Analysis and Prediction of Protein Domains

Russell Leonard Marsden

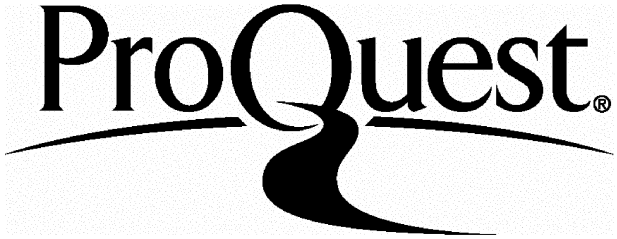
BSc. (Hons) Biological Sciences with Cell Biology

(University of Warwick)

A Thesis Submitted in partial fulfilment of the requirements
for the Degree of Doctor of Philosophy

Department of Computer Science
University College London

April 2003



Abstract

Domains are the structural and functional units of protein structure enabling the formation of multi-functional, multi-domain proteins. The identification of domains within a sequence is an important pre-requisite for many protein analysis techniques and can be achieved from protein tertiary structure or detection of homology to known domain sequences. In the absence of either protein structure or sequence homology a method to delineate domain boundaries from sequence alone is required. In this thesis, a number of domain properties have been considered in order to address the possibility of domain prediction from sequence.

A survey of domain linker characteristics has been made which shows domain linkers to be flexible, exposed and generally unstructured regions of polypeptide, with a high propensity for proline residues which may have repercussions for linker structural independence and folding rate. The surface area and hydrophobicity of protein structures has also been investigated. There appears to be a positive correlation between sequence length and surface area, although a domain prediction method based solely on this characteristic does not seem likely. There was no obvious separation between the percentage of hydrophobic residues in either single or multi-domain proteins.

A domain assignment method based upon the alignment of predicted secondary structure to proteins of known structure has been developed and implemented. The top-hit prediction of continuous domain boundaries achieved a sensitivity of 31% with a selectivity of 32% (±20 residues). The domain number and corresponding boundaries were correctly predicted for 25% of the multi-domain test set (±20 residues). A further method that assigns domains based on post-processing PSI-BLAST alignments has also been developed. This method achieved a continuous domain boundary prediction sensitivity of 30% with a selectivity of 56% (±20 residues). These two methods have also been combined for prediction of domains by sequence comparison and from sequence alone.

The formation of protein oligomers by the exchange of identical units of protein structure between subunits is termed 'domain swapping'. A general analysis of domain swapped proteins, including the properties of the swapped-domain linkers has been carried out. The analysis of domain linker of Chapter 2 also enabled a comparison of their characteristics to those of swapped-domain linkers.

Contents

List of Figures

List of Tables

List of abbreviations

3D	Three-dimensional
Å	Angstrom
ATP	Adenosine triphosphate
ADP	Adenosine diphosphate
ASA	Assessable Surface Area
CAFASP	Critical Assessment of Fully Automated Structure Prediction
CASP	Critical Assessment of Structure Prediction
CATH	Class Architecture Topology Homology
DDD	Dali Domain Dictionary
DGS	Domain Guess by Size
DNA	Deoxyribonucleic acid
DPS	Domains Parsed by Sequence
DSSP	Dictionary of Secondary Structure Definition
FSSP	Families of Structurally Similar Proteins
HMM	Hidden Markov Model
kDa	Kilodalton
mRNA	Messenger ribonucleic acid
NMR	Nuclear Magnetic Resonance
PDB	Protein Data Bank
PSSM	Position-Specific Scoring Matrix
RCSB	Research Collaboratory for Structural Bioinformatics
RSA	Relative Solvent Accessibility
s.d.	Standard deviation
s.e	Standard error
SCOP	Structural Classification of Proteins
SSEA	Secondary Structure Element Alignment
C-terminal	Carboxy-terminal
N-terminal	Amino-terminal
%	Percent

List of amino acid abbreviations

A	Ala	Alanine
C	Cys	Cysteine
D	Asp	Aspartate
E	Glu	Glutamine
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

Acknowledgements

In the three years (plus) a PhD inevitably takes, you meet many people, the majority of which tend to appear on Finchley Road tube station around 8.15am, however a number have done more than make me take a more vertical orientation than perhaps I would have liked at this time in the morning, and for this (and other reasons) they deserve special thanks:
(all for much encouragement and support, in no particular order).

Thanks to all members of the Bioinformatics Group past and present.

Thanks to David Jones for giving such enthusiastic, knowledgeable and good-humoured supervision and who despite the upheaval of three universities in three years, managed to keep me focused on the job in hand. Caroline Hadley for listening to my science/code/PhD rants and always coming up with constructive advice, for so many email laughs, also without whom my programming learning curve would have required climbing gear rather than a good pair of boots... Claire Chivers for giving me a place in Leam even when she knew I wasn't going to be around for as long as expected. Michael Tress for putting so much time into organising football sweep stakes (especially World Cup 2002) and giving much helpful (less sport related) advice. Gabrielle Reeves for helping make us feel so welcome at UCL, and sharing the pains of writing up/finishing off/new job. Marialuisa Pellegrini-Calace for being a great deskmate and often pointing out that the weather in England isn't normal! Kevin Bryson for being such an entertaining person to have in the lab at Warwick and UCL, and who unwittingly timed his return across the channel to coincide with a constant barrage of requests to read chapters – thanks so much for taking on this task so conscientiously and offering so much valuable advice. Liam McGuffin for providing such a positive approach to science especially on those less motivated days... Jonathon Ward for getting in early, so I had someone to talk to in the mornings. Brian Marsden for thesis chats and suggestions, reading chapters, and showing that it could be done. Catherine Marsden for giving so much help and being a continual source of motivation to get this thesis written, and repeatingly telling me to **get on with it!** Daniel Smith for the same and putting up with Catherine putting up with me. Richard Mills and Rachel Agg for being supportive flat mates and still managing to look interested when forgetting that the answer to, just what is it that you do? may take a bit longer than expected... Finally, thanks to my parents for setting such high standards and giving me the opportunity to try and reach them.

A Bioinformatics Production @

Warwick Brunel UCL

Declaration

All research presented in this thesis is the candidate's own work. The content of Chapter 4 was previously published as:

Marsden, R.L., McGuffin, L.J. & Jones, D.T. 2002. Rapid protein domain assignment from amino acid sequence using predicted secondary structure. *Protein Sci.* **11**, 2814-2824.

No portion of the work referred to in this thesis has been submitted in support of an application for another degree or qualification of University College London or any other university or institute of learning.

Chapter 1

Introduction

This year marks the fiftieth anniversary of Watson and Crick's determination of the 3D structure of DNA (1953). This fundamental discovery was followed by a series of theoretical and technical advances in molecular biology, including the identification of the genetic code in the 1960's and the advent of DNA sequencing and polymerase chain reaction in the 1970's and 80's. These have paved the way for the sequencing of whole genomes including the human genome in the last decade.

Genes contain the blueprint for the manufacture of proteins, the essential active agents of biochemistry, that play a key role in almost all metabolic processes that are associated with life. Understanding protein structure is of great importance because the function of a protein is ultimately determined by its structure. The interpretation of genes in terms of the structure and function of the proteins they encode is of crucial importance to fully understand their role in the cell.

1.1 The hierarchy of protein structure

In their folded three-dimensional (3D) state proteins can appear irregular and complex, with no obvious underlying mechanism to explain the packing of the amino acid chain. This structural complexity has arisen over millions of years to perform a diverse range of functions that dominate the biochemistry of our cells. In solution proteins often form globular structures where the peptide chain runs back and forth across the protein core. Protein structure can be explained by an underlying hierarchy that ranges from primary to quaternary structure.

1.1.1 Primary structure

The primary structure, the first level in the hierarchy, is the sequence of amino acids along the polypeptide chain. Each amino acid is joined to the next by a peptide bond, formed by a condensation reaction between the COOH of the N-terminal amino acid with the NH_2 group of the C-terminal amino acid. Each amino acid varies in its chemical properties depending on its side chain. These properties can be grouped into three main classes (Branden and Tooze, 1999): amino acids with strictly hydrophobic side chains (Ala, Val, Leu, Ile, Pro, Phe and Met), those with charged side chains (Asp, Glu, Arg and Lys), and the amino acids with polar side

chains (Ser, Thr, Cys, Asn, Gln, His, Tyr and Trp). Gly is an exception in that it contains only a hydrogen atom as its side chain, 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 3D structure (Anfinsen, 1961), so bestowing the huge variety of protein structures and functions in nature.

1.1.2 Secondary structure

Protein folding is driven by the hydrophobic effect, in which hydrophobic amino acids become buried within the protein core, shielded from the surrounding aqueous environment by hydrophilic residues (Kauzmann, 1959). Globular proteins can therefore be seen as having a hydrophobic inner core, surrounded by a hydrophilic outer shell (Waugh, 1954). However, the burial of hydrophobic side chains is also accompanied by the burial of their main chain atoms which include polar N-H and C=O groups. The solution to this inclusion of the polar polypeptide backbone within the hydrophobic core is to neutralise these groups by the formation of hydrogen bonds between them. This gives rise to regular patterns of hydrogen bonding that form protein secondary structure. The route taken by the polypeptide chain in 3D space to create the secondary structure can be described by the relative positions of three atoms linked within the backbone: the C-alpha, carbonyl carbon (C') and amide nitrogen (N) atoms. The relative positions or angles of rotation between these atoms are described as the phi angle (around the N-C-alpha bond), and the psi angle (around the C-alpha-C' bond). Secondary structure is defined as two main types.

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

Beta-sheet: This secondary structure type is made up of two or more continuous regions of chain called beta-strands, which are found in a fully extended conformation. Beta-strands line up in such a way as to allow hydrogen bond

formation between adjacent C=O and N-H groups. Beta-sheets are built up of beta-strands arranged in parallel, anti-parallel, or a mixture of both.

Secondary structures can in turn combine together to form motifs or super-secondary structure. Examples include beta-hairpins, made up of two anti-parallel strands joined by a short loop, that can exist as an isolated ribbon or form part of a larger sheet structure. Another example is the beta-alpha-beta motif, again made up of two strands, but this time connected by a helix, that packs against the sheet formed by the adjacent strands. This packing shields hydrophobic residues in the beta-strands and the alpha-helix from solvent.

1.1.3 Tertiary structure

The packing together of secondary structure elements or the larger motifs results in the tertiary structure of a protein, that may contain one or more domains. Protein domains are often described as the fundamental units of protein structure, forming high-order building blocks of the protein polypeptide chain (Hubbard and Argos, 1996). Domains can be considered as local, semi-independent units (Richardson, 1981). Each domain contains an individual hydrophobic core that is made up of secondary structure elements which are often conserved across protein families. Secondary structure elements are connected by more 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 and Chothia (1976).

All-alpha domains, built up of mostly alpha-helices are often small folds in which the helices are usually arranged in bundles packing against one another to form a globular core.

All-beta domains, consisting almost entirely of beta-sheets normally in an anti-parallel arrangement within the domain core. Beta-sheets can pack against one another, with the hydrophobic side-chains located at the interface, forming beta-sandwiches.

Alpha-beta domains are built up of a repeating beta-alpha-beta motif that results in the outer layer of the structure being composed of amphipathic alpha-helices, that pack around the central core of beta-sheets.

Alpha and beta domains, like alpha-beta domains, contain alpha-helices and beta-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).

1.1.4 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 evolution of oligomeric complexes has conferred advantages compared to monomer subunits including 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 and Eisenberg, 2002). While the fusion of domains at the genetic level has evolved a permanent interaction between domains, in some cases interactions at the quaternary level can be reversible.

A specialised mechanism for the association of identical protein monomers is called domain swapping (Bennett *et al.*, 1995). In domain swapping, a single secondary structure, or even a whole domain of a monomeric protein is replaced by an identical domain of an identical protein chain which can produce an intertwined oligomer. This process may represent a rapid pathway for the formation of protein oligomers, making use of evolutionarily optimised intramolecular interactions as intermolecular interactions (Heringa and Taylor, 1997).

1.2 The role of domains in proteins

Today, many areas of biology routinely use the term 'domain', although the actual definition of this word can vary. The concept of the protein domain was first used to describe structural units within proteins. As more protein structures were determined, such as hen egg-white lysozyme (Blake *et al.*, 1965) and ribonuclease

(Kartha *et al.*, 1967) it became apparent that proteins contained distinct structural regions. A study by Wetlaufer (1973) analysed 18 protein structures and assigned domains on the basis of globular structural units expected to fold autonomously. In a review in 1981, Richardson defined domains as compact, local, semi-independent units - now a frequently used and cited definition. Interestingly the study by Richardson (1981) yielded a more conservative domain assignment procedure, and the structure of hen egg-white lysozyme previously described as two-domain was classed as single domain. Domains can also be defined in an evolutionary context, where sequence comparison may reveal sequence-similar homologues that are found in different proteins, or they can be described in terms of functional regions within a protein, that does not necessarily take a structural viewpoint into account. Though each of these definitions may appear distinct, they are all valid, and in many cases are compatible. Domains may be viewed as independent folding units, especially those that are found in a range of different proteins, an indication that they may have started out as single domain proteins. Multi-domain proteins may contain domains with different functions, each working separately, or adjacent domains might work together cooperatively. Such functions may include catalysis and substrate binding. Domains can play an important structural role in the cell, acting as building blocks, or modules that build up large repetitive molecular assemblies such as muscle fibres (Campbell and Downing, 1994).

The structure of pyruvate kinase (Larsen *et al.*, 1994) provides an example of a multi-domain protein in which the function of each domain is different. This three-domain enzyme, shown in Figure 1.1, catalyses the last step in glycolysis, where phosphoenolpyruvate is converted to pyruvate via phosphorylation of ADP to ATP. There are two alpha/beta domains, one of which acts as a catalytic domain and the other as a nucleotide binding region and a third all-beta regulatory domain. Each of these domain structures is found in a wide variety of different proteins, with the central alpha/beta substrate binding domain being one of the most commonly recurring enzyme domains. This large alpha/beta catalytic domain forms a barrel structure, also known as a TIM barrel. This fold was named after the structure of triosephosphate isomerase in which it was first observed (Banner *et al.*, 1975). The eight-stranded TIM barrel is one of the most common folds to be observed in protein structures, often functioning as enzymes (Nagano, 2002). The TIM barrel structure in pyruvate kinase is an interesting case, as the all-beta domain has been inserted within

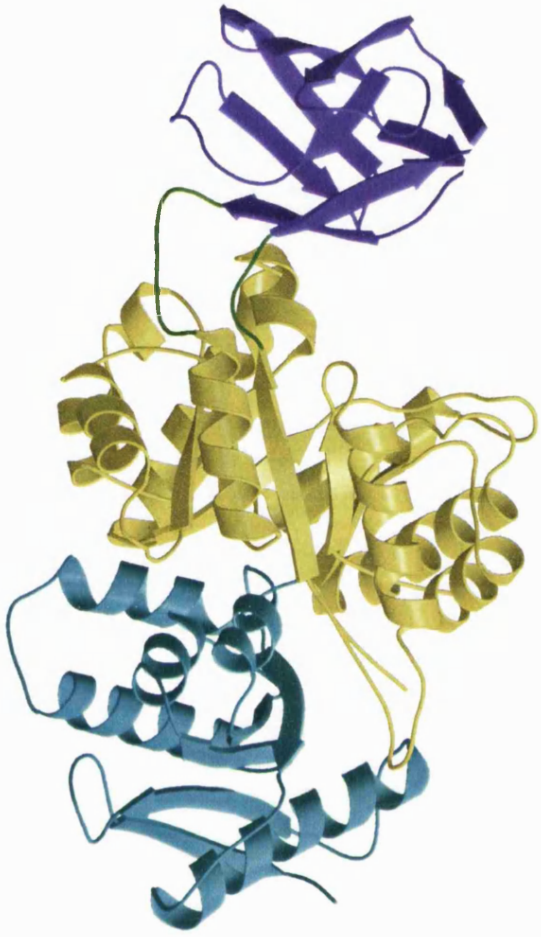


Figure 1.1 Pyruvate kinase – a three-domain multi-functional enzyme

The structure of pyruvate kinase, PDB code 1pkn (Larsen *et al.*, 1994), an example of a multi-domain protein in which the function of each domain is different. The large catalytic TIM barrel domain (yellow) has an all-beta domain (purple) inserted within its sequence, which is connected to the TIM barrel by means of two domain linkers (green). Cartoon figure prepared using MOLSCRIPT (Kraulis, 1991) and Raster3D (Merritt and Bacon, 1997).

its sequence. The sequence of the TIM barrel is therefore split into two regions within the structure, and is therefore described as a discontinuous domain, unlike the all-beta domain which is a continuous domain. This separation of the TIM barrel in sequence is most likely to have occurred by a domain insertion event, within the catalytic domain gene sequence during the proteins evolution (Russell, 1994). Figure 1.2 shows examples of continuous and discontinuous domains, in order of complexity.

The TIM barrel offers an interesting perspective on domain evolution. Convergent evolution has long been thought to explain the occurrence of these frequently occurring folds that often have no discernible sequence homology between them. It is thought that such folds may offer favourable properties for folding and stability (Orengo *et al.*, 1994). However, with the advent of highly sensitive sequence comparison methods, homologous relationships have been proposed between these structures, suggesting divergent evolution from an ancient gene (Copley and Bork, 2000).

The evolution of multi-domain proteins from the fusion of smaller domains has conferred a number of functional and structural advantages. Ghelis and Yon (1979) suggested that the covalent association of domains resulted in a more stable complex compared to oligomeric proteins made up of several subunits. It can be envisaged that ancient biochemical pathways that utilised single domain structures in sequential pathway reactions may have benefited from the fusion of these domains giving a fixed stoichiometric ratio of catalytic sites for any given time (Ostermeier and Benkovic, 2000). Multi-domain proteins have also evolved functional or catalytic sites between domains that are often situated in the inter-domain cleft. Movements between adjacent domains enables an induced-fit binding of the substrate, creating a catalytic environment that protects the substrate from solvent, a model first introduced by Koshland (1958). An example of changes in domain orientation is found in the iron transportation protein lactoferrin that is made up of two similar lobes, each containing two domains. Upon binding iron, the two domains rotate by 53° with the axis of rotation passing through two beta-strands that link the domains (Gerstein *et al.*, 1993). This domain closure is used to recognise and sequester iron in the iron binding site situated within the inter-domain cleft.

The fusion of domains within single structures may have covalently fixed them into a single polypeptide chain but this does not mean that they must form rigid

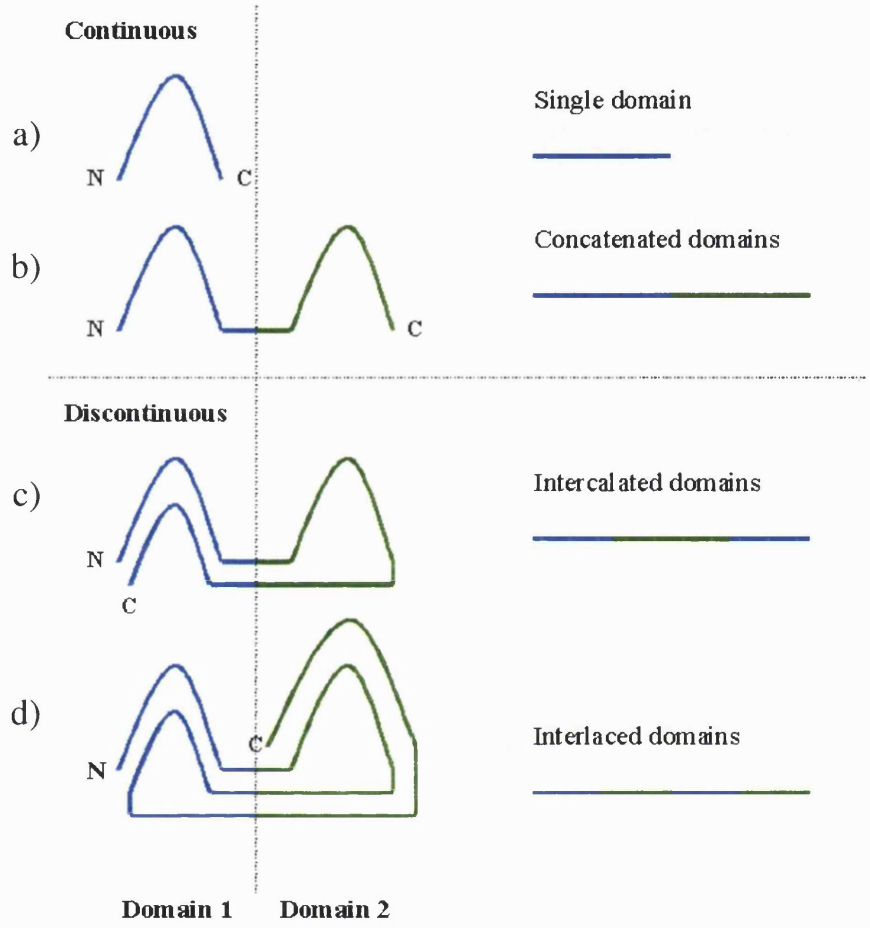


Figure 1.2 Continuous and discontinuous domains

Schematic diagram showing different connectivity between multi-domain proteins (adapted from Siddiqui and Barton (1995) and Das and Smith (2000)). **a)** A single domain chain. **b)** A simple two-domain chain with one connection. **c)** A two-domain chain with two connections, where the chain runs from the first domain into the second, and then back to complete the first domain. This is likely to have been the result of the insertion of the second domain into the first. **d)** A two-domain chain with three connections. This is similar to **c**, except that after completing the first domain, the chain returns to complete the second domain.

associations. As has been described above, the domains of multiple domain proteins are capable of a large degree of conformational flexibility that has been shown to be important for regulatory activity, metabolite transport, formation of protein assemblies and cellular locomotion (Gerstein *et al.*, 1994). The large-scale flexibility in proteins between essentially rigid domains may be attributable to only small segments of chain such as the inter-domain linker (Janin and Wodak, 1983; Bennett and Huber, 1984). Domain movements can be classed into two main types: hinge motions and shear motions (Gerstein *et al.*, 1994). Domain movements may be facilitated by hinge or shear motions alone, or a combination of both motions. Much of the information regarding these movements is inferred from X-ray crystal and NMR structures of the open and closed conformations.

Hinge motions are defined as movements in the polypeptide chain that involve only a few large changes in a localised region and are not constrained by side-chain packing interactions. The simplest hinge motions can be found in beta-strands which can also effect large conformational changes. Beta-sheets are also able to act as hinges though their movements are more limited. Alpha-helices acting as hinges give more of a bending motion because they have a more constrained pattern of hydrogen bonding than extended strands and so the deformations are more spread out. Alpha-helices may also undergo more radical hinge motions in which kinks in helices, often populated by proline residues, can convert between helical and more extended conformations. For example, large torsion angle changes have been observed in the middle residues of the all-helical domain linker in calmodulin, resulting in two shorter separate helices almost perpendicular to one another (Meador *et al.*, 1992; Ikura *et al.*, 1992). Hinge movements enable large torsional changes between domains, however between closely associated domains such movements are not always possible since hinge movements require few packing constraints of the polypeptide chain.

Shear movements involve motions that occur parallel to the plane of the interface (unlike hinge that are usually perpendicular). This means that shear motions are limited to the degree of interaction between side-chains within the interface, and many shear movements between domains are thought to be accommodated by these side-chains, rather than involving their repacking (Gerstein *et al.*, 1994). Domain

motions may include a number of these small shear motions such as the burial of substrate within the active site of citrate synthase (Remington *et al.*, 1982).

1.3 Domains as units of structure

The observation of a recurrent structure within a different protein context is a strong indication of a protein domain. However, domains that share a similar structure are not necessarily related by significant sequence identity. Many similar structures have diverged beyond the detection of any sequence similarity, whilst their function has been retained. The conservation of protein tertiary structure over primary structure demonstrates the importance of the relationship between protein structure and function. Common ancestry between such remote homologues can be determined by searching for conserved features such as catalytic residues. It is also possible that some structures are more favoured because they represent stable folds. Unrelated proteins may have converged towards these folds and therefore do not always have the same function. For example, in the recent version of the CATH domain database (Orengo *et al.*, 1997) the TIM barrel has been documented as being functionally diverse, and has been classed into 21 superfamilies (Nagano *et al.*, 2002).

1.3.1 Identification of domains from structural coordinates

Because domains may function individually within a protein, with discrete functional and structural roles, the comparison of proteins at the domain level can give a comprehensive view of structural biology that would not be possible using entire proteins. At present there are over 20,000 structures deposited in the Protein Data Bank (PDB; Bernstein *et al.*, 1977), however this is a highly redundant collection. The comparison of these structures at the domain level enables a more representative view of protein structure. Such comparisons often begin by classifying structures with similar folds and functions, that can then be further compared on the basis of sequence similarity.

The structures in the PDB are made up of both single and multiple domains. These domains must be identified before comparisons can be made. Though identification and delineation of domains in protein structure can be achieved by

visual inspection it is not an easy process to automate. Automation of this process is necessary to keep up with the ever increasing number of deposited structures. Difficulties arise from the fact that there is a continual progression from proteins that slightly divide into lobes to those that form clearly distinct folded regions separated by an inter-domain linker (Taylor, 1999). For this reason the assignment of domains can be subjective, and because there is no standard definition of a protein domain, automated domain assignment methods have varied enormously, with each researcher using a unique set of criteria (Swindells, 1995). A number of algorithms for identifying structural domains from atomic coordinates have been proposed. The majority of these methods are based on the original concepts of Wetlaufer and Richardson, who observed that the interactions within domains are stronger and more numerous than between domains. Several methods are described below.

Crippen (1978) developed one of the first domain detection methods using a C-alpha-C-alpha distance map combined with a hierarchical clustering routine to group 10 residue protein segments. The identification of small C-alpha distances between the protein segments in the distance map was used to combine them into domains. The stepwise clustering faced difficulties in determining when a cluster had reached a size that constituted a domain.

Rose (1979) considered a protein as a rigid body where three perpendicular axes passing through the centroid (the coordinate centre of the protein), were used to identify continuous chain segments corresponding to compact domains. The method was limited because it could only identify continuous subunits and also required an initial input of the rough domain boundary positions.

Sowdhamini and Blundell (1995) continued the theme of identifying compact domain cores by clustering secondary structure elements on the basis of their intra-chain C-alpha distances. Specific features within the clusters could then be identified as regions corresponding to super-secondary structure and domains.

Siddiqui and Barton (1995) developed a method that located domains by maximising the ratio of internal contacts to the number of external contacts. The protein is divided up and a split value is calculated from the number of contacts, where the value is high when split regions are structurally distinct.

A series of methods have used solvent accessibility to calculate compactness (Rashin, 1985; Zehfus and Rose, 1986; Islam *et al.*, 1995). Wodak and Janin (1981) repeatedly cleaved proteins at varying residue positions and calculated interface areas

by comparing the surface area of the cleaved protein to that of its native structure. Calculations identified potential domain boundaries as sites where the interface area was at a minimum.

The search for domain boundaries by identifying a solution from a large number of possibilities can be computationally expensive (Wernisch *et al.*, 1999). Holm and Sander (1994) attempted to bypass this by developing a method based on inter-domain dynamics that uses principal component analysis to partition structurally rigid regions with a low number of contacts between them. Compactness criteria were used to recursively divide a protein into a series of successively smaller and smaller substructures. Recurrence criteria were then used to select an optimal size level of these substructures.

Swindells (1995) devised a method for the identification of domains based on the idea that domains have a hydrophobic interior. The method tended to fail for domains that did not have well defined hydrophobic cores.

Taylor (1999) developed a domain identification method based only on C-alpha coordinates, where each residue in the protein was given a numeric identifier. If a residue is surrounded by neighbours with on average a higher label, then its label is increased, otherwise it is decreased. This process is repeated for each residue resulting in compact regions evolving towards the same number.

1.3.2 Structural databases

Once domains have been identified, they can be classified into domain structural databases. Such databases exist to order the large volume of information in the PDB. The development of the automated domain identification methods in the early 1990s was accompanied by the development of a series of these classification systems (Table 1.1). Some of the most widely used and comprehensive databases are SCOP (Murzin *et al.*, 1995), CATH (Orengo *et al.*, 1997) and FSSP (Holm and Sander, 1997), each of which represents different methods of classifying protein structure. Hadley and Jones (1999) found that despite the fact that these three databases base their systems on different rules of protein structure and taxonomy, they often agree on assignments. Most disagreements were in instances where domain assignment or structural similarity was ambiguous. SCOP is almost completely manually derived, CATH employs an intermediate process, using

Database	URL	Reference
Structure		
SCOP	http://scop.mrc-lmb.cam.ac.uk/scop	(Murzin *et al.*, 1995)
CATH	http://www.biochem.ucl.ac.uk/bsm/cath_new	(Orengo *et al.*, 1997)
FSSP/ Dali	http://www.ebi.ac.uk/dali	(Holm and Sander, 1997)
Sequence		
BLOCKS	http://blocks.fhcrc.org	(Henikoff *et al.*, 2000)
COGS	http://wwwncbi.nlm.nih.gov/COG	(Tatusov *et al.*, 2001)
DOMO	http://wwwinfobiogen.fr/services/domo	(Gracy and Argos, 1998)
InterPro	http://www.ebi.ac.uk/interpro	(Apweiler *et al.*, 2000)
Pfam	http://www.sanger.ac.uk/Pfam	(Bateman *et al.*, 2000)
PRINTS	http://www.bioinf.man.ac.uk/dbrowser/PRINTS	(Attwood *et al.*, 2002)
ProDom	http://www.toulouse.inra.fr/prodom.html	(Corpet *et al.*, 1998)
PROSITE	http://www.expasy.ch/prosite	(Sigrist *et al.*, 2002)
SBASE	http://www3.icgeb.trieste.it/~sbaserv	(Murvai *et al.*, 2001)
SMART	http://SMART.embl-heidelberg.de	(Schultz *et al.*, 2000)

Table 1.1 Domain structure and sequence databases

automated procedures along with manual inspection and FSSP is a purely automated process. Each is described below in more detail.

1.3.2.1 SCOP (Structural Classification of Proteins)

SCOP (Murzin *et al.*, 1995) organises proteins in a hierachy from class to fold, superfamily and family. The main classes include all-alpha, all-beta, alpha/beta and alpha + beta. The classification of proteins into SCOP is essentially a manual process, using visual and structure comparison. Although most of the classification is based on visual inspection and structural comparison, some degree of automation is involved in the clustering protein chains based on sequence comparison. Most proteins are separated into domains prior to classification, though not all. Domain assignment in SCOP is based on the recurrence of domains as this is a strong indicator that the structure may have existed as an independent unit. Domains sharing common structural and functional features are grouped into superfamilies, where they are thought to share a common evolutionary origin. Proteins with significant sequence similarity or similar enough structure/function characteristics to imply an evolutionary link are grouped into families. SCOP therefore describes near and far evolutionary relationships as family and superfamily.

1.3.2.2 CATH (Class, Architecture, Topology, Homology)

CATH (Orengo *et al.*, 1997) is also a hierarchical system that differs from SCOP in that it incorporates a larger degree of automation in its classification of protein structures. Structures are first taken from the PDB and are divided into domains using a consensus approach that incorporates a number of domain-assignment techniques. Domains are assigned to one of four classes (C-level; mainly alpha, mainly beta, alpha and beta, and few secondary structures) on the basis of composition, secondary-structure contacts and proportion of parallel and anti-parallel sheets. Structure comparisons are made to produce fold groups (T-level) and homologous superfamilies (H level). Superfamiles are further divided on the basis of sequence similarities to produce sequence families (S-level). The final stage involves the assignment of architecture (A-level) unique to CATH. This attempts to summarise shape revealed by the orientation of secondary structure.

1.3.2.3 FSSP

FSSP (Holm and Sander, 1997) is based on the structure-structure comparison of proteins. This forms the basis of the Dali Domain Dictionary (Dietmann *et al.*, 2001) which is a numerical taxonomy of structures in the PDB. The database is split into four levels corresponding to super-secondary motifs, the topology of domains, remote homologues and sequence families. The top of the hierarchy assigns domains in five classes; all-alpha, all-beta, alpha/beta, alpha-beta meander and beta-barrels. If a domain cannot be assigned to one of these classes it is described as mixed class. The Dali Domain Dictionary offers a useful adjunct to SCOP and CATH in that it is wholly automated classification, and does not place protein structures into classes, fold families or superfamilies.

1.4 Domains as units of evolution

Although a huge variety of different domains with different folds have evolved, it is unclear as to what the precursors to these structural units might have been. Lupas *et al.*, (2001) suggested that domains may be descended from the conglomerates of short peptide segments that are seen today as internal repeats and structure-integrated motifs, such as the beta-trefoil fold. The domains we see today are those combinations that were capable of folding and provided a beneficial function. Some domains are widely spread between the three kingdoms of life, archaea, bacteria and eukarya, suggesting they have a common ancestor (Gerstein, 1997). The fact that they have persisted through evolution suggests they are either highly adaptable or they have an essential role in cellular processes. Such domains are often called modules and appear to be highly independent folding units, a characteristic that is important for survival if transferred to a new position in its original genome or to an entirely new genome. Protein modules are continuous regions of sequence, unlike domains which may not be. Mosaic proteins, often associated with multicellularity, are made up from one or a few different modules repeated many times. Examples include constituents of the extracellular matrix, growth factor receptors, cell-cell communication, and clotting mechanisms (Campbell and Baron, 1991; Campbell and Downing, 1994). Extracellular proteins in particular are notable for containing combinations of multicopy tandem arrays of

different modules. These extracellular modules have acquired diverse functions in different proteins, for example some EGF modules bind specific receptors, whilst others mediate interactions through calcium binding (Bork *et al.*, 1991). The immunoglobulin (Ig) module is abundant in proteins. In the case of the muscle protein titin, one of the largest polypeptides known, a total of 244 copies of Ig and fibronectin type III domains account for most of the 30,000 residue muscle protein (Labeit and Kolmerer, 1995). Modules can frequently display different connectivity relationships, for example the motor domain in kinesin can be found at either end of the polypeptide chain (Moore and Endow, 1996). ABC transporters consist of four domains, made up of two unrelated modules, ATP binding cassettes and integral membrane modules that can display different connectivities (Dean and Allikmets, 1995).

Multi-domain proteins are likely to have arisen by a process of domain accretion, where over time an ancestral gene acquires DNA encoding new domains, and so the protein product becomes more complex with only part of the gene being related to the original. In this way domains can be seen as genetic building blocks, that can be combined to create new combinations with new functions. Various mechanisms are thought to be responsible for the movement of domains within and between different organisms. Such processes include recombination events, including exon shuffling, insertions by transposable elements and domain deletions. Domain fusion or separation events can also occur by the alteration of start and stop codons within nucleotide sequences and changes in gene splicing can resulting in the dispersion of domains. It is apparent that some modules are more prolific than others. The ABC transporters are one of the most common module families, appearing to be present in all organisms and occurring at a high frequency in many of them. Other modules may be as widely spread, but show a much lower proliferation within organisms including metabolic enzymes and components of the translational apparatus.

Domain duplication can result in the simplest multi-domain proteins, though such repeated structures may have diverged to the extent where sequence similarity no longer exists between them (Heringa and Taylor, 1997). Domain duplication does not always result in domain insertion at the beginning or end of genes, sometimes the insertions can take place within a domain. In this case, the polypeptide backbone of

the original domain is interrupted by an excursion from an external loop to the inserted domain before returning to complete the parent domain fold (Russell, 1994).

An insight into the evolution of multi-domain proteins can be gained from the observation that many eukaryotic multi-domain proteins exist as independent units in prokaryotes (Davidson *et al.*, 1993). For example, the purine metabolism domains, GAR synthetase, AIR synthetase and GAR transformase, are found in a single polypeptide chain in vertebrates (GARS-AIRS-GART). In insects however, the polypeptide appears as GARS-(AIRS)2-GART, whilst in yeast, GARS-AIRS is encoded separately from GART. In bacteria, each domain is found separately (Henikoff *et al.*, 1997).

Domains exist that are frequently found in eukaryotes but rarely in prokaryotes suggesting that they have an ancient origin, but have been lost in other lineages over time. Conversely, domains that are found in prokaryotes but seldom in eukaryotes may have been passed on by horizontal gene transfer (the transfer of genes between organisms), and may therefore be recent arrivals in multicellular organisms. Much evidence has been amassed that among bacteria in particular, modules have passed horizontally between genomes (Gerstein, 1997). It is possible that some modules have swapped between species by a process of horizontal transfer. In cases where no domain homologues can be detected in prokaryotes, it is possible that they are eukaryotic inventions (Ponting and Russell, 2002).

1.4.1 Domain sequence databases

As an ever increasing number of genome sequences are being published it is of great importance to accurately annotate each gene in order to understand the structure, functional diversity, interactions and evolution of proteins within and between different organisms. The vast amount of sequence data produced by genome sequencing initiatives means that automated annotation methods are required to keep up. As many proteins contain more than one domain, and therefore may have more than one function, sequence characterisation is based at the domain level. A variety of domain, repeat and motif sequence databases are available and have been developed for this purpose, a number of which are described below.

Sequence cluster databases (see Table 1.1) are derived automatically from sequence databases using different clustering algorithms. ProDom (Corpet *et al.*,

1998) is a database of domain families generated from the clustered alignments of sequences in the SWISS-PROT and TrEMBL databases (Bairoch and Apweiler, 2000) using recursive PSI-BLAST (Altschul *et al.*, 1997). ProDom-CG (complete genome) holds data relating only to completed genome sequencing projects. DOMO (Gracy and Argos, 1998) contains multiple alignments of domains generated from the SWISS-PROT and PIR (Wu *et al.*, 2002) sequence databases. The relative N- or C-termini positions of homologous segment pairs within or between proteins are used for domain delineation. SBASE domains (Murvai *et al.*, 2001) are regions of sequence with known structure and/or function. Domain boundaries are taken from the literature or determined by homology to domains with known boundaries within databases including Pfam (Bateman *et al.*, 2000).

