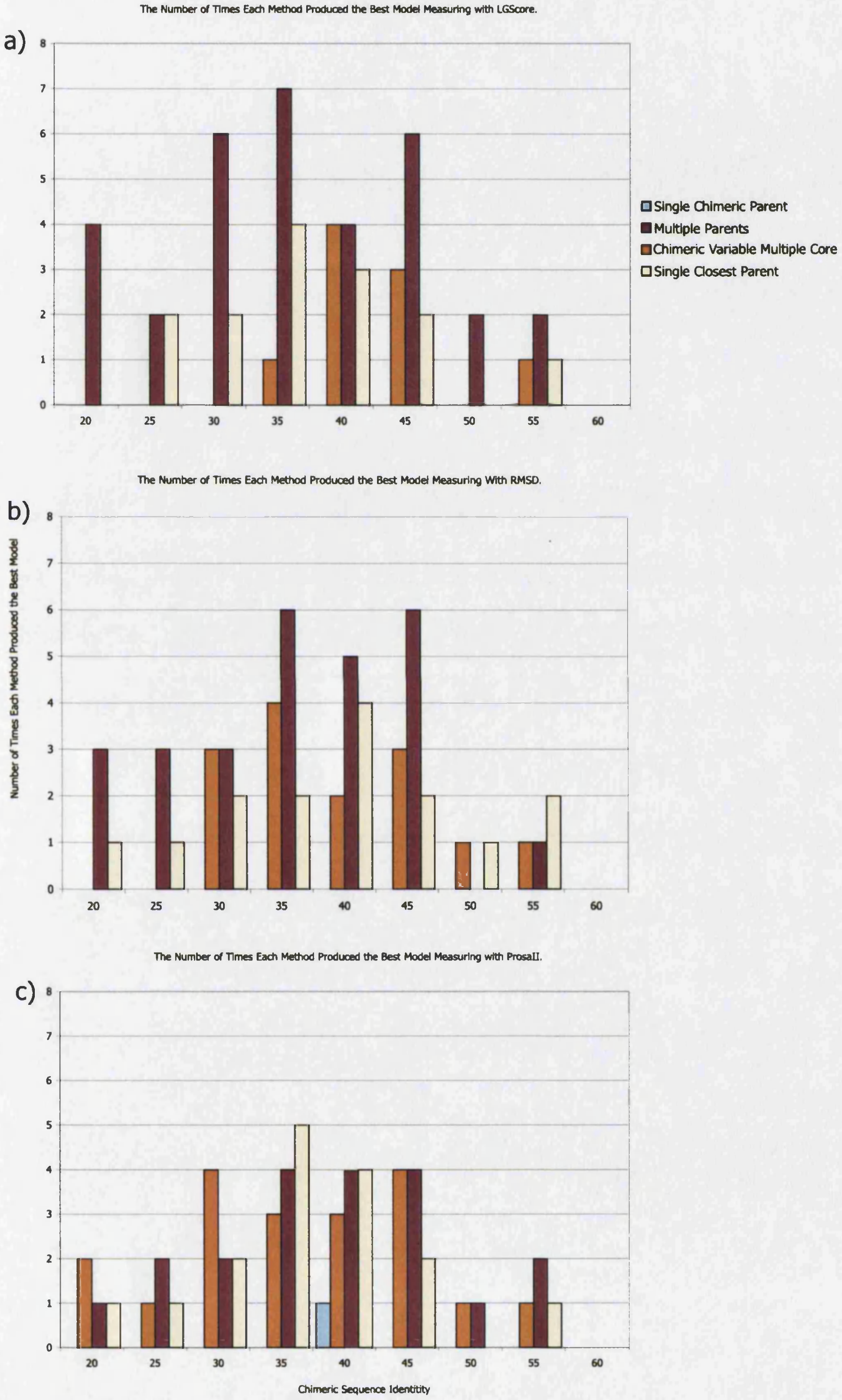


Figure 5.21: The histograms show how many times each of the four methods produced the best model at each sequence identity. The LGScore Q-Value (a), RMSD (b) and ProsaII Z-Score (c) were used as the methods of assessment.

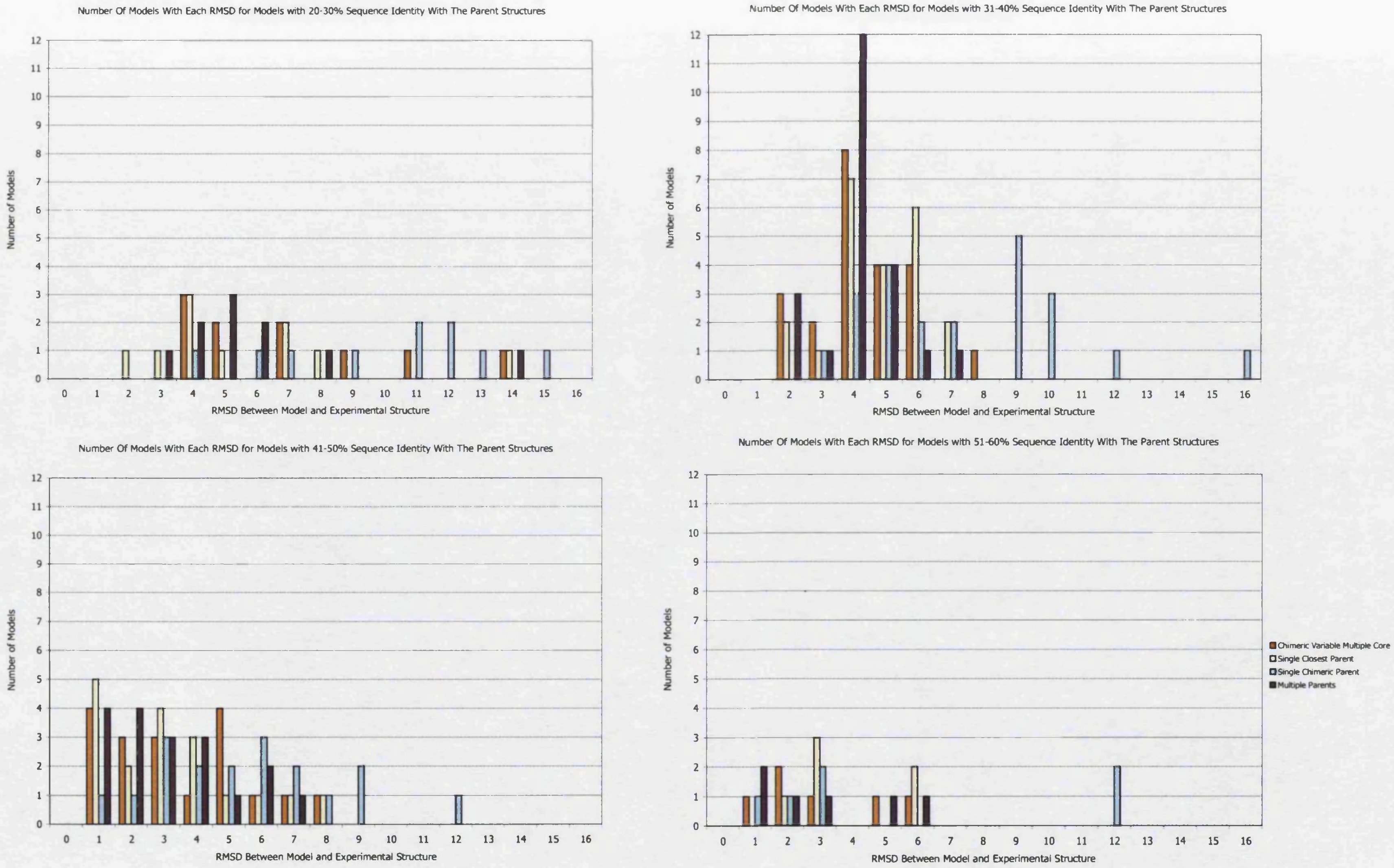


Figure 5.22: Root Mean Square Deviation between model and experimental structure for chimeric sequence identities of 20–60%. The sequence identities are separated into four graphs, models with 20-30% sequence identity to their parent structures, 31-40%, 41-50% and 51-60%. Models are only considered if they have 30% or more correctly aligned residues.