PRINTS (Attwood *et al.*, 2002), PROSITE (Sigrist *et al.*, 2002), BLOCKS (Henikoff *et al.*, 2000), Pfam (Bateman *et al.*, 2000) and SMART (Schultz *et al.*, 2000) all store sequence motifs that are based on searches of SWISS-PROT and TrEMBL. PRINTS is a database of protein fingerprints that aims to characterise a protein family by searching for a set of conserved motifs. PROSITE offers a high quality source of domain family annotation, where each family is represented by a pattern or profile. Each profile provides a means by which the sensitive detection of common protein domains in new protein sequences can be accomplished. BLOCKS holds multiply aligned sequence segments that correspond to the most conserved regions of proteins. These are generated by searching for highly conserved regions in groups of proteins found in various domain databases. Pfam is a collection of domain family alignments and hidden Markov models (HMMs). It is divided into two parts: PfamA is a manually curated set of alignments and HMMs whilst PfamB families are those sequences in ProDom that are not found in PfamA. SMART (a Simple Modular Architecture Research Tool) also contains HMMs and alignments for each domain family. The manually checked alignments are based on known tertiary structure or homologues identified by PSI-BLAST sequence analysis (Altschul *et al.*, 1997). The alignments are annotated including tertiary structures, functional class and functional residues. InterPro (Apweiler *et al.*, 2000) is a collaboration between many of the curators of the domain databases aiming to reduce the amount of duplication between them. Entries are aimed at the functional classification of new sequences and comparative analysis of whole genomes (Rubin, 2000).

1.5 Domains as units of protein folding

Structural domains can be defined as autonomous folding units. It is thought that such units may be able to fold into a native structure if cleaved from the rest of the protein. Anfinsen (1961) developed a 'thermodynamic hypothesis' to explain the native conformation that is adopted by protein structures based on the study of ribonuclease. By denaturing this single-chain protein it was shown that spontaneous refolding occurred when the protein was freed from the denaturant. This refolding requires that only one of an immense number of possible polypeptide conformations is found within a finite time as described by Levinthal (1968) and now known as the Levinthal paradox. This must mean that renaturation is not random and that structure is determined by sequence. Anfinsen's thermodynamic hypothesis stated that the native state of a protein is that at which the free energy of the protein is at a global minimum (Anfinsen, 1973). This lead to the suggestion that there must be specific pathways for folding.

More recently a model of protein folding has been described which views folding in terms of an energy landscape (Dill and Chan, 1997; Dobson and Hore, 1998). Such folding pathways can be viewed as a folding 'funnel' in which the unfolded polypeptide has a large number of available conformations, whilst the folded state has much fewer. The funnel shape implies as the protein folds, the landscape narrows, such that there is a decrease in energy as more native contacts are made than non-native and a decrease in entropy with the increase in tertiary structure formation. Finally, as the conformational options become fewer and fewer, the protein achieves its native structure. A number of experimental studies of protein folding have suggested that folding begins with the formation of secondary structures, which then assemble into the tertiary structure, driven by hydrophobic interactions (Dobson and Karplus, 1999).

The folding of multi-domain proteins *in vitro* has been observed to occur with low efficiency, possibly due to unproductive interactions between domains during the folding process. It is thought that the evolution of larger multi-domain proteins must have been accompanied by the co-evolution of mechanisms that improve the efficiency of folding of these chains (Ellis and Hartl, 1999). Work by Netzer and Hartl (1997; 1998) suggested that such a mechanism may be the co-translational folding of adjacent domains that are connected by a flexible linker, where one

domain folds before the next domain is synthesised. After being exposed to denaturant it was found that a two-domain fusion protein made up from two single-domain proteins, H-Ras and dihydrofolate reductase, was unable to refold, unlike both single domain proteins which folded efficiently. However, it was found that the fusion protein *did* fold efficiently when synthesised in rabbit reticulocyte extract, where it is thought the sequential folding of the domains avoided inter-domain misfolding. Such co-translational folding may involve small chaperones such as hsp70 and hsp40. Though this explains the folding of continuous domains, the folding of discontinuous domains may involve a post-translational system, perhaps requiring sequestering into chaperonin folding cages to allow efficient folding (Netzer and Hartl, 1998).

Despite these possible difficulties for large proteins the evolution and accretion of protein domains may represent an advantage for protein folding, where several folding units may achieve the native state faster than a single larger folding unit. The slowest step in the folding of a multi-domain protein is the pairing of adjacent domains. This may be a result of incorrectly folded domains, or that the small adjustments that are required to attain the optimal interaction between the domains are energetically unfavourable (Creighton, 1992; Frydmann *et al.*, 1999).

Whilst there is little doubt that most biologically functional proteins fold spontaneously into stable structures, there is growing evidence that this is not the case for all proteins. It is now recognised that many protein domains are intrinsically unstructured and non-globular (Wright and Dyson, 1999). For example many transcriptional activators have been found to only fold when binding to their protein targets (Donaldson and Capone, 1992; O'Hare and Williams, 1992). Such observations pose challenges for genome annotation as it is generally assumed that a protein's function is closely linked to its structure.

1.6 The importance of domain prediction

1.6.1 Genomic sequence analysis

Today, the technology of sequencing has developed to the point where the sequencing of an entire genome is not only a practical proposition but is almost routine in its application. The amount of sequence data produced by these projects is huge, and many of the predicted gene products lack any experimentally determined biological function. The challenge of the 'post genomic' era is the high-throughput examination of the genes and gene products of an organism, with the aim of assigning their functions. In turn, the genome-wide comparison of proteins will be an important tool in identifying the functional linkages between proteins and reconstructing their evolution.

A huge number of functions within the cell are controlled by coordinated interaction networks between proteins (Marcotte *et al.*, 1999). However, only a fraction of these networks has been deduced through biochemical, genetic or structural experimentation. Comparison of sequences between genomes can reveal orthologous (closely related gene sequences between species) and paralogous (closely related gene sequences within species) relationships between proteins of known function. The availability of fully sequenced genomes has enabled the development of computer methods to construct protein interaction networks. Phylogenetic profiling aims to describe the pattern of presence or absence of proteins across a set of genomes, where proteins with the same profile are likely to act in the same cellular process (Eisenberg *et al.*, 2000). The Rosetta Stone method (Marcotte *et al.*, 1999) determines functional linkages between proteins by analysing the fusion pattern of protein domains. Single domain proteins that are found in isolation in one organism may be present in a multi-domain protein in another. The linkage between these domains may demonstrate a linkage in function.

The detection of sequence similarity can be of great use for the characterisation of new proteins. The concept of protein domains is critical to the analyses of sequences (Russell and Ponting, 1998). Because domains often have individual functions, it is important to consider genomic sequence data at this level as this will allow maximum information to be extracted. Assignment of function to predicted proteins by sequence similarity searching using protein family based

resources is gaining more and more importance (Kriventseva *et al.*, 2001) The organisation of sequences into families can provide evolutionary, functional and structural data to organise the huge volumes of information into more structured hierarchies. Difficulties in sequence alignment can occur in cases where proteins consist of multiple domains. For example, similarity searching may assign a single function to a multi-functional, multi-domain protein, resulting in annotation errors that may be passed on to other sequences. Another source of error can occur when using clustering algorithms that are designed to assign large collections of sequences into families. Difficulties can arise where multi-domain sequences contain domains that are related to different families. For example, Figure 1.3 shows three multi-domain proteins, *a*, *b* and *c*, that could be incorrectly assigned to the same family. Protein *a* contains an src homology 2 (SH2) domain and an src homology 3 (SH3) domain, whilst protein *b* contains an SH3 domain followed by a phosphotyrosine-binding domain (PTB). Comparison of these proteins would show significant homology between them due to the shared SH3 domain, and as such will be clustered into a single family. Protein *c* contains a PTB domain followed by a fibronectin type III (FN3) domain. All three proteins may be clustered into the same family, though proteins *a* and *c* are not related, and have been incorrectly associated through association with protein *b*. The identification of individual domains would help avoid these potential errors.

Today there are many ways to detect an evolutionary relationship between two proteins. Pairwise sequence comparison methods such as BLAST (Altschul *et al.*, 1990) and FASTA (Pearson and Lipman, 1988) measure similarities using mutation data matrices such as the BLOSUM or Dayhoff matrices (Henikoff and Henikoff, 1992; Dayhoff *et al.*, 1978). These matrices give the likely variation between homologues and with this information, the extent of similarity between two protein sequences can be measured. The similarity score between the target and template sequences is then used to determine the likelihood of the two proteins having that similarity by chance. A low probability of a chance relationship suggests that the two sequences are related by a common ancestor.

Simpler pairwise alignment methods can also be used to compare two protein sequences. It is generally believed that above a sequence identity of 25-30% (over approximately 100 or more residues) two protein sequences will share a common

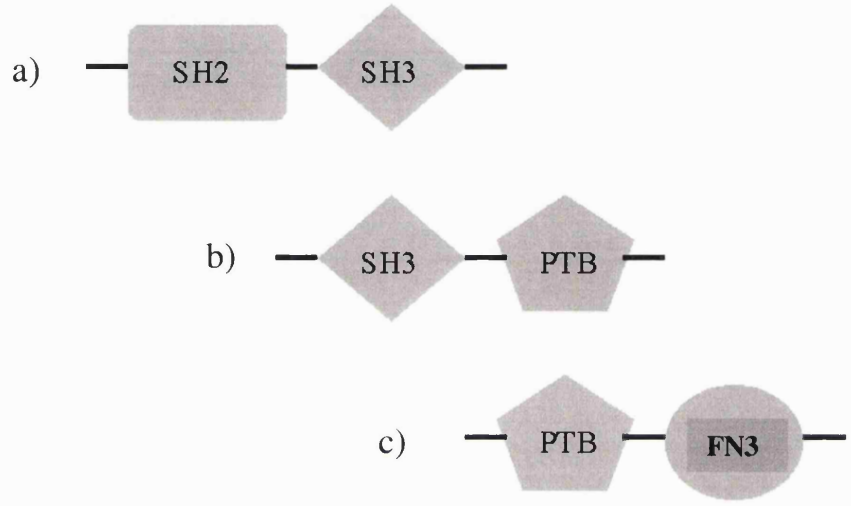


Figure 1.3 Errors in sequence comparison

The three proteins **a**, **b**, and **c** may all be clustered into the same family, though proteins **a** and **c** are not related, and have been incorrectly associated since they both, independently, share a domain with protein **b**.

ancestor. However, below this threshold (the twilight zone), homology can be more difficult to detect (Rost, 1999).

Improvements in database search methods have come from the use of multiple sequence alignments. Park et al., (1998) demonstrated that the detection of remote homology could be extended if intermediate sequences were used. If the relationship between two sequences cannot be detected by sequence-based methods, their relationship may be detected by a third sequence that is intermediate to them both. This method increases the ability to detect remote homology amongst proteins with low sequence identity. Park et al., (1998) noted the importance of using multiple sequences in detecting remote homologues. Sequence profile methods reflect the conservation of amino acids in groups of homologous sequences. It can be extremely difficult to define homology beyond the twilight zone with any certainty. Profiles incorporate more information about a protein sequence, extending the limits of sequence based techniques. The creation of profile based search methods such as PSI-BLAST was a major step in the progression of search algorithms. PSI-BLAST searches a sequence database, like its predecessor BLAST, but with each pass it combines the homologous sequences it detects above a pre-defined threshold to form a sequence profile. This profile is then used to search the database again, and this process is repeated until no new sequences above this threshold are detected, or until it has reached a pre-defined number of iterations. By incorporating information from multiple sequences, the method is more sensitive to variations between related proteins and can therefore detect remote homologues more easily. PSI-BLAST has been used as an important tool in genome annotation (Huynen *et al.*, 1998; Teichmann *et al.*, 1998; Muller *et al.*, 1999; Salamov *et al.*, 1999).

Multi-domain proteins can be a problem when using iterative search algorithms, where common domains may mask weak but significant matches to other domains, or the search profile only allows matches to proteins with a similar domain organisation. Multiple sequence alignment can also be hindered by the presence of proteins containing similar domains, but in different orders. For example, when trying to align a two-domain protein with domain *a* followed by *b*, to a protein with the same domains, but *b* followed by *a*.

1.6.2 Structural genomics

The growth in sequence data has been phenomenal, especially with the completion of genome sequencing projects. Whilst the rate of protein structure determination is also increasing, the gulf between the number of sequences and structures is vast. However knowing the 3D structures of proteins is of huge importance in understanding function, evolution and enabling drug design. The logical step after sequencing a genome is to identify the structure and function of the gene products. Such annotation can be achieved in part by the detection of significant sequence similarity between a novel sequence and a protein of known structure. If such a match is found, a 3D model of the novel protein may be constructed by comparative, or homology modelling. Comparative modelling creates a model of the protein of unknown structure using the backbone conformation of the known structure. Knowledge of the structure of a protein can be used to infer function since function is related to structure. This has given rise to structural genomics which aims to experimentally determine the structures of most, if not all the protein families (Blundell and Mizuguchi, 2000). A wide ranging set of representative structures could then enable comparative modelling to assign structures to a huge number of sequences. Estimates of the number of folds that need to be determined range from 10,000 to 100,000 (Skolnick *et al.*, 2000), and is dependent on whether protein structure space is discrete, with a finite number of folds, or continuous, where structures can morph into each other.

Amongst the major challenges in such initiatives is the high throughput expression, purification and structure determination of target sequences. X-ray crystallography can be used to solve those structures that crystallise and defract well, whilst NMR can be used to determine structures of those protein that express well, but do not crystallise (Montelione and Anderson, 1999). Structure elucidation by both of these methods is often bound by size constraints, especially in the case of NMR which has an upper limit of ~30 kDa. The structural determination of larger proteins is often best achieved by a 'divide and conquer' approach (Campbell and Downing, 1994), where the protein is divided up into its individual folding units. The delineation of proteins into their individual domains will be an important stage in these projects, especially as the easier targets are solved.

Comparative modelling approaches are unsuitable in cases where sequence similarity cannot confidently locate a protein of known structure, for example when sequence identity drops down into the twilight zone between 25-30% (Rost, 1999). Threading methods (Jones *et al.*, 1992; Bryant and Lawrence, 1993) aim to predict the structure based on the observation that apparently unrelated proteins may adopt the same tertiary structure. The sequence is optimally aligned to each structure in a library of folds, and the compatibility of each pair is evaluated using a scoring scheme. The most probable fold is then identified as the template with the lowest energy. The prediction of protein structure by threading techniques is best approached at the domain level since fold libraries are usually made up of domains and computational expense can be minimised (Siddiqui and Barton, 1995).

1.7 Methods for domain prediction

1.7.1 Limited proteolysis

Limited proteolysis is the specific fission of only one or a few peptide bonds in a folded protein chain. Because domains are generally compact and stable, they are more resistant to proteolytic cleavage than other less-structured regions (Hubbard, 1998). Domains can therefore be detected by subjecting natively folded protein to a series of digests, that can be monitored by methods such as high performance liquid chromatography or gel electrophoresis. Mass spectrometry or circular dichroism can then be used to analyse the sequence fragments. The method can fail when the reaction conditions do not ensure a native structure is present, which can result in proteolytic cleavage at sites that would normally be protected by tertiary structure (Hubbard, 1998).

1.7.2 Prediction of domains from RNA

The identification of exons lead to the theory that these DNA regions corresponded to units of protein structure including domains (Blake, 1978; Doolittle, 1978). Such a belief was perpetuated by observations that intron positions may be linked to tertiary structure and function (Traut, 1988; de Souza *et al.*, 1998). There are two main theories to explain the role of introns in evolution. The introns-early theory argues that introns are the relics of the ancestral form of all living organisms.

According to this view, whilst eukaryotes retained introns, prokaryotes lost all introns over the course of time gaining increased efficiency thereby speeding up replication times (Gilbert and Glynias, 1993). The introns-late theory argues that introns are a eukaryotic invention, and that introns are continually inserted, as well as removed from their genes (Patthy, 1999). However a study by de Souza *et al.*, (1998) showed that though many introns appear to have been added to pre-existing genes, some appear to be ancient and related to the boundaries of modular protein structures. Ancient introns were found in phase with codons, whereas new introns are out of phase, appearing to be randomly inserted. Recently Pathy (1999) has argued that exon-shuffling has had a key role in the evolution of extracellular proteins, and correspondingly the evolution of multicellular organisms. Many of these proteins show correlation between domain boundaries and the location of introns, especially those in phase with codons. Introns enable a large variety of proteins to be produced from a single gene by alternative splicing which may explain their prevalence in eukaryotes. It is possible that they have had a similar important role in allowing significant exon shuffling (Pathy, 1999). There is clearly much debate as to the exact stage introns appeared, or alternatively disappeared in evolution. Accordingly, the reliable prediction of domains from pre-mRNA from the analysis of intron location does not at present appear to be possible.

Pausing during mRNA translation plays some role in ensuring proper folding of newly synthesised sequences. Such pausing occurs when rare triplets are encountered in the mRNA because it takes additional time for the corresponding rare species of tRNA to be delivered to the ribosome (Varenne *et al.*, 1984). It has been suggested that the positions of these pause regions may correspond to domain boundaries, facilitating co-translational folding of multi-domain proteins by slowing the rate of polypeptide synthesis at domain boundaries (Thanaraj and Argos, 1996). The profiles of these rare codons may be of use in the prediction of domain boundaries within mRNA.

1.7.3 Domain identification from sequence analysis

Many methods that predict domains use alignment information where a query sequence is searched against a domain database. Such search methods are reliant upon a number of issues. The sequence must have a match within the chosen database. Domain boundaries can only be assigned if domain reordering has occurred at some point in evolution, where similar domains are found in related proteins but in a different context. If domains are located in the same order within their sequences there are no positional constraints with which to identify them. This can lead to false identification of domains, a situation that can also occur due to incorrect domain annotation within domain databases. A variety of methods have been designed to automatically assign domain sequences based on comparative sequence analysis.

ProDom uses MKDOM version 2 (Servant *et al.*, 2002) which attempts to stack all the sequences in SWISS-PROT and TrEMBL. An iterative approach is then taken to identify the shortest sequence in the stack, using PSI-BLAST to find related sequences, generating a domain family and finally to remove all the clustered sequences from the stack. This process is repeated until no sequences are left in the database.

DOMO (Gracy and Argos, 1998) is constructed from an all-against-all comparison of sequences in SWISS-PROT (Bairoch and Apweiler, 2000) based on amino acid and dipeptide composition. Local and multiple alignment similarities are used to cluster the domains. Clustering is based on anchors defined by sets of sequence similar proteins that are then used to identify domain boundaries.

The PASS (prediction of autonomous folding units based on sequence similarities; Kuroda *et al.*, 2000) method bases its domain identification on the stacking of sequences located in a BLAST search of the query sequence. The regions along the query sequence can have varying numbers of matches leading to abrupt increases or decreases in sequence numbers. Regions with greater than 20% increase in the number of BLAST hits, followed by a slope following a change of less than 10% over 30 residues are used to infer domain boundaries.

GEANFAMMER (Park and Teichmann, 1998) is a suite of programs that divide a set of protein sequences into families. There are three main parts: A comparison method that uses an implementation of the Smith-Waterman algorithm (Smith and Waterman, 1981). A clustering algorithm is then used to connect

sequences based on the sequence comparisons. Each cluster is then inspected by a domain cutting algorithm DIVCLUS (Park and Teichmann, 1998) which tests sequence pairs for significance and the extent of alignment between them. This stage is iterative; initially three sequences are accepted as belonging to a domain family if the shorter pair overlap by 70% and all three overlap by 30 residues. The common overlap segment is then used to test successive sequence pairs, where accepted sequences are merged. Non-matching segments are removed from the clusters.

GeneRAGE (Enright and Ouzounis, 2000) is aimed at clustering sequences within and between genomes. BLAST is used in an all-against-all comparison of sequences, that are deemed as similar or dissimilar. In cases of opposing results such as protein *a* matches protein *b* but the reverse is not true, the Smith-Waterman algorithm is used. Multi-domain proteins are detected in cases where protein *b* matches proteins *a* and *c*, but *a* and *c* do not match. Similarity between *a* and *c* is checked by a Smith-Waterman alignment. This process is used to cluster sequences, allowing multi-domain sequences to be shared between clusters without linking them.

1.7.4 Statistical measures to predict domains

Wheelan *et al.*, (2000) described a method to predict domain boundaries based on the observation that domains appear to have limits on their sequence length. Their method, Domain Guess by Size calculated the likelihood of a given chain length having one or more domains and the position the domain boundaries. Probability values are calculated from the distribution of chain and domain lengths for a representative set of proteins with known structure. A ranked list of predictions is provided for each target protein. The method has shortcomings as its first prediction is usually single domain. This is described in more detail in Chapter 4 of this thesis.

1.7.5 Domain prediction based on physical principals

Though many algorithms have been designed to assign domains from coordinate data or by sequence comparison, only a few methods have been published that aim to assign domains from physical principles alone. The lack of a comprehensive theory of protein structure based on physical and chemical principles has limited such attempts, and none has yet appeared to be accurate in their domain assignments.

Busetta and Barrans (1984) aimed to predict domain class and boundaries by simulating the folding of the polypeptide chain. The method first attempts to predict secondary structure using energies derived from a data set of 52 proteins of known structure. Each of the 20 amino acids is assigned a structural character, including conformation, specific contacts, strand direction, and estimated association energies between residues in different strands. These propensities are used to predict secondary structure. Each predicted secondary structure, considered as a nucleation point, is then assigned a 'formation energy' that is calculated for each pentapeptide within the secondary structure. The nuclei with the highest formation energy is then used as the starting point for folding, whilst the remainder are successively considered. The preferred positions of the secondary structures and strength of interaction is calculated according to their separation in sequence, accessibility, and 'decision constraints' for each residue. Interacting nuclei form larger nuclei with greater formation energy. When all the secondary structures are assigned to a domain, the domain boundaries are assigned to positions where the interaction energy between secondary structures is lowest. The multi-domain test set consisted of two-domain chains each with a single domain boundary, as assigned by Wodak and Janin (1981). The method was able to predict 5 out of 11 boundaries within 10 residues. In this method the prediction of secondary structure was based only on a single sequence. Recent developments in secondary structure prediction have used sequence profile based methods which has considerably improved prediction accuracy (Rost and Sander, 1993; Jones, 1999).

The method of domain assignment by Vonderviszt and Simon (1986) is based on finding low strength, short range interactions in plots of statistically determined short range preferences between amino acids along the polypeptide chain. This is based upon previous observations that short-range regularities exist in primary

sequence between pairs of nearest and next nearest amino acids. They concluded that every amino acid has a characteristic sequence environment defined by short-range interactions that play a dominant role in the stabilisation of protein structure. The interactions between consecutive amino acids at domain boundaries would be weak in comparison to short-range interactions within domains. For a given protein, propensities are assigned to each amino acid and then smoothed using a moving window along the chain. Predictions were made for 4 multi-domain proteins, where boundaries were predicted from local minima in the scans. However, other significant minima also appeared, often corresponding to alpha-helices and beta-sheets.

Kikuchi *et al.*, (1988) have described a method to assign domains both in proteins of known structure and those of unknown structure. In each case predictions are made by analysing the density of occurrence of contacts in various regions of a contact map. This contact map is constructed from a real distance map for 3D structures or an average contact map for amino acid sequences which is generated from statistical data. Interactions between residues within a protein are subdivided into a series of ranges according to their separation in amino acid sequence; short, medium and long. The long range interactions are further divided into a number of groups. Within each range (except the short-range) the average spatial distances between every pair of amino acids is calculated from a set of 42 proteins of known structure. If the average distance for a given pair of residues is less than a cut-off distance then this pair is predicted to form a contact. An average distance map is then constructed by plotting all pairs in contact. Determination of domain boundaries of compact regions are based on changes of density that occur when the contact map is scanned horizontally, in other words domain boundaries are based on changes in density. The length of the sequence assigned to a domain by this method was generally shorter than other methods as it was more tuned to identifying the compact core of a domain.

1.8 Aims

Domains are an important aspect of molecular biology as they can be used to describe structural, functional and evolutionary relationships between proteins. The identification of domain boundaries within a protein sequence is an important initial step in the preparation of proteins for structure determination, protein engineering and the efficient use of structure prediction methods. In addition to this, comparative sequence analysis is more successful when using an individual domain, rather than a whole sequence. Domain boundaries can be assigned if a given protein has a known tertiary structure, either by visual inspection or by automated assignment from the co-ordinate data. Alternatively, domain boundaries may be assigned if a given sequence has homology to a known domain sequence. In the absence of either a known structure, or sequence homology, a method to delineate domain boundaries from sequence alone is required.

The major aim of this study is to predict structural protein domains from amino acid sequence using computational methods. Such domain boundary assignment might be based upon a number of previous observations. For example, domain linking peptides appear to have distinct characteristics. A statistical analysis of 51 domain linking peptides by Argos (1990) found that the amino acids threonine, serine, proline and aspartate were desirable linker constituents. Compact globular structures in proteins have been shown to be determined by amino acid sequences of high informational complexity (Wooton, 1994) whilst low complexity regions may be found between domains (Gouzy *et al.*, 1999). The length of domains can vary, although a large majority are less than 200 residues (Siddiqui and Barton, 1995) with an average length of approximately 140 residues (Xu and Nussinov, 1997). Small domains, below 40 residues are likely to be stabilised by disulphide bonds (Dill, 1985), whilst sequence lengths greater than 300 residues often consist of more than one hydrophobic core (Garel, 1992). Finally, domains have been defined as autonomous folding units (Welaufer, 1973). The underlying mechanism for the formation of a single hydrophobic core by an individual domain, that is independent of neighbouring domains might be identifiable from sequence. The packing together of secondary structural elements results in the tertiary structure or fold of a protein domain. Recent developments in secondary structure prediction (Jones, 1999) have greatly improved prediction accuracy thereby increasing the usefulness of predicted

secondary structure including the recognition of protein folds (McGuffin *et al.*, 2001). The analysis of patterns of predicted secondary structure elements within a multi-domain sequence may enable the assignment of protein domains.

The work in this study aims to develop computational methods for domain assignment. A number of domain properties are considered to address the possibility of domain prediction from sequence. These include analyses of domain linking peptides, investigation of the surface area and the average hydrophobicity of protein domains, and the use of secondary structure content of domains for prediction. Two domain prediction methods are developed and tested on proteins of known structure. An analysis of protein domain-swapping is also made, where a simple method is developed to identify putative domain swapped structures in protein tertiary structure. An investigation into the properties of swapped-domain linkers is also made.

Chapter 2

A survey of domain linking peptides

2.1 Introduction

Many large proteins contain two or more domains and appear to have been created as a result of the combination of single domain proteins (Apic *et al*, 2001). These variable multi-domain architectures are thought to have evolved from gene fusion and recombination events within the genome (Ponting and Russell, 2002). The combination of domains within a single structure may have conferred a number of selective advantages to the cell, for example, multi-functional proteins with a fixed concentration of active sites able to efficiently catalyse a series of pathway steps. Substrate binding or catalytic sites of multi-domain proteins are often situated in a cleft between domains where movements between them allows the active site to close and protect the substrate from solvent via a mechanism of induced fit (Gerstein and Krebs, 1998).

The region of polypeptide chain that connects neighbouring domains is often termed the domain linker. These inter-domain regions of polypeptide are thought to play a vital role in the evolution of multi-domain systems, where their connectivity has enabled the assembly of domain 'building blocks' into multi-functional proteins (Gokhale and Khosla, 2000). A number of studies have shown the importance of the domain linker in the activity of multi-domain proteins by alterations in amino acid composition or length of the linker peptide.

Mattison *et al.*, (2002) demonstrated the importance of the domain linker in response regulator protein OmpR where it was found to play a key role in communication between the N- and C-terminal domains. Phosphorylation of the N-terminal domain enhances DNA binding affinity of the C-terminal domain, and conversely, DNA binding by the C-terminal domain increases the phosphorylation of the N-terminal domain. The activity of OmpR was decreased on alteration of the length or amino acid composition of the linker peptide.

Spitzfaden *et al.*, (1997) investigated the interactions between the ninth and tenth type IIII modules of fibronectin and demonstrated that the activity of the module pair could be altered by changing the length of the linker. Other studies have similarly revealed the importance of domain linkers, each showing its role in allowing the correct orientation and distance between domains to achieve a stable structure, capable of wild-type levels of activity (Hegvold and Gabrielsen, 1996; Sauer *et al.*, 2001).

The advent of gene-fusion techniques has enabled the construction of chimeric proteins by combining domains from unrelated proteins. This has provided a means to increase the expression of soluble proteins, facilitate protein purification, and produce new protein domain and functional combinations (Arai *et al.*, 2001). The construction of fusion proteins involves linking protein domains by means of a peptide linker, the correct selection of which can be crucial for the engineering of a functional fusion protein (Sorensen *et al.*, 2002). For example, work by Gokhale and Khosla (2000) demonstrated the importance of inter-domain linkers in modular polyketide synthases. These multi-enzyme proteins contain a series of domains which form an assembly line for the biosynthesis polyketides. By engineering new combinations of these domains, it was found that the transfer of polyketides between fused domains was dependent on the use of a linker that could provide the correct module connectivity, and as such play a crucial role in the catalysis of this protein.

A study of domain linker peptides was made by Argos (1990) on 32 multi-domain proteins, which contained a set of 51 linker peptides. Though this analysis made use of as many protein structures as were available at the time, the last ten years has seen a substantial increase in the number of solved protein structures (almost a 30 fold increase in the number deposited). The work in this chapter aims to update this analysis, by making use of an enlarged data set, and to see if the conclusions made by Argos (1990) are still applicable to a larger set of domain linking peptides.

The study by Argos (1990) observed that preferred linker amino acids tended to be polar or hydrophilic, with large and bulky side-chains being avoided. Threonine, serine, proline and aspartate were the most preferred amino acids with linker peptides tending to be on average as flexible as other protein regions.

Here a number of domain linker characteristics are a surveyed, including their structural characteristics, flexibility, amino acid propensities and solvent accessibility. These properties are considered in relation to the role domain linkers have been shown to play within protein structure, and are also used to define a number of linker-classes which may be of benefit for construction of gene fusion chimeric proteins.

2.2 Methods

2.2.1 Data set

The analysis of domain linkers was made on a representative set of multi-domain protein structures. Representatives in the non-redundant set were taken from the CATH database version 1.7 (Orengo *et al.*, 1997). The data set was constructed to contain proteins with well determined structures, for this reason structures with resolutions > 2.5 Å and those solved by NMR were excluded. Also excluded were domains formed by more than one chain. A pair-wise alignment of the CATH sequences generated a non redundant set of 1177 chains, sharing a sequence identity of no more than 30%. Sequence alignments were performed using CLUSTALW (Thompson *et al.*, 1994). Domain assignments for each chain were taken as those given by CATH. The resulting set consisted of 1177 chains. Of these 786 were single domain chains, which were used along with the remaining proteins to calculate characteristics (such as amino acid propensities) found in all proteins. The set also contained 391 multi-domain chains, 214 of which contained only continuous domains, and 177 chains contained at least one discontinuous domain. This gave an analysis data set of 747 domain linking peptides.

Secondary structure assignments for each chain were assigned by the DSSP program (Kabsch and Sander, 1983). The eight secondary structure states given by DSSP were converted to three secondary structure states (Rost and Sander, 1993). Residues assigned as H and G were taken to denote helical residues, E and B to be strand residues, and the remainder were considered as coil. Furthermore, a strand was defined as a consecutive run of three or more residues assigned as strand, and a helix, five or more residues consecutively assigned as helix. This simpler scheme was used to clarify secondary structure element assignments, and in turn, allow more decisive delineation of the domain and linker regions.

2.2.2 Identification of domain linker sequences

Initial domain boundary definitions were taken from the CATH database for consistency. Before a multiple domain protein structure can be classified in CATH, it must be parsed into its separate domains. This process involves an automated consensus method, together with manual intervention when agreement

cannot be found (Jones *et al.*, 1998). Once domains have been located, a cut-point is recorded along the chain spanning between adjacent domains in sequence. However, although this method does give the approximate location of a putative domain linker, it does not delineate its boundaries within the sequence. Therefore domain linker regions were identified by visual inspection, using the CATH domain boundary cut-points as a starting point to their delineation. Protein structure co-ordinates taken from the Protein Data Bank (PDB; Bernstein *et al.*, 1977) were viewed through the Rasmol protein structure viewer (Sayle and Milner-White, 1995).

To maintain as consistent a linker assignment process as possible over the representative set, an assignment protocol was built from initial observations of the domain linking peptides. The linker peptide was defined as the region of chain spanning between domains. First domains were outlined and helix, strand and coil elements were partitioned to the appropriate globular domain unit assigned by CATH. The linker sequence was then assigned as those residues found between the last helix or strand belonging to the C-terminal linked domain and the first helix or strand belonging to the N-terminal linked domain. In other words, the linker boundaries were taken to be those residues adjacent to helices or strands belonging to the linked domains. Any secondary structure elements that did not appear to belong to either adjacent domain were assigned as all or part of the linker. Assignments of domains containing few secondary structures involved a slightly different protocol. As such domains rarely contained helices or sheets that bordered the linker peptide, the linker terminus were taken as the residue found on the boundary of the outlined domain. All other regions or residues in the protein not assigned as part of the domain linker are referred to as 'non-linker' in this chapter.

2.2.3 Amino acid composition

Amino acid frequencies were calculated from the sequence records in the corresponding PDB files for: linker peptides, non-linker coil regions and for all residues in the representative set. From these frequencies linker amino acid propensities could be calculated for residues in linker peptides and residues in non-linker coil. For example residue propensity values in the linker peptides, compared to the protein as a whole were calculated as follows:

$$P_{linker} = \frac{\left(Nr_{i,l} / \sum_i Nr_{i,l}\right)}{\left(Nr_{i,p} / \sum_i Nr_{i,p}\right)}$$

where P_{linker} is the propensity for residue i, $Nr_{i,l}$ and $Nr_{i,p}$ are the number of amino acid i in the domain linker set (l) and in all proteins (p) respectively. $\sum_i Nr_{i,l}$ and $\sum_i Nr_{i,p}$ are the total number of amino acids in the domain linker set and in the full protein set respectively.

Residue propensities give the relative importance of each amino acid. In this example, values greater than one indicate a residue type is more favoured compared to all non-linker protein regions.

2.2.4 Linker peptide flexibility

Residue flexibility was measured by crystallographic temperature factors (B-values) taken from PDB files. The B-values or atomic displacement parameters give a measurement of the flexibility of a given residue within the protein structure. However, the average B-values observed in different protein structures can vary to a large degree, complicating comparisons between different protein structures (Parthasarathy and Murthy, 1997). This can be overcome by expressing values in terms of standard deviation about the mean B-value of a given protein. For each protein in the representative set the mean B-value and standard deviation over all C-alpha positions were calculated. The B-values for C-alpha positions in a given protein were then normalised as follows, where B is a given B-value, $\overline{B}$ is the mean B-value and $\sigma(B)$ is the associated standard deviation.

$$\text{Normalised B-value} = \left(B - \overline{B}\right)/\sigma(B)$$

Frequency distributions for normalised B-values were calculated for linker, helix, strand and non-linker coil residues. These were calculated in intervals (or bins) of normalised B-values of size 0.5 over the data set.

2.2.5 C-alpha extension

The C-alpha extension for each linker peptide was calculated by dividing the distance (in Å) between the N-terminal and C-terminal C-alpha coordinates of the linker peptide, by the length of the peptide minus one.

The C-alpha extent of non-linker helix, strand and coil elements was also calculated for twenty randomly selected pentapeptides (corresponding to each secondary structural state) as described in the study by Argos (1990). In this case the mean C-alpha extensions for the pentapeptides were given by division of the distance between the pentapeptide terminal C-alpha residues by 4 (i.e. length minus one).

2.2.6 Solvent accessibility

The degree of burial of a given residue was described by its solvent accessibility within the protein structure as calculated by the program DSSP. Percentage relative solvent accessibility (RSA) was calculated by normalising the accessible surface area with the maximum values found in a GLY-x-GLY conformation given by Rose *et al.*, (1985). Thresholds were used to define different states of solvent accessibility including buried (RSA≤10%) and exposed (RSA>10%). The mean accessible surface area values were also calculated for residues in linkers, non-linker coil and all non-linker positions.

2.2.7 Residue conservation

Each residue in the non-redundant set was assigned a score relating to its conservation in evolution as given in the database of homology-derived secondary structure of proteins (HSSP; Sander and Schneider, 1991). Sequence conservation, described in HSSP as variation, for a given residue is derived from its position in a multiple sequence alignment based on the Dayhoff exchange matrix. The variability score ranges from a value of 0, extremely conserved, to 100, indicating no residue conservation at this position. This measure is derived from a multiple sequence alignment, where the values are scaled depending on how many sequences there are in the alignment. It is possible that residues with low occupancy in the alignment, i.e. few alignments span that position, may have unreliable variability scores. For this

reason, only amino acids in positions with an occupancy of greater than 5 residues were taken into account (Sander and Schneider, 1991). To avoid any contribution that solvent accessibility might have on the degree of residue conservation, each residue in the protein was assigned to an interval or bin according to its relative solvent accessibility. Mean variability scores were calculated over the non-linker region of the protein for five bins of RSA; [0,5), [5,10), [10,15), [15,20), [20,100) $Å^2$. A linker residue was deemed to be significantly less variable if its score was greater than 1 standard deviation below the corresponding non-linker mean for a similar accessible surface area.

2.2.8 Hydrogen bonding

The HBPLUS algorithm (McDonald and Thornton, 1994) was used to define hydrogen-bonds within the linker peptides. The average number of hydrogen-bonds per linker residue was calculated.