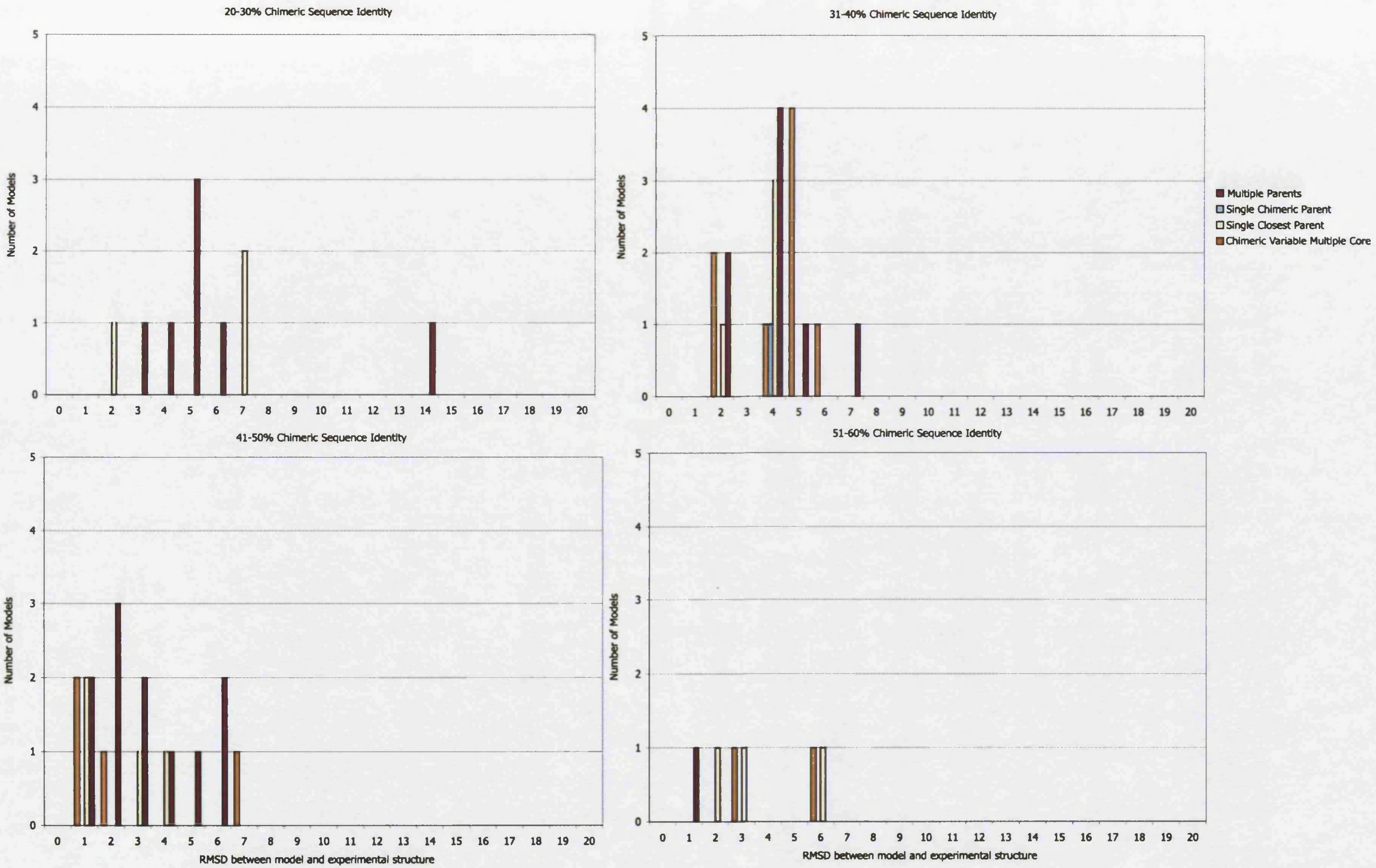


Figure 5.23: Root Mean Square Deviation between model and experimental structure for chimeric sequence identities of 20–60%. The RMSD is plotted for the method which creates the best model only.

5.3.8.1 Reviewing the Single Chimeric Parent Method

Figures 5.14 to 5.16 demonstrate that the Single Chimeric Parent method is the least successful of all the methods assessed. In fact, the Single Chimeric Parent only features once as the best model as measured by ProsaII (Figure 5.21c) and many of the models have high RMSD. This could be due to the gaps or junctions that may exist between adjacent parent segments, making it difficult to create a continuous structure. Many of the models created from single chimeric structures showed a lack of defined secondary structure and instead they appeared to be made up of large expanses of coil, suggesting that the restraints in MODELLER were not defined well enough, a possible effect caused by the junctions between each segment.

5.3.8.2 Reviewing the Use of the Single Closest Parent and Multiple Parent Methods

As the structural distance between the experimentally determined structure and its closest relative increases, so does the usefulness of the Multiple Parent procedure, producing a slightly better model. At greater structural similarity between experimental structure and its closest relative, models using single parents are increasingly better. Figure 5.24 shows that in general, at above 50% sequence identity, the use of the Single Closest Parent method should be considered although there is not a significant difference.

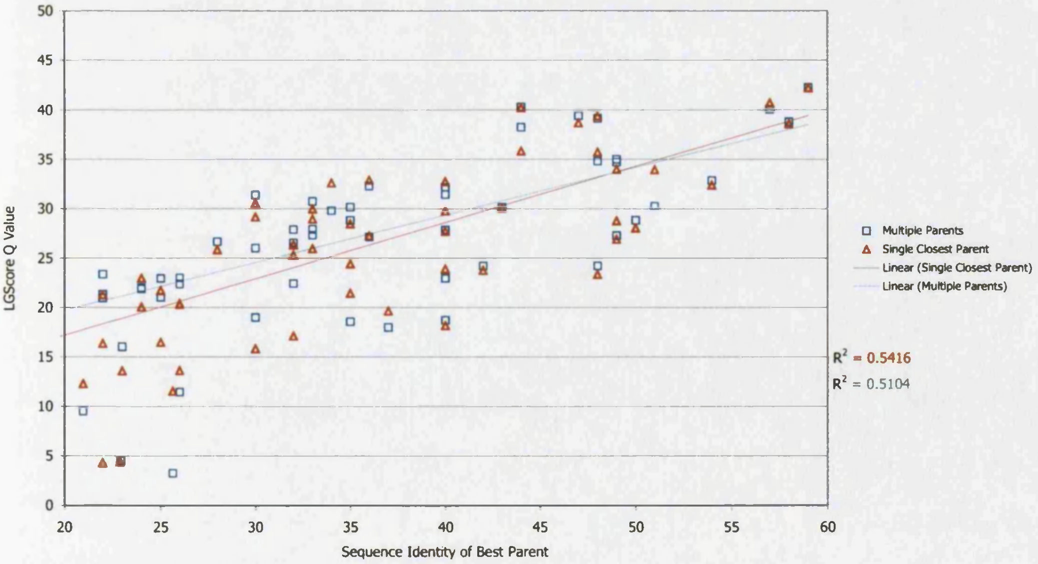


Figure 5.24: When sequence identity between the best parent and model is low, multiple parents generate better models but as the closest parent becomes more similar to the target, the quality of the single parent models begin to be slightly more successful than the multiple parents.