2.3 Results

2.3.1 Secondary structure of domain linkers

The frequency of protein secondary structure assignments given for linker residues is shown in Table 2.1. These frequencies are shown for the eight different secondary structure states as assigned by DSSP and also for the frequencies of helical, sheet and coil residues given by the simplified secondary structure assignment scheme (section 2.2.1). It can be seen that the majority of linker residues are found in a coil conformation. The DSSP eight state assignments show 85.5% of linker residues to be in a coil, turn or bend region of structure. This bias towards an unstructured coil conformation becomes more pronounced using the simplified scheme with nearly 94% shown to be coil residues - in fact nearly 84% of the linkers were assigned as all-coil regions of chain, whilst just over 9% were predominantly coil regions, containing short helical or strand elements.

The nature of a discontinuous domain requires that two or more linkers are needed to connect the discontinuous domain segments (which together form the discontinuous domain region) to adjacent domains in the multi-domain protein. It became apparent when visually delineating such linkers that many were closely

Secondary structure state (DSSP)	DSSP assignment	Simplified scheme
	Percentage of residues	
C	52.10	93.90
S	16.80	0.00
T	16.60	0.00
E	3.90	2.30
B	3.30	0.00
H	3.60	3.80
G	4.70	0.00
I	0.00	0.00

Table 2.1 Secondary structure assignments of linker residues

The eight secondary structure states assigned by the DSSP programme are defined as; C, coil, S, bend, T, turn, E, beta-strand, B, isolated beta-bridge, H, alpha-helix, G, 3_{10}-helix, I, alpha-helix. The percentage of domain linker residues assigned to each secondary structure state by DSSP is shown as well as these percentages using the simplified secondary structure assignment scheme (section 2.2.1).

associated in the structure, such that beta-sheet regions of structure could be formed. These sheet regions were most often composed of two linkers. An example can be seen in Figure 2.1, which shows the structure of thioredoxin reductase from *E. coil* (Waksman *et al.*, 1994). This enzyme is made up of two domains - an FAD and NADPH binding domain, which are connected by two linkers which form an anti-parallel beta-sheet between them. It is interesting to note that Waksman *et al.*, (1994) propose that the orientation of the two domains can change by nearly 66° on catalysis, which must either be permitted by the linker sheet or may result in the dissociation of the beta-sheet.

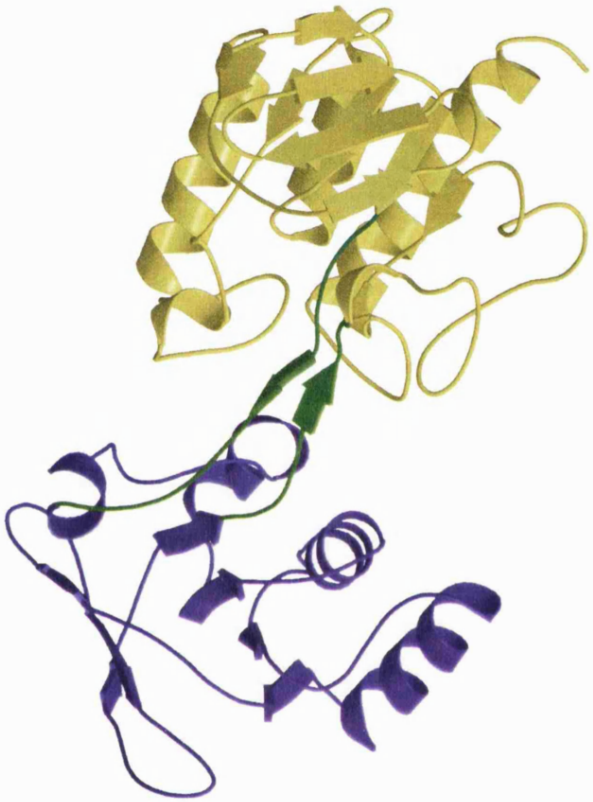
A large proportion of linker residues assigned as beta-strand can be attributed to beta-sheet formation between adjacent linkers associated with discontinuous domains. In many cases, all residues of such linkers formed part of the beta-sheet, and as such were considered as all-beta linkers. Nearly 80% of these all-beta linkers appeared to be as a result of sheet formation with adjacent linkers, as described above. Altogether 7% (52 linkers) of all the domain linkers identified were found to be all-alpha or all-beta secondary structural linkers. In other words, the linker sequence joining adjacent domains consisted solely of a continuous stretch of residues forming a helix or strand. These linkers were considered separately from the remaining linkers analysed in this study (section 2.3.7).

2.3.2 Length distribution of domain linkers

The length distribution of the domain linking peptides outlined in this study (section 2.2.2) is shown in Figure 2.2. It can be seen that though the distribution of lengths is large, between 2 to 33 residues, the majority of linkers were found at the lower end of this range. The mode of the distribution is 8 residues, with the mean calculated as 9.8 residues. Observations made of the linker structures and the associated length distribution made it possible to class linker according to their length. The ease of assignment of domain linkers ranged in difficulty, often depending on the degree to which adjacent domains were associated, i.e. it was often more difficult to definitively assign domain linkers to tightly associated domains. These linkers were often short, between 2 to 4 residues, and are considered as a small linker class, representing just over 17% of all linkers. Around 65% of the linkers were of an intermediate length, between 5 and 12 residues, with the remaining longer

Figure 2.1 Example of beta-sheet formation between domain-linkers

The structure shows the two-domain enzyme, thioredoxin reductase from *E. coli*, with the discontinuous FAD domain coloured in yellow and the continuous NADPH domain coloured in purple. Two linkers connect the domains and form an anti-parallel sheet between them (coloured green).



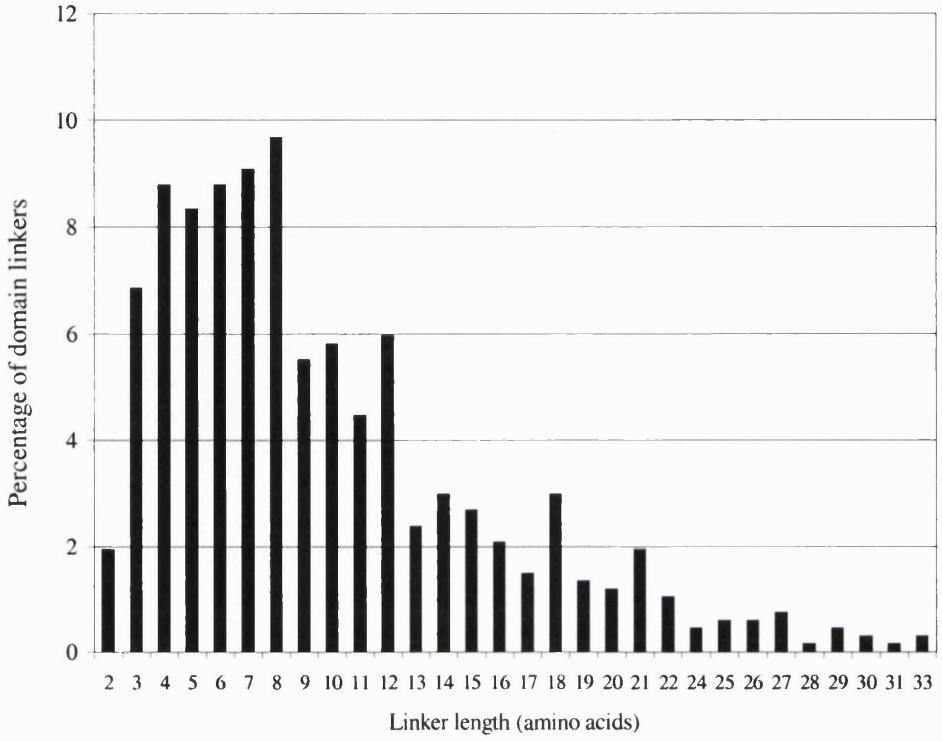


Figure 2.2 Length distribution of domain linking peptides.

linkers being 13 or more residues in length. In general, the longer the linker, the less frequently it occurs, although the linkers of length 18 and 21 residues seem to be outliers in this distribution, being more frequent than might be expected. In general, those linkers containing small regions of alpha or beta secondary structure tended to be longer.

2.3.3 Linker amino acid propensities

The frequency of each amino acid found in the linker peptides and in all proteins, together with the associated linker residue propensity values are shown in Table 2.2 (section 2.2.3). Propensity values greater than one indicate a prevalence of the corresponding amino acids in linker regions compared to those found in proteins in general. Proline is clearly favoured, being nearly twice as prevalent in linkers as it is in all protein regions (propensity of 1.94). This is followed by glycine (propensity of 1.19), and then asparagine, threonine, aspartate and lysine, all having similar propensity values (1.18 to 1.12). It can be seen that the majority of polar residues have linker propensities greater or very close to one. There appears to be no overall preference for charged residues, with a propensity value for all charged residues of 1.01. Hydrophobic and bulky residues including the aromatic polar residues tend to be disfavoured. The overall propensity for hydrophobic residues is calculated as 0.95, however this includes proline, which is highly favoured (the overall hydrophobic residue propensity is 0.8 excluding proline). It can be seen that hydrophobic residues such as valine, isoleucine and methionine have low propensity values. Tryptophan is easily the most disfavoured residue, occurring nearly half as many times as it is in proteins generally. The high proline content contradicts this rule although it is one of the smallest of the hydrophobic residues. The observation that proline and glycine are both favoured is perhaps no surprise, as the combination of these residues will enable sudden changes in direction of the linker, which may be necessary to link adjacent domains. This is discussed to a greater extent in section 2.4.

When rank correlation (Spearman's) is calculated between the propensities obtained for the domain linkers in this study and those given in the study by Argos (1990), a correlation r=0.74 is found, demonstrating fair agreement in rankings between the two analyses. However, ranked correlation may overstate the

Amino acid	Percentage		Propensity
	In Linkers	In all Proteins	
PRO	9.79	5.04	1.94
GLY	9.17	7.68	1.19
ASN	5.42	4.58	1.18
THR	6.64	5.74	1.16
ASP	6.80	5.90	1.15
LYS	6.32	5.64	1.12
SER	6.48	6.06	1.07
HIS	2.39	2.44	0.98
GLN	3.53	3.73	0.95
PHE	3.70	4.10	0.90
ARG	4.40	4.86	0.90
TYR	3.22	3.71	0.87
GLU	5.44	6.35	0.86
LEU	7.11	8.36	0.85
ALA	6.75	8.00	0.84
CYS	1.23	1.53	0.80
VAL	5.16	6.95	0.74
ILE	4.05	5.62	0.72
MET	1.49	2.10	0.71
TRP	0.91	1.59	0.57
Polar	48.78	42.12	1.16
Charged	22.96	22.75	1.01
Hydrophobic*	28.26	35.15	0.80

Table 2.2 Amino acid propensities in domain linkers

* excluding proline (see section 2.4)

relationship, and if one calculates the correlation between the actual values as shown in Figure 2.3a, a lower correlation, r=0.64, is found. It can be seen from this figure that several amino acids, (though ranked similarly) are found in different proportions in each analysis, including proline, cysteine and tryptophan.

A comparison of the propensities obtained in this work to propensities found for residues in all non-linker coil regions in the representative set used in this analysis is shown in Figure 2.3b. In this case, a rank correlation of r=0.68 is calculated, (r=0.78 using actual values). Figure 2.3b shows a reasonably tight distribution between the residue propensities of domain linker and non-linker coil, though proline is a clear outlier in the group, being much more highly favoured in domain linkers.

Figure 2.4 shows the amino acid propensities calculated for residues in the small, intermediate and large linker classes. The intermediate and large sized linkers appear to have similar residue propensities to those calculated for the entire linker data set. Small linker propensities however differ to a greater extent from those for all linkers, most especially for cysteine and methionine which are more favoured and aspartate which is less favoured. The largest differences are for glycine and tryptophan. Small linkers have a much higher preference for glycine (propensity of 1.66) compared to all the linkers (propensity of 1.19), and a much lower preference for tryptophan which has a propensity of 0.09 in small linkers compared to the propensity of 0.57 for all linkers.

2.3.4 Flexibility of linker residues and C-alpha extension

Crystallographic temperature values (B-values) were taken from the PDB and used as an indication of residue flexibility within the associated protein structure, (section 2.2.4). Values were normalised according to the mean and standard deviation B-values for each corresponding protein. Values above the normalised mean of zero show a higher than average flexibility in the protein as a whole, whilst those below zero are considered to be more structurally constrained. Flexibility measurements of the linker residues were compared to those calculated for non-linker coil, helical, and strand residues (for all proteins in the representative set).

Figure 2.5 shows the distribution of these values together with tabulated values for the mean, associated standard error and standard deviation of each

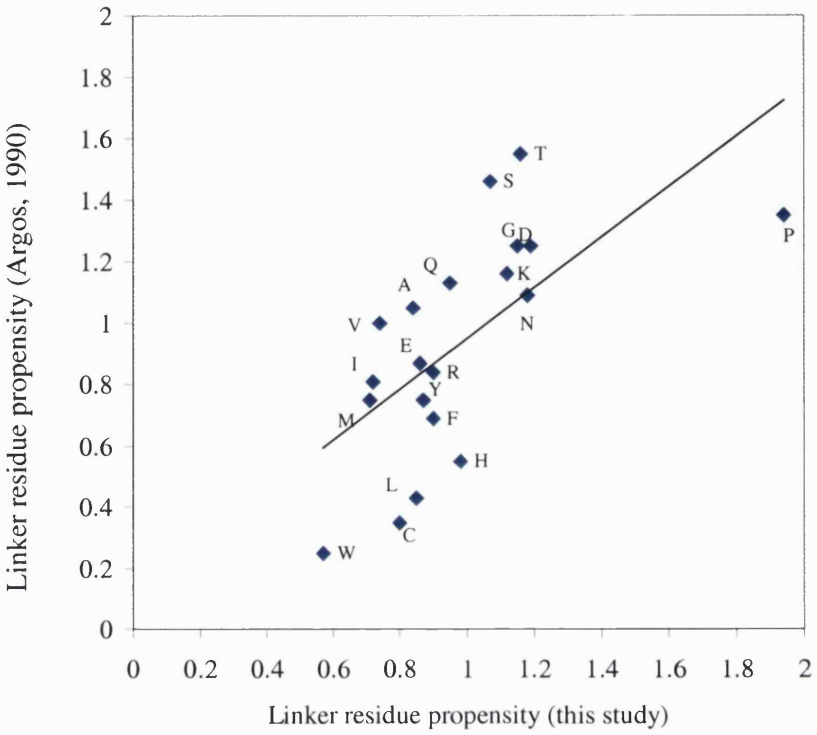
Figure 2.3 Comparison of domain linker amino acid propensities

Domain linker amino acid propensities calculated for the linkers in this study were compared to:

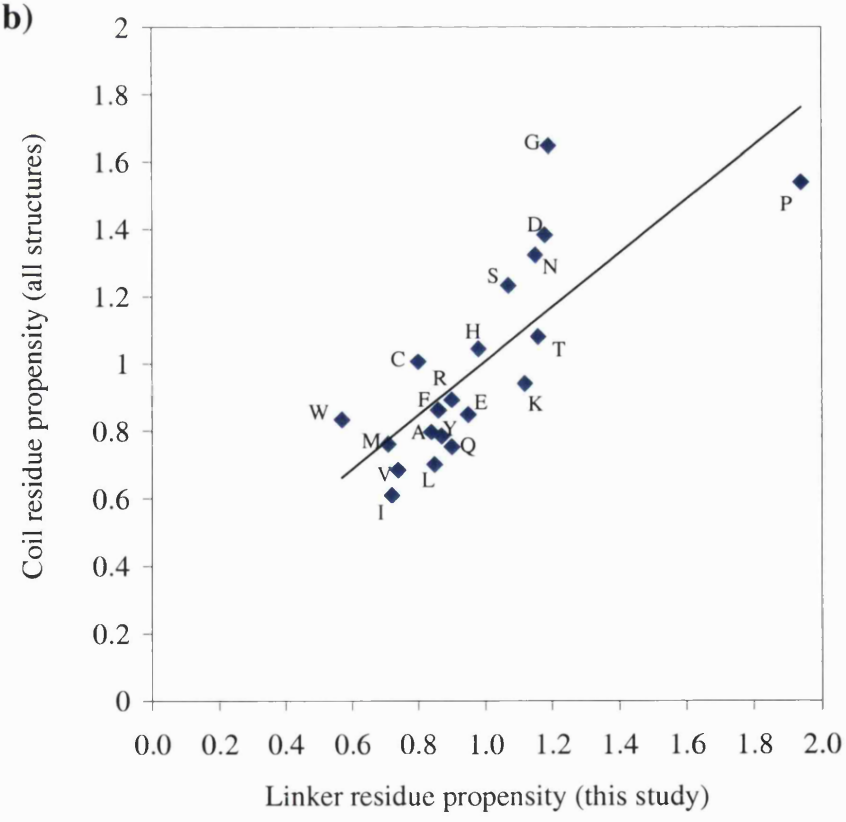
a) linker propensities given in the study by Argos (1990).

b) amino acid propensities calculated for non-linker coil residues (over all structures in the non-redundant set).

a)


Linker residue propensity (Argos, 1990)
Linker residue propensity (this study)
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
2
T
S
G
D
P
Q
K
A
N
V
E
I
R
Y
M
F
H
L
C
W

b)


Coil residue propensity (all structures)
Linker residue propensity (this study)
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
2
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
1.8
2.0
G
P
D
N
S
H
T
C
R
K
F
E
W
A
Y
M
Q
V
L
I

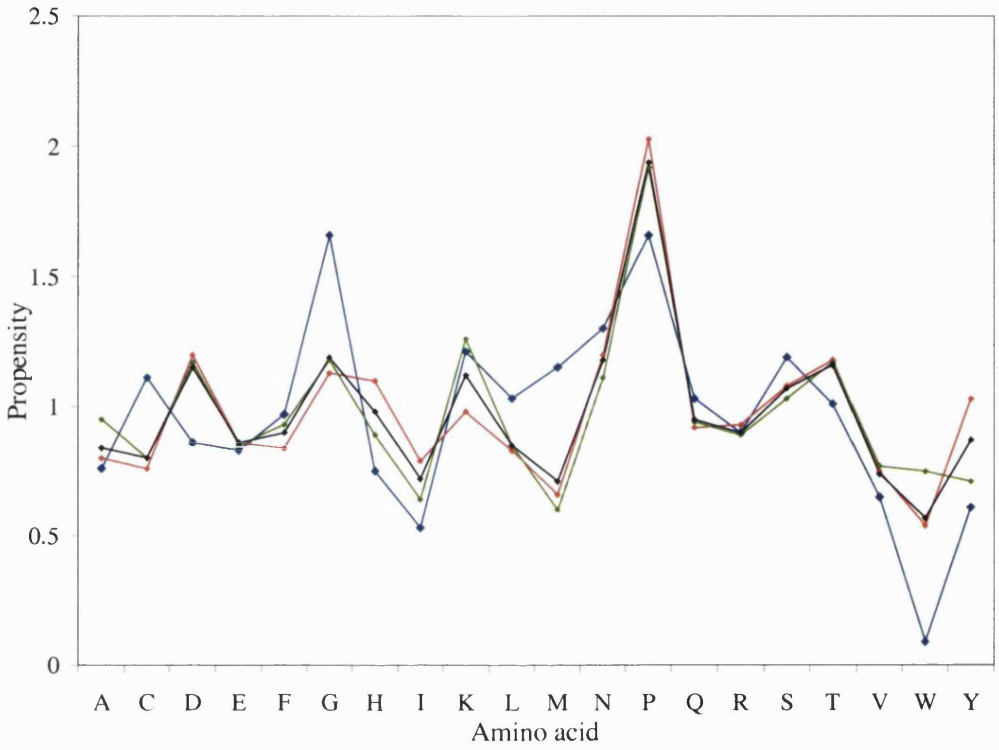


Figure 2.4 Amino acid propensities for small, intermediate, large and all linkers

Propensities for small linkers shown in blue, those for intermediate linkers in red, those for large linkers in green and for all linkers in black.

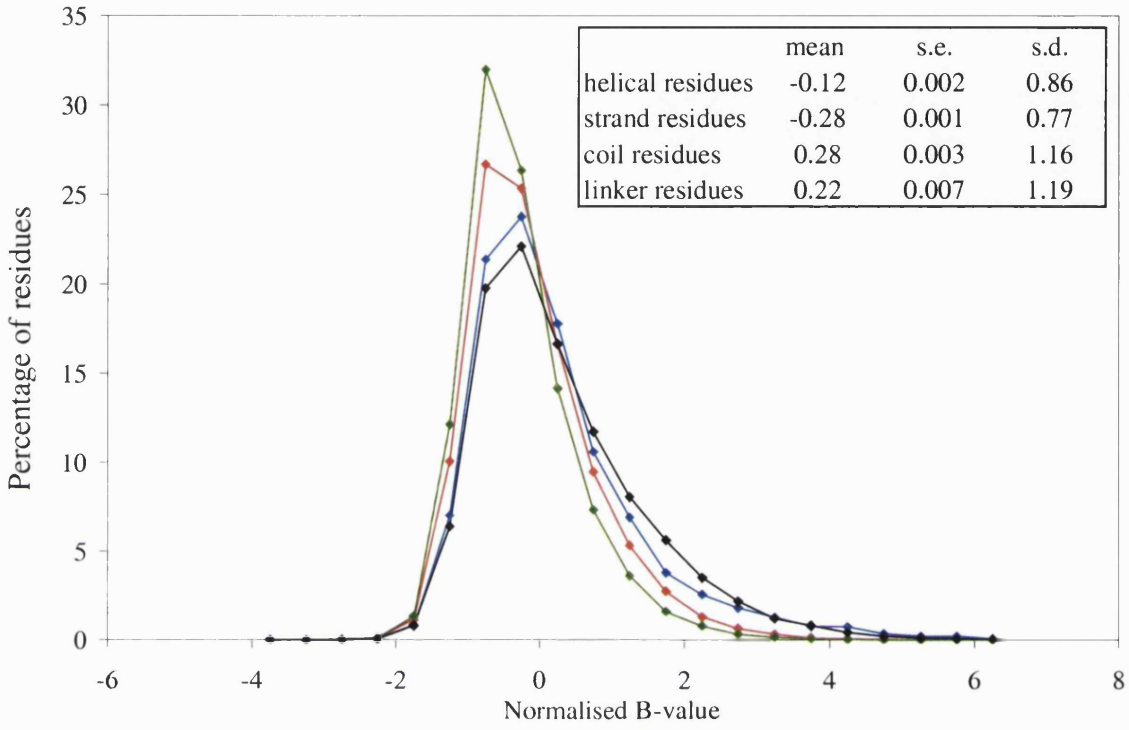


Figure 2.5 Distribution of normalised B-values for linker, non-linker coil, helix and strand residues

The distribution for linker residues is shown in blue, for helical residues in red for strand residues in green and non-linker coil resides in black. Inset is the mean, associated standard error of the mean and standard deviation for each of the distributions.

distribution. Though the distributions of B-values can be seen to overlap to a considerable extent the means of each analysis set are significantly different at the 5% level given the calculated standard errors. This was confirmed by the calculation of the non-parametric Wilcoxon ranked test, which showed the means to be significantly different (data not shown). Analysis of the distributions and corresponding standard deviations show strand and helical residue values to be distributed closer to the normalised mean, i.e. are less flexible, than coil residues perhaps to be expected due to their more constraining hydrogen-bond patterns. The distribution of values for non-linker coil residues can be seen to be skewed towards the flexible end of the plot with fewer residues having values close to the normalised mean. Overall the distribution for linker residue are similar to those shown for coil residues, demonstrating a similar level of flexibility. For values above the above the normalised mean, the frequency distribution for linker residues tends to be higher than the distribution for helices and strands, but below the distribution for coil. However, there appears to be a small 'bump' at normalised values above 4, where the linker distribution overtakes the coil distribution. Such a peak may indicate that a greater number of linker resides are capable of extreme flexibility compared to coil residues, although it is quite possible that this may be a result of small number statistics. There are many more observations for non-linker coil residues, giving a smoother distribution, whilst extreme values in the linker distribution may be open to more fluctuations due to a smaller amount of available data.

The distance between the N-terminal and C-terminal C-alpha residue of each linker was calculated (section 2.2.5). The average distance between adjacent residues, together with associated standard deviations are shown in Table 2.3. Also shown are the values for helical, strand and non-linker coil regions. It can be seen that the mean C-alpha extension of linker residues is 2.21 Å, smaller than the extension calculated for general coil residues, 2.84 Å, but more extended than the mean extension found for helical structures, 1.54 Å.

	% total exposed residues	RSA		C-alpha extension Å	
		mean	s.d.	mean	s.d.
helical residues	56.4	23.3	24	1.54	0.07
strand residues	42.6	15	18.8	3.21	0.18
coil residues	72.3	33.6	27.8	2.84	0.53
continuous domain linker residues	77.4	31.6	26.5	2.26	0.52
discontinuous domain linker residues	66.7	28.8	26	2.16	0.82
all linker residues	71.3	31.8	26.5	2.21	0.78

Table 2.3 Residue solvent exposure and C-alpha extension

Percentage of exposed residues for each structural element, mean relative accessible surface area and standard deviation for helical, strand, coil and linker residues. Linkers connecting only continuous domain regions are termed continuous domain linkers and linkers connecting only discontinuous domain segments are termed discontinuous domain linkers. The mean C-alpha extension per residue pair and associated standard deviation is also shown.

2.3.5 Solvent accessibility and sequence conservation of linker residues

The relative solvent accessibility (RSA) was calculated for each residue in the data set (section 2.2.6). Residues with an RSA greater than 10% were defined as exposed, whereas those with less than or equal to 10% were defined as core residues. The mean percentage of exposed residues found in all the linker peptides, the linkers joining only continuous domains, those joining only discontinuous linker segments is shown in Table 2.3. It can be seen that linkers joining discontinuous domain segments appear to be less exposed than linkers joining continuous domains, with a total of 66.7% of residues being exposed compared to 77.4% respectively. Overall, linkers tend to have a similar percentage of exposed residues (71.3%) compared non-linker coil (72.3%).

The calculations of mean RSA values for non-linker helical, strand, coil residues and linker residues can also be seen in Table 2.3. Linkers connecting discontinuous domain segments are less exposed (with a mean RSA of 28.8%) than continuous linker residues, mean RSA 31.6%. Calculations for all the linker residues show they are as exposed as coil residues whilst residues in strand elements which have a mean RSA 15.0% and helical elements with a mean RSA 23.3% are less exposed. The large standard deviation associated with these mean RSA values make it difficult to place too much significance on such comparisons, and reveals the difficulties in measuring such variable values.

The values in Table 2.4 further demonstrate that domain linking residues are more exposed than residues found in proteins generally. Residues within linkers expose over 33% more surface area to solvent than do average residues. Linker residues however, expose just over 7% more surface area than non-linker coil residues found in the remainder of the protein structures.

The comparison of mean variability values, for 5% intervals of RSA, between linker residues and non-linker coil residues is shown in Table 2.5. As described in section 2.2.7, residues were assigned to bins according to their respective solvent exposure, to take into account the effect that residue exposure to solvent may have on the rate of mutation. It can be seen that the variability between residues, where a value of 100 indicated a totally conserved residue, is similar between linker and coil residues. Again large standard deviation values makes it difficult to read too much

	Average accessibility (Å^2)			% change	
	all non-linker	non-linker coil	linker	all vs linker	coil vs linker
Ala	21.68	32.24	31.11	48.71	-3.50
Arg	12.95	18.81	15.93	45.25	-15.31
Asn	61.13	67.03	70.93	9.65	5.82
Asp	79.42	91.53	88.73	15.25	-3.06
Cys	24.14	36.83	33.48	52.57	-9.10
Gln	22.77	28.88	25.89	26.83	-10.35
Glu	49.14	57.70	58.61	17.42	1.58
Gly	19.10	34.15	29.80	78.80	-12.74
His	93.74	106.39	103.49	13.49	-2.73
Ile	21.14	33.56	30.07	58.75	-10.40
Leu	26.12	40.76	36.03	56.05	-11.60
Lys	57.14	65.24	58.41	14.18	-10.47
Met	45.40	48.11	44.21	5.97	-8.11
Phe	68.35	81.51	74.23	19.25	-8.93
Pro	78.71	90.90	86.56	15.49	-4.77
Ser	35.87	43.18	40.76	20.38	-5.60
Thr	38.01	48.17	48.51	26.73	0.71
Trp	18.86	32.97	34.45	74.81	4.49
Tyr	34.32	47.32	35.26	37.88	-25.49
Val	37.20	49.65	41.65	33.47	-16.11
average				33.55	-7.28

Table 2.4 Average amino acid accessibility values

Solvent accessibility of amino acid residues calculated over all non-linker, non-linker coil and linker positions, together with the percentage difference in solvent accessibility between all-residues and linker residues, and non-linker coil and linker residues.

	Variability scores			
	linker residues		coil residues	
% RSA interval	mean	s.d.	mean	s.d.
0<RSA<5	19.21	14.67	18.97	15.55
5<RSA<10	23.45	16.30	22.69	17.06
10<RSA<15	25.66	16.43	23.99	16.28
15<RSA<20	27.67	16.33	25.89	16.01
20<RSA<100	32.61	14.14	32.49	14.78

Table 2.5 Sequence variability in domain linker and non-linker coil residues

Mean variability scores for linker residues and non-linker coil residues, for normalised intervals of relative solvent accessibility.

into any of the small differences between the respective values, and this again demonstrates the difficulty in such analysis measurements.

Linker residues with a variability scores that were one standard deviation below the mean non-linker coil value for the rest of the protein (for a given accessible surface area bin) were considered to be more conserved than might be expected on average. The frequencies of such linker amino acids were calculated, to examine any relationship between linker residue propensity values and their ranked position corresponding to their overall mean variability score. A fairly low correlation coefficient was found ($r = 0.54$). The same calculation was also made for completely conserved residues i.e. those with a variability score of 0. A similar correlation was calculated ($r = 0.57$), although in this case the sample data is far more scarce. In both calculations glycine was found to be the most frequently conserved residue, followed by proline and asparagine.

2.3.6 Hydrogen bonding in domain linkers

As outlined in section 2.2.8, the average number of internal hydrogen bonds per linker residue was calculated. The calculation for hydrogen-bonds assigned by HBPLUS gave a value of 0.39 internal hydrogen-bonds per linker residue.

2.3.7 Assignment of all-alpha and all-beta linkers

As already discussed in section 2.3.1, 52 linkers were found to be all-helical or all-strand in structure. Of these linkers, 29 were found to be all-helical and 23 all-strand forming beta sheets with surrounding domain strands or other linkers in the structure.

The all-helical linkers were assigned only if the N- and C-termini of the helix appeared to start and end within the core regions of the linked domains. Their lengths ranged from 9 to 34 residues, and were found in both two and three-domain proteins. The amino acid propensities of the helical linkers were compared to those calculated for all residues participating in helical structures in the data set. A correlation co-efficient of $r=0.88$ was calculated between the two distributions, showing a strong similarity between the residue sets. A well studied example found

within the data set was the calcium binding protein Calmodulin, shown in Figure 2.6a. Here each globular domain is joined by a long alpha-helix.

Throughout the linker delineation process, great care was taken not to split beta sheets. In fact, one of the rules of domain delineation described by Taylor (1999) is that sheet elements should not be split between domains. However, several cases were found where clear domains existed, but were linked by strand elements. These linkers ranged in size from 9 to 23 residues. As already discussed in section 2.3.1, several of these cases were found between linkers joining discontinuous domain segments. However, cases for single linkers were found. Such an example is shown in Figure 2.6b which shows the structures of domains one and two of the T-cell surface glycoprotein, CD4. Here the two immunoglobulin folds are joined by a strand that runs continuously from the N- to the C-terminal domain.

2.4 Discussion

The substantial increase in the number of structures in the protein databank in the last ten years has enabled an update of the domain linker analysis made by Argos (1990) to be carried out. In this study, 391 protein structures were analysed, compared to the 32 by Argos (1990), resulting in a set of 747 domain linking peptides, compared to 51 in the Argos study (1990).

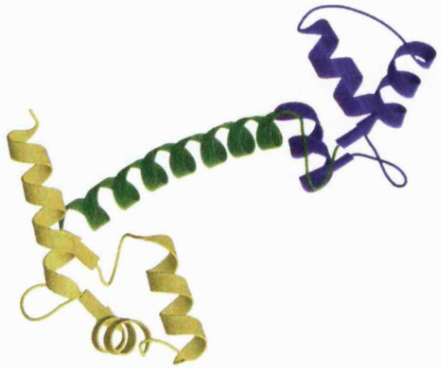
In this study the overall length distribution of domain linkers shows a wide spread, between 2 and 33 residues, with a mean length of 9.8 residues, longer than the average length of 6.5 residues calculated by Argos (1990). In both this study and the Argos study (1990), secondary structure was assigned by the DSSP algorithm. The percentage of linker residues that were assigned as forming coil, bend or turn regions in this study was 85%, compared to 75% in the Argos study. Just over 8% of the linker residues adopted a helical conformation (assigned as H, G or I) compared to 13% by Argos and just over 7% a strand conformation (assigned as E or B) compared to 12% in the Argos study (1990). Both studies find that the majority of linker residues adopt a coil conformation, though more linker residues are assigned as coil in this study (especially true in this study when the simplified secondary structure assignment scheme is used). In fact, a predominant number of the linkers consisted of *all* coil residues (nearly 84%) or mainly coil residues (9%).

Figure 2.6 All-helical and all-strand linking peptides

a) A well studied example found within the data set was the calcium binding protein Calmodulin (PDB code 1cll, Chattopadhyaya *et al.*, 1992). Here each globular domain is joined by a long alpha-helix, (coloured green) which spans from the interior of the N-terminal domain (coloured yellow) to the interior of the C-terminal domain (coloured purple).

b) The structure of the T-cell surface glycoprotein, (PDB code 1cyd, Wu *et al.*, 1996), domains one (coloured yellow) and two (coloured purple). Here the two domains are joined by a strand that runs continuously from the N- to the C-terminal domain (coloured green).

a)



b)

As described above, the Argos study found that more linker residues adopt a helical or strand conformation, though the values calculated in this study do not include the all-helical or all-strand linkers, whereas 2 of the 51 linkers in the Argos set were all-helical (1990). These differences are most likely due to the enlarged data set and the different assignment processes used. Observations of the linker structures in this study showed many of the linker residues adopting a strand conformation can be attributed to beta-sheet regions that were observed to form between a number of domain linking peptides, especially those associated with discontinuous domain segments.

The amino acid propensities calculated for the linker peptides showed the most preferred residue to be proline followed by glycine, aparagine, threonine and aspartate. In general, the favoured amino acids tended to be hydrophilic (apart from proline) or charged residues, whilst residues with bulky side-chains including the aromatic polar residues were avoided. This bias towards proline and glycine residues will provide a high degree of flexibility to linker peptides; both proline and glycine are often found together in unstructured regions of protein – proline is known to be a helix breaker (perhaps a reason why so few linkers were found to contain helical segments) due to its main-chain nitrogen being unavailable for hydrogen bonding and glycine, with no side chain is small and suffers very few steric constraints. Therefore proline and glycine confer turn and bend propensities to domain linkers. The lack of the amide hydrogen will also mean that these residues may isolate the linkers from the domains, as hydrogen-bonds cannot be formed with adjacent structure. Proline rich peptides will have many non-interacting connections, making proline rich linkers independent of surrounding domains. Though proline is hydrophobic, its special properties make it a valuable constituent of linker peptides.

The linker propensities calculated in this study were compared to those given in the Argos study (1990). A Pearson correlation of r=0.60 was calculated, showing a significant correlation at the 5% level. However, plotting the propensity values of the two distributions (Figure 2.3a) shows the values to be fairly different, especially for residues such as proline, histidine, leucine and cysteine. Such differences could be attributable to small sample size in the Argos study (1990) – though there is no way of saying that the values in either study are the definitive answer.

As linkers were found to be mainly coil regions of peptide it was of interest to determine whether there was a difference in the composition of amino acids in linkers as compared to non-linker coil regions. A correlation coefficient was calculated between the propensities of linker residues and non-linker residues giving a value of r=0.78 (Pearson). These figures show that there is a relatively high correlation between the two amino acid distributions, although differences also exist. These differences can be seen more clearly in Figure 2.3b; whilst there is a similarity in residue propensity for many residues in linkers and non-linker coil structures, glycine and proline are the largest outliers. Interestingly, although glycine is the second most preferred linker amino acid (propensity of 1.2) it is found at a lower frequency in linkers than in general coil (propensity of 1.65). These values emphasise the proline-rich nature of domain linkers.

The conservation of linker residues was found to be similar to other exposed coil regions found in the protein, and although no real correlation was found between residue conservation and linker propensities, glycine and proline were found to be the most frequently conserved residues, showing a selective pressure towards retention of these favoured residues. True comparisons between mean RSA and residue conservation measurements were complicated by the large standard deviations that accompanied these values. This reveals the variation of residue accessibility and conservation even for well defined regions of structure such as coil, helix and strand elements. This applies even when variation due to other factors, such as solvent exposure is taken into account when measuring conservation.

The analysis of temperature factors revealed linker residues to be similar in flexibility to non-linker coil residues, though it is possible that some linker residues have more conformational freedom than might be expected non-linker coil. Argos (1990) also concluded that linkers are average in flexibility compared to other protein regions, although no distinction was made between non-linker coil and general protein regions. Calculations showing the mean C-alpha extension (normalised by dividing the distance between the N- and C-terminal linker residues by the number of residues in the linker minus one) showed a lower value of 2.31 Å compared to 2.84 Å for non-linker coil peptides. This lower extension value could be attributed to domain linker regions having a more bent conformation perhaps conferred by the high proline content.

The conformational freedom of linker regions is also indicated by accessible surface area measurements, which show linker residues to be amongst the most exposed regions of the protein. Interestingly, comparison of linkers connecting only continuous domain regions, to linkers connecting only discontinuous domain segments, showed that continuous linkers tend to have more exposed residues (defined by a RSA greater than 10%) than discontinuous linkers. This difference may be due in part to the observation made above, whereby linkers joining discontinuous domains regions are often paired-up with adjacent linker peptides, forming sheet structures, and therefore becoming more buried within the protein structure.

Structural flexibility is an essential attribute to many multi-domain proteins enabling catalysis, regulation of activity and metabolite transport (Gerstein *et al.*, 1998). Many NMR studies have suggested that there are wide variations in the flexibility of domain-domain pairs (Spitzfaden *et al.*, 1997). The findings that most domain linkers are unstructured, solvent exposed and disfavour hydrophobic residues supports these observations.

Analysis of hydrogen-bonding, as assigned by HBPLUS gave an average of 0.39 internal linker hydrogen-bonds per linker residue similar to the value of 0.4 internal linker hydrogen-bonds per linker residue given in the study by Argos (1990). It is worth noting that the Argos study found a high frequency of linker serine and thronine residues forming hydrogen bonds between their main chain amino group and gamma oxygen. Such a pattern of hydrogen-bonding was not found in this study as the HBPLUS algorithm does not assign such strained hydrogen-bonds in these residues. However, such interactions would satisfy the hydrogen-bonding of these residues and confer some structural independence to linkers containing these residues.

The high propensity for proline in domain linkers may also have implications for protein folding. The amino and carbonyl groups of the polypeptide usually point in opposite directions, known as the *trans* form, which represents the most stable form of the peptide group. Another form is also possible, known as the *cis* form where the amino and carbonyl groups point in the same direction. The *trans* form of the peptide bond is approximately 1000 times more stable than the *cis* form except in cases where the second residue is a proline (Branden and Tooze, 1999). For most

residues the steric hindrance between the functional groups attached to C-alpha atoms will be greater in the *cis* configuration, however the cyclic nature of the proline side-chain means the *cis* and *trans* state have more equivalent energies. Whilst most prolines are found in the *trans* form, *cis*-prolines are found in tight bends in the peptide chains which results in conformational flexibility. In the native unfolded state, *cis* and *trans*-proline exist in equal proportions. As the protein folds, a substantial number of the proline peptide bonds will be in the incorrect configuration, and the higher the number of proline residues, the greater the number that must swap to the correct configuration. The *cis-trans* isomerisation of proline peptides is intrinsically slow, and has often been found to be the rate limiting step of folding *in vitro* (Branden and Tooze, 1999). It can therefore be envisaged that a high proline content in domain linkers may be an important property for efficient folding in multi-domain proteins. A proline rich linker might be the final part of the protein to fold, enabling the linked domains to fold first and then assemble inter-domain interactions as the linker peptide adopts its folded conformation. *In vitro* folding studies by Frydman *et al.*, (1999) suggested that the slowest step in the folding of a multi-domain protein is the pairing of adjacent domains. This may be a result of incorrectly folded domains, or that the small adjustments that are required to attain the optimal interaction between the domain are energetically unfavourable.

As has been described in Chapter 1, the folding of multi-domain proteins can occur co-translationally (most often in eukaryotes) or post-translationally (most often in prokaryotes). In post-translational folding separate domains fold concurrently, whilst co-translationally the domain folding is sequential. This may explain why many multi-domain eukaryotic proteins missfold when expressed in bacterial systems (Ellis and Hartl, 1999). It would certainly be of interest to see if domain linking peptides in prokaryotic proteins are more likely to have a high proline content than eukaryotic multi-domain proteins. Proline-rich linkers may be more beneficial to bacterial post-translational protein folding, slowing the rate of folding and preventing non-native interactions to form between domains in the early stages of folding.

Whilst assigning the domain linkers in this study, it became clear that not all linkers were equal and, in fact, they can be grouped into separate classes according to their properties. The most obvious classes involve those linkers that adopt an all-

helical or all-strand conformation, and those that do not. Of the 747 linkers in the data set 52 (or 7%) were found to be all-helical or all-strand elements that spanned between adjacent domains. The low frequency of these linking peptides suggests they have a specialised structural role that is not suitable for the majority of multi-domain proteins. For example, the sharing of an all-strand linker between domains may play an important role in forming a rigid association between the adjacent structures. In the case of CD4 (Figure 2.6b) is thought give a near 180 degree twist angle between the domains (Wang *et al.*, 1990). Domain fusion experiments generally attempt to fuse contiguous units of protein structure together, however it is also of interest to note that the linkers joining inserted domains (i.e. discontinuous domains) often associate to form beta-sheets. Such interactions, whilst satisfying the hydrogen-bonding of the linkers, may also add stability to these inserted regions. In a some examples, the linker between two domains was found to consist entirely of an alpha-helix. Comparison of the residue frequencies in these helical-linkers compared to the frequencies in all helical elements showed a high correlation between the residue distributions (r=0.88). All-helical linkers can facilitate both small and large movements between domains. Small deformations, spread out over the helix can produce bending or stretching motions (Gerstein *et al.*, 1994). Much larger movements can occur when the helix contains kinks which often involve proline residues. For example, calmodulin can assume two different conformations depending on a calcium-induced conformational change. In the substrate bound form, the helical linker breaks into two separate shorter helices which move almost perpendicular to one another (Ikura *et al.*, 1992). These properties of the linker the enable it to act like a molecular switch that regulates the calcium induced conformational change and therefore activity of calmodulin (Sorensen *et al.*, 2002). Gene fusion studies generally use a flexible linker to combine domains (Nixon *et al.*, 1997). However work by Arai *et al.*, (2001), found that by engineering a bifunctional fusion protein with a helical linker retained the activity of the domains, whereas the use of a flexible linker did not. The use of the all helical linker was thought to be important in allowing the correct spatial distance between the hetero-functional domains to allow them to work independently.

Linkers that are mainly unstructured can also be sub-divided by length, as described in section 2.3.2. Such observations may be important for linker design.

Nixon *et al.*, (1998) suggest that the correct domain linker must be chosen when engineering chimera proteins, to optimise the inter-domain contacts and also the orientation of the two domains for catalysis. Such improvement may come about by varying the linker length or residue content to improve the stability and catalytic efficiency of such constructs.

Over 25% of the mainly coil linkers were between 2 to 4 residues in length, forming a small-linker class. The assignment of domain boundaries is not an insignificant problem even when using 3D structural data (Hadley and Jones, 1999). In turn the assignment of domain linking regions in this study was not a trivial issue. In fact just over 12% of the CATH domain cuts were found in helical or strand elements, when they could have been assigned to adjacent coil regions, subsequently assigned as the linking peptide. Using visual inspection, CATH boundary assignments were identified and the corresponding linker termini regions recorded. It is clear from the manual visualisation of the linking regions that the identification of domain linkers can be as subjective a matter as domain assignment. A protocol for domain linker assignment (section 2.2.2) was therefore used to keep the linker assignments consistent. For a number of linkers, it was difficult to easily pinpoint which part of the protein chain exactly constituted the linker region, even though the cut point between the adjacent domains was given in CATH. Such linkers make up the majority of the class of small-linker peptides.

Analysis of the amino acid propensities in this small-linker class showed a high propensity for glycine whilst a very low propensity for tryptophan. A propensity for residues with small (or no) side chains such as glycine in these small linkers, and the low propensity for residues with bulky side chains such as tryptophan may be a result of these linkers being found between closely packed domains, where there must be close packing of atoms within a limited space. Whether such short peptide regions can truly be described as domain linkers is perhaps open to debate. They clearly have a role in joining domains to one-another, however, in some cases they may appear as almost a continuation of structure from one domain to the next. The recurrence of the linked domains would demonstrate that these structural units may have existed independently, and as such, the linker region would have had an important role in the fusion of domains into larger proteins.

A large proportion of the linkers (65%) were classed as an intermediate length (between 5 and 12 residues), whilst the remainder, the large-linker class, had

lengths between 13 and 33 residues. Whilst many of these linkers adopted a primarily coil structure, several additionally adopted short secondary structure conformations. It is possible that when designing linkers the use of peptides that form a mixture of helical and coil structure may give a flexibility between domains that is intermediate to all-coil and all-helical linkers. However, the high frequency of coil linkers may mean that the fusion of domains into multi-domain proteins may have been most successful when linker sequences formed mainly coil regions. The connecting of adjacent domains by one or more linkers that do not possess a rigid secondary structure will permit conformational freedom, and allow the relative orientation of the domains to vary. Such unstructured conformations may also cause the least interference with the folding of the adjacent domain regions. Helical linkers may have a similar role in keeping domains apart as they fold. Unlike proline-rich linkers however, this may be achieved by these areas folding first, rather than last. Aurora *et al.*, (1997) suggested that helical conformations may be amongst the first regions of the protein to fold. If the helical linker folded rapidly, it may act as a rod, holding neighbouring domains away from each other, stopping the misfolding and aggregation of the folding units.

The additional linker analysis carried out in this study was appropriate since the last was carried out over ten years ago (Argos, 1990). Comparison of the results has shown a number of points of disagreement between the analyses, including residue preferences and secondary structure content, perhaps necessarily so, as a far larger data set was used in this study. Overall both studies found linker peptides to be flexible, exposed, generally unstructured coil-like regions of structure. The high propensity for proline residues may relate to linker flexibility, structural independence and folding.

Though the linker residue propensity values showed an imperfect correlation to non-linker coil residues, it is unclear whether the propensities are different enough to be able to predict linker regions from sequence, for example by using a neural network. The search for proline rich coil regions is a possibility. A valuable extension to this study may be an analysis of the pair-wise ordering of amino acids within domain linkers, providing more information on linker properties and design.

Chapter 3

A survey of structural characteristics of protein domains

3.1 Introduction

It is widely accepted that a fundamental principal underlying protein folding is the burial of hydrophobic side chains, away from surrounding solvent, and into the protein core. This concept was first proposed as being important factor in protein structure by Langmuir (1938). It was then discussed to a greater extent by Kauzmann (1959), who described the thermodynamic properties of the hydrophobic effect. The understanding of the hydrophobic effect is an important step in unravelling the complexities of protein folding, and as such, much work has been carried out on this principal.

This formation of a hydrophobic core, surrounded by hydrophilic surface residues, lead Waugh (1954) to describe the creation of a non-polar inner volume and polar outer volume. Studies by Fisher (1965) postulated a 'limiting law of protein structure', where the sequence length and overall conformation of a protein structure may be constrained by the ratio of polar and non-polar residues. Fisher describes how there must be a limited number of non-polar residues required such that they could be surrounded by the remaining polar residues. For example, too few non-polar residues, and the protein would be unable to form a spherical unit of structure, whereas too many would cause protein aggregation, as a result of the necessity to shield the non-polar residues from solvent. Work by Chothia (1975) proposed that the surface area of a protein is a function of its molecular weight. It has also been suggested that proteins with fewer than 70 amino acids are unlikely to fold to a stable end product, if their folding is solely determined by the hydrophobic forces (Dill, 1985) and that possible stabilising mechanisms are required for these proteins such as disulphide-bond formation (Miller *et al.,* 1987; Creighton, 1988).

A number of studies are carried out in order to see if such characteristics might be used to predict domains from sequence. Are there differences in characteristics, between single and multi-domain proteins, that can be exploited to distinguish single domain sequences from multi-domain sequences, and in turn to allow possible prediction of domain-linker regions? In this chapter the number of exposed residues present on the protein surface is determined, for single and multi-domain proteins. The amino acid propensities of residues found in the protein core, surface and at interfaces between domains is also examined. Further, an analysis of the percentage of hydrophobic residues within single and multi-domain proteins is made.

3.2 Methods

3.2.1 Data set

The non-redundant set of protein chains described in section 2.2.1 of this thesis were used as the representative set of protein structures in this study. As previously described in section 2.2.1, domain definitions were taken from the CATH database. The set consisted of 1177 chains; 786 single domain chains and 391 multi-domain chains.

3.2.2 Relative solvent accessibility

Each residue was assigned a relative solvent accessibility (RSA) value, exactly as described in section 2.2.6. Cut-offs were used to define different states of relative solvent accessibility where residues with an RSA $\leq$ 10% were considered as buried and those with an RSA > 10% were considered as exposed. Residues in domain interfaces were defined as those that lost >1 $Å^2$ accessible surface area on complexation with the adjacent domain in the multi-domain structure as described by Jones *et al.*, (2000). The total surface area of each protein structure in the data set was calculated by the DSSP algorithm.

3.2.3 Amino acid propensity

Amino acid propensities were calculated for buried (or core) residues, exposed (surface) residues and residues in domain interfaces. Propensities were calculated exactly as described in section 2.2.3.

3.2.4 Assignment of polar and non-polar amino acids

Residues were assigned to three classes – hydrophobic, charged and polar as follows:

Hydrophobic: Gly, Ala, Val, Leu, Ile, Pro, Phe, Met.

Charged: Asp, Glu, Thr, Lys.

Polar: Ser, Thr, Tyr, His, Cys, Asn, Gln, Trp.

3.3 Results

3.3.1 Sequence length and percentage of exposed residues

If a globular protein domain is considered as an idealised sphere containing closely packed residues, one may calculate the surface area for a single domain and two-domain chain as:

The volume of a sphere $= 4/3\ \pi\ r^3$ (where r is the radius)

and surface area $= 4\ \pi\ r^2$

So, for example, consider a spherical single-domain protein of 300 residues. Given that the mean volume of an amino acid is approximately 109Å^3 (Creighton, 1992), such a protein will have a volume of 32700Å^3, with radius 19.8Å. The surface area can therefore be calculated to be 4945.2Å^2. The same volume, equally divided into two 150 residue spherical domains, might represent a two domain protein where each domain has a volume 16350Å^3 and a radius 15.7Å. The two domain protein will therefore have an total surface area of 6230.5Å^2. It might therefore be supposed, given that a multi-domain chain sequence has a larger surface area than an equivalent sized single domain protein, it will contain a greater number of exposed residues.

The percentage of exposed residues calculated for each protein in the data set (section 3.2.2) is shown in Figures 3.1a and b. The above calculation shows that in principal (for an idealised example) an increase in domain number should give an increase in the percentage of exposed residues for proteins of similar length. There is considerable overlap between the distribution of single and multi-domain chains with many single and multi-domain chains of similar length having a similar percentage of exposed residues. However, although the separation is not clear, it is possible to see a trend in the distribution where multi-domain proteins above 200 residues have a higher percentage of exposed residues. In general, it can be seen that shorter chains have a higher percentage of exposed residues than longer chains.

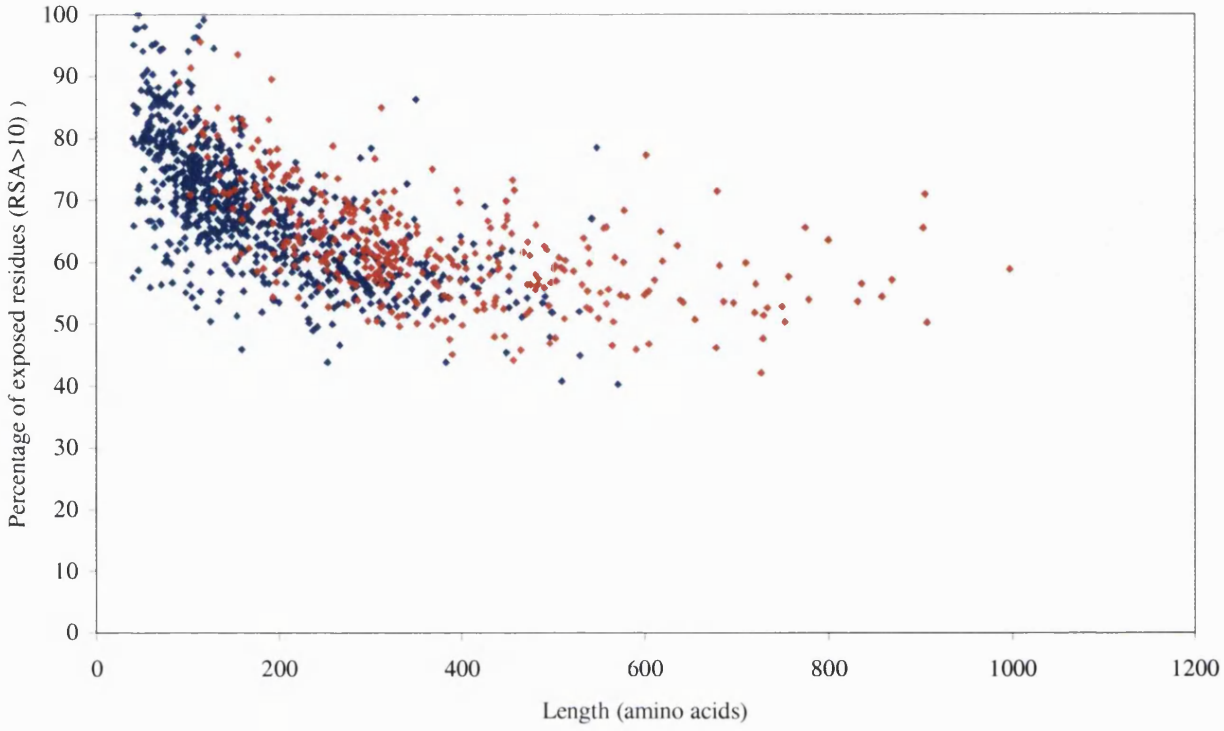
The overall surface area of the single and multi-domain structures was calculated (section 3.2.2). Figure 3.2 shows the relationship between sequence length and overall surface area (measured in Å^2). It is possible that the surface area of multi-domain chains of lengths < 200 residues appears to be higher than the single domain chains of similar length, though the small sample size of these multi-domain chains (compared to the number of single domain chains) does not make such observations

Figure 3.1 Percentage of solvent exposed residues for single and multi-domain proteins

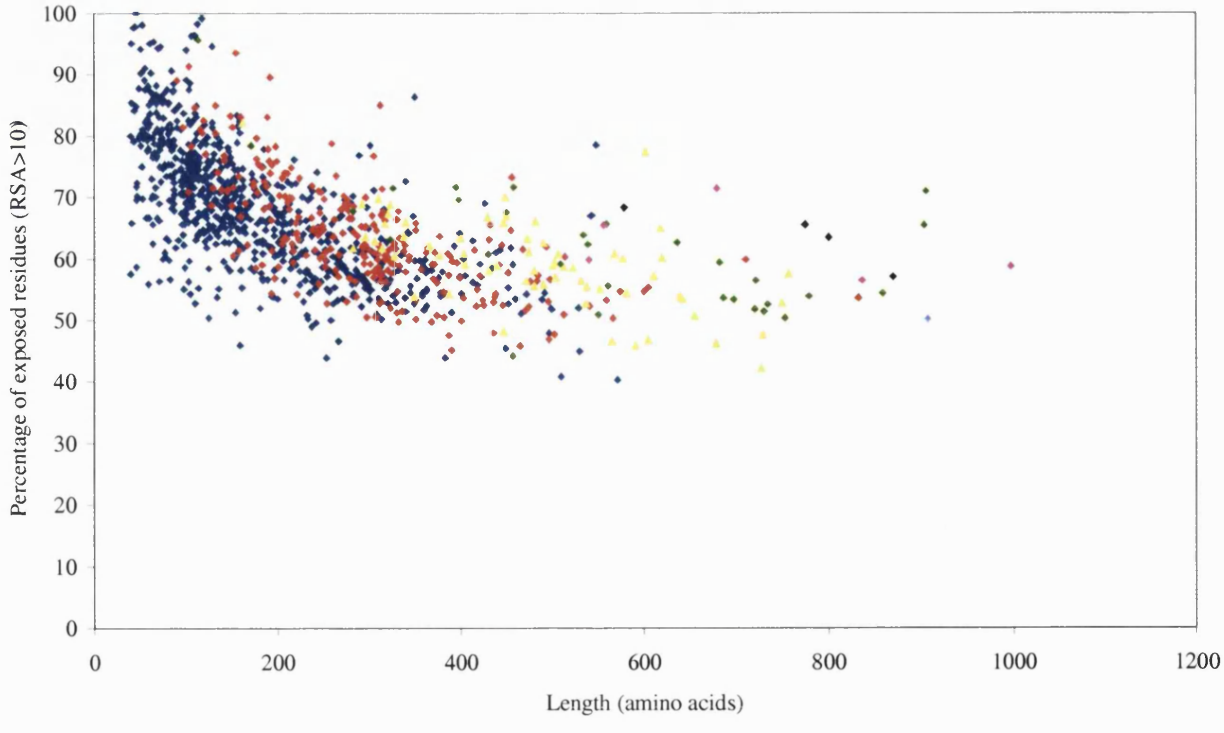
a) Distribution for single domain (blue) and multi-domain (red) chains.

b) The same distribution, showing number of domains for multi-domain chains. Single domain (blue), two-domain (red), three-domain (yellow), four-domain (green), five-domain (pink), six-domain (black), seven-domain (light blue).

a)


Percentage of exposed residues (RSA>10))
100
90
80
70
60
50
40
30
20
10
0
0
200
400
600
800
1000
1200
Length (amino acids)

b)


Percentage of exposed residues (RSA>10)
100
90
80
70
60
50
40
30
20
10
0
0
200
400
600
800
1000
1200
Length (amino acids)

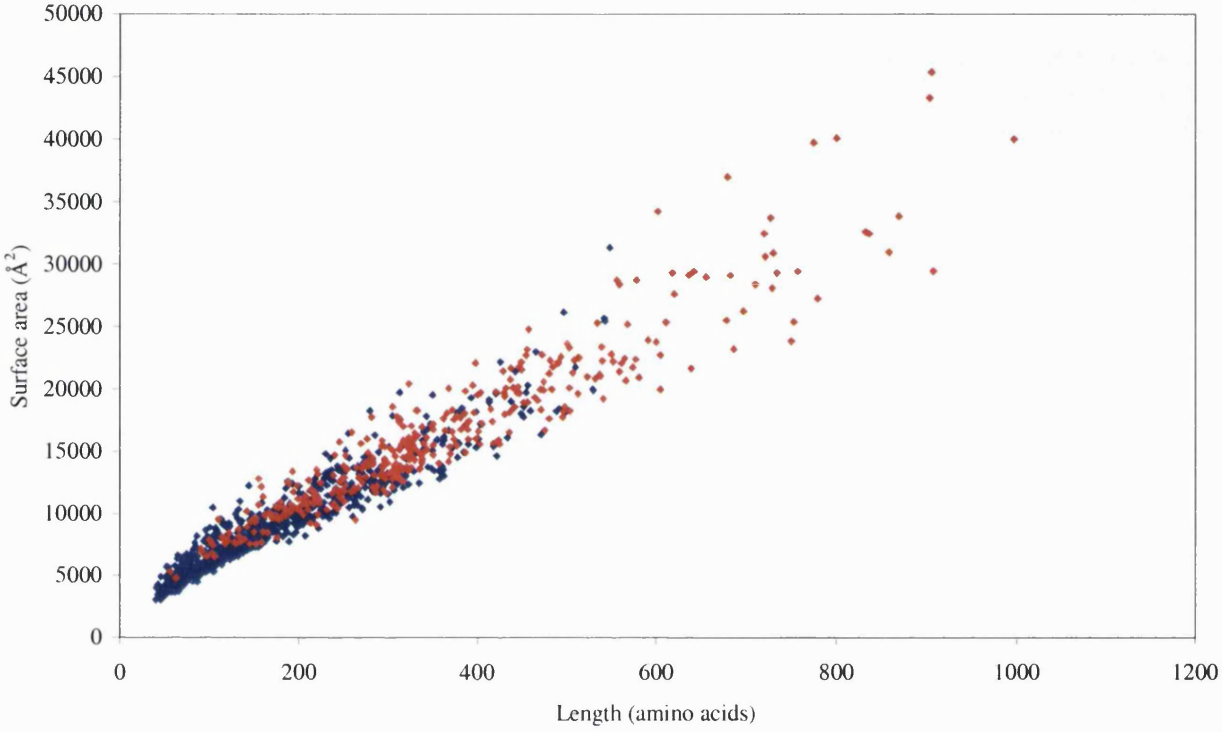


Figure 3.2 Surface area of single and multi-domain chains

The total surface area (measured in $Å^2$) for single and multi-domain chains was calculated by the DSSP algorithm. Single domain chains are shown as blue diamonds and multi-domain chains as red diamonds.

overly conclusive. There is a clear link between protein length and surface area, with the distribution becoming more scattered at higher chain lengths. This result indicates that it may be possible to predict the surface area of a protein given that you know its sequence length.

Figures 3.3a and b show the percentage of buried hydrophobic residues (sections 3.2.3 and 3.2.4) calculated for each protein in the data set. Such residues may represent the core of a given protein. Again, though the values show significant overlap the distribution of the multi-domain chains seems to be slightly separated from the single chains, where the multi-domain chains tend to have fewer buried hydrophobic residues.

3.3.2 Sequence length and percentage of hydrophobic residues

The percentage of hydrophobic residues (section 3.2.4) in each protein in the representative set was calculated, and plotted against their corresponding sequence length, Figures 3.4a and b. Again it can be seen that there is a large overlap between the two distributions, the mean percentage of hydrophobic residues found for single domain proteins being 47.1% and the mean for multi-domain proteins being 48.3%. Sequences shorter than ~ 150 residues, representing mainly single-domain proteins, tend to have a wider distribution. This might be partly due to small number statistics, where the values for shorter chains are more open to variation from just a small change in the number of hydrophobic residues. Nevertheless the overall mean percentage of hydrophobic residues seems to be consistent over the distributions at approximately 48%.

3.3.3 Amino acid propensities: surface, core and domain interface

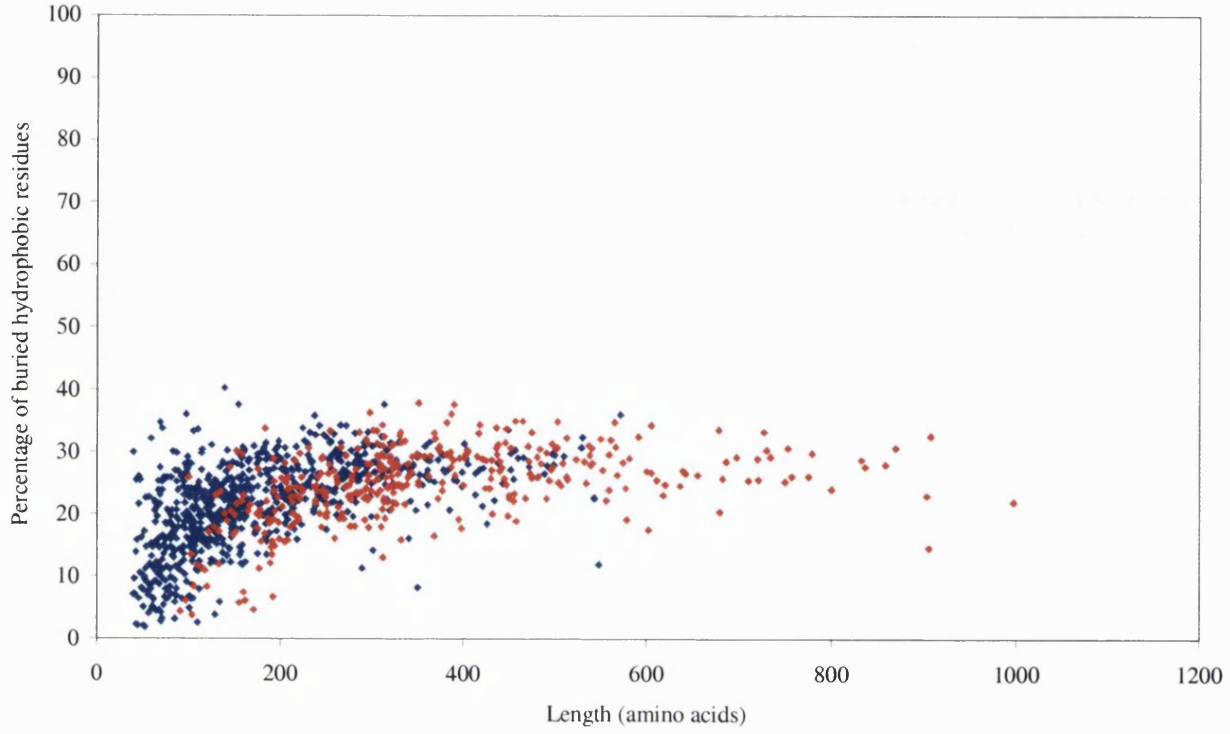
The amino acid propensities for residues in the protein core, surface and domain interfaces were calculated as described in sections 3.2.2 and 3.2.3. Figure 3.5 shows the residue propensity distributions. As might be expected, hydrophobic residues are preferred in the core of proteins, whilst most polar and all the charged residues are disfavoured. The interface shows the strongest preferences for lysine, arginine and tyrosine (propensity values of 1.28, 1.31 and 1.29 respectively). The interface residue propensities seem most similar to the surface residue propensities,

Figure 3.3 Percentage of hydrophobic buried residues in single and multi-domain proteins

a) Distribution for single domain, (blue diamonds) and multi-domain, (red diamonds) chains.

b) The same distribution, showing individual domain numbers. Single domain (blue), two-domain (red), three-domain (yellow), four-domain (green), five-domain (pink), six-domain (black), seven-domain (light blue).

a)


Percentage of buried hydrophobic residues
100
90
80
70
60
50
40
30
20
10
0
0
200
400
600
800
1000
1200
Length (amino acids)

b)

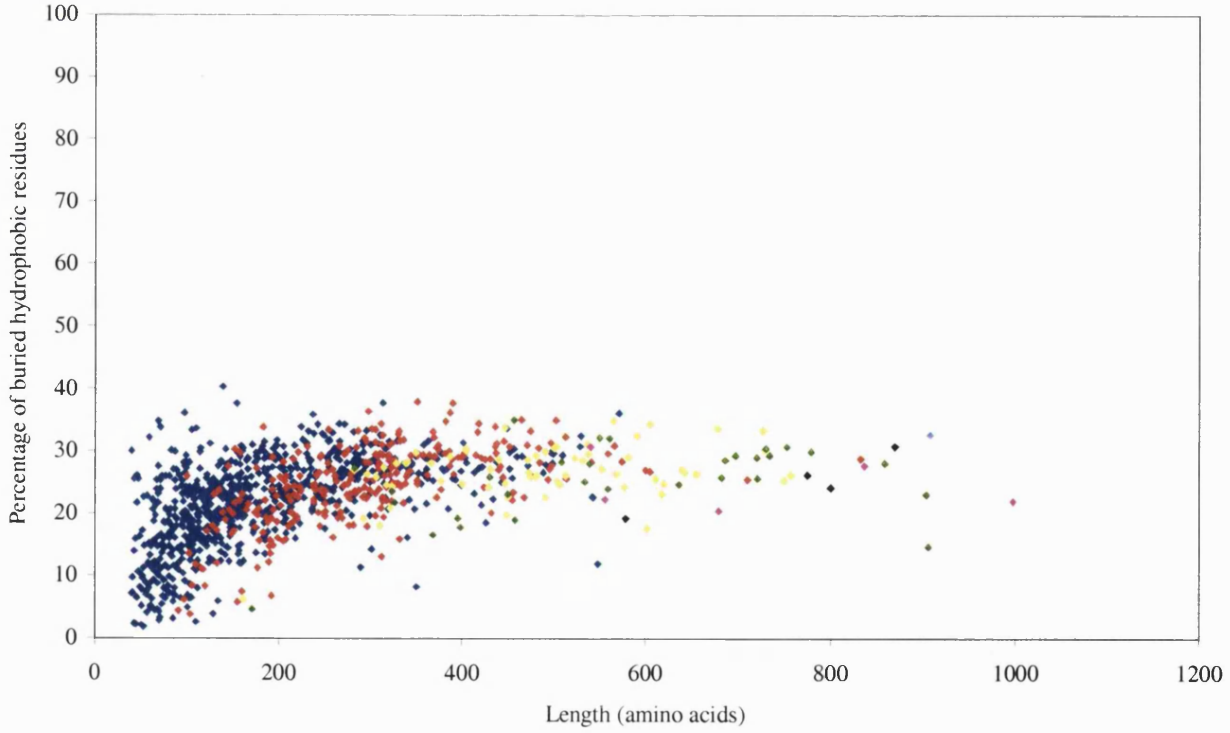
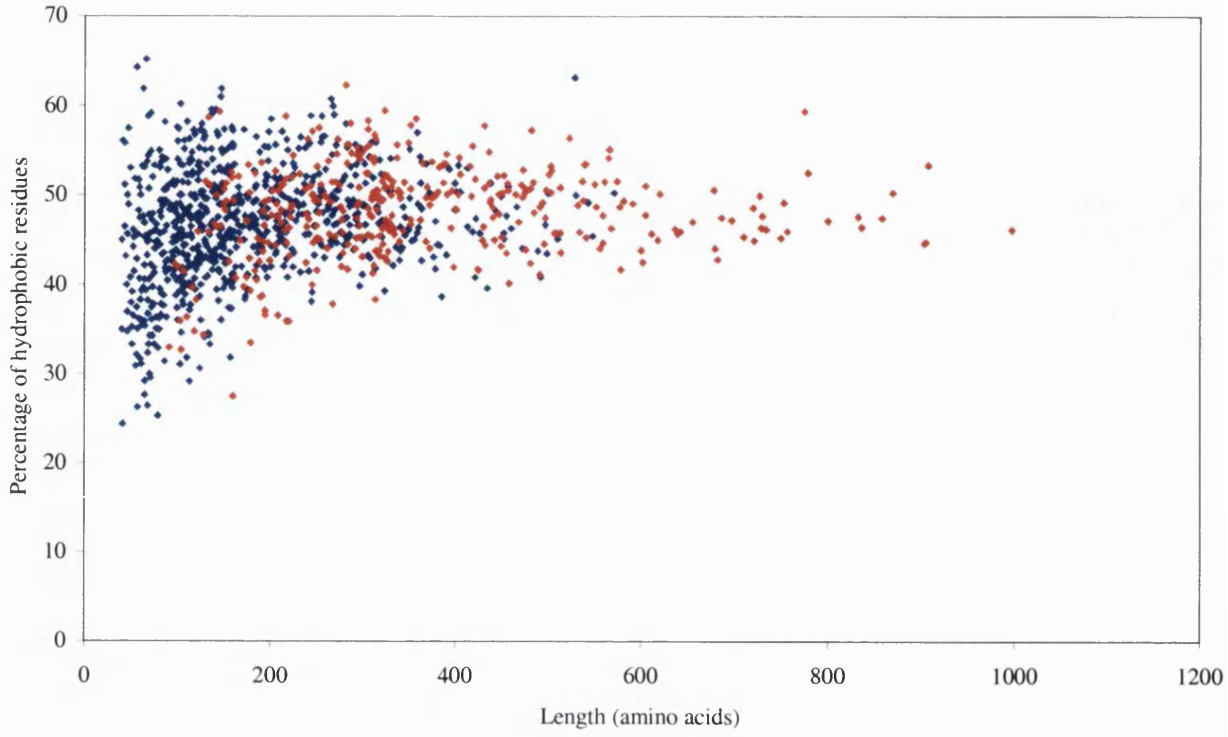

Percentage of buried hydrophobic residues
100
90
80
70
60
50
40
30
20
10
0
0
200
400
600
800
1000
1200
Length (amino acids)

Figure 3.4 Percentage of hydrophobic residues in single and multi-domain proteins

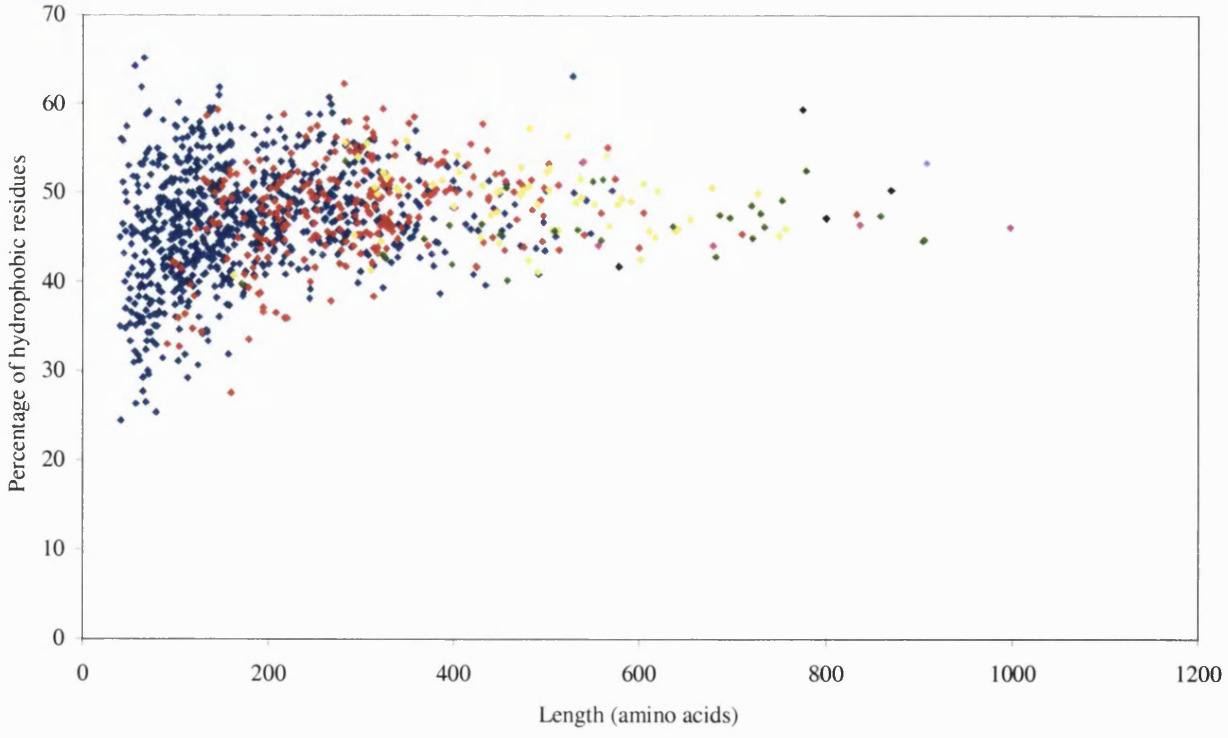
a) For single domain, (blue diamonds) and multi-domain, (red diamonds) chains.

b) The same distribution, showing individual domain numbers. Single domain (blue), two-domain (red), three-domain (yellow), four-domain (green), five-domain (pink), six-domain (black), seven-domain (light blue).

a)


Percentage of hydrophobic residues
0
10
20
30
40
50
60
70
0
200
400
600
800
1000
1200
Length (amino acids)

b)


Percentage of hydrophobic residues
0
10
20
30
40
50
60
70
0
200
400
600
800
1000
1200
Length (amino acids)

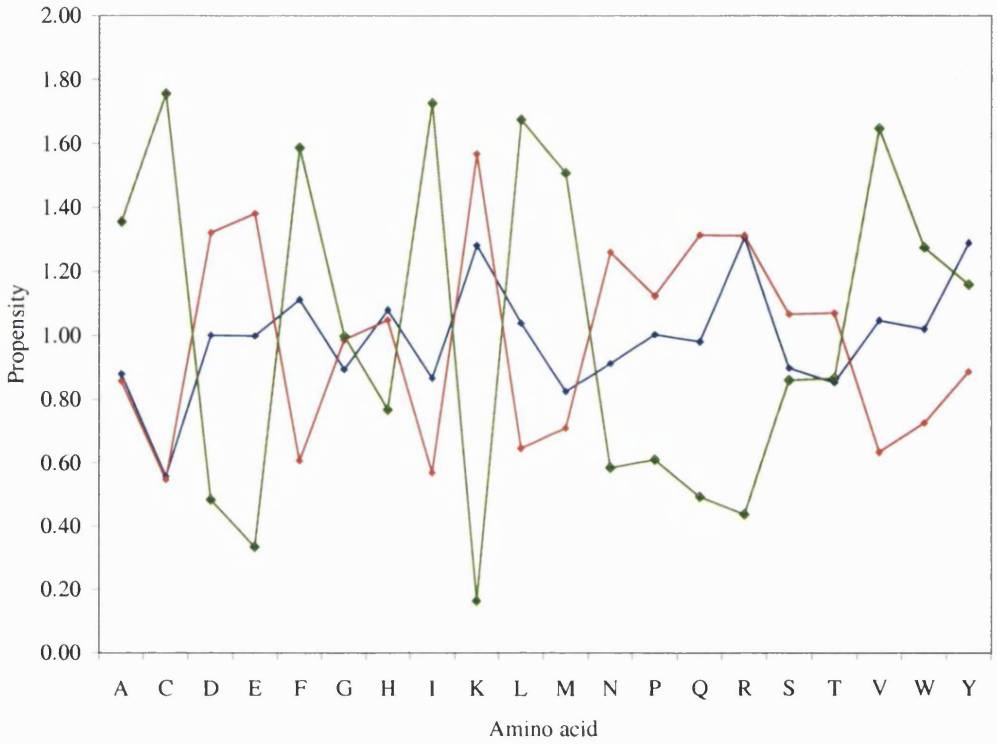


Figure 3.5 Propensity values of residues in protein surfaces, cores, and domain interfaces

Residue propensities for surface residues (red), core residues (green), domain interface residues (blue).

although aspartate and glutamate have no preference for domain interfaces (both having propensities of 1.0), whilst favouring the protein surface (propensities of 1.32 and 1.38). Also phenylalanine and valine are disfavoured on the protein surface (propensities of 0.61 and 0.63) whilst slightly favoured in domain interfaces (propensities of 1.1 and 1.05 respectively). Polar and charged residues are generally favoured both on the protein surface and within the domain interface. The propensity values for glycine are interesting as they are close to 1 in all of the three domain locations analysed. This indicates that there is no real preference (or aversion) for glycine on the domain surface, in the core or at domain interfaces, possibly due to its lack of side chain.