Figure 5.25 shows that as the structural diversity between closest parent and experimentally determined structure decreases, the quality of the single parent model increases at a faster rate than the multiple parent model (a steeper gradient to the line). However, the models created using Multiple Parents are better quality overall.

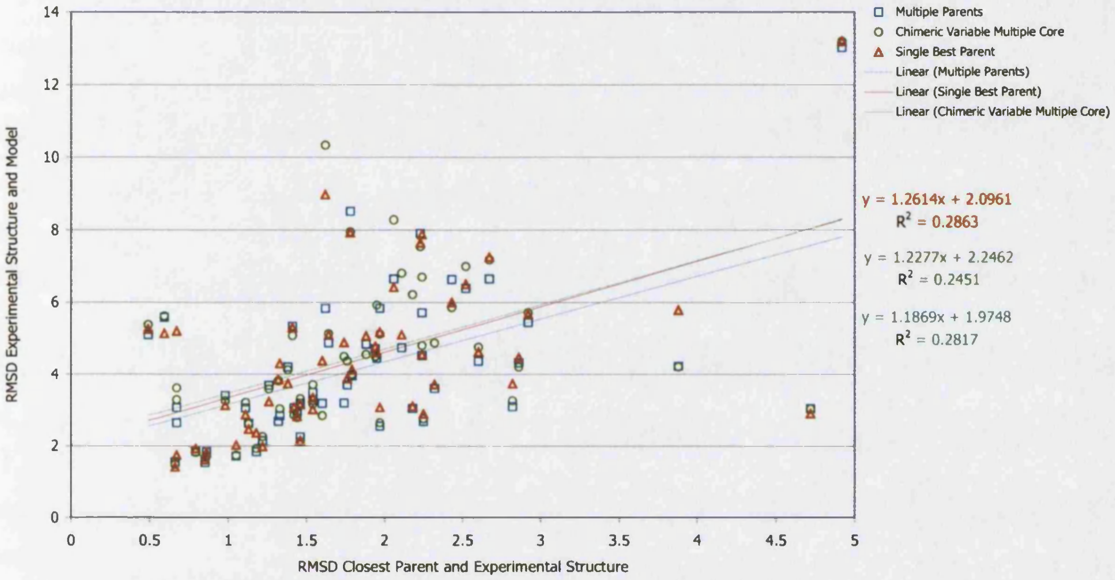


Figure 5.25: The RMSD between the experimental structure and its closest relative versus the RMSD between the experimental structure and its model. As structures become more diverse from their closest parent it is more effective to use multiple parents to create a model.

5.3.8.3 Assessing the Use of a Chimeric Parent in the Variable Regions

The chimeric parent was designed to take advantage of any local areas of similarity between any of the parents and the target sequence. Generally, the models are better when all positions in the alignment are considered rather than creating a chimeric structure in the variable regions (Figure 5.25). Using the Multiple Core Chimeric Variable method becomes slightly more effective at lower structural diversity between closest parent and experimentally determined structure, that is, when the closest parent becomes more similar to the model. This is because variable regions in the closest parent are then more similar to the experimentally determined structure. The implementation of a score cut–off would help in situation where there are no structures which match with high enough similarity. In these cases all of the structures could be used to model the variable region instead. Additionally, if the variable region is large enough to contain secondary structures, a secondary structure prediction and alignment method could indicate which parent should be used in that particular region.

Figure 5.25 shows that in many cases the Chimeric Variable Multiple Core and Single Closest Parent methods produced models with a very similar RMSD relative to their experimental structure. This may well be expected as the Chimeric Variable Multiple Core may often select the segments from the closest parent. In such cases, the quality of the Single Closest Parent model and the Chimeric Variable Multiple Core model would be either identical or may even produce a better model i.e. if there is a segment which is closer to the experimentally determined structure than to the closest parent. An example of a successful model created with the Chimeric Variable Multiple Core method is shown in Figure 5.26. The region on the four parent structures, shown in green, highlights a particularly variable region. The Chimeric Variable Multiple Core method has predicted a much closer structure for this region than the Multiple Parent model.

If selection of a particular parent depends on the sequence, the correct region of the sequence must be used to make the selection which means that the alignment between target and parents must be accurate. The lower the sequence identities and greater the structural diversity between parent structures and target, the more the chimeric parent methods will benefit the model, but, there may also be a corresponding loss in alignment accuracy.

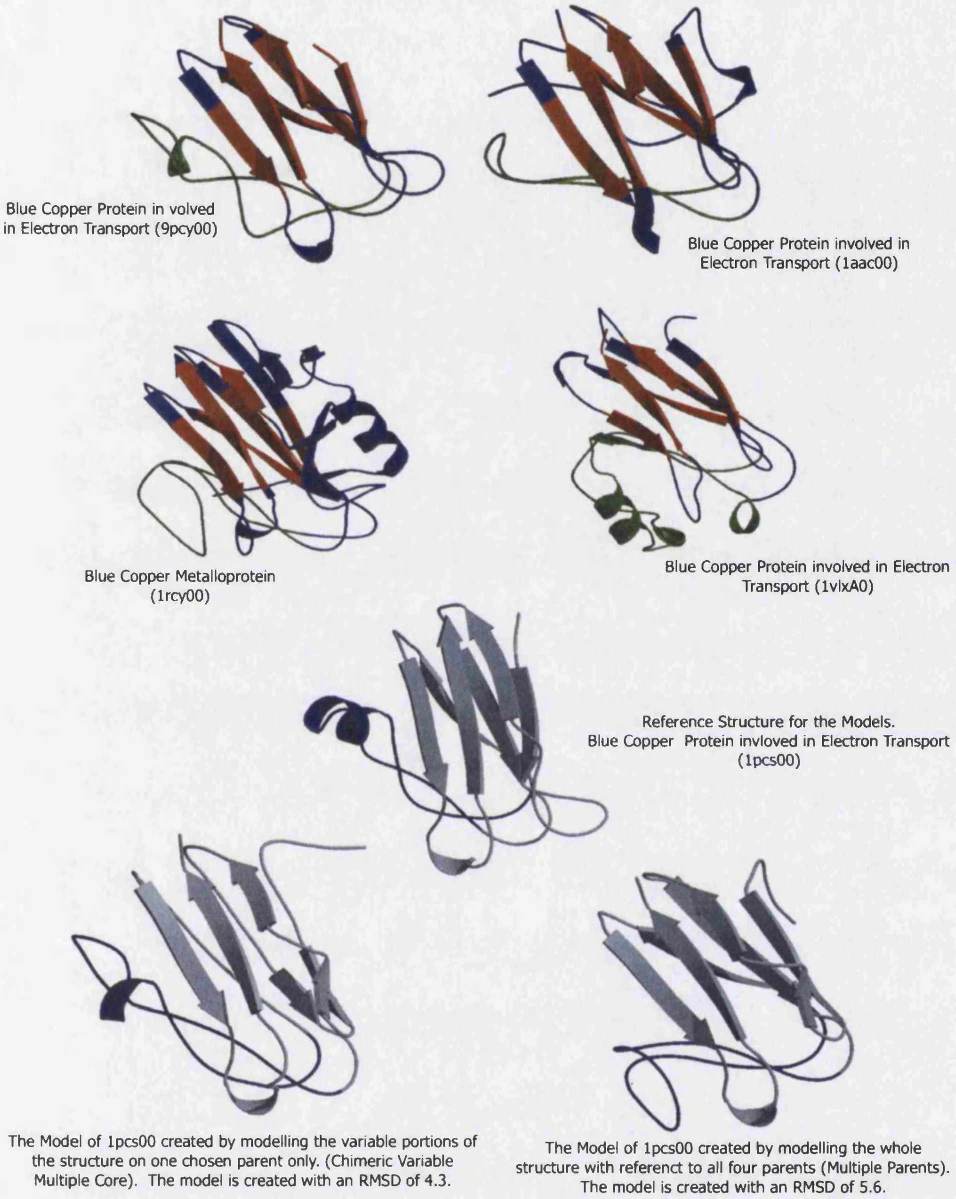

Blue Copper Protein in volved
in Electron Transport (9pcy00)
Blue Copper Protein involved in
Electron Transport (1aac00)
Blue Copper Metalloprotein
(1rcy00)
Blue Copper Protein involved in Electron
Transport (1vlxA0)
Reference Structure for the Models.
Blue Copper Protein invloved in Electron Transport
(1pcs00)
The Model of 1pcs00 created by modelling the variable portions of
the structure on one chosen parent only. (Chimeric Variable
Multiple Core). The model is created with an RMSD of 4.3.
The Model of 1pcs00 created by modelling the whole
structure with referenct to all four parents (Multiple Parents).
The model is created with an RMSD of 5.6.

Figure 5.26: The diagram shows the modelling of the blue copper protein involved in electron transport (1pcs00) with 49% chimeric sequence identity. The parent structures are shown in red and blue. The red indicates those regions, identified by Findcore, that are less than 3 Å from equivalent atoms when superposed. The blue regions indicate regions of variability. The experimentally determined structure, 1pcs00, is shown in blue and below, also in blue, are two of the models created. Bottom right is the model created using Multiple Parents. Bottom left is the model created by using multiple parents for the core regions of the structures but selecting the most similar parent for each variable region individually. The most variable region on the the parents is indicated in green and the modelling of this region is indicated in dark blue. Using multiple parents in this region has averaged out the diversity in the structure, producing an inaccurate result. Selecting one single parent to model this variable region has produced a more accurate result. The parent this variable region is modelled on is 9pcy00.

5.4 Conclusions

5.4.1 Providing Models with Quality Evaluation

A result of this analysis has been the development of a protocol for the analysis of model quality. A future goal is to model protein structures for whole genomes systematically. The ability to give an indication of model quality will be an important factor in the success of these project. A model quality profile could include the following factors:

1. The sequence identity between aligned target sequence and parents will indicate the likely overall quality of the model. This can also be an indication on whether to accept the model created with multiple parents or the closest single parent.

2. Methods which use a reference structure for their analysis, LGScore and RMSD, can be used to profile the family from which the structure has been modelled. Variable SSGs will reveal a profile of low LGScores and high RMSDs

3. A Findcore analysis could be used to highlight those regions defined as core ($\leq$3Å). Also the percentage of the structural parents represented as core can indicate the structural variability present between the parents.

4. ProsaII Z-Scores and Energy plots can reveal which areas of the model may be problematic.

From this information, users would be able to detect if the model is good enough for their needs. For example, if the user is most interested in the active site of the protein and has identified key residues on the sequence and mapped them to the structure they can see if these regions are in a structurally sound region of the model.

This chapter has included the following:

- GenMod: the development of a pipeline for automatic comparative modelling using the alignments from SAM–T99 profiles and the comparative modelling program MODELLER has been presented.

- An analysis of this method on a large dataset of 140 sequences from structurally diverse families has been undertaken.

- Three assessment methods, RMSD, LGScore and ProsaII have been used to benchmark the results.

- Four different parent selection methods have been analysed.

- Areas for improvement in the automatic comparative modelling pipeline have been identified, notably alignment quality.

This study selected structurally diverse CATH superfamilies in order to assess model building. The results, therefore, describe the quality of models which could be produced for diverse superfamilies in CATH, highlighting a number of key areas which need to be addressed.

The quality of the alignments need to be improved, throughout the range of chimeric sequence identities studied (20–60%) there were some sequences which aligned well, despite having <30% chimeric sequence identity. This suggests that, in some families, there are some highly conserved anchor residues which help to align sequences, whereas, in other families, these common residues are not so prominent. A study on different sequence alignment methods would be the next step. It may be the case that different methods should be used at different sequence identities.

Chapter 6

Conclusions

The work in this thesis has provided insights into structural variability within CATH domain superfamilies. The CATH domain database provides clusters of homologous domains, sharing significant sequence, functional and/or structural similarity at the H-level. Secondary structural similarity between members of the same homologous superfamily is shared in the core arrangement and connection of these elements, whilst in some superfamilies there may also be additions or subtractions of peripheral secondary structures. Chapter 2 introduced a novel tool (2DSEC) for the visualisation and analysis of these secondary structure insertions. 2DSEC provides a description of the consensus secondary structures found across the domains within a superfamily and the location and extent of any secondary structure insertions or embellishments. 2DSEC enabled the identification of superfamilies with extensive secondary structural embellishments and the characterisation of the embellished positions within the peptide chain and locations on the three-dimensional structure.

The analysis in Chapter 3 sought to characterise the types of structural variation found in CATH homologous superfamilies. Firstly, the relationship between sequence and structure was measured using the SSAP structural alignment program and calculation of percentage sequence identity. The results were in agreement with Chothia & Lesk (1986), finding that there are two trends. Below 25–35% sequence identity the amount of structural change to sequence change is much greater than above this threshold. Above 25–35% the change is more gradual and linear. However, this rate of structural change to sequence change (structural mutation sensitivity, SMS) is dependent on the superfamily. For example, this rate of change ranged from from 0.19 in the most versatile superfamilies to 0.06 in the most conserved. These findings are in agreement with those given by Wood & Pearson (1999). Dividing these data into the structural classes (mainly α, mainly β, and $\alpha\beta$) revealed the mainly β class to be less tolerant to structural change in the region of >35% sequence identity, with fewer SMS values above the mean value.

A continuation of the work undertaken by Pascarella & Argos (1992) was presented in Chapter 3. Although this analysis was carried out on a much larger dataset, it revealed similar findings. It was observed that indels prefer to be short, about six residues in 0%–20% sequence identitiy, four residues in 20%–40% sequence identitiy and two residues in 40%–95% sequence identitiy. The percentage of indels more than 10 residues was less than 10 % at all sequence identities for most structural pairs. These observations may provide a useful adjunct to traditional scoring schemes used to calculate gap penalties. This would be especially useful for the alignment of target sequences to parent structures in comparative modelling.

The study was then extended to examining superfamilies with insertions of whole secondary structure units as identified by 2DSEC. The superfamilies adopting $\alpha\beta$ sandwich architectures appear to be more tolerant to structural change than other superfamilies. It is possible that the high number of these superfamilies that are identified as being structurally variable could be attributable to a biased dataset since over 30% of N95Reps in the database are $\alpha\beta$ sandwich architectures. However, it may also be because they are more tolerant to structural change, and therefore have been modified during evolution to fulfil more functions. Additionally, it was found that whilst secondary structure embellishments are often inserted into a number of non-localised regions in the peptide chain they are often co-localised on the three-dimensional structure.

These variations in structure between domains of the same homologous superfamily may provide a useful insight into their differing functional roles or protein-protein interactions. Therefore, a method for identifying those core secondary structures and those secondary structures which are only present in one or a few superfamily members is key to the elucidation of the structural evolution of the superfamily. In Chapter 4 a protocol was developed for examining these secondary structural insertions in three homologous superfamilies: the cupredoxin superfamily, the ATP-grasp superfamily and the thioredoxin-like superfamily. The analysis combined the identification of the secondary structural embellishments present in the superfamily with functional information and information on the orientation of the domains and subunits within the whole biological unit of each superfamily. It was found that the embellishments were often located in the regions between the domains or subunits and were involved in promoting the domain-domain or subunit-subunit interactions. This is the case for the multi copper oxidase cluster of the cupredoxins, for biotin carboxylase and phosphoribosylaminoimidazole carboxylase of the ATP-grasp superfamily and glutathione peroxidase of the thioredoxin-like superfamily. Additionally, in some cases these embellishments were found to directly modify the geometry of the active site. In the ATP-grasp superfamily, those members containing a large

C-terminal embellishment have enclosed active sites.