Overall comparison of the distributions show that domain interface residues most closely correspond to surface residues than core residues, with confidence values of 0.94 between the interface and surface distributions and 5.28×10^{-06} between the interface and core residue distributions (χ^2 test).

3.4 Discussion

The analysis of a number of domain characteristics in this chapter was carried out in order to assess how useful they could be for domain prediction. Calculating the percentage of exposed residues for single and multi-domain chains showed that as sequence length increases, so does the percentage of exposed residues. Though there is not a clear separation between the single and multi-domain distributions, in general multi-domain proteins tend to have a larger percentage of exposed residues. It is possible that a method could be developed to distinguish between single and multi-domain proteins, based on these results, using a combination of chain length and percentage of solvent exposed residues. However such a method would rely on the prediction of residue solvent accessibility which would presumably result in a greater overlap between the distributions of percentage solvent accessible residues for the single and multi domain proteins, due to prediction errors.

The distribution shown in Figure 3.2 shows a strong relationship between chain length and percentage of solvent exposed residues as observed by Chothia (1975) and Miller *et al.*, (1987) for datasets containing 15 and 46 protein structures respectively. Rost (1999b) described a method that predicts the globularity of protein sequences, based on the prediction of accessible surface area and sequence length.

The study aimed to distinguish sequences corresponding to known protein domains, from random sequence fragments. Though the method went some way towards achieving this goal, it was concluded that the measure was not sufficiently reliable to predict domains from sequence. Sequence fragments of varying lengths were often found to have a 'globularity' score similar to that expected for true domains.

The percentage of hydrophobic residues in single and multi-domain chains was calculated (Figure 3.4). The mean proportion of hydrophobic residues in single and multi-domain chains appears to be consistent, around 48%. Overall, there appears to be no obvious separation between the single and multi-domain distributions, on which a prediction method might be based.

The distribution of hydrophobic and hydrophilic residues within protein sequences was also considered. As protein domains contain a hydrophobic core, it may be possible that their distribution within the sequence may indicate regions corresponding to domains. If long, continuous stretches of hydrophobic residues were present in protein sequences that corresponded to the hydrophobic core, domain assignment could be based on such distributions of residues. However observations of residue distributions appeared random - plotting a window of hydrophobic to hydrophilic ratio along protein sequences in the data set showed that generally, the ratio remained similar (though rather noisy, suggesting random fluctuations along the chain) over the sequence lengths (data not shown). The use of such a calculation to distinguish between single and multi-domain chains does not seem worthwhile - there is no pattern of residues that can be used to predict boundary from sequence. Prediction might be possible if there were a clearly identifiable hydrophobic stretch of residues, surrounded by hydrophilic surface residues in globular proteins. This is in agreement with the study by White and Jacobs, (1990) that came to a similar conclusion, stating that the distribution of hydrophobic residues along the chain cannot be distinguished from that expected from a random distribution for the vast majority of soluble proteins examined.

Analysis of the amino acid propensities for core, surface and domain interface residues showed that domain interface residue compositions are most similar to those for protein surfaces than cores. These observations are in agreement with the study by Jones *et al.*, (2000), which came to similar conclusions. This may not be surprising given that a non-redundant set of chains from the CATH database was used in both studies, though version 1.7 was used here, as compared to version

1.5 in their study. They proposed that the presence of amino acids in surface-like proportions in domain interfaces support a protein folding pathway in which domains first fold, and then collapse into a multi-domain structure.

It is possible that the use of factors such as the percentage of exposed residues and some measure of hydrophobicity may have some use in domain prediction, but not in isolation. The results have shown that there are constraints to the number of surface exposed residues and hydrophobic residue content of proteins as conferred by the hydrophobic effect, as proposed by Fisher (1965). The creation of the hydrophobic core means that as well as the burial of non-polar side chains, the polar main chain must also be buried. The polar groups of the main chain form secondary structures, thus satisfying their hydrogen bond potential. These secondary structures then come together to form the fold of the protein. Whilst the primary sequence between similar folds my not be conserved, the secondary structure pattern is. The use of the more conserved protein secondary structure may be of potential use when designing a domain prediction method.

Chapter 4

Rapid protein domain assignment from amino acid sequence using predicted secondary structure

4.1 Introduction

Difficulties in elucidating the domain content of a given sequence at the structural and sequence homology level arise when the target sequence has no experimentally determined structure and searching the target sequence against sequence domain databases results in a lack of significant matches. In such situations, an *ab initio* approach to domain assignment from sequence is required. Indeed, several attempts have been made, although with limited success, to describe protein domains from sequence alone, including those by Busetta & Barrans (1984), Vonderviszt & Simon (1986) and Kikuchi (1988).

Two of the most recently published algorithms that attempt to overcome this difficulty are Domain Guess by Size (DGS) by Wheelan *et al.*, (2000) and SnapDRAGON (George & Heringa, 2002a). DGS aims to predict the likelihood of putative domains within a given sequence based on probability distributions of chain and domain lengths within a representative set. SnapDRAGON is a much more computationally intensive approach that averages several hundred predictions obtained from *ab initio* simulations of the 3D-structure for a given sequence to assign its domain content. Of the two methods, SnapDRAGON appears to be the most reliable, although the computational requirements (i.e. running hundreds of *ab initio* simulations for each target sequence) render it impractical for routine use, especially for any kind of genome-scale analysis.

The approach described here is based on the idea that a crude fold recognition algorithm based on the mapping of predicted secondary structures to observed secondary structure patterns in domains of known 3-D structure might be reliable enough to parse a long target sequence into putative domains. This is often the way in which a human sequence analyst will attempt to parse a protein into domains when homology-based approaches have been unsuccessful. Automatic analysis of secondary structure is therefore a very logical approach. Also, recent improvements in secondary structure prediction accuracy (Jones, 1999) where methods now routinely achieve 3-state prediction accuracies of 77%, have greatly increased the usefulness of predicted secondary structure in recognising protein folds.

Although many previous approaches to fold assignment using secondary structure attempted to align strings of secondary structure codes, a successful recent approach (Russell *et al.*, 1996) has used a scoring scheme based on the alignment of

secondary structure elements. With the recent advances in secondary structure prediction accuracy, secondary structure element alignment methods (SSEA) have been shown to provide a rapid prediction of the fold for given sequences with no detectable homology to any known structure and have also been applied to the related problem of novel fold detection (McGuffin *et al.,* 2001; McGuffin & Jones, 2002). In this study a new domain prediction algorithm, DomSSEA, is evaluated that uses predicted secondary structure to predict continuous domains, aimed at the automated annotation of genome sequence data. Several other methods that range in their complexity are also assessed.

4.2 Methods

4.2.1 Data set

For this study, a non-redundant set of protein chains, sharing no more than 30% pairwise sequence identity and with X-ray crystallographic resolutions ≤2.5 Angstroms was used. This set was derived from all chains in the CATH domain database, version 2.3 (Orengo *et al.*, 1997), from which corresponding domain assignments were taken. The set used to assess continuous domain assignment consisted of 1137 chains from the non-redundant set, which formed single domain and continuous multi-domain structures. A further 123 discontinuous two-domain chains (again taken from the main non-redundant set) were used to measure the effectiveness of DomSSEA in predicting discontinuous domain boundaries. All domain predictions for a given chain were compared to assignments given in CATH.

The following sections, 4.2.2 to 4.2.7 outline the different methods used to predict protein domain from sequence.

4.2.2 Random prediction

4.2.2.1 Random prediction of domain number

As a baseline measure of domain number prediction, the number of domains was assigned randomly to each chain in the representative set. The random assignments were weighted in terms of the frequencies of single and multi-domain proteins in the representative set. In this study the shortest length permissible for a

domain was 40 residues (over 99% of domains in CATH are greater or equal to this length). In turn the shortest length considered for a two-domain assignment was 80 residues (i.e. an equal division yields two 40 residue domains). Similarly, the shortest length for predictions of three-or-more domains was 120 residues.

4.2.2.2 Random prediction of domain boundaries

If a sequence was randomly predicted to be multi-domain, random assignments were also made for corresponding domain boundaries. Cuts were made such that no boundary was placed within a distance less than 40 residues from the sequence termini. For example, in the case of a two-domain protein, a window within the sequence was considered in which a cut could be made, whereby 40 residues at the C-terminal and N-terminal extremes of the sequence were masked off. A random cut was then made in this window. In cases where the sequence length was exactly 80 residues, an equal partition was made. Similarly, when three-or-more domains were predicted, random cuts were made such as to ensure that again, no domain was less than 40 residues in length.

4.2.2.3 Trivial boundary assignment procedure

Given that the number of domains for the target sequence has been predicted, one of the simplest ways to partition the sequence into domains is to divide it into equal fragments. In other words, given a sequence length L, and the predicted number of domains N, each domain length can be considered as L/N.

For all the random methods, random simulations were carried out 100 times, and the average success rate calculated.

4.2.3 Sequence alignment

An all-against-all alignment of sequences in the non-redundant set was carried out in order to predict both domain number and domain boundaries. FASTA (Pearson & Lipman, 1998) was used to align each target sequence against all other sequences in the representative chain set. The top scoring alignment was used to determine the domain number of the target. In cases where the top-scoring hit was

multi-domain, the cut points were determined by mapping the known cut-points of the template chain onto the target chain.

4.2.4 Absolute difference in length

All chains in the data set were compared such that the similarity of chain pairs was scored according to their absolute difference in sequence length. This was normalised by the maximum length to make values comparable. Domain number and boundaries were taken from the top scoring hit.

4.2.5 Domain Guess by Size (DGS)

The original DGS algorithm was implemented using the probability distributions as outlined by Wheelan *et al.*, (2000) (here, called DGS-W). The algorithm was also implemented using probabilities generated from the representative data set (here, called DGS-M).

4.2.6 Secondary structure element alignment (DomSSEA)

An all-against-all alignment of the secondary structure elements for each chain in the non-redundant set was carried out using a modified version of the dynamic programming algorithm previously developed by McGuffin *et al.*, (2001) with a scoring scheme adapted from Przytycka *et al.*, (1999).

Secondary structure element alignments were made by aligning the secondary structure elements of a target chain to secondary structure of a putative template chain. Alignments were made using a dynamic programming algorithm based on that of Needleman and Wunsch (1970). The method emulates the secondary structure alignment scoring method of Przytycka *et al.*, (1999). Secondary structure elements were aligned, and the alignment scored according to the scheme outlined by Przytycka *et al.*, (1999). Matching elements, were scored by the minimum length of the two elements. Alignment of helix with coil or strand with coil was scored by half of the minimum length of the two elements. Alignment of helix with strand scored 0. No gap penalty was imposed.

The use of both observed and predicted secondary structure was assessed. Top hits were taken as the pair with the highest alignment score. Domain boundaries were taken from the position to which the template domain boundary aligned to the target. Assignments were weighted towards coil regions of chain, as analysis of domain-linking peptides revealed they are most commonly found as unstructured regions of chain (Chapter 2).

Observed secondary structures for all chains were taken from DSSP assignments (Kabsch & Sander, 1983). The eight DSSP structural states were simplified to three as outlined in section 2.2.1.

Secondary structure predictions were made using PSIPRED (Jones, 1999). Five sets of neural network weights are used to train the network, and in cases where a sequence is found to have homology to sequences used to train a sets of weights, the corresponding weight set are excluded, whilst the remainder are used for prediction. Q_3 and Sov (Zemla *et al.*, 1999) scores were calculated to measure the prediction accuracy. The Q_3 score gives the percentage of a protein sequence that is correctly predicted based on a three-state classification, helix, strand and coil. The Sov score also takes into account the location and lengths of the predicted secondary structure segments.

4.2.7 Upper benchmark using the Dali Domain Dictionary

The Dali Domain Dictionary (DDD) (http://www.ebi.ac.uk/dali/) was used as an upper control for benchmarking the methods. The algorithm used by the DDD to assign domains from structural data is PUU (Holm & Sander, 1994). A common set of chains found both in the representative set and in the DDD was compiled, and the domain number and boundary definitions given in the DDD were compared to the CATH assignments.

4.2.8 Homology filter

All top scoring pair-wise matches for the DomSSEA and FASTA alignment methods were further filtered for any possible remaining homology between them that could be detected by PSI-BLAST (Altschul *et al.*, 1997). PSI-BLAST is one of the most successful methods for detecting remote sequence similarity when used in

conjunction with a large non-redundant sequence database (Salamov *et al.*, 1999). The use of sensitive sequence comparison methods is often one of the first steps in locating putative domains in a target sequence with no known structure. In this study it was important to establish a starting point when benchmarking the methods, in which all sequence homology between matched sequences was eradicated so as to simulate cases where sequence searching had been exhausted. In other words, it was important that correct assignments were not attributable to matches at the sequence level.

PSI-BLAST was run with default parameters for five iterations, or until convergence. A large non-redundant sequence database was used, nrdb90 (Holm and Sander, 1998) which was combined with the set of representative CATH chains used in this study. Each chain in the representative set was scanned against the sequence database and all significant pairwise matches (E-value <0.01) found within the CATH representative set were recorded. This list was used to filter the top hits generated by the alignment methods.

4.2.9 Measuring the accuracy of domain prediction

This study was undertaken to try and measure the usefulness of prediction methods as if they were to be applied as automatic assignment algorithms. In terms of a typical CAFASP (Fischer *et al.*, 2001) assessment where automatic methods for fold recognition are assessed, the fold template with the highest score or 'top hit' is taken to be the predicted fold of a given target. In this study, a similar approach in assessing the domain assignment methods was taken, basing the measurements on the presumption that they could be used to automatically analyze a large number of sequences or even whole proteomes. The accuracy of each domain assignment method was primarily measured by calculating the number of correctly assigned top hits.

4.2.9.1 Accuracy of domain content prediction

The measurement of the success of a particular method in domain content assignment was calculated in two ways. First, the exact domain number, as assigned by CATH was validated i.e. it was simply a question of how often the method

predicted the correct number of domains as assigned by CATH. Second, predictions were scored on a basis of correctly assigning the target chain as single or multi-domain. In cases where two or more hits were found to have the same assignment score for a given target the success rate was calculated to reflect this. For example, if a target was assigned three hits with identical scores, and two were correct predictions and one incorrect, the overall prediction for that particular target was given an accuracy score of 2/3.

4.2.9.2 Accuracy of domain boundary prediction

For measurement of domain boundary prediction accuracy, a correct assignment was given if the predicted cut fell within a given cut-off window around the boundary defined by CATH. A scale of ± 1 to 20 residues either side of the CATH cut was used to assess the accuracy of the boundary prediction. Accuracy was calculated in two ways. First, in cases where the target contained multiple boundaries, the correct number of boundaries had to be given with the assignments falling within the CATH boundaries for a prediction to be regarded as correct (for a given window cut-off). Second, the sensitivity and selectivity of domain boundary prediction was calculated. Sensitivity was defined as the number of correct boundary predictions, divided by the number of boundaries to predict. Selectivity was defined as the number of correctly assigned domain boundaries, divided by the number of boundary predictions made. In cases where more than one hit shared the highest score, a random selection was made from the predictions.

4.2.10 Consensus domain boundary prediction method

A consensus boundary prediction method was used in order to take into account predictions made by several methods used in the study, including DGS, DomSSEA (predicted secondary structure), difference in length and the equal division procedure. Predicted cut points were grouped in terms of regions with most neighbouring predictions and then the average of the most populated group was taken as the cut point. Where no consensus could be reached, the assignment made by DomSSEA was used.

4.2.11 Comparison of the different methods

The comparison of the prediction methods was carried out in three main areas.

1) Correct prediction of the domain content in a target chain.

2) Correct prediction of domain boundaries. Initial observations showed it was difficult to compare the overall top prediction given by DGS with the other methods and easily draw decisive conclusions from the results. In order to analyze the success rate of domain boundary delineation, each algorithm was assessed for its ability to predict the domain boundary for all the two-domain proteins in the non-redundant set (where each method was provided with the knowledge that the target chain was two-domain). This procedure was necessary to provide a more level playing field for comparison of the methods in terms of boundary prediction accuracy. To achieve this for the alignment methods, a pair-wise comparison of the two-domain chains was undertaken.

3) Assessment of overall prediction accuracy, i.e. prediction of domain content and boundaries.

4.3 Results

4.3.1 Length distributions

Figures 4.1 and 4.2 show the length distributions for the chains and domains in the non-redundant set (section 4.2.1) as used in our own implementation of the DGS algorithm. Figure 4.2 shows an overlap between the different chain length distributions, which has implications in domain prediction. If the length distributions of single and multi-domain proteins did not overlap, the prediction of domain content for a given sequence would not be such a complex an issue. As chain length increases, the likelihood of the chain having a multi-domain conformation also increases. The mean domain length was found to be 150 residues. In general, domains from single and multi-domain chains have a similar length distribution.

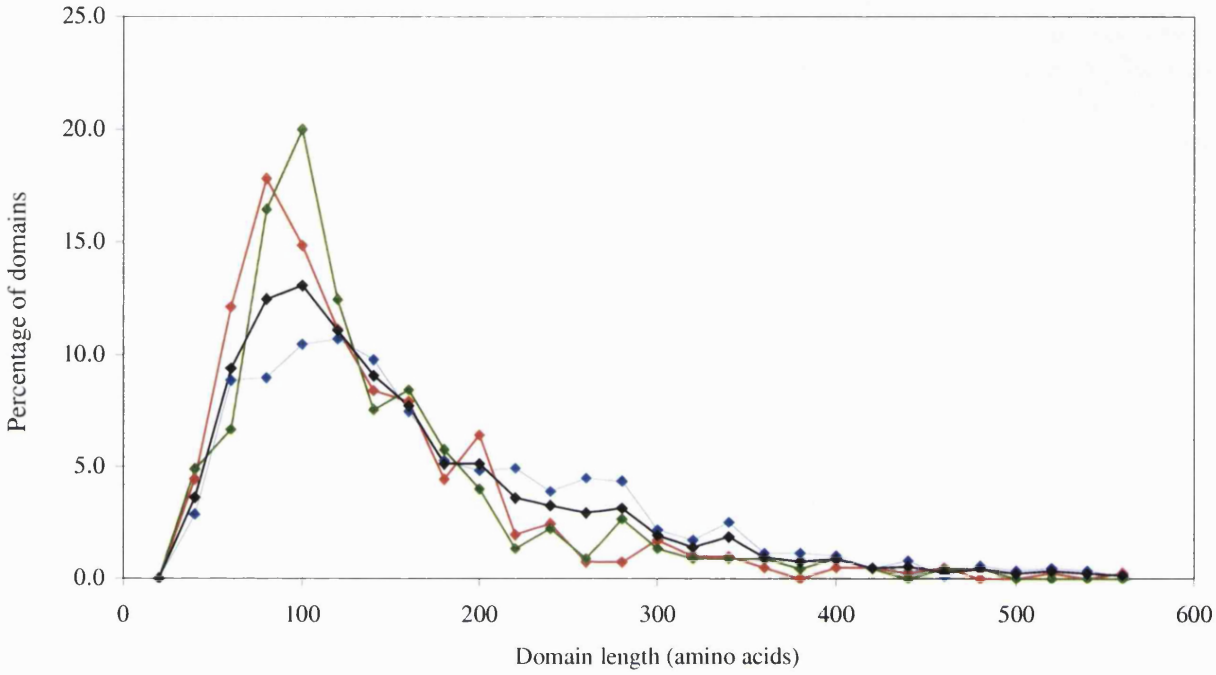


Figure 4.1 Domain length distributions

Domain length distributions as observed in the CATH representative set used in this study. Intervals were calculated with a width of 20 residues. The domain frequencies were used by DGS-M to calculate the probabilities of a predicted domain length. Single domain (blue), two-domain (red), three-or-more domains (green), all-domains (black).

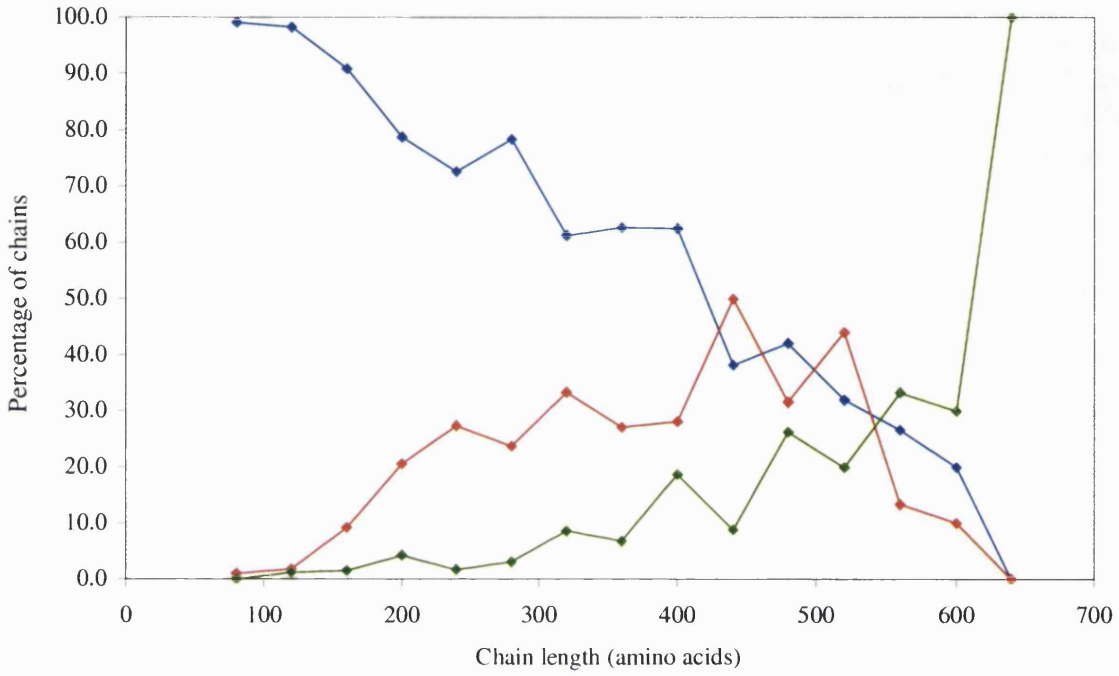


Figure 4.2 Frequency of chain lengths

Frequency of chain lengths of one (blue), two (red), and three-or-more domain (green) chains for a 40-residue length interval. These frequencies were used by DGS-M to calculate the likelihood of the number of domains for a given chain length.

4.3.2 Secondary structure prediction accuracy

PSIPRED secondary structure predictions (section 4.2.6) had a Q_3 accuracy of 76.6% and Sov score of 72.5%.

4.3.3 Domain number prediction

The success rate of each method in predicting the number of domains for each chain in the non-redundant set can be seen in Table 4.1. This was measured as the percentage of one, two and three-or-more domain chains predicted correctly. Also shown is the success rate for domain number prediction for all the chains in the representative set. The simplest method, random-weighted (section 4.2.2.1) set the lower limit of prediction. Domain number was assigned according to the frequencies found in the representative set. Here the overall success rate was 61.4%, with three quarters of the single domains correctly assigned, 16.8% of two-domain and 6.3% of three-or-more domains. These values agree well with the theoretical values of 76%, 17% and 7% for single, two and three-or-more domain chains respectively, calculated from the sum of squares of the frequencies of the single and multi-domain chains in the non-redundant set (last line of Table 4.1).

The comparison of the CATH and DDD assignments set an upper limit for domain prediction (section 4.2.7). The PUU algorithm used by DDD to assign domains is a fully automated method in contrast whilst CATH domain identification is semi-automated. Table 4.1 shows that agreement between the domain databases covers approximately 80% of single domain chains, whereas nearly two-thirds of two-domain and three-or-more domains are given matching assignments. The results of the all-against-all alignment of sequences in the non-redundant set (section 4.2.3) are close to those values generated by the random method, confirming the lack of discernable sequence identity in the benchmarking procedure (section 4.2.8). The top assignments for both DGS-W and DGS-M (section 4.2.5) were most often found to predict the target as a single domain chain. This gives 100% prediction accuracy for single domain chains, but few correct predictions for multi-domain chains. Therefore, here the success rate of DGS top hit domain number prediction reflects the percentage of single domain chains in the test set only. Scoring the all-against-all comparison of the non-redundant set in terms of the absolute difference in length

Method	Percentage correctly assigned domain number			
	All	1 domain	2 domains	3+ domains
Dali Domain Dictionary	79	81	66	65
DomSSEA observed secondary structure	75.4	83.9	47.5	38.1
DomSSEA predicted secondary structure	73.3	82.3	45.6	36.5
DGS-M	76.7	99.8	1.0	0.0
DGS-W	76.7	100.0	0.0	0.0
Absolute difference in length	66.2	78.4	22.3	38.1
Sequence alignment (Fasta)	60.9	74.9	17.3	7.9
Random (weighted)	61.4	75.8	16.8	6.3
Sum of Squares	62.0	76.0	17.0	7.0

Table 4.1 Prediction of domain number

The percentage of chains given a correct domain number prediction (top hit prediction). Values are shown for chains correctly predicted as single, two-domain and three or more domains. Also shown is the domain number prediction success rate for all chains in the representative set.

gave an overall success rate of 66.2%. A large percentage of the single domain chains were predicted correctly, with just over 20% of the two-domain chains and over one-third of the multi-domain chains.

Of all the methods, DomSSEA (section 4.2.6) achieved the highest success when considering multi-domain proteins, especially for two-domain chains. Over 80% of the single domain chains are correctly assigned, with just under one-half of the two-domain chains and one-thirds of three-or-more domain chains predicted correctly. The use of predicted secondary structure over observed does not appear to be overly detrimental to the outcome of the method. Table 4.2 shows the percentage of correct and incorrect domain number prediction given by DomSSEA (predicted secondary structure). The majority of false positive predictions given by DomSSEA tend to be under predictions of domain number (and in turn domain boundary frequencies).

4.3.4 Boundary prediction for two-domain chains

As shown, each method tested predicts domain number with varying levels of success. To provide a more level playing field and facilitate an easier comparison of domain boundary prediction each method was used to predict the domain boundary for the 202 two-domain chains in the non-redundant set (rather than predicting both domain number and boundaries) (section 4.2.11). In other words, given the target was known to be two-domain by a given method, how often could the cut point between the domains be correctly predicted? Figure 4.3 shows a plot for each method for the percentage of correct assignments for windows of ±1-20 residues. Table 4.3 shows the percentage of top hits giving the correct domain boundary within a window of ±20 residues around the CATH assignment. The methods in this table are ranked in order of success. Random boundary assignment provides the baseline in dividing two-domain chains. This random simulation sets the base line of locating the domain boundary in the two-domain chains, with just under 27% of the boundaries correctly located (±20 residues). Again the alignment of sequences resulted in a similar level of prediction accuracy. The most successful method, and upper benchmark are the domain assignments given in the Dali Domain Dictionary. The common set of chains found in CATH and DDD gave an 81.8% agreement in the domain boundary assignments. Interestingly the results from the two

Number of domains predicted by DomSSEA	Number of domains assigned by CATH		
	1	2	3 or more
1	**82.3**	43.1	15.9
2	14.7	**45.6**	47.6
3 or more	3.0	10.9	**36.5**

Table 4.2 Percentage of correct and incorrect domain number predictions given by DomSSEA

The predicted number of domains by DomSSEA (predicted secondary structure) against the number of domains assigned by CATH.

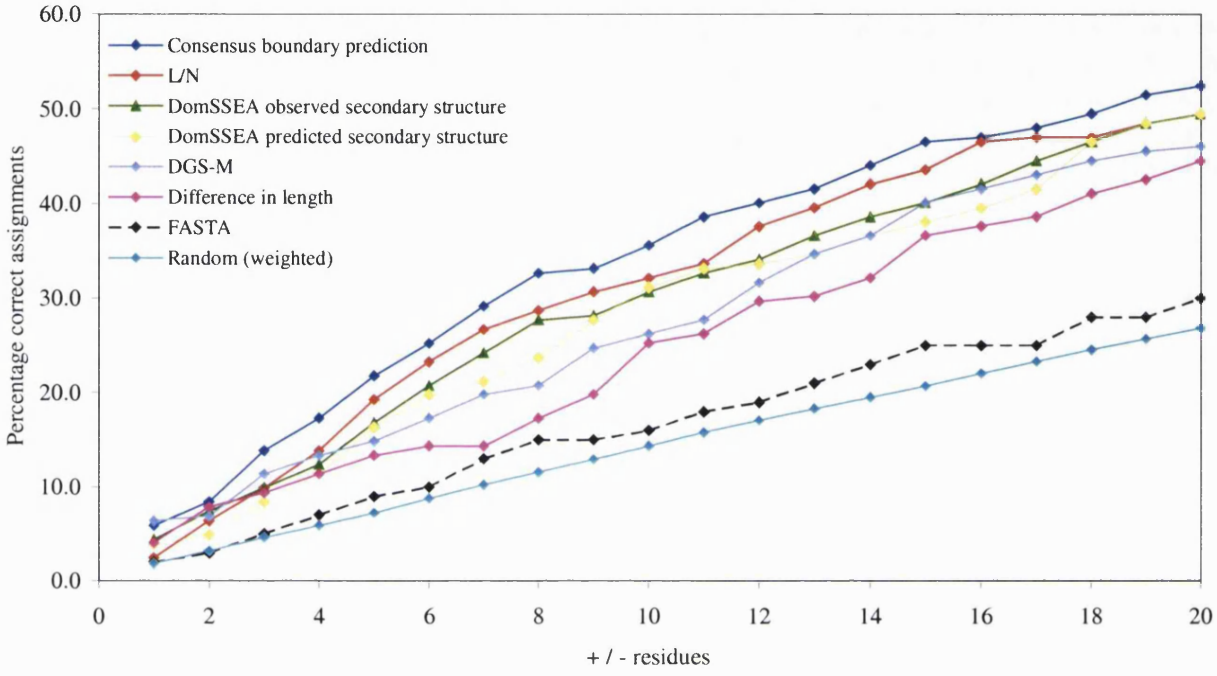


Figure 4.3 Prediction success of two-domain chain boundary assignment

Success rate for the top-hit domain boundary assignment for two-domain chains, for window cut-offs between ±1-20 residues.

Methods	Percentage correctly assigned boundaries
Dali Domain Dictionary	81.8
Consensus	52.5
L/N	49.5
DomSSEA observed secondary structure	49.5
DomSSEA predicted secondary structure	49.0
DGS-M	46.0
Absolute difference in length	44.6
DGS-W	37.1
FASTA	30.0
Random (weighted)	26.8

Table 4.3 Domain boundary prediction for two-domain chains

Prediction of domain boundaries were made for a representative set of two-domain protein chains (± 20 residues).

implementations of DGS differ somewhat. The results generated by DGS-W achieved correct assignments in approximately 37.1% of the two-domain chains, whereas DGS-M, using probabilities generated from our own data set, predicted a higher percentage of 46% correct boundary assignments at this cut-off (±20 residues). The success rate of absolute difference in length falls between DGS-W and DGS-M. Alignment of predicted secondary structure elements by DomSSEA produced some improvement over the DGS-M, with slightly over 49% of the predicted two-domain boundaries being correctly assigned (±20 residues). It is apparent that the division of two-domain chains into equal fragments is a valuable procedure. Just under half of the chains were assigned a correct domain cut. This reflects the degree to which the domain assignment in CATH partitions two-domain chains into equally sized units. Finally, the method that assigned the most cuts correctly in the absence of 3D structure was the consensus method (section 4.2.10) with approximately 52% of the chains assigned a correct cut (±20 residues).

4.3.5 Overall prediction of domain number and domain boundaries

A useful domain identification method must predict domain number and any corresponding domain boundaries with a reasonable degree of reliability. In terms of a fully automated protocol, one must consider the methods as an overall procedure, and the prediction is taken as the top hit assignment. The overall accuracy of top hit predictions for domain number and boundaries for multi-domain chains can be seen in Figure 4.4. Table 4.4 also demonstrates the effectiveness of each method in giving correct assignments for all chains in the representative set, at ±20 residues as well as for solely multi-domain predictions. The use of DomSSEA to both predict domain number and boundaries using predicted secondary structure gives correct assignments for just under 25% of the multi-domain chains, at ±20 residues (four times better than the next best method, difference in length). The simple procedure of dividing the chains into equal length, given the domain number was predicted by DomSSEA, results in a similar success rate to boundary assignments by secondary structure element alignment. Using DomSSEA to predict domain number and the consensus method to locate the corresponding domain boundaries proved to assign the greatest number of correct domain boundaries for the multi-domain chains over the window cut-offs of ±1-20 residues (Figure 4.4). As well as these predictions

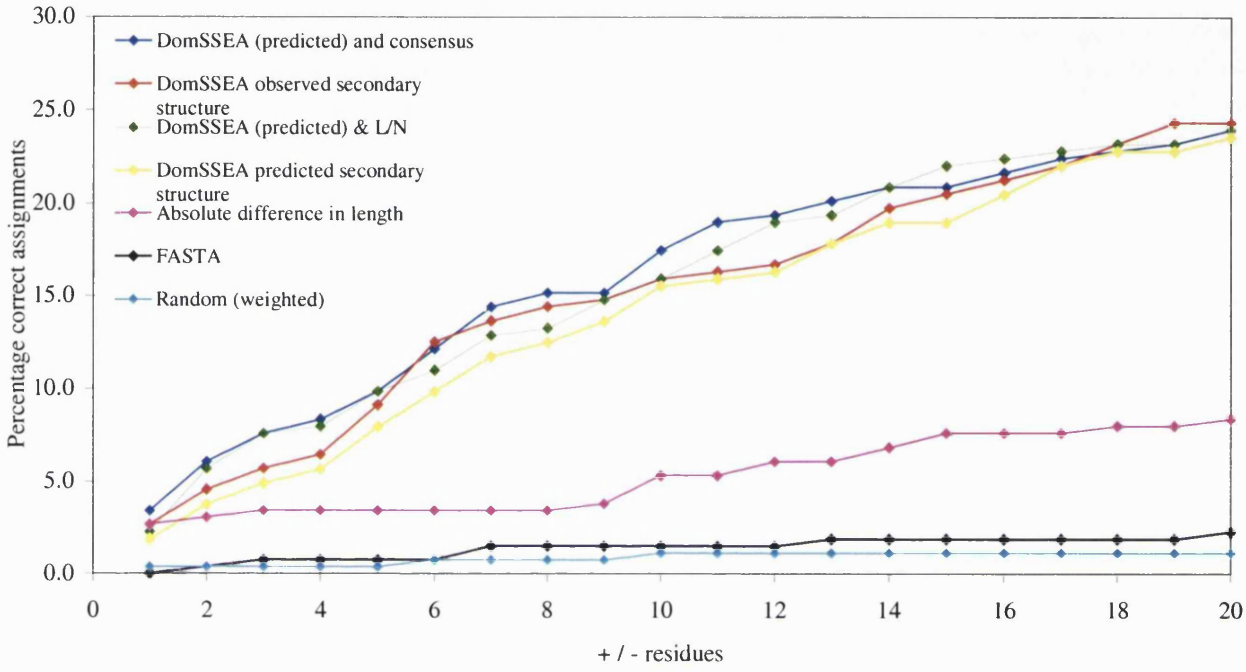


Figure 4.4 Overall prediction of domain number and boundaries for multi-domain chains

Overall predictions for multi-domain chains, for window cut-offs between ±1-20 residues. Correct predictions required both correct domain number and boundary assignments. Success was measured in terms of top-hit assignments.