Chapters 3 and 4 have provided a method for identifying and locating secondary structural embellishments and identifying possible functional implications for their presence. It was found that, often, secondary structural embellishments promote interactions between domains or subunit. The identification of these secondary structural embellishments in superfamilies could be used to help predict the locations of domain–domain or subunit–subunit interactions. Various steps of the protocol could now be automated and used for all structurally embellished superfamilies identified in Chapter 3. For example, KDTree, the algorithm used to measure the proximity of the residues in the embellished secondary structures to other domains and subunits present in the biological unit, could be used on the whole dataset. If active site residues are known for the structurally embellished superfamilies, KDTree could also be used to measure the proximity of the residues in the embellished secondary structures to the active site. A future goal could incorporate the use of GO annotations (Ashburner *et al.*, 2000) to correlate embellishments and changes in function within a superfamily.

The second part of this thesis focused on the development of an automatic comparative modelling pipeline, GenMod. Parents were selected using CATH Structural Sub-Groups (SSGs) and sequences aligned by the SAM-T99 software were modelled using MODELLER. Three methods were used to test the quality of the models, RMSD, LGScore and ProsaII. In addition, a novel method for the selection of template structures was assessed. The parent structures were divided into structurally variable and structurally conserved regions and each region was modelled based on the closest parent for that region (selected on the basis of sequence similarity). Although the method had limitations, it has provided a stepping stone to a number of other possible ideas. Findcore has provided a method of selecting SVRs and SCRs and more rigorous methods of SVR selection, such as fragment databases, which could be built into GenMod.

Methods for identifying and measuring structural variability have been developed and it was found that superfamilies differ in their structural variability. A subset of structurally embellished superfamilies were identified. Methods for measuring residue and secondary structure insertions and deletions present in CATH superfamilies have been developed. However, structural variability could be measured in a number of different ways. To increase the amount of information on the structural variability of the selected CATH superfamilies, the angles between secondary structures and the lengths of the loops between each secondary structure element could be taken into account, providing a total view of the structural variability present in that superfamily. These measurements could then be used as a profile to identify not only the most variable superfamilies, but in which ways these superfamilies vary. Such a profile may include information on secondary structure insertions, shifts in the orientations of the secondary structures or variability in loop lengths. This could be used to create a profile of structural variability specific to a particular superfamily so that it could be used when modelling new members of that superfamily. Work in this thesis has shown that structural variation is not consistent across all superfamilies. Comparative modelling relies on general sequence/structure rules that are thought to apply to all superfamilies. It would be useful to tailor comparative modelling approaches to use superfamily-specific information as described above.

Bibliography

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

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

Anfinsen, C. (1973). Principles that govern the folding of protein chains. *Science*, **181**, 223–30.

Aronson, H., Royer, W., Jr & Hendrickson, W. (1994). Quantification of tertiary structural conservation despite primary sequence drift in the globin fold. *Protein Sci*, **3**, 1706–11.

Artymiuk, P., Poirrette, A., Rice, D. & Willett, P. (1996). Biotin carboxylase comes into the fold. *Nat Struct Biol*, **3**, 128–32.

Ashburner, M., Ball, C., Blake, J., Botstein, D., Butler, H., Cherry, J., Davis, A., Dolinski, K., Dwight, S., Eppig, J., Harris, M., Hill, D., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J., Richardson, J., Ringwald, M., Rubin, G. & Sherlock, G. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*, **25**, 25–9.

Attwood, T. (2002). The PRINTS database: a resource for identification of protein families. *Brief Bioinform*, **3**, 252–63.

Baker, D. & Sali, A. (2001). Protein structure prediction and structural genomics. *Science*, **294**, 93–6.

Barton, G. (2002). The OC Tree Program. http://www.compbio.dundee.ac.uk/Software/OC/oc.html.

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

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

Bates, P. & Sternberg, M. (1999). Model building by comparison at CASP3: Using expert knowledge and computer automation. *Proteins*, **37**, 47–54.

Bates, P., Jackson, R. & Sternberg, M. (1997). Model building by comparison: a combination of expert knowledge and computer automation. *Proteins*, **Suppl 1**, 59–67.

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

Bernstein, F., Koetzle, T., Williams, G., Meyer, E., Jr, Brice, M., Rodgers, J., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). The Protein Data Bank. A computer-based archival file for macromolecular structures. *Eur J Biochem*, **80**, 319–24.

Botuyan, M., Toy-Palmer, A., Chung, J., Blake, R., 2nd, Beroza, P., Case, D. & Dyson, H. (1996). NMR solution structure of Cu(I) rusticyanin from Thiobacillus ferrooxidans: structural basis for the extreme acid stability and redox potential. *J Mol Biol*, **263**, 752–67.

Boutselakis, H., Dimitropoulos, D., Fillon, J., Golovin, A., Henrick, K., Hussain, A., Ionides, J., John, M., Keller, P., Krissinel, E., McNeil, P., Naim, A., Newman, R., Oldfield, T., Pineda, J., Rachedi, A., Copeland, J., Sitnov, A., Sobhany, S., Suarez-Uruena, A., Swaminathan, J., Tagari, M., Tate, J., Tromm, S., Velankar, S. & Vranken, W. (2003). E-MSD: the European Bioinformatics Institute Macromolecular Structure Database. *Nucleic Acids Res*, **31**, 458–62.

Branden, J., C; Tooze (1999). *Introduction to Protein Structure*. Garland.

Brown, K., Tegoni, M., Prudencio, M., Pereira, A., Besson, S., Moura, J., Moura, I. & Cambillau, C. (2000). A novel type of catalytic copper cluster in nitrous oxide reductase. *Nat Struct Biol*, **7**, 191–5.

Burke, D., Deane, C. & Blundell, T. (2000). Browsing the SLoop database of structurally classified loops connecting elements of protein secondary structure. *Bioinformatics*, **16**, 513–9.

Choi, H., Kang, S., Yang, C., Rhee, S. & Ryu, S. (1998). Crystal structure of a novel human peroxidase enzyme at 2.0 A resolution. *Nat Struct Biol*, **5**, 400–6.

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

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

Chothia, C., Levitt, M. & Richardson, D. (1977). Structure of proteins: packing of alpha-helices and pleated sheets. *Proc Natl Acad Sci U S A*, **74**, 4130–4.

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

Collins, F., Patrinos, A., Jordan, E., Chakravarti, A., Gesteland, R. & Walters, L. (1998). New goals for the U.S. Human Genome Project: 1998-2003. *Science*, **282**, 682–9.

Copley, R. & Bork, P. (2000). Homology among (betaalpha)(8) barrels: implications for the evolution of metabolic pathways. *J Mol Biol*, **303**, 627–41.

Cristobal, S., Zemla, A., Fischer, D., Rychlewski, L. & Elofsson, A. (2001). A study of quality measures for protein threading models. *BMC Bioinformatics*, **2**, 5.

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

Di Gennaro, J., Siew, N., Hoffman, B., Zhang, L., Skolnick, J., Neilson, L. & Fetrow, J. (2001). Enhanced functional annotation of protein sequences via the use of structural descriptors. *J Struct Biol*, **134**, 232–45.

Doolittle, W. & Brown, J. (1994). Tempo, mode, the progenote, and the universal root. *Proc Natl Acad Sci U S A*, **91**, 6721–8.

Drickamer, K. (1999). C-type lectin-like domains. *Curr Opin Struct Biol*, **9**, 585–90.