Methods	Percentage correctly assigned	
	All chains	Multi-domain chains
DomSSEA observed secondary structure	70.2	24.7
*DomSSEA predicted & consensus	68.6	24.0
*DomSSEA predicted & L/N	68.0	24.0
DomSSEA predicted secondary structure	68.7	23.6
Difference in length	62.0	8.4
*Average domain length & DGS-M	66.6	6.1
Sequence alignment (FASTA)	57.9	2.3
Random (weighted)	58.3	1.1
DGS-M	76.6	0.0
DGS-W	76.6	0.0

Table 4.4 Overall prediction of domain number and boundary, for single and multi-domain chains (± 20 residues)

DGS achieved the highest overall correct assignments (for all chains) as it most often predicts single-domain as its top hit. Using average domain length to predict domain number also achieves a high overall success rate as any chain less than 300 residues in length (twice the domain average) is predicted as single-domain.

* Combined methods use first method for domain number predictions, and second method for boundary prediction. For example, DomSSEA predicted and consensus, uses the number of domains given by DomSSEA (predicted secondary structure) and then uses the consensus domain boundary prediction method.

including a high number of correct assignments for two-domain chains, several correct assignments were made for chains containing three-or-more domains with just over one-third of domains correctly assigned as three-or-more domains being given at least one correct domain boundary prediction within ±20 residues.

In an attempt to guide the top prediction given by DGS-M, the mean domain length in the representative set (150 residues) was used to predict the number of domains. For example, chain lengths less than or equal to 150 residues were predicted as single domain, between 150 to 450 residues as two-domain and greater than 450 residues (three times the average domain length) as three-or-more domains. DGS-M was then used to predict domain boundaries. This achieved a correct domain number and cut prediction for only 3% of the 265 multi-domain chains. As a further method, the average domain length was also used to predict domain boundaries, for instance a chain length of 320 residues was divided at 150 residues from the N-terminus. However this resulted in fewer correct predictions than using DGS-M to locate domain boundaries.

4.3.6 Sensitivity and selectivity

The method used in benchmarking domain boundary prediction in this study is more stringent than that used in a number of other domain prediction studies. Here, scoring domain boundary predictions for continuous multi-domain chains is dependent on correct assignment of domain number. Domain boundary prediction can also be scored on a basis of the number of boundaries correctly predicted, out of all the boundaries to predict in the multi-domain data set. Such a score is often defined as the sensitivity of prediction. However, this value does not take into account the number of predictions that have been made. For example, a two-domain chain of 100 residues has 21 possible cut points (if the minimum permissible domain size is 40 residues) after residue 40 and before residue 61. If 21 predictions were made, the sensitivity would be 100%, however this would be a misleading figure. A corresponding value, called the selectivity of prediction can be calculated, the number of correct predictions, divided by the number of predictions made. In this example the selectivity would be less than 5%.

Calculating such values for DomSSEA (predicted secondary structure) gives a sensitivity of 31% with an associated selectivity of 31.6% (±20 residues). To make these calculations for DGS-M is made more difficult by the fact, as previously observed, that its top hit prediction is most often single domain. A way to guide the domain number assignment made by DGS is to base the predictions on the average length of a domain (found to be 150 residues for the data set used in this chapter). For example, chains less than 150 residues can be assigned as single domain, those between 150 and 450 residues as two-domain, the remainder above 450 residues as containing three-or-more domains. Using this modified form of DGS-M gives a sensitivity of 21.1% with an associated selectivity of 23.5%. Using a random prediction method (where domain number is predicted and any corresponding cuts are randomly predicted) gave a sensitivity of 3.3% with an associated selectivity of 5.2% (±20 residues).

4.3.7 Discontinuous domain assignment

This analysis has so far only focused on the assignment of continuous domains. However, the use of DomSSEA for delineation of discontinuous domain boundaries was also addressed. Pairwise alignments were made by DomSSEA for a representative set of two-domain chains containing only continuous domains or at least one discontinuous domain (section 4.2.1). Predicted secondary structure was used for the target sequence. Two random baseline measurements were also implemented. Baseline 1 predicted discontinuous domain boundaries by equally partitioning the target protein into 3 equal fragments thus predicting two linker regions (found to be the most common number of linker regions in two-domain discontinuous chains). Baseline 2 randomly predicted the position of two linkers regions (Table 4.5).

The percentage of the discontinuous two-domain proteins with all linkers correctly predicted was 13% (±20 residues). Linker assignment accuracy was also calculated on a linker basis, in other words, the number of linkers in the two-domain discontinuous test set correctly identified. Again, such prediction accuracy can be measured in two ways: sensitivity, the number of linkers correctly predicted divided by the total number of linkers to predict and selectivity, the number of correct predictions divided by the total number of predictions made. Table 4.5 shows both

Method	Sensitivity		Selectivity	
	± 10	± 20	± 10	± 20
DomSSEA	16.4	33.1	24.6	49.7
Baseline 1	13.4	24.4	17.7	32.3
Baseline 2	11.0	24.1	14.6	31.9

Table 4.5 Prediction of domain boundaries for a representative set of two-domain chains containing discontinuous domains

Baseline 1 predicts domain boundaries by assigning two linkers to each chain by equally dividing the chain into three fragments. Baseline 2 also assigns two linkers, but randomly.

sensitivity and selectivity values for boundary cut-offs of ±10 and ±20 residues for DomSSEA and the two baseline methods. Baseline 1 gives a sensitivity of 13.4% followed by Baseline 2 with 11.0% at ±10 residues. DomSSEA gives a slightly higher success rate if 16.4% of the discontinuous linkers assigned correctly at the same cut-off. The selectivity measurements give higher values for the two baseline methods as well as DomSSEA, reflecting its tendency to under predict discontinuous domain linkers.

4.4 Discussion

In this study a domain identification method has been implemented using the alignment of predicted secondary structures of target sequences against observed secondary structures of chains with known domain boundaries. Whilst mutations at the sequence level can obscure similarity between homologues, their secondary structure patterns remain more conserved because changes at the structural level are less tolerated. The secondary structure alignment methods used here aim to exploit these conserved features to locate domain regions within secondary structure strings. The increase in accuracy in secondary structure prediction methods in recent years has also made such attempts worthwhile. The overall aim was to evaluate how well domain number and boundaries can be assigned to a given sequence using the new method as well as other methods, in a situation where homology searches to sequences with known domain assignments has been exhausted.

The similarity of the sequence alignment methods to random methods confirmed that sequence homology was eliminated from the representative set by the PSI-BLAST filter. In terms of distinguishing between one, two and three-or-more domain chains, DomSSEA gave the most reliable results. Analysis of the two-domain chains as a simple means to measure boundary prediction showed some improvement of DomSSEA over the next best method DGS, in predicting domain boundaries. However the advantage of this method can be best seen when used as an overall. It achieves the highest number of correct domain number and boundary assignments, for 25% of the multi-domain chains (±20 residues, see Figure 4.4).

The comparison of the methods evaluated in this study to DGS was not trivial. Taking only the top assignment from each prediction exposes the limitations of DGS in providing a reliable top guess. This issue was addressed in two main

ways; first by evaluating the ability of each method to predict the domain boundary for a set of two-domain proteins thus making a fairer comparison and secondly by using average domain length (calculated from the representative set) to guide the DGS-M domain number prediction and therefore top predictions. Using domain length to guide the DGS top prediction gave a linker prediction sensitivity of 21.1% by DGS-M compared to a sensitivity of 31% for DomSSEA, with a selectivity of 23.5 compared to 31.6 for DomSSEA (±20 residues). The top hit given by DGS becomes a lot more valuable with this approach and shows that domain prediction based on protein length distributions can be effective. It should perhaps be noted that the use of length cut-offs to guide domain number prediction for DGS only is not necessarily a fair comparison, as other methods are not treated in the same way. If the published form of DGS were to be used automatically, it is less useful than DomSSEA. DGS is more useful as a guide to human experts, as it produces a selection of likely possibilities from which a decision can be made. Automatic methods would have to decide on a single answer without human intervention.

A interesting observation from this analysis is the frequency with which multi-domain chains contain domains of similar length. Figure 4.3 shows that at a cut-off of ±10 residues around the CATH cut, 33% of the representative two-domain chains contained a domain boundary at the mid-point of the sequence. In order to verify that this equal partitioning of chains was not just a feature of the CATH assignment algorithm, the CATH non-redundant set of chains was compared to a common set of chains found in DDD, and a further common set of chains found in SCOP. These common sets were searched for the chains assigned with two equally sized domains by CATH, ±10 residues. Of these chains found in DDD, 88% were also assigned as two-domain with a boundary midpoint in sequence, whilst 97% of these chains found in SCOP had similar assignments, ±10 residues. Furthermore of all the chains assigned as continuous two-domain in the DDD common-set, and all those assigned as continuous two-domain in the SCOP common-set, 33% and 34% were given domain cuts midpoint in the sequence respectively. The tendency to partition chains into equal fragments therefore does not appear to be solely a feature of CATH. Although domain number and boundary assignments differ to varying degrees, depending on which two classifications are compared, all three classifications assign over 30% of their two-domain proteins with a boundary midway between the C- and N-termini of the sequence. Indeed, as shown, the equal

division of multi-domain chains is a successful method in determining domain boundaries given the correct domain number is known. Analysis of predictions given by DGS (with domain number prediction fixed for two-domain chains (section 4.2.9.2), or guided by domain length (section 4.3.6)) showed that many two-domain assignments predicted a cut midway in the target sequence – many of which were correct. The relatively high incidence of proteins containing similar sized domains may in some cases be attributable to domain duplication events at the gene level, forming domain repeats within a single structure (Heringa & Taylor, 1997). It may be possible that the folding of multi-domain chains is best achieved when domains are of a similar length, and may fold at a similar rate thereby avoiding aggregation.

Although DomSSEA (using predicted secondary structure) and the equal partition method predicted domain boundaries with a similar success rate, to what extent do their predictions overlap? If the top two predictions given by DomSSEA are evaluated, 28% of the multi-domain chains are given correct domain number and boundary assignments (±10 residues). Alternatively if the top prediction given by DomSSEA is taken, but a second prediction is based on the number of domains predicted by DomSSEA (second prediction) and the domain boundary predicted by the equal division method, 34% of the multi-domain chains are given correctly assigned boundaries (±10 residues). This increase demonstrates fewer overlapping predictions between DomSSEA and the equal division method. (A similar procedure for the partition of two-domain chains gives 41% correct hits for the top two DomSSEA predictions, and 53% if both DomSSEA and the equal division predictions are considered). Although these boundary prediction methods do overlap the secondary structure element alignment procedure is able to predict more complex domain arrangements than the simple subdivision method. Such a combination of methods is worthy of consideration.

The assessment of the top ten assignments given by a prediction method has advantages, allowing correct predictions further down the list to be taken into account. In terms of predicting domain number however, benchmarking such a top set of assignments could be a rather meaningless measure; in cases where several different domain number predictions are given, it is likely one is going to be correct. Perhaps more valuable is a top set of predictions for cases where a multi-domain chain has been predicted. Here different boundary assignments could be checked and used accordingly. This would most likely be a manual procedure and would be

difficult to integrate into an automated annotation method. For example, for a given target with no detectable sequence homology to a known structure or domain sequence, one could take the domain number prediction given by DomSSEA. If the target was predicted as two-domain, the top three two-domain predictions could be considered. This would thus give six putative domains to be threaded. For the two-domain chains in the representative set (±20 residues), one of the top-three predictions by DomSSEA gave a correct boundary assignment for over 60% of the targets. Nevertheless, care would have to be taken benchmarking such a list of hits as the likelihood of a correct assignment increases as more domain cuts are considered, especially for shorter chains. This however would be at the expense of the number of domains that would need to be tested by threading methods. Predicting all the domain boundaries correctly within chains of three-or-more domains has been found to be a difficult problem for all the methods analyzed. The most successful method was dividing the chains into equal domain lengths. This reflected the observed frequency of those multi-domain chains having similar sized domains. However there are many more multi-domain chains having dissimilar sized domain combinations.

A two-domain protein test set containing continuous and discontinuous domains was used to measure the potential of DomSSEA in predicting discontinuous domain boundaries. Although such an all-against-all alignment of two-domain chains does not give an indication of how introducing discontinuous domains into the DomSSEA library alters domain number prediction and overall assignment accuracy, it does give an insight into boundary prediction given that the correct domain number has been predicted. With just over 13% of the two-domain discontinuous chains given correct assignments for all domain linkers, (±20 residues) the boundary prediction accuracy is not high. The calculation of boundary assignment on a per linker basis showed some increase in assignment accuracy of DomSSEA over the baseline random methods. The selectivity measure of nearly 50% of linkers correctly predicted (±20 residues) appears encouraging, but must be tempered by the fact that this value is partially attributable to the observation that DomSSEA tends to under-predict discontinuous domain linkers. This is due in part to the false positive alignment of chains composed of continuous domains against target chains containing discontinuous domains. Interestingly although the equal division of continuous chains gave a similar percentage of correct domain assignments to

DomSSEA, the same is not so for baseline 1, where the success rate was lower. Whilst the addition of discontinuous domains to the DomSSEA library would make discontinuous domain assignment possible to some degree, it would also have a detrimental affect on the reliability of continuous domain assignment, introducing a greater number of false positive boundary predictions. One would have to weigh up the advantage of assignment of discontinuous domains, with the trade off in reducing continuous assignment accuracy. If methods such as DomSSEA are to be applied to genomes of higher organisms, as is intended, one must take into account the modularity of higher eukaryotic gene products, especially for larger proteins. A large frequency of multiple-domain proteins in higher eukaryotes are made up of continuous domain units, a result of gene duplications and fusion events making proteins containing continuous modular regions of structure the predominant class. Furthermore, the usefulness of discontinuous domain assignment must also be considered in terms of structure prediction. At present, the ability to predict the structure of such domains using fold recognition, given that fold libraries consist of continuous domains is extremely limited.

Recently the SnapDRAGON method developed by George & Heringa (2002a) has been published which uses *ab initio* folding simulations to predict the domain boundaries within a given amino acid sequence. Direct comparison of success rates between SnapDRAGON and DomSSEA is not easy due to the different calculations that have been used to measure the accuracy of the methods. However the success in assignment of domain number appears to be similar, with DomSSEA (using predicted secondary structure) giving correct predictions for 73.3% of protein chains compared to 72.4% by SnapDRAGON. Comparison of domain linker prediction can be carried out on the basis of sensitivity and selectivity of linker prediction. The selectivity values of the two methods are relatively similar, shown to be 39.1% for continuous chains in the SnapDRAGON study and 31.6% here. Comparison of the sensitivity values between the methods shows a more substantial difference, with SnapDRAGON having a higher sensitivity of 55% compared to DomSSEA having 31%. The data sets used in the two method varies somewhat as do the linker boundary assignments. Perhaps more important is the computationally intensive aspect of SnapDRAGON, for example, 100 processors are required to make a prediction of length <400 residues in approximately an hour. It can be seen that this leads to a trade-off between the increase in accuracy of SnapDRAGON versus

DomSSEA, and the far greater time required to obtain a SnapDRAGON prediction compared to DomSSEA.

The use of DomSSEA alignments is based on the premise of a crude fold recognition method, as previously explained. At its best, the alignment of similar secondary structure strings between the template and target should hopefully provide a successful domain boundary prediction (or at least domain number prediction). This study has attempted to show what can presently be achieved by using relatively simple methods to predict protein domains from sequence in the absence of homology. The results have shown that the alignment of secondary structure elements is the most reliable of the methods analyzed here for domain number assignment and overall domain number and boundary prediction. A given method may perform well at predicting domain number or domain boundary, but it is when accuracy in both is combined that the most useable results will be achieved. The methods in this study were tested on a non-redundant set of chains taken from the CATH structural database. Although this is not a full set of genomic sequences, it enables a reliable insight into the effectiveness of these methods in comparison to one another. A future stage will be applying DomSSEA to such genomic data, in order to gauge its usefulness in larger scale genome annotation applications. Although it must be conceded that methods such as DomSSEA are still somewhat limited in their overall reliability, there is certainly room for such fast procedures to act as a pre-filtering stage in automatic genome annotation and threading methods, where domain boundaries cannot be located purely from comparative sequence analysis.

Chapter 5

A combined approach to domain assignment using PSI-BLAST sequence alignment and DomSSEA

5.1 Introduction

One of the best approaches to domain assignment, and the procedure that should be used first in any domain assignment strategy for a target protein sequence is to search sequence data banks. The transfer and duplication of whole domain sequences at the genetic level has enabled the evolution of modular proteins containing combinations of domains in varying orders and numbers (Rigden, 2002). The recurrence of domains in different protein sequences, where they may be found at different proximities to the N- or C-termini or as duplications, may be used for the assignment of domain boundaries. The comparative analysis of such proteins can identify domains common to many sequences (Sonnhammer and Kahn, 1994).

The alignment of domain sequences such as those found in SMART (Schultz *et al.*, 1998), Pfam (Bateman *et al.*, 2002) and ProDom (Servant *et al.*, 2002) can used to infer domain locations in sequence. However, such searches may be limited by the number of domain family representatives made available in these databases.

Sequence database search methods that use a position-specific score matrix (PSSM) or profile, generated from related sequences, can find more remote homologous matches than simpler pair-wise comparisons. One of the most widely used profile methods is PSI-BLAST (Altschul *et al.*, 1997), which iteratively builds a PSSM and uses it to search the sequence database. This continues until no more sequences can be detected or the program reaches the specified number of iterations. Assignment of domain boundaries by sequence comparison can be achieved in a number of ways. A number of methods with differing levels of complexity have been developed to assign domains from similarity searches made using the BLAST sequence alignment algorithm to large non-redundant sequence databases. These include BALLAST (Plewniak *et al.*, 2000), PASS (Kuroda *et al.*, 2000) and more recently DOMAINATION (George and Heringa, 2002b).

In this chapter a sequence based domain assignment algorithm is developed, Domains Parsed by Sequence (DPS), based on PSI-BLAST sequence alignments. The DPS method attempts to predict the domain boundaries within a multi-domain target sequence based on the observation that domain sequences can exist independently or may have been shuffled into different sequence contexts within different proteins. The alignment of such sequences to a putative multi-domain target sequence may provide enough information to infer domain boundary locations. The

approach used in DPS is that a domain boundary search is made by calculating a termini-profile over the query sequence, generated from the N- and C- termini alignment positions of all significant PSI-BLAST aligned sequences from a large non-redundant database of sequences. As such DPS is designed to be a simple method in which domain boundaries can be inferred from sequence alignments. In this chapter domain predictions are again made for the representative set of chains described in section 4.2.1. However, the DPS method is used as a pre-filter for domain assignment prior to the use of DomSSEA. Similarly to the previous chapter, the general strategy for domain assignment is focused towards delineating continuous domains, however the assignment of discontinuous domains is also considered.

5.2 Methods

5.2.1 The data set of chains used in this study

The data set used for this study was required for two main applications; first, to address the outcome of varying the parameters used by the sequence alignment method, DPS on its domain predictions, and second, to assess the use of DPS for domain prediction and the combined approach using DPS and DomSSEA. The set was made up of single and multi-domain sequences as assigned by the CATH domain database and were selected from the non-redundant set of chains described in section 4.2.1. This gave 369 multi-domain sequences, of which 263 contained only continuous domains, and 106 sequences contained one or more discontinuous domains. The discontinuous domains were selected on the basis that they contained no more than two domain-linking regions between adjacent domains, similar to the study by George and Heringa (2002b). A further set of 369 single-domain chains were also selected from the non-redundant set used in Chapter 3. These chains were selected as decoys to the multi-domain chains for the optimisation of the DPS algorithm. Further, they were compiled to assess domain prediction by DPS and the combined use of DPS and DomSSEA (described in sections 5.2.4 and 5.2.5). It is highly likely that a sequence of length less than 120 residues will be single domain (Jones *et al.* 1998, section 4.3.1). Therefore all single-domain chains used here were of a length greater than 120 residues, and were therefore of similar lengths to known multi-domain chains.

Reference domain assignments were taken from the CATH database. Corresponding domain boundary assignments were compared to domain boundary predictions, with a margin of error of ±20 residues about the CATH assignment. This is the error margin for boundary assignment that is used for most analysis of domain boundary assignments (Kuroda *et al.*, 2000; Wheelan *et al.*, 2000; George and Heringa, 2002b).

5.2.2 Domains Parsed by Sequence (DPS) – outline of the method

Each query sequence is searched against nrdb90 (Holm and Sander, 1998) using PSI-BLAST for detection of related sequences. PSI-BLAST parameters were set as described by Altschul and Koonin (1998) with an E-value of 0.001 for inclusion into the next round of searching (-h option). Each search was carried out for no more than 4 iterations, or until convergence. Identical matches made to the query sequence were removed. Each pair-wise sequence alignment generated by PSI-BLAST with an E-value less than 0.01 (see 5.3.1.1) was then analysed.

Next, the positions to which the aligned termini of the database sequence have matched to the query sequence were recorded. This was repeated for all significant PSI-BLAST matches, such that a count is made at a residue position in the query sequence whenever a database sequence termini residue has been matched to it. For example, if the first residue (i.e. N-terminal aligned residue) of an aligned database sequence alignment was matched to residue 100 of the query chain, the count for residue 100 was incremented by 1. In cases where alignment termini residues were within 15 residues of the real N- or C-terminus of the database sequence, the match was counted twice, as such aligned residues represent genuine database sequence domain. Two separate counts were made; one for database sequence N-termini matched residues and one for matched C-termini residues.

Once all the significantly aligned sequences have been counted, a smoothing window of 15 residues (section 5.3.1.1) was moved over the query sequence, one for the N-terminal distribution and another for the C-terminal distribution. This procedure involved averaging the values of all the residues that lay within the window, and giving the average value to the central residue. This smoothing effect was used in order to take into account variability in the length of homologous sequences. Next, the subsequent N-terminal and C-terminal smoothed distributions

were combined, taking into account those regions in which both N-terminal and C-terminal residues had been found to align. Such regions were more likely to indicate the end and beginning of adjacent domains and therefore a domain boundary in the query sequence. In such situations it might be expected that the C- termini region of a homologous domain sequence may be closely followed by the N-termini aligned region of another homologous sequence. Regions that contained counts for both N- and C-terminal aligned residues were given a weighting, such that half the maximal count at the N- or C-terminal residue position was added to the sum of the N- and C-terminal counts to form a combined value. In addition, for positions in which *only* N- or C-terminal (or neither) alignments had been observed, the combined profile was the sum of the two distributions. The combined smoothed distribution now represented a profile of the aligned database sequences. The elevated regions of the profile should represent putative domain boundaries, i.e. regions in which a number of database domain termini have been found. To assign domain boundaries from such elevated regions, or profile peaks, their significance was compared to the distribution over the remaining query chain. In order to do this a mean and standard deviation was calculated over the distribution and a Z-score was calculated for each residue in the query sequence; where the Z-score is given by:

$$\text{Z-score} = \left(X - \overline{X}\right) / \sigma\left(X\right)$$

where X is the profile value for each residue position.

The alignment of nrdb90 sequences can result in termini hits that are equivalent to the N- and C- termini of the query chain. The aligned positions of these termini hits may not exactly match those of the query sequence, a result of the natural variation in lengths of homologous sequences. This can lead to profile 'edge effects' in both the first and last 40 residues of the query sequence. It is important to avoid incorrect boundary predictions due to such edge effects which may appear as significant peaks in the profile. In order to address this issue, the first and last 40 residues of the boundary profile are flattened, i.e. they are given a value of 0. In cases where edge effects are found to extend beyond the first or last 40 residues of the sequence, the corresponding residues are also flattened. The mean and standard

deviation of the smoothed N- and C-termini alignment counts were therefore calculated for all residues, *excluding* those considered to be part of an edge effect.

All peaks in the profile that were found to have a Z-score greater or equal to 1.5 (5.3.1.1) were assigned as putative domain boundaries. Domain boundaries were assigned to the highest peak first. In cases where a domain boundary was assigned, residues within ± 30 amino acids of the central residue of the profile boundary peak were assigned a Z-score of 0. This was carried out in order to avoid domain boundaries being assigned within 30 residues of one another, since 30 residues is the smallest domain size permitted in the CATH database (Orengo *et al.*, 1997).

5.2.3 A study of DPS parameters

The different parameters used in the DPS algorithm were varied in order to investigate the effect on domain prediction. This was carried out by varying the values of three major parameters: The PSI-BLAST E-value threshold, with which a matched sequence was deemed a significant match, the length of the smoothing window, and the Z-score cut-off with which to assign domain boundaries from peaks in the termini-profile. Two main issues were addressed: The first is to reliably distinguish between single-domain and multi-domain protein chains and the second is that prediction of multi-domain chains must be accompanied with reliable domain boundary assignments. The percentage of correctly and incorrectly assigned single and multi-domain predictions was calculated for different values of each of these parameters. Also the assignment of domain boundaries by DPS was considered by again varying each of these values. Reasonable values were considered to be those that gave the highest number of correct assignments whilst maintaining the lowest number of incorrect assignments for both domain content *and* boundary prediction.

5.2.4 Benchmarking domain prediction by DPS

DPS was used to predict the domain content and corresponding domain boundaries of the single and continuous multi-domain chains in the non-redundant test set (section 5.2.1). In cases where a boundary was predicted, the chain was predicted as multi-domain, whilst in cases where no boundaries were found, the chain was predicted as single-domain. Predicted boundaries were compared to

boundaries assigned by CATH, and those within ±20 residues of a CATH boundary were recorded as correct predictions.

5.2.5 Domain prediction using both DPS and DomSSEA

As outlined in Chapter 4, DomSSEA was designed to predict protein domains in cases where domain content could not be inferred by sequence comparison techniques. Using DPS together with DomSSEA aimed to provide a predictive solution in cases where DPS was unable to assign domain boundaries to a target protein, due to a lack of available homologous sequences. In such cases a structural approach to domain assignment, as used by DomSSEA can often be useful

DomSSEA was used in two ways. First, in cases where DPS assigned a query sequence as single-domain, it was not definitive as to whether this was a correct prediction, or was a result of a lack of homologues to form a termini-profile. DomSSEA was therefore used to verify the single-domain prediction and if verified, the query sequence was assigned as a single domain chain. However, if DomSSEA did not verify this single-domain prediction i.e. it made a multi-domain prediction then it was this multi-domain prediction that was assigned to the query. In other words, DomSSEA was used to assign both domain number and corresponding domain boundaries. The second case in which DomSSEA was used was when DPS predicted a chain to be multi-domain. In such a situation, a further multi-domain prediction was made by DomSSEA, and the boundary predictions of both the DPS and DomSSEA predictions were combined by overlaying the predictions. In cases where predictions were within 30 residues, the DPS boundary prediction was used. This combination of boundary predictions was used to address the possibility that domain boundaries, undiscovered by sequence searching, might exist, and therefore be found by DomSSEA.

5.3 Results

5.3.1 Variation of DPS parameters

By varying these parameters, a default set could be chosen that gave a reasonable trade-off between domain number and domain boundary prediction. Choosing parameters can be rather subjective, as the selection made can depend

largely on what you want to use the method for. In this case a choice of parameters was made to assign domain boundaries as accurately as possible, whilst also retaining a useful domain number prediction accuracy.

5.3.1.1 Domain content prediction

The DPS algorithm was developed in order to assign domain boundaries from sequence comparison using PSI-BLAST local alignments. Three main parameters were varied, E-value, Z-score and the length of the smoothing window (section 5.2.3). The analysis was carried out using the data set of single domain and multi-domain chains (section 5.2.1). Whilst the study was mainly focused upon the assignment of continuous domains, discontinuous domain assignment has also been addressed.

Table 5.1 shows the results for the prediction of the domain content of the test sequences for Z-score cut-offs of 1.0, 1.5 and 2.0, together with E-value cut-offs of 1, 10^{-5} and 10^{-10}. All values were calculated with an initial fixed smoothing window length width of 15 residues. For the multi-domain chains, the percentage of multi-domain sequences correctly predicted to contain more than one domain is shown, i.e. the percentage of true-positives. The true-positive multi-domain prediction values are also shown for continuous multi-domain and discontinuous multi-domain chains. The number of single domain chains incorrectly predicted to be composed of more than one domain is shown as the percentage of false-positives. By increasing the Z-score and therefore the distance by which a peak in the termini-profile must deviate from the mean, the number of correctly assigned multi-domain chains decreases, but with a corresponding decrease in the false positive prediction of single-domain chains as multi-domain. Furthermore, it can be seen from Table 5.1, that by decreasing the E-value and therefore increasing the significance of a PSI-BLAST alignment hit permitted to be included in the termini-profile, appears to have a smaller effect on the true and false-positive rate of domain content prediction. However the general trend seems to be smaller E-values giving fewer false-positives with fewer true-positives. From these results, it would seem that a reasonable trade-off between the true and false-positive prediction rate of domain content prediction for this study is given by a Z-score of 1.5.

		% predicted as multi-domain			
		single domain	all multi-domain	continuous multi-domain	discontinuous multi-domain
Z-score	E-value	(FP's)	(TP's)	(TP's)	(TP's)
1	1	23.8	55.3	62.0	38.7
	10^{-5}	21.1	52.6	57.4	38.7
	10^{-10}	19.5	49.3	54.0	37.7
1.5	1	14.6	45.8	50.2	34.9
	10^{-5}	13.6	43.1	48.7	29.2
	10^{-10}	12.5	40.7	45.6	28.3
2	1	11.7	37.7	43.0	24.5
	10^{-5}	10.0	34.7	39.5	22.6
	10^{-10}	8.9	31.7	36.9	18.9

Table 5.1 Prediction of domain content by DPS varying Z–score and E-value cut-offs

Domain content was predicted using DPS with varying Z-score and E-value cut-offs. Predictions were made for 369 multi-domain chains (263 continuous domain chains, 106 containing one or more discontinuous domains). Results are shown as percentage correct multi-domain predictions (TP = true positives), and percentage incorrect multi-domain predictions (FP = false positives), i.e. single-domain chains predicted to be multi-domain.

Figure 5.1 considers the effect of a wider range of E-values from 10^{-10} to 1, with a fixed Z-score of 1.5 and a fixed smoothing window size of 15 residues. It can be seen, similar to the values in Table 5.1, that the percentage of correctly assigned multi-domain chains only increases slightly over this range, by 5%, with an increase of false-positive multi-domain assignments of 3%. Therefore, an E-value of 0.01 was chosen as a cut-off for use in this analysis, since it gives the highest multi-domain prediction accuracy, whilst retaining a false-positive assignment rate similar to that given by smaller E-values cut-offs.

Correspondingly, Figure 5.2 shows the multi-domain prediction accuracy for an E-value of 0.01, and Z-scores ranging from 1 to 5 with a fixed smoothing window size of 15 residues. Although increasing the Z-score decreases the number of false-positive multi-domain assignments, it also gives a rapid decrease in the number of true-positives. The use of a Z-score of 1.5 appears to give a reasonable trade-off in the prediction rate and reliability, giving as high a true-positive prediction rate, with as few single-domain chains assigned domain boundaries as possible.

Finally, in order to assess the effect that different window sizes would have on prediction accuracy window smoothing sizes between 7 and 19 (with an interval of 2 residues, as the window must be an odd number, to allow a centralised residue) were used (Figure 5.3). Figure 5.3 shows domain content prediction for the different smoothing window lengths with a fixed Z-score of 1.5 and E-value of 0.01. A length of 15 was chosen as it gives as few false-positive multi-domain predictions, without decreasing the number of correct assignments to too large an extent.

5.3.1.2 Domain boundary prediction

So far, the parameters have been optimised for domain content prediction. For a perfect domain assignment method, in all cases where a correct multi-domain assignment is made, the corresponding boundary predictions will be true domain boundaries (in this case as assigned structurally by CATH). However, as has been previously found, structural domain boundaries are not always equivalent to those found in sequence (Marchler-Bauer *et al.*, 2002). Table 5.2 shows the effect that different Z-score cut-offs have on the success of domain boundary assignment for multi-domain chains. Shown separately are the results for continuous and chains containing discontinuous domains. The sensitivity and selectivity (the calculation of

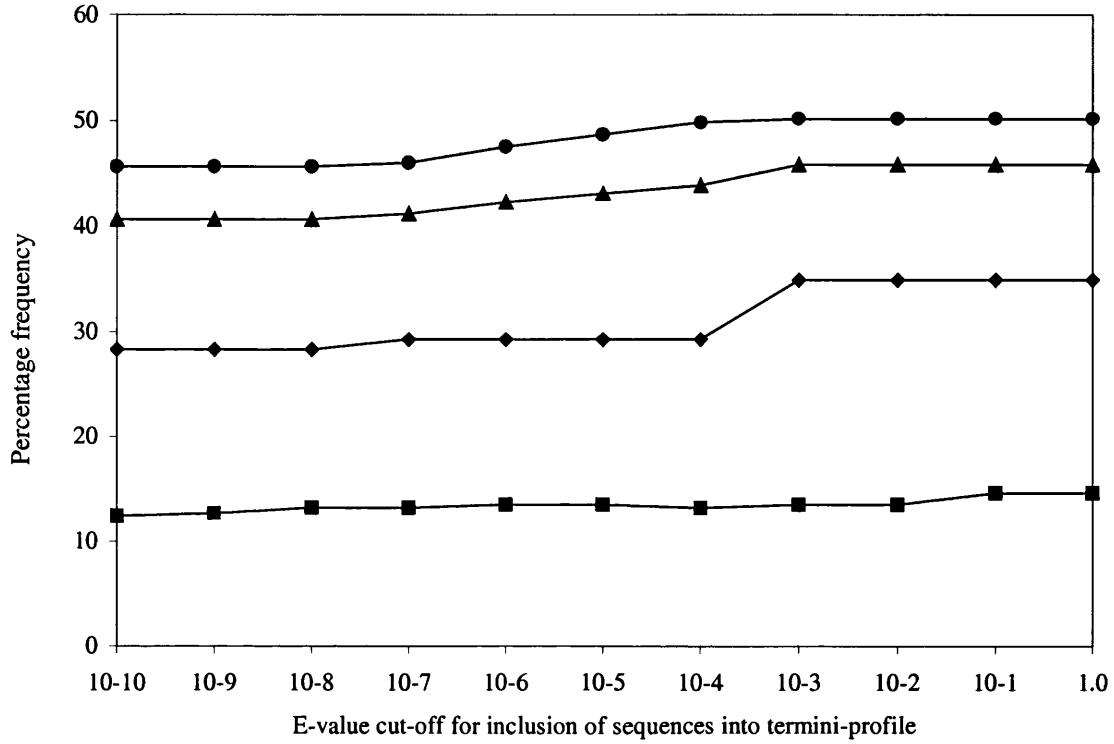


Figure 5.1 Prediction of domain content by DPS varying the E–value threshold

Domain content was predicted using DPS varying the E-value thresholds between 10^{-10} to 1.0, for inclusion of sequences into termini-profile distribution. A fixed Z-score of 1.5 and smoothing window length of 15 residues was used. Results are shown as the percentage of multi-domain predictions made, for both the multi-domain and single-domain chains where; multi-domain chains correctly predicted as multi domain (triangle), continuous multi-domain chains correctly predicted as multi-domain (circle), discontinuous multi-domain chains correctly predicted as multi-domain (diamond), single domain *incorrectly* predicted as multi-domain (square).

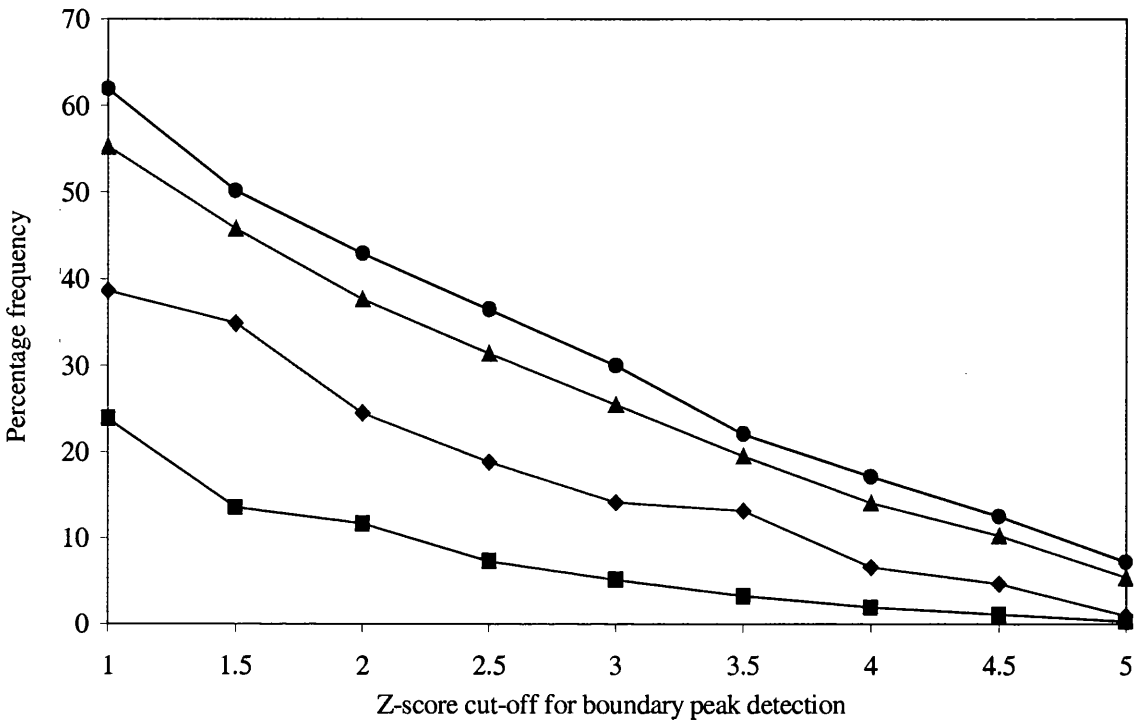


Figure 5.2 Prediction of domain content by DPS varying Z–score cut-off

Domain content was predicted using DPS varying the Z-score between 1 and 5, for assignment of domain boundaries form termini-profile peaks. A fixed E-value of 0.01 and smoothing length of 15 residues was used. Results are shown as the percentage of multi-domain predictions made, for both the multi-domain and single-domain chains where; multi-domain chains correctly predicted as multi domain (triangle), continuous multi-domain chains correctly predicted as multi-domain (circle), discontinuous multi-domain chains correctly predicted as multi-domain (diamond), single domain *incorrectly* predicted as multi-domain (square).

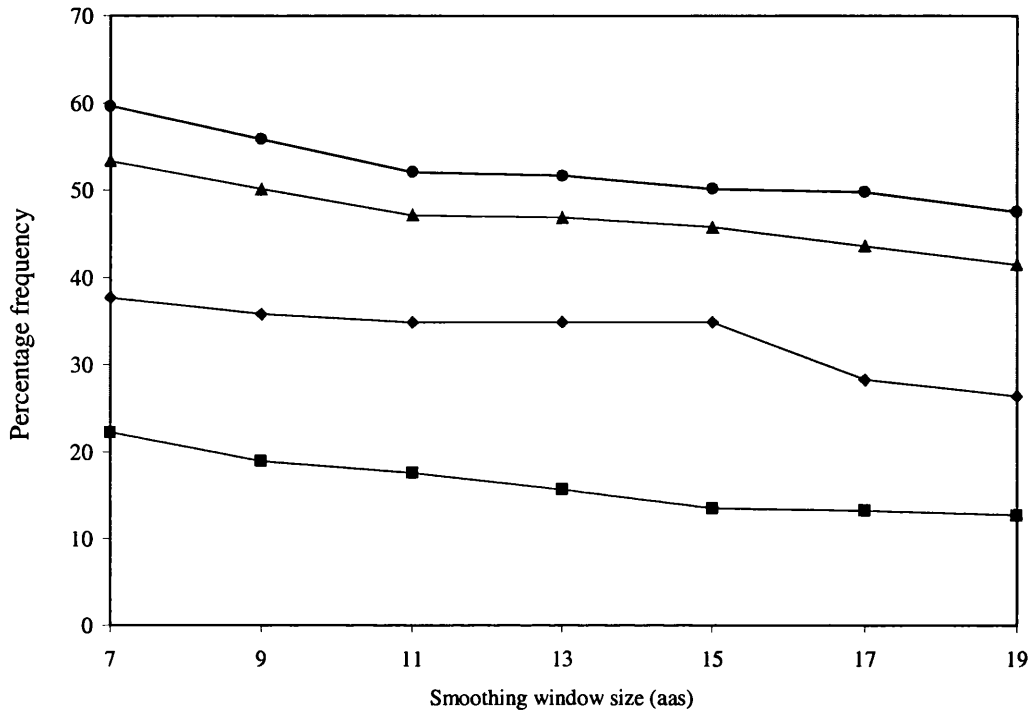


Figure 5.3 Prediction of domain content by DPS varying the smoothing window length

Domain content was predicted using DPS varying the smoothing window length between 9 and 19 residues, for assignment of domain boundaries form termini-profile peaks. A fixed E-value of 0.01 and Z-score of 1.5 was used. Results are shown as the percentage of multi-domain predictions made, for both the multi-domain and single-domain chains where, multi-domain chains correctly predicted as multi domain (triangle), continuous multi-domain chains correctly predicted as multi-domain (circle), discontinuous multi-domain chains correctly predicted as multi-domain (diamond), single domain *incorrectly* predicted as multi-domain (square).

	multi		continuous		discontinuous	
Z-score	**sensitivity**	selectivity	**sensitivity**	selectivity	**sensitivity**	selectivity
1.0	**34.4**	51.6	**39.3**	53.6	**12.6**	44.1
1.5	**25.4**	54.0	**29.7**	56.0	**8.4**	45.5
2.0	**21.7**	59.9	**25.0**	59.5	**7.5**	62.1
2.5	**18.9**	68.8	**21.4**	66.7	**7.1**	81.0
3.0	**14.9**	72.8	**16.8**	69.3	**5.9**	93.3
3.5	**11.7**	78.7	**12.4**	73.8	**5.9**	100.0
4.0	**8.5**	81.1	**9.9**	78.3	**2.9**	100.0
4.5	**7.0**	92.1	**8.0**	87.9	**2.5**	100.0
5.0	**3.8**	95.0	**4.9**	94.7	**0.4**	100.0

Table 5.2 Domain boundary predicition by DPS with varying Z-score

Domain boundaries were predicted using DPS with varying Z-score cut-offs. Predictions were made for 369 multi-domain chains (263 continuous domain chains, 106 containing one or more discontinuous domains). Results are shown as sensitivity and selectivity values where sensitivity is defined as the number of correct boundary predictions divided by the number of boundaries to predict. The selectivity is defined as the number of correct boundary predictions made, divided by the total number of boundary predictions made.

these values is described in section 4.2.9.2) of boundary assignments are shown, for Z-scores of 1.0 to 5, with a fixed E-value cut-off of 0.01, and smoothing window of 15 residues. It can be seen that by increasing the Z-score cut-off, sensitivity is decreased, with an according increase in selectivity. Using a Z-score of 1.0 looks reasonable for assignment of domain boundary since predictions using this value give nearly 9% more correct boundary assignments than when a Z-score of 1.5 is used, with only a slightly smaller selectivity of 51.6% vs. 54%. However, as discussed in section 5.3.1.1, a Z-score of 1.5 gives a fair trade off between false-positive and true positive domain content prediction and as such, this value should be retained.

5.3.1.3 The discontinuous domain problem

The prediction of discontinuous domains has also been addressed and results are shown in Tables 5.1 and 5.2 and Figures 5.1 to 5.3. It can be seen that there is a common trend in all these optimisation results regarding the assignment of discontinuous domains. That is - both assignment of domain content, and domain boundaries is lower for chains containing discontinuous domains as compared to those consisting of only continuous domains. These discontinuous domain prediction values represent the prediction rate that can be obtained from PSI-BLAST local sequence alignments, and illustrates the overall complexity in assigning such domains. This is examined to a greater extent in section 5.3.5.

5.3.1.4 The consequence of flattening "edge effects"

As described in section 5.2.2, peaks in the termini-profile can occur at the query N- and C-terminus as a result of a large number of homologous sequence matches of similar length to the query. It is possible that such peaks or "edge effects" in the profile may be assigned as false positive domain boundaries. The DPS algorithm was therefore designed to flatten such edge effect regions as described in section 5.2.2. Although this flattening procedure may result in real domain boundaries within the flattened region being missed, such boundary profile peaks may be obscured by the edge effects anyway. To assess the effect this flattening may have on prediction of domain boundaries the cumulative frequency distribution of

continuous and discontinuous domain boundary cuts within a given distance of the N- or C- termini of their sequences is shown in Table 5.3. For the multi-domain chains used in this study it can be seen that continuous domains boundaries are only found above a distance of 50 residues from the N- or C-sequence termini. In contrast, discontinuous domain boundaries are found in relatively high numbers close to the N- or C- termini. For example, nearly 10% are found within the first or last 60 residues as compared to just over 3% of continuous domains. This flattening procedure (where 40 residues at the both ends are the profile are given a value of 0) therefore has more of an effect upon discontinuous boundary prediction than continuous domain boundary prediction.

5.3.2 Domain content prediction by DPS using the default parameters

Now that a choice of default parameters for the Z-score, E-value and smoothing window length had been made, domain predictions for the 263 single and 263 continuous multi-domain protein chains from the test set were carried out using DPS (section 5.2.4). The results of these predictions are summarised in Table 5.4. In this table it can be seen that 50% of the CATH multi-domain chains (131) were predicted to contain more than one domain. This prediction rate was achieved with a high selectivity of nearly 80%. The incorrect assignment of many multi-domain chains as single domain is attributable to the insufficient number of sequences within nrdb90 that enabled domain boundary delineation. This method is limited in cases where homologous sequence segments of query sequence domains are found in an insufficient quantity and/or context within the alignment database to calculate a termini-profile peak of enough significance to allow domain boundary assignment. It is also possible that homologous domain sequences may have existed in the sequence database, but the same domain arrangement as the query sequence. It can also be seen from Table 5.4 that the success rate in correct assignment of single-domain chains is 87%. As shown the selectivity of single-domain prediction is somewhat lower than the high sensitivity value. This is mainly due to those multi-domain target chains, shown to have no domain boundaries in the termini-profile, and therefore predicted to be single-domain. Finally, the overall sensitivity value for domain content prediction by DPS is 68%.

Distance from N- or C- termini (residues)	Cumulative frequence of CATH domain boundaries	
	continuous	discontinuous
0-10	0.0	0.7
0-20	0.0	2.7
0-30	0.0	4.3
0-40	0.0	6.1
0-50	0.0	6.8
0-60	3.3	9.5
0-70	6.9	11.3
0-80	11.0	12.9
0-90	17.9	17.5
0-100	25.6	20.0
0-110	30.5	23.8
0-120	34.9	26.3

Table 5.3 Cumulative frequencies of the distance of the N- or C-termini of the CATH domain boundaries within the non-redundant test data set

The cumulative number of boundaries of continuous and discontinuous domains of the multi-domain test set that fell within a given distance of their sequence N- or C-termini are shown. If a boundary was found within the distance cut-off from both the N- and C- termini, it was counted only once.

	Domain content prediction								
	All	Single				Multi			
	Sensitivity (%)	TP	FP	**Sensitivity (%)**	Selectivity (%)	TP	FP	**Sensitivity (%)**	Selectivity (%)
DPS	68	228	132	87	63	131	35	50	79
DPS and DomSSEA	73	181	63	68	74	201	81	76	71

Table 5.4 Domain content prediction by DPS and DPS together with DomSSEA

Domain content was predicted using DPS and DPS together with DomSSEA. Results are shown for all chains, single-domain chains, and multi-domain chains (*containing only continuous domains*). Sensitivity and selectivity are shown for the single and multi-domain chain predictions. Sensitivity values alone are shown for all chains since in this case sensitivity is equivalent to selectivity. The number the number of true and false positive predictions from which these were calculated are also shown where; TP denotes true positives e.g. single-domain chain predicted to be single-domain, FP denotes false positives e.g. single-domain chains predicted to be multi-domain. Sensitivity is defined as TP/N, where N = the number of single, or multi-domain chains to predict, in this case N=263. Selectivity is defined as TP/(TP+FP)

5.3.3 False positive multi-domain content predictions by DPS

Next it was important to consider those single-domain chains incorrectly assigned as multi-domain by DPS. Figure 5.4 shows the distribution of chain lengths for the 263 single-domain chains used in assessing domain prediction in this study, The average single domain length is 148 residues. Also shown as points in Figure 5.4 are the chain lengths of the 35 single-domain chains incorrectly assigned as multi-domain by DPS. It can be seen that all of them have lengths greater than the average domain length, and a majority of them have lengths which are unlikely for single-domain chains. Thus, the rate of false positive multi-domain predictions increase with chain length - perhaps not surprising as larger chains tend to be multi-domain.

To check that these chains were not in fact multi-domain chains miss-assigned as single-domain by CATH, the domain assignments given by SCOP were checked. Of the 35 chains, 33 were also assigned as single-domain by SCOP. Only two were assigned as multi-domain. In both cases, DPS correctly assigned the SCOP domain boundaries, (1nub chain A predicted cut at 143, SCOP cut at 151, and 1ank chain A, predicted by SCOP to contain a discontinuous domain, DPS correctly assigned the N-termini linker of the continuous domain inserted within the discontinuous domain, DPS predicted cut at 123, SCOP cut at 135).

5.3.4 Domain content prediction by DPS together with DomSSEA

For protein sequences for which no reliable domain assignment can be made by sequence comparison based methods, a structural approach is required. For this reason DPS was used together with DomSSEA for such sequences (section 5.2.5). Table 5.4 shows domain prediction results using this combined sequence and structural approach when applied to the 526 single and continuous multi-domain sequences in the test set. This combined approach and the results obtained are also outlined in Figure 5.5.

In cases where DPS assigns a query chain as putative single-domain (i.e. finds no domain boundary), the domain content prediction is taken from the top scoring assignment made by DomSSEA. For chains predicted to be multi-domain by DPS, DomSSEA was used to make a further multi-domain prediction, and the results combined for boundary assignment as described in section 5.2.5. In Table 5.4 it can

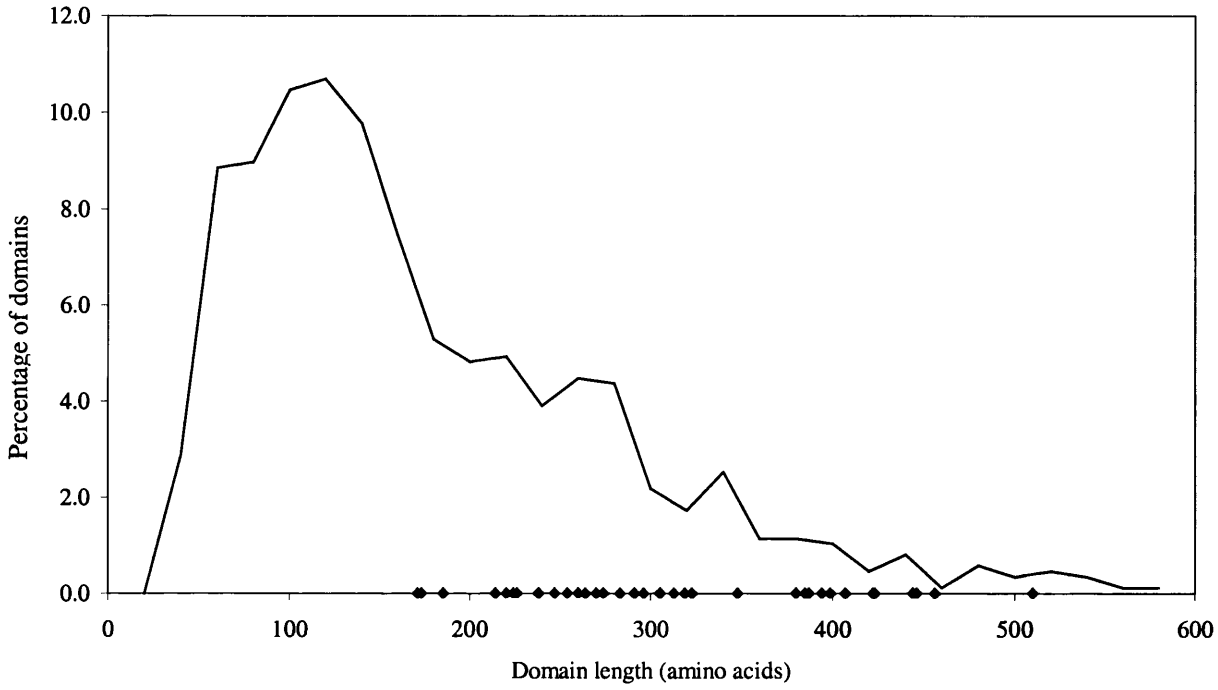


Figure 5.4 Chain length distribution of false-positive multi-domain predictions made by DPS

The lengths of those single-domains predicted as multi-domain by DPS are shown against the distribution of chain lengths for the single-domain chains in the test set. Each dot on the X-axis represents a single domain chain (35 in all) assigned a domain boundary by DPS and therefore considered a false positive multi-domain prediction.

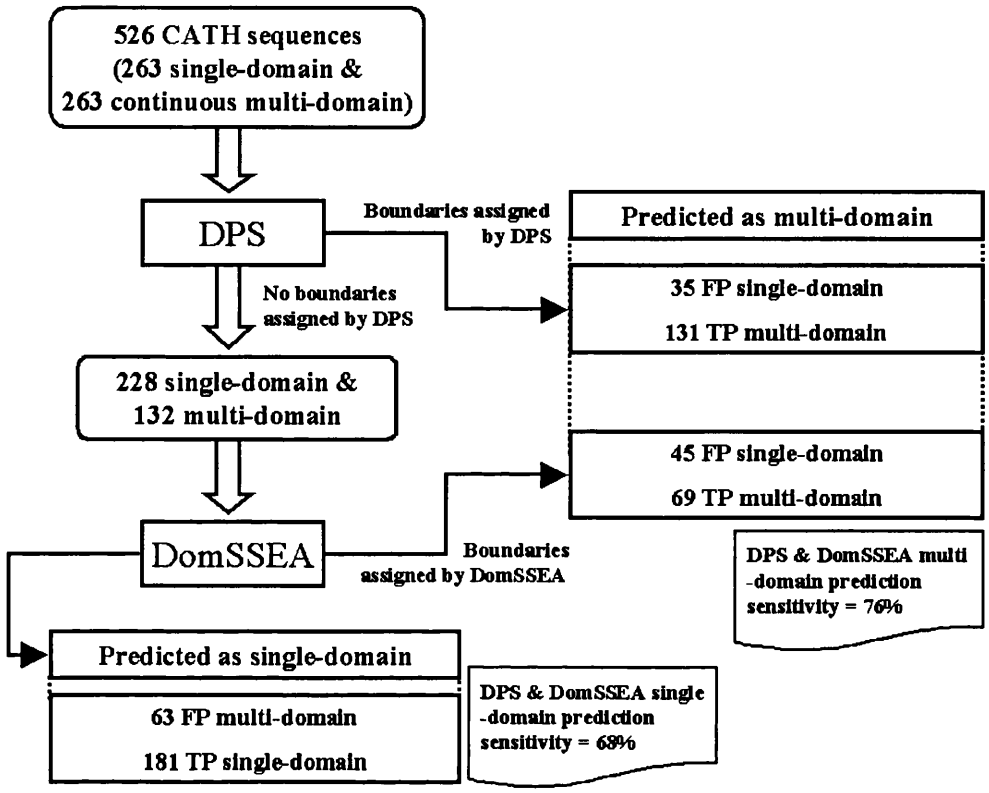


Figure 5.5 Combined approach to domain assignment using DPS and DomSSEA

The flow-chart shows domain prediction (and results) using the combined approach when applied to the 526 single and continuous multi-domain sequences in the test set. The results obtained are shown in Table 5.4. TP denotes true positive, FP denotes false positive.

be seen that the selectivity of multi-domain prediction by DPS is nearly 80%. It is therefore a reasonable assumption that in most cases that DPS assigns a boundary, the given sequence can be said to be multi-domain. However in cases where no domain boundary is assigned, it is still unclear whether the sequence is single or multi-domain and therefore a prediction for the sequence is made by DomSSEA.

The combined approach to domain assignment increases the sensitivity of multi-domain prediction of the 263 multi-domain chains by 26% from 50% to 76%, though selectivity decreases from 79% to 71% (Table 5.4). The sensitivity of single-domain prediction however can be seen to drop from 87% to 72%. The sensitivity decreases since some single-domain chains (correctly assigned by DPS) are incorrectly reassigned as multi-domain by DomSSEA. However, the selectivity of single-domain assignment goes up since some chains assigned as single-domain by DPS are correctly reassigned as multi-domain by DomSSEA. Overall the sensitivity of domain number prediction increases from 68% using DPS, to 73% using the combined methods.

5.3.5 Domain boundary assignment by DPS and DPS together with DomSSEA

Table 5.5 shows the sensitivity and selectivity of domain boundary prediction by DPS and DPS combined with DomSSEA. As shown, 30% of the multi-domain linkers are correctly assigned within ±20 residues of the corresponding CATH cut by DPS. The combined methods show an increase in correct boundary prediction by 25% to 55% of the domain linkers being located. Although with this increase in sensitivity there is also a drop in selectivity (from 56% to 45%), this is to be expected with more predictions made, in cases where DPS and DomSSEA boundary predictions were combined.

5.3.6 Discontinuous domain assignment and comparison of DPS to other methods

The use of a non-redundant set of 369 multi-domain chains containing the 263 continuous domain chains, and an additional 106 discontinuous multi-domain chains enabled an assessment of the use of DPS in the prediction of target chains containing discontinuous domains. Furthermore, it enabled a comparison of domain

	Continuous domain boundary prediction (±20)			
	TP	FP	Sensitivity(%)	Selectivity (%)
DPS	108	85	30	56
DPS and DomSSEA	148	161	55	45

Table 5.5 Domain boundary prediction by DPS and DPS together with DomSSEA

Domain boundaries were predicted using DPS and DPS together with DomSSEA. Results are shown for multi-domain chains containing only continuous domains as sensitivity and selectivity values. The number of true and false positive predictions from which these were calculated are also shown where; TP denotes true positives e.g. correct domain boundary prediction, FP denotes false positives e.g. incorrect domain boundary prediction.

prediction by DPS with predictions made by the methods PASS (Kurodat, *et al.*, 2000) and DOMAINATION (George and Heringa, 2002b). Of the 369 multi-domains, 46% were predicted to have more than one domain by DPS, with an associated selectivity of 77% (calculated by making additional domain number predictions for 369 'decoy' single domain chains). This compares to the study by George *et al.*, (2002b) in which their method DOMAINATION correctly predicts 56% of a similar set of 452 multi-domain chains as containing more than one domains. It is difficult to directly compare these values as no selectivity value for domain content prediction is given in the study by George and Heringa, (2002b). Whilst DPS makes no attempt to specifically tackle the problem of delineating discontinuous domain boundaries, DOMAINATION attempts to identify and join discontinuous fragment sequences in order to identify their boundaries. No single-domain chains are included in their test set so the false positive rate of multi-domain prediction was not addressed.

The overall boundary prediction by DPS for sequences in the multi-domain set (including continuous and discontinuous domains) gave a sensitivity of 26% with an associated selectivity of 54%. The sensitivity of boundary assignment by DOMAINATION (George *et al.* 2002b), was show to be just over 23% with an associated selectivity of 42%. This value was obtained from the first iteration of DOMAINATION. Subsequent iterations were showed some increase in the sensitivity of boundary prediction, but with a resulting loss in selectivity.

The method by Kuroda *et al.*, (2000) PASS (Prediction of Autonomous folding units based on Sequence Similarities) was used to predict the domain boundaries for a set of 52 specially selected SCOP multi-domain chains, shown to have clear globular structures by visual inspection. Here, the sensitivity of linker prediction was found to be just over 14%, with a selectivity of over 52%, showing whilst being less sensitive than DPS it is reasonably selective.

5.4 Discussion

The aim of this chapter has been to develop a simple sequence based approach to protein domain assignment, using PSI-BLAST local sequence alignments. When attempting to assign domains to query sequences, it is important to establish if domain delineation can be achieved from sequence comparison. This will

allow relatively obvious hits to be discovered, before the use of methods that may give a less accurate prediction. The observation of domain rearrangement or shuffling within protein sequences has enabled a domain boundary identification method to be implemented by post-processing PSI-BLAST sequences alignments made to a large non-redundant sequence database. The distribution of aligned sequence termini to the query sequence and subsequent analysis is used to delineate putative domain boundaries.

It has been shown that the use of DPS in assigning domains to protein sequence provides a reliable approach to domain boundary annotation. In this study, DPS has been shown to correctly assign the domain content of 68% of the single domain and multi-domain test set (containing only continuous domains) used in this study, including 50% of the multi-domain chains with a selectivity of 79%. As such DPS has also been shown to act as a useful pre-filter for sequences for which domain predictions can be made from multiple sequence alignments, before a structural approach is required. A valuable extension to this analysis would be to address how the performance of the DPS corresponds to the number of domains to predict. For example, Table 5.1 in this thesis shows the success rate of domain content prediction by DomSSEA for chains containing one, two and three-or-more domains. This in turn would facilitate a more complete comparison between DPS and DomSSEA.

The main analysis of DPS and DPS combined with DomSSEA has focused on the reliable assignment of continuous domain boundaries. The use of DPS to assign boundaries to discontinuous domains has been considered, however, the reliability of such assignments are below continuous boundary assignments. Reliable assignment of continuous domains is of great importance, especially for structure prediction by homology modelling or fold recognition methods, where prediction methods are constrained to continuous domain structures.

In most cases where DPS did not detect a CATH domain boundary, it was because of insufficient homologous sequences to generate a boundary profile peak. The prediction of boundary regions that are not assigned at a structural level by CATH occurred for several chains correctly predicted to be multi-domain, i.e. were assigned a domain boundary. This was also seen for many of the false-positive multi-domain assignments. Such sequence defined regions may represent sequence repeats or domains found within sequence databases, that are less or not at all apparent when analysing 3D protein structures. Similar observations were made in the study by

Kuroda *et al.* (2000) which they define as Autonomous Folding Unit's corresponding to 'boundaries' in sequence that do not correspond to boundaries assigned in structure. Furthermore, the analysis by Marchler-Bauer *et al.*, (2002), found that though structural domain databases compare well with carefully curated sequence domain databases, such as PfamA, any discrepancies were often due to domain regions identified by sequence comparison being shorter than those identified from structure comparison.

DPS was also compared to the results given in the study by George and Heringa (2002b) for their method DOMAINATION. Both methods have been tested on a similar data set for boundary prediction. DPS showed a slightly higher boundary prediction accuracy of 26% compared to 23% by DOMAINATION, with a selectivity of 54% compared to 42%.

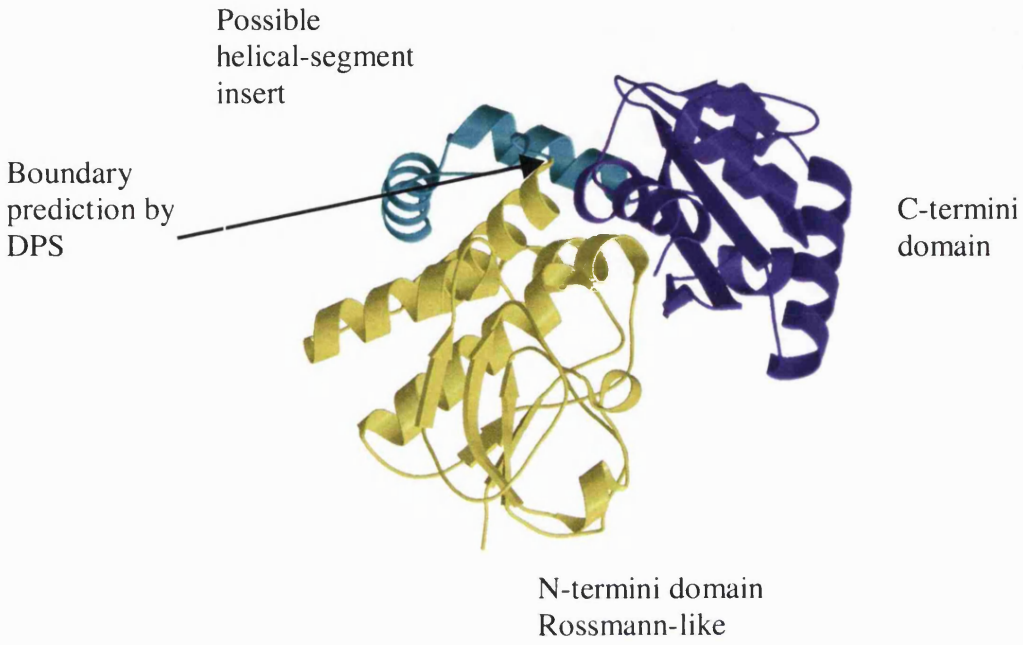
The lower level of discontinuous domain (compared to continuous domain assignment) assignment accuracy was not really surprising. It is a well known that assignment of discontinuous domain boundaries from sequence comparison is not a trivial issue. Discontinuous domain boundaries are most likely to be found by DPS if bordered by a continuous domain, whose sequence may then be repeated within the sequence database.

By combining of DomSSEA with DPS, predictions can be made for domains that cannot be delineated by sequence comparison. This combination provides an approach by which 'obvious' sequence matches can be filtered out, leaving tougher assignment cases to be predicted by DomSSEA. As shown the combination of methods gave a measurable increase in domain assignment with 73% of the test set correctly predicted to be single or multi-domain, with 76% of the multi-domain chains correctly predicted as such. Furthermore, by combining DPS and DomSSEA predictions, the sensitivity of domain linker prediction is increased to over 55% (± 20 residues from the CATH cut). The domain prediction accuracy by DomSSEA alone can be calculated, and compared to the previous study described in Chapter 3 of this thesis – it compares well. Over 50% of the multi-domain chains passed on to DomSSEA were correctly predicted as such, with nearly 79% of the single domain chains given correct assignments.

The inherent difficulties in domain assignments from structure and sequence can be illustrated by the domain predictions for the multi-domain CASP4 target (T0087 in the fold recognition and new fold category) by DPS and DomSSEA.

Figure 5.6 Domain assignment for CASP4 target T0087

Difficulties in domain assignments from structure, and sequence for the multi-domain CASP 4 target (T0087 fold recognition and new fold category. From analysis of the structure, this protein can be split into two domains, where the N-terminal domain (yellow) has similarities to a Rossmann fold and the C-terminal domain (purple) is a novel alpha-beta fold. However it may be considered that domain 1 also contains a helical subdomain (coloured cyan) (Koretke *et al.*, 2001, Lesk *et al.*, 2001) appended to its C-terminus.


Possible
helical-segment
insert
Boundary
prediction by
DPS
C-termini
domain
N-termini domain
Rossmann-like

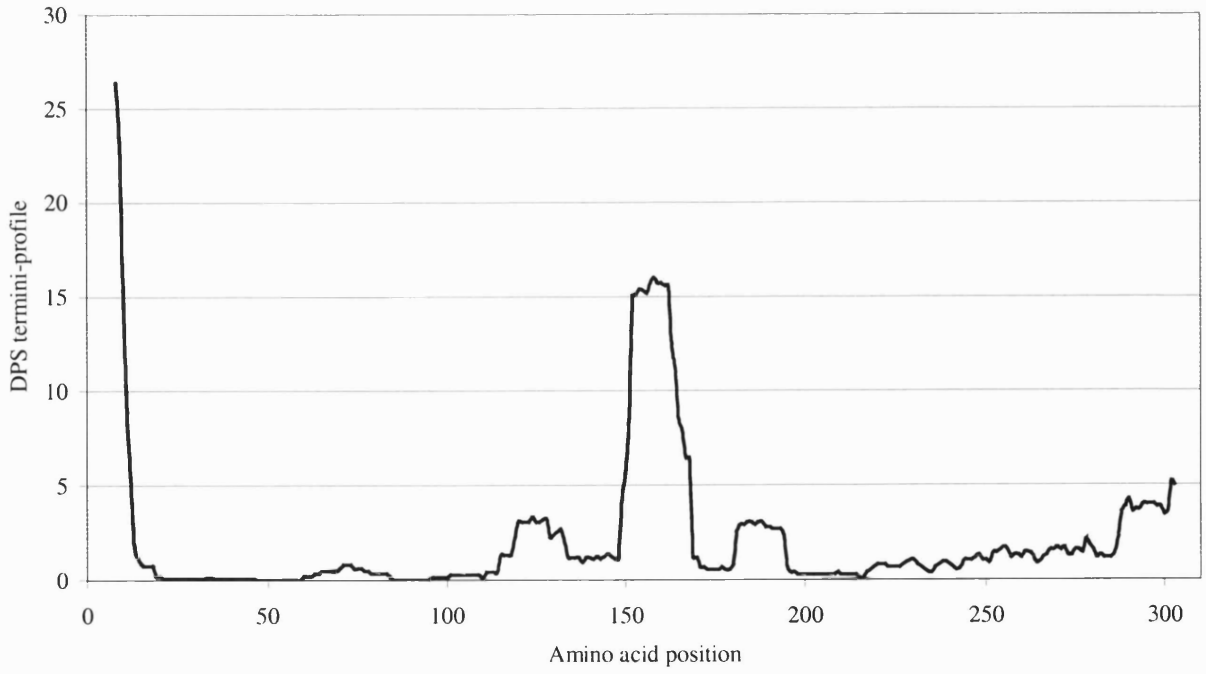

DPS termini-profile
30
25
20
15
10
5
0
0
50
100
150
200
250
300
Amino acid position

Figure 5.6 shows the structure of the CASP 4 target pyrophosphatase from *Streptococcus mutans*, sequence length 310 residues (Merckel *et al.*, 2001). Domain assignment from 3D structure revealed that this protein can be split into two domains, where the N-terminal domain has similarities to a Rossmann fold and the C-terminal is a novel alpha-beta fold. However it may be considered that domain one also contains a helical subdomain (coloured cyan in Figure 5.6) (Koretke *et al.*, 2001; Lesk *et al.*, 2001) appended to its C-terminus.

Figure 5.6 also shows the associated termini-profile calculated by DPS for target T0087. The X-axis shows the residue position, whilst the y-axis represents the smoothed termini frequency profile calculated from the frequency of N- and C-termini aligned database sequences. A clear and significant peak can be seen to be centred around residue position 170. This peak has a Z-score of 5.4 and is therefore well above the cut-off of 1.5 required for a boundary assignment. Two further peaks can be seen towards the N- and C-termini of the major peak, both of which have Z-scores below 1, and therefore are not significant enough to be assigned as domain boundaries. The domain prediction results given by DomSSEA also predict the sequence to be two-domain, with a boundary at residue position 160. However as mentioned, even using 3D structural data for this protein, there is some degree of difficulty in assigning domain boundaries. Although a boundary is given at residue position 190 (Merckel *et al.*, 2001) there is also a possible domain insertion from residue 155 to 190 containing helical elements. DPS predicts the N-terminal domain boundary at 150, which does not take into account this possible insertion domain, and furthermore, DomSSEA assigns the first domain as a Rossman fold from 2-155, correct as shown by the published data. This example shows the difficulties in domain assignment (and benchmarking prediction methods). Domain assignment, even with structural data can be a subjective process. Although the sequence method gives a clear signal, this is not close enough to be considered a correct assignment in the benchmarking procedure outlined in this study.

Chapter 6

A study of protein domain swapping

6.1 Introduction

The evolution of oligomeric proteins has enabled a number of biologically advantageous properties to be conferred upon these complexes compared to their monomeric counterparts. Such advantages include new possibilities in allosteric control, higher local concentrations of active site residues, the formation of larger binding surfaces or new active sites between the monomers, and is an economic way to evolve protein interaction networks and molecular machines from the monomer subunits (Liu and Eisenberg, 2002). The formation of protein oligomers is thought to have come about through the accumulation of random mutations in the interfaces between the monomers involved in oligomeric complexes, allowing a stable association to form (Bennett *et al.*, 1995). The interactions formed at the interface site must be favourable enough to overcome the loss in entropy that results from the oligomerisation event (Bennett *et al.*, 1995). Protein domain swapping has been proposed as a alternative evolutionary mechanism enabling protein oligomers to form from identical protein subunits (Heringa and Taylor, 1997). Protein domain swapping was first described by Crestfield *et al.*, (1962) where the swapping of the N-terminal peptide of bovine pancreatic ribonuclease was proposed. In 1990 Bax *et al.*, reported domain swapping in the X-ray structure of beta B2-crystallin, suggesting that the domain-swapped interface may have been conserved through evolution. The first use of the term 'domain swapping' was used by Bennett *et al.*, (1994), however several structures of domain-swapped proteins had been solved before this time.

An overview of terms used to describe domain swapping is shown in Figure 6.1. Domain swapping can be considered as a mechanism in which a dimer (or higher homo-oligomer) uses an intramolecular domain interface (occurring in the monomer) by exchanging one of the domains (or part of a domain) of each monomer such that each domain packs on the other monomer using the old *intra*molecular interactions as *inter*molecular interactions (Taylor and Heringa, 1997). The swapping of domains involves breaking the non-covalent bonds between the primary interfaces within each original monomer, the movement of the swapped-domains, and the reconstruction of the complete original interface between domains, now from either chain in the oligomer, where the non-covalent interactions are reformed (Taylor and Heringa, 1997).

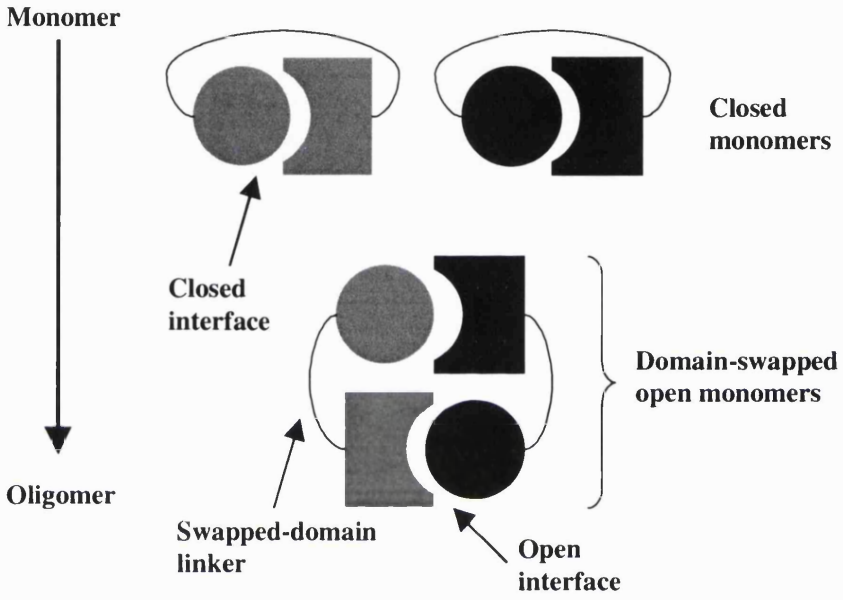


Figure 6.1 An overview of domain swapping

Schematic diagram illustrating domain swapping definitions adapted from a figure given by Bennett *et al.*, (1995) and Liu and Eisenberg (2002). The swapped domains, represented by circles, can be entire globular domain or a single helix or strand that extends into the main domain, represented by squares of a neighbouring subunit. The interaction formed in the oligomer is identical to that formed in the closed monomer. The swapped-domain linker joins the main domain and swapped domain. The closed interface is found between the main domain and swapped domain in the monomer, whilst the open interface is found in the domain swapped oligomer.