Ducros, V., Brzozowski, A., Wilson, K., Ostergaard, P., Schneider, P., Svendson, A. & Davies, G. (2001). Structure of the laccase from Coprinus cinereus at 1.68 A resolution: evidence for different 'type 2 Cu-depleted' isoforms. *Acta Crystallogr D Biol Crystallogr*, **57**, 333–6.

Dunbrack, R., Jr & Cohen, F. (1997). Bayesian statistical analysis of protein side-chain rotamer preferences. *Protein Sci*, **6**, 1661–81.

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

Esser, L., Wang, C., Hosaka, M., Smagula, C., Sudhof, T. & Deisenhofer, J. (1998). Synapsin I is structurally similar to ATP-utilizing enzymes. *EMBO J*, **17**, 977–84.

Fan, C., Moews, P., Walsh, C. & Knox, J. (1994). Vancomycin resistance: structure of D-alanine:D-alanine ligase at 2.3 A resolution. *Science*, **266**, 439–43.

Fan, C., Moews, P., Shi, Y., Walsh, C. & Knox, J. (1995). A common fold for peptide synthetases cleaving ATP to ADP: glutathione synthetase and D-alanine:d-alanine ligase of Escherichia coli. *Proc Natl Acad Sci U S A*, **92**, 1172–6.

Fidelis, K., Stern, P., Bacon, D. & Moult, J. (1994). Comparison of systematic search and database methods for constructing segments of protein structure. *Protein Eng*, **7**, 953–60.

Fischer, D., Elofsson, A., Rychlewski, L., Pazos, F., Valencia, A., Rost, B., Ortiz, A. & Dunbrack RL, J. (2001). CAFASP2: the second critical assessment of fully automated structure prediction methods. *Proteins*, **Suppl 5**, 171–83.

Fiser, A., Do, R. & Sali, A. (2000). Modeling of loops in protein structures. *Protein Sci*, **9**, 1753–73.

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

Fraser, M., James, M., Bridger, W. & Wolodko, W. (1999). A detailed structural description of Escherichia coli succinyl-CoA synthetase. *J Mol Biol*, **285**, 1633–53.

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

Gaudet, R., Savage, J., McLaughlin, J., Willardson, B. & Sigler, P. (1999). A molecular mechanism for the phosphorylation-dependent regulation of heterotrimeric G proteins by phosducin. *Mol Cell*, **3**, 649–60.

Gerstein, M. & Altman, R. (1995a). Average core structures and variability measures for protein families: application to the immunoglobulins. *J Mol Biol*, **251**, 161–75.

Gerstein, M. & Altman, R. (1995b). Using a measure of structural variation to define a core for the globins. *Comput Appl Biosci*, **11**, 633–44.

Godden, J., Turley, S., Teller, D., Adman, E., Liu, M., Payne, W. & LeGall, J. (1991). The 2.3 angstrom X-ray structure of nitrite reductase from Achromobacter cycloclastes. *Science*, **253**, 438–42.

Greer, J. (1981). Comparative model-building of the mammalian serine proteases. *J Mol Biol*, **153**, 1027–42.

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

Grishin, N. (2001). Fold change in evolution of protein structures. *J Struct Biol*, **134**, 167–85.

Gronwald, W., Loewen, M., Lix, B., Daugulis, A., Sonnichsen, F., Davies, P. & Sykes, B. (1998). The solution structure of type II antifreeze protein reveals a new member of the lectin family. *Biochemistry*, **37**, 4712–21.

Guddat, L., Bardwell, J. & Martin, J. (1998). Crystal structures of reduced and oxidized DsbA: investigation of domain motion and thiolate stabilization. *Structure*, **6**, 757–67.

Guss, J., Merritt, E., Phizackerley, R. & Freeman, H. (1996). The structure of a phytocyanin, the basic blue protein from cucumber, refined at 1.8 A resolution. *J Mol Biol*, **262**, 686–705.

Hammann, C., Messerschmidt, A., Huber, R., Nar, H., Gilardi, G. & Canters, G. (1996). X-ray crystal structure of the two site-specific mutants Ile7Ser and Phe110Ser of azurin from Pseudomonas aeruginosa. *J Mol Biol*, **255**, 362–6.

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

Hart, P., Nersissian, A., Herrmann, R., Nalbandyan, R., Valentine, J. & Eisenberg, D. (1996). A missing link in cupredoxins: crystal structure of cucumber stellacyanin at 1.6 A resolution. *Protein Sci*, **5**, 2175–83.

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

Henrick, K. & Thornton, J. (1998). PQS: a protein quaternary structure file server. *Trends Biochem Sci*, **23**, 358–61.

Heringa, J. & Taylor, W. (1997). Three-dimensional domain duplication, swapping and stealing. *Curr Opin Struct Biol*, **7**, 416–21.

Hirotsu, S., Abe, Y., Okada, K., Nagahara, N., Hori, H., Nishino, T. & Hakoshima, T. (1999). Crystal structure of a multifunctional 2-Cys peroxiredoxin heme-binding protein 23 kDa/proliferation-associated gene product. *Proc Natl Acad Sci U S A*, **96**, 12333–8.

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

Holm, L. & Sander, C. (1997). New structure–novel fold? *Structure*, **5**, 165–71.

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

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

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

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

Jones, T., Zou, J., Cowan, S. & Kjeldgaard (1991). Improved methods for building protein models in electron density maps and the location of errors in these models. *Acta Crystallogr A*, **47 (Pt 2)**, 110–9.

Kabsch, W. & Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–637.

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

Katti, S., LeMaster, D. & Eklund, H. (1990). Crystal structure of thioredoxin from Escherichia coli at 1.68 A resolution. *J Mol Biol*, **212**, 167–84.

Katti, S., Robbins, A., Yang, Y. & Wells, W. (1995). Crystal structure of thioltransferase at 2.2 A resolution. *Protein Sci*, **4**, 1998–2005.

Kelley, L. & Sutcliffe, M. (1997). OLDERADO: on-line database of ensemble representatives and domains. On Line Database of Ensemble Representatives And DOmains. *Protein Sci*, **6**, 2628–30.

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

Kendrew, J., Bodo, G., Dintzis, H., Parrish, R., Wyckoff, H. & Phillips, D. (1958). A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. *Nature*, **181**, 662–6.

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

Lattman, E. & Rose, G. (1993). Protein folding–what's the question? *Proc Natl Acad Sci U S A*, **90**, 439–41.

Leonidas, D., Vatzaki, E., Vorum, H., Celis, J., Madsen, P. & Acharya, K. (1998). Structural basis for the recognition of carbohydrates by human galectin-7. *Biochemistry*, **37**, 13930–40.

Levitt, M. & Chothia, C. (1976). Structural patterns in globular proteins. *Nature*, **261**, 552–8.

Levitt, M. & Gerstein, M. (1998). A unified statistical framework for sequence comparison and structure comparison. *Proc Natl Acad Sci U S A*, **95**, 5913–20.

Liu, Y. & Eisenberg, D. (2002). 3D domain swapping: as domains continue to swap. *Protein Sci*, **11**, 1285–99.

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

Lupas, A., Ponting, C. & Russell, R. (2001). On the evolution of protein folds: are similar motifs in different protein folds the result of convergence, insertion, or relics of an ancient peptide world? *J Struct Biol*, **134**, 191–203.

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

Makarova, K. & Grishin, N. (1999). Thermolysin and mitochondrial processing peptidase: how far structure-functional convergence goes. *Protein Sci*, **8**, 2537–40.

Marti-Renom, M., Stuart, A., Fiser, A., Sanchez, R., Melo, F. & Sali, A. (2000). Comparative protein structure modeling of genes and genomes. *Annu Rev Biophys Biomol Struct*, **29**, 291–325.

Martin, A., MacArthur, M. & Thornton, J. (1997). Assessment of comparative modeling in CASP2. *Proteins*, **Suppl 1**, 14–28.

Martin, J. (1995). Thioredoxin–a fold for all reasons. *Structure*, **3**, 245–50.

Messerschmidt, A. & Huber, R. (1990). The blue oxidases, ascorbate oxidase, laccase and ceruloplasmin. Modelling and structural relationships. *Eur J Biochem*, **187**, 341–52.

Messerschmidt, A., Ladenstein, R., Huber, R., Bolognesi, M., Avigliano, L., Petruzzelli, R., Rossi, A. & Finazzi-Agro, A. (1992). Refined crystal structure of ascorbate oxidase at 1.9 A resolution. *J Mol Biol*, **224**, 179–205.

Michel, G., Chantalat, L., Duee, E., Barbeyron, T., Henrissat, B., Kloareg, B. & Dideberg, O. (2001). The kappa-carrageenase of P. carrageenovora features a tunnel-shaped active site: a novel insight in the evolution of Clan-B glycoside hydrolases. *Structure (Camb)*, **9**, 513–25.

Mizuguchi, K. & Blundell, T. (2000). Analysis of conservation and substitutions of secondary structure elements within protein superfamilies. *Bioinformatics*, **16**, 1111–9.

Momany, F. & Rone, R. (1992). Validation of the general purpose QUANTA 3.2/CHARMm forcefield. *J. Comp. Chem.*, 888–900.

Mulder, N., Apweiler, R., Attwood, T., Bairoch, A., Bateman, A., Binns, D., Biswas, M., Bradley, P., Bork, P., Bucher, P., Copley, R., Courcelle, E., Durbin, R., Falquet, L., Fleischmann, W., Gouzy, J., Griffith-Jones, S., Haft, D., Hermjakob, H., Hulo, N., Kahn, D., Kanapin, A., Krestyaninova, M., Lopez, R., Letunic, I., Orchard, S., Pagni, M., Peyruc, D., Ponting, C., Servant, F. & Sigrist, C. (2002). InterPro: an integrated documentation resource for protein families, domains and functional sites. *Brief Bioinform*, **3**, 225–35.

Murphy, M., Lindley, P. & Adman, E. (1997a). Structural comparison of cupredoxin domains: domain recycling to construct proteins with novel functions. *Protein Sci*, **6**, 761–70.

Murphy, M., Turley, S. & Adman, E. (1997b). Structure of nitrite bound to copper-containing nitrite reductase from Alcaligenes faecalis. Mechanistic implications. *J Biol Chem*, **272**, 28455–60.

Murzin, A. (1998). How far divergent evolution goes in proteins. *Curr Opin Struct Biol*, **8**, 380–7.

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

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

Ogata, K. & Umeyama, H. (2000). An automatic homology modeling method consisting of database searches and simulated annealing. *J Mol Graph Model*, **18**, 258–72, 305–6.

Ollis, D., Cheah, E., Cygler, M., Dijkstra, B., Frolow, F., Franken, S., Harel, M., Remington, S., Silman, I., Schrag, J. & et al. (1992). The alpha/beta hydrolase fold. *Protein Eng*, **5**, 197–211.

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

Orengo, C., Jones, D. & Thornton, J. (1994). Protein superfamilies and domain superfolds. *Nature*, **372**, 631–4.

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

Orengo, C., Jones, D. & Thornton, J. (2003). *Bioinformatics: genes, proteins and computers*. BIOS Scientific Publishers.

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

Pascarella, S. & Argos, P. (1992). Analysis of insertions/deletions in protein structures. *J Mol Biol*, **224**, 461–71.

Pearl, F., Martin, N., Bray, J., Buchan, D., Harrison, A., Lee, D., Reeves, G., Shepherd, A., Sillitoe, I., Todd, A., Thornton, J. & Orengo, C. (2001). A rapid classification protocol for the CATH Domain Database to support structural genomics. *Nucleic Acids Res*, **29**, 223–7.

Pearson, W. & Lipman, D. (1988). Improved tools for biological sequence comparison. *Proc Natl Acad Sci U S A*, **85**, 2444–8.

Petratos, K., Dauter, Z. & Wilson, K. (1988). Refinement of the structure of pseudoazurin from Alcaligenes faecalis S-6 at 1.55 A resolution. *Acta Crystallogr B*, **44 (Pt 6)**, 628–36.

Pieper, U., Eswar, N., Stuart, A., Ilyin, V. & Sali, A. (2002). MODBASE, a database of annotated comparative protein structure models. *Nucleic Acids Res*, **30**, 255–9.

Poget, S., Legge, G., Proctor, M., Butler, P., Bycroft, M. & Williams, R. (1999). The structure of a tunicate C-type lectin from Polyandrocarpa misakiensis complexed with D -galactose. *J Mol Biol*, **290**, 867–79.

Ponstingl, H., Henrick, K. & Thornton, J. (2000). Discriminating between homodimeric and monomeric proteins in the crystalline state. *Proteins*, **41**, 47–57.

Ptitsyn, O. (1998). Protein folding and protein evolution: common folding nucleus in different subfamilies of c-type cytochromes? *J Mol Biol*, **278**, 655–66.

Ptitsyn, O. (1999). Protein evolution and protein folding: non-functional conserved residues and their probable role. *Pac Symp Biocomput*, 494–504.

Ptitsyn, O. & Ting, K. (1999). Non-functional conserved residues in globins and their possible role as a folding nucleus. *J Mol Biol*, **291**, 671–82.

Ren, B., Huang, W., Akesson, B. & Ladenstein, R. (1997). The crystal structure of seleno-glutathione peroxidase from human plasma at 2.9 A resolution. *J Mol Biol*, **268**, 869–85.

Ren, B., Tibbelin, G., de Pascale, D., Rossi, M., Bartolucci, S. & Ladenstein, R. (1998). A protein disulfide oxidoreductase from the archaeon Pyrococcus furiosus contains two thioredoxin fold units. *Nat Struct Biol*, **5**, 602–11.

Rice, D. & Eisenberg, D. (1997). A 3D-1D substitution matrix for protein fold recognition that includes predicted secondary structure of the sequence. *J Mol Biol*, **267**, 1026–38.

Richardson, J. (1981). The anatomy and taxonomy of protein structure. *Adv Protein Chem*, **34**, 167–339.

Romero, A., De la Cerda, B., Varela, P., Navarro, J., Hervas, M. & De la Rosa, M. (1998). The 2.15 A crystal structure of a triple mutant plastocyanin from the cyanobacterium Synechocystis sp. PCC 6803. *J Mol Biol*, **275**, 327–36.

Rost, B., Sander, C. & Schneider, R. (1994). PHD–an automatic mail server for protein secondary structure prediction. *Comput Appl Biosci*, **10**, 53–60.

Rowland, P., Basak, A., Gover, S., Levy, H. & Adams, M. (1994). The three-dimensional structure of glucose 6-phosphate dehydrogenase from Leuconostoc mesenteroides refined at 2.0 A resolution. *Structure*, **2**, 1073–87.

Russell, R. & Barton, G. (1992). Multiple protein sequence alignment from tertiary structure comparison: assignment of global and residue confidence levels. *Proteins*, **14**, 309–23.

Russell, R. & Barton, G. (1994). Structural features can be unconserved in proteins with similar folds. An analysis of side-chain to side-chain contacts secondary structure and accessibility. *J Mol Biol*, **244**, 332–50.

Ryden, L. & Hunt, L. (1993). Evolution of protein complexity: the blue copper-containing oxidases and related proteins. *J Mol Evol*, **36**, 41–66.

Sali, A. & Blundell, T. (1990). Definition of general topological equivalence in protein structures. A procedure involving comparison of properties and relationships through simulated annealing and dynamic programming. *J Mol Biol*, **212**, 403–28.

Sali, A. & Blundell, T. (1993). Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol*, **234**, 779–815.

Sali, A. & Overington, J. (1994). Derivation of rules for comparative protein modeling from a database of protein structure alignments. *Protein Sci*, **3**, 1582–96.

Sanchez, R. & Sali, A. (1998). Large-scale protein structure modeling of the Saccharomyces cerevisiae genome. *Proc Natl Acad Sci U S A*, **95**, 13597–602.

Sanchez, R. & Sali, A. (1999). ModBase: a database of comparative protein structure models. *Bioinformatics*, **15**, 1060–1.

Sayle, R. & Milner-White, E. (1995). RASMOL: biomolecular graphics for all. *Trends Biochem Sci*, **20**, 374.

Scapin, G., Blanchard, J. & Sacchettini, J. (1995). Three-dimensional structure of Escherichia coli dihydrodipicolinate reductase. *Biochemistry*, **34**, 3502–12.

Schmidt, R., Gerstein, M. & Altman, R. (1997). LPFC: an Internet library of protein family core structures. *Protein Sci*, **6**, 246–8.

Schwede, T., Kopp, J., Guex, N. & Peitsch, M. (2003). SWISS-MODEL: an automated protein homology-modeling server. *Nucleic Acids Res*, **31**, 3381–5.

Shih, H., Brady, J. & Karplus, M. (1985). Structure of proteins with single-site mutations: a minimum perturbation approach. *Proc Natl Acad Sci U S A*, **82**, 1697–700.

Sibanda, B. & Thornton, J. (1985). Beta-hairpin families in globular proteins. *Nature*, **316**, 170–4.

Sigrist, C., Cerutti, L., Hulo, N., Gattiker, A., Falquet, L., Pagni, M., Bairoch, A. & Bucher, P. (2002). PROSITE: a documented database using patterns and profiles as motif descriptors. *Brief Bioinform*, **3**, 265–74.

Sillitoe, I. (2002). *Consensus Templates for Protein Structure Prediction*. Phd thesis, University of London.

Sinning, I., Kleywegt, G., Cowan, S., Reinemer, P., Dirr, H., Huber, R., Gilliland, G., Armstrong, R., Ji, X., Board, P. & et al. (1993). Structure determination and refinement of human alpha class glutathione transferase A1-1, and a comparison with the Mu and Pi class enzymes. *J Mol Biol*, **232**, 192–212.

Sippl, M. (1993). Recognition of errors in three-dimensional structures of proteins. *Proteins*, **17**, 355–62.

Smith, T. & Waterman, M. (1981). Identification of common molecular subsequences. *J Mol Biol*, **147**, 195–197.

Sutcliffe, M., Haneef, I., Carney, D. & Blundell, T. (1987a). Knowledge based modelling of homologous proteins, Part I: Three-dimensional frameworks derived from the simultaneous superposition of multiple structures. *Protein Eng*, **1**, 377–84.

Sutcliffe, M., Hayes, F. & Blundell, T. (1987b). Knowledge based modelling of homologous proteins, Part II: Rules for the conformations of substituted sidechains. *Protein Eng*, **1**, 385–92.

Swindells, M. (1995). A procedure for the automatic determination of hydrophobic cores in protein structures. *Protein Sci*, **4**, 93–102.

Taylor, W. (1986a). The classification of amino acid conservation. *J Theor Biol*, **119**, 205–18.

Taylor, W. (1986b). Identification of protein sequence homology by consensus template alignment. *J Mol Biol*, **188**, 233–58.

Taylor, W. (1987). Multiple sequence alignment by a pairwise algorithm. *Comput Appl Biosci*, **3**, 81–7.

Taylor, W. (1999). Protein structural domain identification. *Protein Eng*, **12**, 203–16.

Taylor, W. & Orengo, C. (1989). Protein structure alignment. *J Mol Biol*, **208**, 1–22.

Thoden, J., Kappock, T., Stubbe, J. & Holden, H. (1999). Three-dimensional structure of N5-carboxyaminoimidazole ribonucleotide synthetase: a member of the ATP grasp protein superfamily. *Biochemistry*, **38**, 15480–92.

Thoden, J., Blanchard, C., Holden, H. & Waldrop, G. (2000). Movement of the biotin carboxylase B-domain as a result of ATP binding. *J Biol Chem*, **275**, 16183–90.

Todd, A. (2001). *Evolution of Function in Protein Superfamilies*. Phd thesis, University of London.

Todd, A., Orengo, C. & Thornton, J. (2001). Evolution of function in protein superfamilies, from a structural perspective. *J Mol Biol*, **307**, 1113–43.

Tramontano, A., Leplae, R. & Morea, V. (2001). Analysis and assessment of comparative modeling predictions in CASP4. *Proteins*, **Suppl 5**, 22–38.

Tsukihara, T., Aoyama, H., Yamashita, E., Tomizaki, T., Yamaguchi, H., Shinzawa-Itoh, K., Nakashima, R., Yaono, R. & Yoshikawa, S. (1996). The whole structure of the 13-subunit oxidized cytochrome c oxidase at 2.8 A. *Science*, **272**, 1136–44.

Van Driessche, G., Koh, M., Chen, Z., Mathews, F., Meyer, T., Bartsch, R., Cusanovich, M. & Van Beeumen, J. (1996). Covalent structure of the flavoprotein subunit of the flavocytochrome c: sulfide dehydrogenase from the purple phototrophic bacterium Chromatium vinosum. *Protein Sci*, **5**, 1753–64.

Van Roey, P., Rao, V., Plummer, T., Jr & Tarentino, A. (1994). Crystal structure of endo-beta-N-acetylglucosaminidase F1, an alpha/beta-barrel enzyme adapted for a complex substrate. *Biochemistry*, **33**, 13989–96.

Waldrop, G., Rayment, I. & Holden, H. (1994). Three-dimensional structure of the biotin carboxylase subunit of acetyl-CoA carboxylase. *Biochemistry*, **33**, 10249–56.

Wallace, A., Borkakoti, N. & Thornton, J. (1997). TESS: a geometric hashing algorithm for deriving 3D coordinate templates for searching structural databases. Application to enzyme active sites. *Protein Sci*, **6**, 2308–23.

Walter, R., Ealick, S., Friedman, A., Blake, R., 2nd, Proctor, P. & Shoham, M. (1996). Multiple wavelength anomalous diffraction (MAD) crystal structure of rusticyanin: a highly oxidizing cupredoxin with extreme acid stability. *J Mol Biol*, **263**, 730–51.

Wang, S., Trumble, W., Liao, H., Wesson, C., Dunker, A. & Kang, C. (1998a). Crystal structure of calsequestrin from rabbit skeletal muscle sarcoplasmic reticulum. *Nat Struct Biol*, **5**, 476–83.

Wang, W., Kappock, T., Stubbe, J. & Ealick, S. (1998b). X-ray crystal structure of glycinamide ribonucleotide synthetase from Escherichia coli. *Biochemistry*, **37**, 15647–62.

Williams, P., Blackburn, N., Sanders, D., Bellamy, H., Stura, E., Fee, J. & McRee, D. (1999). The CuA domain of Thermus thermophilus ba3-type cytochrome c oxidase at 1.6 A resolution. *Nat Struct Biol*, **6**, 509–16.

Wood, T. & Pearson, W. (1999). Evolution of protein sequences and structures. *J Mol Biol*, **291**, 977–95.

Xue, Y., Okvist, M., Hansson, O. & Young, S. (1998). Crystal structure of spinach plastocyanin at 1.7 A resolution. *Protein Sci*, **7**, 2099–105.

Yamaguchi, H., Kato, H., Hata, Y., Nishioka, T., Kimura, A., Oda, J. & Katsube, Y. (1993). Three-dimensional structure of the glutathione synthetase from Escherichia coli B at 2.0 A resolution. *J Mol Biol*, **229**, 1083–100.