Domain-swapped proteins whose tertiary structure has been solved can be classified in three ways; first, those domain-swapped proteins that have also have a known tertiary structure in the monomeric, non-domain-swapped form are known as *bona fide* domain-swapped structures. Secondly, those cases where only the structure of a monomeric non-swapped *homologue* has been solved are described as *quasi* domain-swapped proteins. Finally, cases of domain-swapped structures for which no monomer non-swapped form of the protein has been found are termed domain swapping candidates (Liu and Eisenberg, 2002).

Domain swapping nearly always occurs at the N- or C-terminus. However occasionally this is not the case; for example, in the blood coagulant factor IX/X-bp (Mizuno *et al.*, 1997) a loop found in the central region of the protein is swapped. The term 'domain' swapping can be misleading since swapped-domain structures often do not comply with a generally accepted definition of a structural domain – a compact, local, semi-independent unit of structure (Richardson *et al.*, 1981). Although some swapped structures are domains, for example Cyanovirin-N (Yang *et al.*, 1999), where 48 residues of the 101 residue protein are swapped, they may also contain only a single secondary structural element, such as the cell cycle regulation protein, suc1, where only a single strand is exchanged in the homo-dimeric oligomer (Khazanovich *et al.*, 1996).

Though domain swapping may act as a rapid evolutionary route to protein oligomer formation it may also have less advantageous effects within the cell. It has been suggested that domain swapping may provide a possible mechanism for protein aggregation or fibre formation, for example the formation of amyloid fibril deposits (Newcomer, 1997).

In this chapter, a general analysis of domain-swapped proteins is made, focusing on the swapped-domain linker region that joins the swapped-domain to the main domain structure. However, before this analysis is undertaken, a search for domain-swapped proteins is made in order to extend the data set of structures that have been listed in the paper by Liu and Eisenberg (2002), and those that can be obtained through literature searches. A domain swapping search algorithm is developed and implemented to identify potential domain-swapped structures, by searching for proteins in the Protein Data Bank (PDB; Bernstein *et al.*, 1977) that contain extended, non-globular segments of polypeptide at the N- or C-termini of

their structure. Such exposed elements would not normally be tolerated in general globular proteins, however they may represent structures that are swapped between subunits, and are buried in the oligomeric complex. The findings of the swapped-domain linker analysis are compared to the results of the analysis of domain-linking peptides (Chapter 2) in order to assess if any similarities exist between them.

6.2 Methods

6.2.1 Data set

A set of domain-swapped structures, and if possible, their closed monomer counterparts were obtained, using the structures listed in the review by Liu and Eisenberg (2002) as a starting point. Further domain-swapped structures were found by literature search and keyword searches (using 'domain' and 'swapped', 'swapping', 'swap') in the PDB and the CATH (Orengo *et al.*, 1997) and SCOP (Murzin *et al.*, 1995) structural classification databases. The resultant list consisted of 39 domain-swapped structures, of which 14 were *bona fide* domain-swapped proteins, with their closed monomer conformations available in the PDB, 11 *quasi* domain-swapped, with a corresponding homologue in the closed monomer conformation available in the PDB, and 14 putative domain-swapped structures with no known or solved structure (or homologue) in the closed monomer state within the PDB.

6.2.2 General search for domain-swapped oligomers

Assembling a list of domain-swapped proteins by literature search alone is both time consuming, and is also dependent upon domain-swapped proteins being described as such. The search for domain-swapped structures focused on identifying non-globular N- or C-terminal regions of polypeptide chain that were extended and exposed and as such may be in a swapped state. The aim was to design a method that measured the percentage of exposed residues found within a given cut-off distance of the N- or C-terminus of a given protein structure. Termini-regions found with a high percentage of exposed residues were considered as putative swapped-domain structures. The degree of a given residues exposure to solvent was measured by calculating the number of neighbouring amino acid C-alpha contacts inside a sphere

of radius 10 Å. A cut-off of 10 Å was used, as this was found to give the best correlation between contact number and relative solvent accessibility in the study by Farisellii and Casadio (2000). The use of contact number to describe a residues exposure to solvent, rather than solvent accessibility as assigned by, for example, DSSP, meant the search algorithm could be run solely on PDB files, giving speed and simplicity. A residue contact number of 12 or less was used to define residues exposed to solvent (section 6.3.1). The structure of a given protein chain was considered in isolation when calculating contact number for each residue in order to exclude counting residues from neighbouring chains that may be present in the PDB file.

To take into account the variation in swapped-domain lengths (section 6.3.5), different N- and C-terminal cut-off lengths were used to measure the number of exposed residues found at the termini of a given protein structure. The use of a single length cut-off could be problematic. For instance, using a 20 residue cut-off (i.e. the first or last 20 amino acids of a sequence) would not be suitable for shorter swapped-domains. Here, though residues closest to the structure termini may be assigned as exposed, those closer to the main domain are likely to be more buried due to there only being a small swapped element. The overall percentage of exposed residues in the 20 residue peptide may therefore be no different to a randomly chosen peptide. To address this problem, seven different length cut-offs, from the appropriate N- or C- structural terminus were used, from 8 to 20 residues, (using an interval of 2 residues). The percentage of exposed residues within each cut-off was measured, and an average of the 7 measurements calculated. This simple procedure takes into account the distribution of exposed residues. For smaller length exposed chain termini, the percentage of exposed residues in the short length cut-offs would be high, whilst the same measurements for the longer cut-offs might give lower values. However, the overall mean percentage calculated over all cut-offs should be high enough to identify this structure as putatively domain-swapped.

The search method was benchmarked (section 6.3.1.2) on the data set of domain-swapped structures described in section 6.2.1. A mean percentage of 80% exposed residues calculated over the seven distance cut-offs was used to identify a potential domain-swapped structure. As an additional attempt to identify short swapped-domains, if the mean value between the 8 and 10 residue cut-off lengths

was 90% or higher, whilst the overall mean for all lengths was 70% or greater, the corresponding region was additionally assigned as a potential hit.

6.2.3 Protein Quaternary Structure (PQS)

Domain-swapped oligomers are made up of two or more identical protein subunits. Therefore a search for domain-swapped structures in the PDB would ideally be focused on structures known to, or predicted to form homo-oligomeric complexes. The protein quaternary structure (PQS) database provides the co-ordinates for the likely quaternary states for structures found in the protein databank 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 download all protein structures described as forming homo-oligomeric complexes of all quaternary structure types, eg dimeric, trimeric, tetrameric etc. These structures were then searched for potential domain-swapped structures using the domain-swapped search method (section 6.2.2).

Additionally the PQS database was used to obtain co-ordinates representing the oligomeric state of domain-swapped proteins in the analysis data set in cases where the PDB file only contained a single chain copy of their open monomer, so enabling an analysis of the interaction site of the swapped-domain structures in the oligomeric state.

6.2.4 Sequence alignment using FASTA

FASTA (Pearson & Lipman, 1998) was used with default parameters for three separate stages in the search for domain-swapped proteins. First, FASTA was used to remove any redundancy in the initial list of putative swapped-domains that were obtained by the search strategy. A pair-wise comparison of the sequences was made using an E-value of 10^{-5} to remove any clearly homologous sequences. Any remaining homologous sequences could then be checked manually, in case of potential differences in their conformations. Second, FASTA sequence alignments were made between the putative domain-swapped sequences, and the sequences of the domain-swapped structures found in the initial data set. Any matches to already listed swapped structures i.e. those in the initial data set, could then be identified, and the corresponding structures removed from the data set generated by the search. Finally, FASTA was used to search the sequences of putative domain-swapped structures identified by the swapped-domain search method, against PDB sequences in order to locate any closed monomeric counterparts.

6.2.5 Secondary structure assignments

Secondary structure assignments were taken from definitions given by the DSSP algorithm (Kabsch and Sander, 1983). A simplified secondary structure scheme, converting the eight DSSP states to three was used, as described in section 2.2.1 of this thesis. The percentage assignments of the eight DSSP structural states, as well as the three simplified states, were calculated for the swapped-domain linkers in the open monomer conformation. Residue frequencies assigned to these secondary structure states were also calculated for the *bona fide* swapped-domain linkers, for both the open and closed conformations.

6.2.6 Identification of swapped-domain linker peptides

Assignments of the domain-swapped regions of each chain, taken from the relevant literature, were verified by viewing each structure in Rasmol (Sayle and Milner-White, 1995). The region of chain joining the swapped-domain to the main domain is defined here as the domain-swapped linker. The N- and C-termini of these

linker peptides were identified by visual inspection, using protein structure co-ordinates taken from the PDB viewed through the Rasmol protein structure viewer. Secondary structure was cross-validated by residue assignments made by DSSP (according to the simplified form (section 2.2.1)). In a similar procedure to that outlined in section 2.2.2 of this thesis, the domain-swapped linker sequence was assigned as those residues found between the last helix or strand belonging to the main domain and the first helix or strand belonging to the swapped-domain. The assignment of linkers connecting swapped structures containing few secondary structures was slightly different. In this case where no helices or sheets bordered the linker peptide, the linker termini residues were taken as those found on the boundary i.e. on the edge of what might be visualised as the outline of the domain that the swapped region extends into. All other regions of the protein not assigned as part of the domain-swapped linker are referred to as 'non-linker' in this chapter.

6.2.7 Amino acid composition

Amino acid propensities were calculated exactly as described in section 2.2.3 of this thesis. Propensities were calculated for all swapped-domain linker residues.

6.2.8 C-alpha extension

The C-alpha extension for each swapped-domain linker peptide was calculated exactly as described in section 2.2.5. The mean C-alpha extension was calculated for all swapped-domain linkers in the analysis data set. Further, the mean C-alpha extension was calculated for the monomeric and swapped conformations of the *bona fide* swapped-domain linkers in the data set.

6.2.9 Solvent accessibility

The degree of burial of a given residue was described by its solvent accessibility within the protein structure as described in section 2.2.6. RSA values were calculated for all residues in the domain-swapped linker data set by the DSSP algorithm.

The surface area for the *bona fide* swapped-domain interface was calculated by subtracting the solvent accessible surface area of the closed monomer from that of the open monomer (one subunit of structure). Interface calculations were only made for the *bona fide* swapped-domains (where a closed and open monomer structure was available) as differences in surface area could then be attributed to the exchanged domain.

6.2.10 Hydrogen bonding

The mean number of internal hydrogen bonds in the domain-swapped linkers was measured in addition to the number of hydrogen bonds found in the *bona fide* swapped-domains in their monomer and swapped state. Hydrogen bonds were calculated using the HBPLUS algorithm (McDonald and Thornton, 1994), in which hydrogen bonds were defined according to standard geometric criteria. Hydrogen bonds made to solvent were excluded from the measurements.

6.3 Results

6.3.1 Development of a domain-swapped protein search method

The search for domain-swapped regions of chain in this study focused on an initial search for regions of N- or C-terminal chain segments that appear highly exposed when considered as a single chain subunit (i.e. not in its oligomeric form). First a subset of proteins with this type of swapped-domain region was taken from the initial data set of domain-swapped chains that had been acquired through literature and domain classification database searches. A total of 25 proteins were found, ranging from highly extended regions of chains to slightly more globular but still exposed exchanged sections of peptide chain. The degree of solvent accessibility of these 25 domain-swapped units of structure can be seen in Figure 6.2, where the percentage frequency of swapped residues, within RSA intervals of width 5%, is shown. Also shown is the distribution of relative solvent accessibility for residues taken from the first and last 20 residues of a random selection of 100 single domain chains (taken from the representative set described in Chapter 4). The relative solvent accessibility of amino acids in these 20 residue N-and C-terminal peptides was measured to assess the degree of exposure that might be expected for such residues

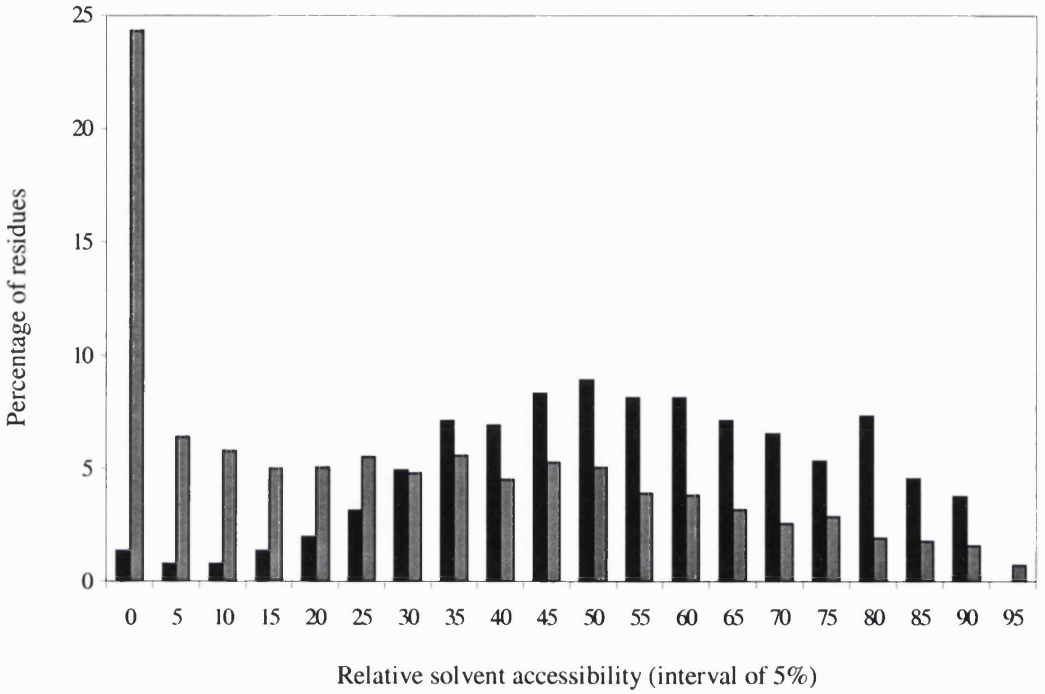


Figure 6.2 Distribution of relative solvent accessibility values

The distribution of relative solvent accessibility values were calculated for residues in the structures with highly exposed swapped domains (black columns) as well as the first and last twenty residues in a representative set of 100 single domain chains (grey columns).

in known globular protein domains that are not domain-swapped. A peptide length of 20 was used since this was close to the mean length of the 'non-globular' swapped-domain structure subset. As is expected the two distributions differ with the globular domain N-and C- terminal residues having a large proportion of buried residues (RSA≤10%), whilst the vast majority of the domain-swapped residues have been assigned as exposed (RSA>10%). A similar analysis is shown in Figure 6.3, although in this case the percentage frequency of C-alpha contacts for domain-swapped residues, and non-domain-swapped residues is shown. Again, as expected, the distribution of residue contacts numbers for domain-swapped residues, when compared to globular N-and C-terminal residues shows a distinct skewed distribution towards the lower end of the contact number scale. It can be seen that most non-globular swapped-domain residues have a contact number of less than or equal to 12. The correlation of contact number to relative solvent accessibility (data not shown) shows a contact number of 12 corresponds to a mean relative solvent accessibility of approximately 45% i.e. highly exposed residues. A contact number of 12 or less was therefore chosen as the criteria to define a residue as exposed enough to be part of an exposed domain-swapped structure. Figure 6.3 also shows that there are a number of the residues taken from the single domain chains have contact numbers of less than 12, though these are likely to correspond to individual exposed residues, for example in loop regions, rather than the consistent exposure of the N- or C-terminal 20 residue peptides. In general it was important to distinguish globular N- and C-terminal regions of protein structure which though exposed, still interacted with the main domain fold.

The domain-swapped search algorithm was benchmarked on three sets of structures. First, the algorithm was run against the subset of 25 structures with known extended non-globular swapped structures. Of these 25 proteins, 21 (nearly 85%) were correctly identified as domain-swapped. Missed structures were on the border-line of forming more folded, globular swapped regions. An example output of the search method for the 25 non-globular swapped-domains is shown in Table 6.1. The percentage of residues with a contact number below or equal to 12 for cut-offs between 8 and 20 is shown, together with the overall mean of these values for each protein. For highly exposed swapped regions all residues within each length cut-off (i.e. 100%) have a contact number of less than or equal to 12 and are therefore

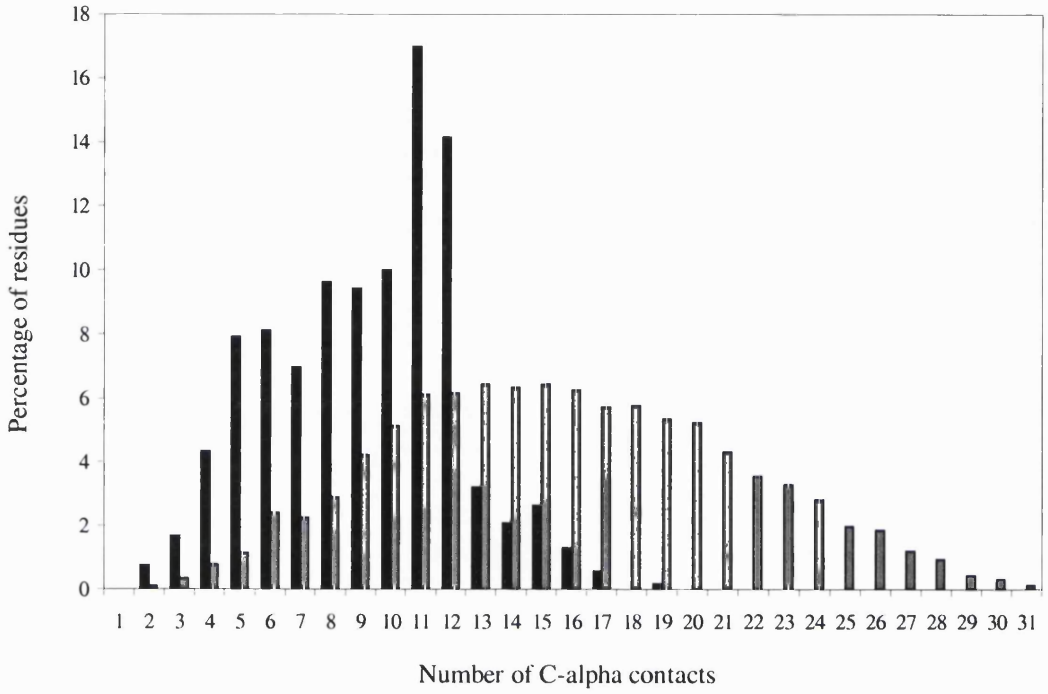


Figure 6.3 Distribution of residue contact numbers (10 Å cut-off)

The number of contacting C-alpha atoms within a cut-off of 10 Å was calculated for residues in the structures with highly exposed swapped domains (black columns) as well as the first and last twenty residues in a representative set of 100 single domain chains (grey columns).

Protein	Swapped-domain length	Percentage of residues with contact number < 12, at different cut-off distances: 8	10	12	14	16	18	20	Mean percentage over all cut-off distances	Type of swapped domain
1yvsN	30	75.0	70.0	58.3	64.3	68.8	66.7	65.0	66.9	2
1ht9A	38	62.5	60.0	58.3	57.1	50.0	55.6	60.0	57.6	2
1cdcA	45	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	2
1i4mA	27	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	1
1jmlA	16	100.0	100.0	100.0	100.0	100.0	94.4	95.0	98.5	1
1a2wA	12	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	1
1f0vA	9	100.0	100.0	100.0	100.0	87.5	77.8	70.0	90.8	1
1dz3A	12	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	1
1sndA	20	100.0	80.0	83.3	78.6	81.2	83.3	85.0	84.5	1
1pucN	13	75.0	70.0	75.0	64.3	68.8	61.1	60.0	67.7	1
1wwaX	9	100.0	100.0	100.0	100.0	93.8	88.9	80.0	94.7	1
1cksA	14	100.0	100.0	100.0	100.0	100.0	100.0	95.0	99.3	1
1bh5A	13	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	1
1hulA	23	100.0	90.0	91.7	92.9	93.8	94.4	95.0	94.0	1
1ilkN	41	100.0	100.0	100.0	100.0	93.8	94.4	95.0	97.6	2
1obpA	35	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	2
1fyrA	25	62.5	70.0	75.0	64.3	62.5	55.6	50.0	62.8	2
2spcA	31	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	1
1bylA	9	100.0	90.0	75.0	64.3	56.2	54.2	52.0	70.3	1
1buoA	5	100.0	100.0	100.0	100.0	93.8	83.3	75.0	93.2	1
1g5cA	9	100.0	100.0	100.0	100.0	93.8	83.3	80.0	93.9	1
1g6uA	14	100.0	100.0	100.0	100.0	93.8	88.9	80.0	94.7	2
1dudN	10	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	1
1fzrA	29	87.5	90.0	83.3	85.7	87.5	88.9	90.0	87.6	2
1k04A	36	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	1

Table 6.1 Output for swapped-domain finding algorithm

The output generated by the swapped domain search algorithm for the 25 domain swapped proteins in the database which exchange non-globular, or predominantly non-globular structures. The PDB code (and chain identifier) is shown for each protein, followed by the length of the swapped domain. The percentage of residues within the different length cut-offs, with a residue contact number less than or equal to 12 is also shown followed by the mean percentage value for the seven cut-offs. The final column represents the type of structure swapped, where 1 indicates very extended, and 2 indicates less extended, may contain buried residues within the swapped domain. Swapped domains were assigned to these classes by visual inspection.

considered as exposed enough to be a putative swapped structure, for example, the Human prion homo-dimer (Knaus *et al.*, 2001), PDB code 1i4m, that consists of a highly extended alpha-helix. Those swapped-domains that were missed tended to be those slightly more globular swapped-domains.

The search method was also used to search the N- and C-termini regions of the monomer structures (i.e. closed conformations) of the *bona fide* and *quasi* domain-swapped proteins. Here the swapped-domain forms part of the main domain, and as such is no more exposed than would be expected for the termini regions of general protein structures. Of these 24 structures, no false positive identifications were made, i.e. all were considered to have terminal peptide regions with solvent accessibility similar to that expected for general globular termini structures.

Finally, this algorithm was run against a set of 100 single domain proteins, visually inspected to verify their compactness. Again, no false positive predictions were made (data not shown). This was repeated using a contact number cut-off of 14 rather than 12, allowing residues with a slightly higher degree of burial into the calculations. This gave no change in the prediction results. However, a cut-off of 12 was retained for the remainder of the analysis, as it was felt that a more potentially more restrictive cut-off was necessary to keep any false positive predictions as infrequent as possible.

6.3.2 Searching for domain-swapped oligomers in the Protein Data Bank

As a preliminary search for potential domain-swapped structures in the PDB, a list of proteins assigned as forming homo-oligomeric structures was obtained from the PQS (http://pqs.ebi.ac.uk) database, giving a set of 4824 protein structures, (section 6.2.3). A run of the search method generated a list of 292 putative domain-swapped proteins. A manual approach, by visual inspection of these structures, was then used to remove structures that did not appear to fit the criteria of a domain swapped protein. Such structures included those with exposed N- or C-terminal regions of chain that appeared extend out to, and interact with adjacent subunits, but did not appear to make contacts in the oligomer that would be identical to those made in the monomer. Other structures included leucine-zipper helices, and fragments of larger structures, that would be monomeric in their complete form, but appear to swap structures in their truncated form.

A pair-wise alignment of the remaining putative domain-swapped sequences was then carried out using FASTA to remove any redundancy in the set (section 6.2.4). The resulting list of chains was aligned to those in the initial domain-swapped data set and 19 matches to these known domain-swapped chains were found all having E-values above 10^{-10} (data not shown). The final set of possible domain-swapped structures consisted of 8 chains, and is summarised in Table 6.2. A search for the structures of non domain-swapped monomeric forms of these structures, was made using FASTA against all the sequences of all structures in the PDB. Of the 8 putative swapped structures, potential closed monomer matches were found for 2. All are described in more detail below.

6.3.3 Domain-swapped structures found by the search algorithm

The putative domain-swapped structures (shown in Figure 6.4a-j) are described below and are also listed in Table 6.2.

Flavin mono-nucleotide binding protein (PDB code – 1eje)

The domain-swapped structure of flavin mononucleotide (FMN) binding protein from *Methanobacterium Thermoautotrophicum* (Christendat *et al.*, 2000) is shown in Figure 6.4a. The crystal structure of this dimer appears to swap the first 20 residues of the N-terminal of the polypeptide chain, exchanging two helical elements of structure. The N-terminal helix does not appear to form a close interaction to the main domain of the neighbouring subunit.

6-Phosphogluconate dehydrogenase from *T.brucei* (PDB code – 1pgj) and sheep liver (PDB code – 2pgd)

The three-dimensional structures of 6-phosphogluconate dehydrogenase (6PGDH) (an enzyme forming part of the pentose phosphate pathway) from both the protozoan *Trypanosoma brucei* (Phillips *et al.,* 1998) and sheep liver (Somers *et al.*, 1992) are shown in Figures 6.4b and c. Though both structures share low sequence identity (Phillips *et al.,* 1998) (and therefore both representatives were retained when filtering the set of putative domain-swapped proteins), it can be seen that they both have a C-terminal tail extending from the main domain into the neighbouring subunit forming a homo-dimeric complex. The N-terminal domains of both two-domain

Table 6.2 Data set of domain-swapped structures used in this analysis

The domain swapped proteins used in the analysis data set are listed together with their closed monomer counterpart in cases where a structure is known.

The sequence length of each protein and corresponding PDB code is shown. This is followed by additional data for the domain swapped structures describing the length and position in sequence of the swapped structure and the number of helices or strands exchanged in the open monomer structures. The class of domain swapped protein is also shown, where BF = *bona fide*, Q = *quasi* and C = candidate for domain swapping. Next, the oligomeric state in the PDB file is shown where N = no oligomeric co-ordinates given (and were therefore downloaded from the PQS web site), and Y = co-ordinates show structure as oligomer. Functions for each structure are listed (if known), and finally the reference for each structure is shown.

[#] These structures were identified by the swapped domain search algorithm.

Protein (PDB code)	No. residues Subunit	Swapped-domain	Linker position	Structure exchanged	Type	Oligomer in PDB?	Function	Reference
*Barnase monomer (1brn)	110						Endonuclease	Buckle & Fersht, 1994
*Barnase trimer (1yvs)	110	30	33-41	N-term helices (2)	BF	N	Endonuclease	Zegers et al., 1999
*Calbindin D9k wild-type monomer (4icb)	76						Calcium binding	Svensson et al., 1992
*Calbindin D9k dimer (1ht9)	76	38	38-47	C-term helices (2)	BF	Y	Not known	Hakansson et al., 2001
*CD2 monomer (1hng)	177						T lymphocyte adhesion protein	Jones et al., 1992
*CD2 dimer of N-terminal binding domain (1cdc)	99	45	44-54	N-term strands (4)	BF	Y	Not known	Murray et al., 1995
*Cyanovirin-N monomer (2ezm)	101						HIV inactivating protein	Bewely et al., 1998
*Cyanovirin-N dimer (3ezm)	101	44 or 49	50-57	N or C term domain strands (5)	BF	N	Not known	Yang et al., 1999
*Diptheria toxin monomer (1mdt)	535						Toxin (ADP-ribosylating)	Bennet & Eisenberg, 1994
*Diptheria toxin dimer (1ddt)	535	147	377-388	Large N-term domain	BF	N	Putative receptor binding	Bennet et al., 1994
*Human prion monomer (1qlx)	108						Prion protein	Zahn et al., 2000
*Human prion dimer (1i4m)	108	27	190-199	N-term helix	BF	N	Possible receptor	Knaus et al., 2001
*B1 domain of protein L monomer (1hz5)	62						IgG binding	O'Neill et al., 2001
*B1 domain of protein L dimer (1jml)	61	16	45-48	N-term strands (4)	BF	N	Not known	Kuhlman et al., 2001
*Ribonuclease A monomer (5rsa)	124						Ribonuclease	Wlodawer et al., 1982
*Ribonuclease A dimer, N-term swap (1a2w)	124	12	13-24	N-term helix	BF	Y	Ribonuclease	Liu et al., 1998
*Ribonuclease A dimer, C-term swap (1fov)	124	9	112-115	C-term strand	BF	Y	Ribonuclease	Liu et al., 2001
*Single-chain antibody NC10 monomer (1nmc)	246						Antigen binding	Malby et al., 1998
*Diabody (1lmk)	243	110	121-228	N- or C-term domain	BF	Y	Antigen binding	Perisic et al., 1994
*Phosphorylated-SPO0A monomer (1qmp)	129						Sporulation response regulator	Lewis et al., 1999
*Phosphorylated-SPO0A dimer (1dz3)	129	12	106-112	N-term helix	BF	N	Sporulation response regulator	Lewis et al., 2000
*Staphylococcal nuclease monomer (1snc)	149						Nuclease	Loll & Lattman, 1989
*Staphylococcal nuclease dimer (1snd)	143	20	112-121	C-term helix	BF	Y	Not known	Green et al., 1995
*suc1 monomer (1sce)	113						Cell cycle regulation	Bourne et al., 1995
*suc1 dimer (1puc)	113	13	86-93	C-term strand	BF	Y	Not known	Khasanovich et al., 1996

continued...	No. residues							
Protein (PDB code)	**Subunit**	**Swapped-domain**	**Linker position**	**Structure exchanged**	**Type**	**Oligomer in PDB?**	**Function**	**Reference**
*TrkA monomer (1www)	120						Nerve growth factor binding	Wiesmann et al., 1999
*TrkA dimer (1wwa)	109	9	291-298	C-term strand	BF	Y	Not Known	Ultsch et al., 1999
*Human cyclin dependent kinase type 1 (CksHs1) monomer (1dks)	79						Cell cycle regulation	Arvai et al., 1995
*Human cyclin dependent kinase type 2 (CksHs2) dimer (1cks)	79	14	60-65	C-term strand	Q	Y	Cell cycle regulation	Parge et al., 1993
*Gamma-B crystallin monomer (4gcr)	174						Eye lens protein	Najmudin et al., 1993
*Beta B2 crystallin trimer (1blb)	204	81 or 88	81-88	N- or C-term domain	Q	Y	Eye lens protein	Nalini et al., 1994
*Chicken cystatin monomer (1cew)	108						Protease inhibitor (Cysteine)	Bode et al., 1988
*Human cystatin C dimer (1g96)	120	34	55-59	N-term helix & strands (2)	Q	N	Not known	Janowski et al., 2001
*Glyoxalase 1 of E.coli (dimer)(1fa5)	135						Lyase	He et al., 2000
*Human Glyoxalase 1 (swapped-dimer) (1bh5)	183	13	20-32	N-term helix	Q	Y	Lyase	Ridderstrom et al., 1998
*Human granulocyte macrophage colony stimulating factor monomer (1gmf)	127						Granulocyte macrophage stimulating factor	Diederichs et al., 1991
*Interleukin-5 dimer (1hul)	113	23	85-88	C-term helix	Q	Y	B and T cell stimulating factor	Milburn et al., 1993
*Interferon-beta monomer (1rm1)	160						Interferon	Senda et al., 1992
*Interleukin-10 dimer (1ilk)	160	41	108-118	C-term helix	Q	N	Cytokine synthesis factor	Zdanov et al., 1995
*Mannose binding protein (1msb)	115						Mannose binding	Weis et al., 1991
*IX/X-binding protein coagulation factor dimer (1ixx)	129	17	73-76 93-98	Middle loop	Q	Y	Coagulation factor	Mizuno et al., 1997
*Major urinary protein monomer (1mup)	166						Rodent-pheromone transport	Bocskei et al., 1992
*Odorant-binding protein (1obp)	159	35	121-124	C-term helix & strand	Q	Y	Odorant-binding	Bianchet et al., 1996
*Grb2 adaptor (SH2 + SH3) (1gr1)	217						Signal transduction	Maignan et al., 1995
*Grb-SH2 domain dimer (1fyr)	93	25	120-127	C-term helix	Q	Y	Binding phosphorylated peptide	Schiering et al., 2000
*Fyn-SH3 monomer (1fyn)	62						Signal transduction	Musacchio et al., 1994
*SH3 domain of Eps8 (1aoj)	65	26	34-39	C-term strands (2)	Q	Y	Proline-rich sequence	Kishan et al., 1997

continued...	No. residues							
Protein (PDB code)	**Subunit**	**Swapped-domain**	**Linker position**	**Structure exchanged**	**Type**	**Oligomer in PDB?**	**Function**	**Reference**
*Type 3 secretion chaperone SigE monomer (1k3s)	113						Chaperone	Luo et al., 2001
*Type 3 secretion chaperone CesT dimer (1k3e)	156	33	34-37	N-term helix & strands (2)	Q	N	Chaperone	Luo et al., 2001
*Bleomycin resistance protein (1bly)	122	9	7-9	N-term strands (4)	C	N	Bleomycin resistance	Dumas et al., 1994
*BTB domain of PLZF (1buo)	120	5	11-13	N-term strand	C	Y	Protein-Protein interaction motif	Ahmad et al., 1998
*Cab type beta carbonic anhydrase (1g5c)	170	9	11-26	N-term helix	C	Y	Lyase	Strop et al., 2001
*Citrate synthase (1cts)	433	14	417-423	C-term helix	C	N	Citrate synthase	Remington et al., 1982
*Designed domain-swapped dimer (1g6u)	48	15	33-34	C-term helix	C	Y	not known	Ogihara et al., 2001
*dUTPase trimer (1dud)	136	10	125-126	C-term strand	C	N	dUTPase hydrolase	Larsson et al., 1996
*HSP33 dimer (1hw7)	255	51	177-183	C-term helices (3)	C	N	Molecular chaperone	Vijayalakshmi et al., 2001
*Recombination endonuclease VII dimer (1en7)	157	74	75-83	N-term helices (2) & strands (2)	C	Y	Mismatch repair	Raaijmakers et al., 1999
*Phage T7 GP4D helicase hexamer (1e0j)	326	23	283-305	N-term helix	C	Y	Helicase	Singleton et al., 2000
*E.coli RecA protein hexamer (2reb)	352	24	27-39	N-term helix	C	N	Recombination	Story et al., 1992
*Simian virus 40 oligomer (1sva)	361	61	296-300	C-term helix & strands (2)	C	N	Virus coat protein	Stehle et al., 1996
Bacteriophage T7 endonuclease dimer (1fzr)	129	29	46-48	N-term helix & strands (2)	C	Y	Endonuclease	Hadden et al., 2001
Focal adhesion kinase targeting domain dimer (1k04)	142	36	944-946	N-term helix	C	N	Transferase	Arold et al., 2002
Histidyl-tRNA sythetase dimer (1kmn)	449	95	320-329	C-term domain	C	Y	tRNA-synthetase	Arnes et al., 1997
#Flavin mononucleotide binding protein dimer (1eje)	192	20	24-27	N-term helices (2)	C	N	FMN-binding protein	Christendat et al., 2000
#6-Phosphogluconate dehydrogenase #T.brucei dimer (1pgj)	478	33	437-442	N-term helix	C	Y	Oxidoreductase	Phillips et al., 1998
#6-Phosphogluconate dehydrogenase sheep liver dimer (2pgd)	473	36	434-439	N-term helix	C	Y	Oxidoreductase	Somers et al., 1992

continued...	No. residues							
Protein (PDB code)	**Subunit**	**Swapped-domain**	**Linker position**	**Structure exchanged**	**Type**	**Oligomer in PDB?**	**Function**	**Reference**
#Bovine interferon-gamma dimer (1d9g)	143	34	82-84	N-term helices (2)	C	Y	Interferon	Randal & Kossiakoff, 2000
#Mannose specific agglutinin from snowdrop (1msa)	109	9	96-99	C-term strand	C	Y	Agglutinin	Hester et al., 1995
#Metallo-beta-lactamase *B. fragilis* (1znb)	230						Antibiotic resistance	Concha et al., 1996
#L1 metallo-beta-lactamase *S. maltophilia* (1sml)	266	13	2-14	N-term strand	Q	Y	Antibiotic resistance	Ullah et al., 1998
#Feline HIV DUTP pyrophosphatase trimer (1f7o)	115	8	504-508	C-term strand	C	Y	Phosphatase	Prasad et al., 2000
#Binding domain Phage P22 tailspike protein (1lkt)	103	19	25-30	C-term strand	C	Y	Tailspike protein	Steinbacher et al., 1997

structures form Rossman folds, whilst the all helical C-terminal domain forms the dimer interface and forms the swapped-domain structure. As summarised in Table 6.2, the swapped region of T.brucei 6PGDH (33 residues) is slightly shorter than that of the sheep enzyme (36 residues). However, in both cases the main swapped element is an alpha helix.

Bovine interferon–gamma (PDB code – 1d9g)

The homo-dimeric crystal structure of interferon-gamma (IFN-g) (Randal and Kossiakoff, 2000) is shown in Figure 6.4d. The dimeric structure is composed of the two chains, both made up of six helices, in which the last two C-terminal helices of each subunit swap over to form the domain of the adjacent dimer subunit. The search for homologous monomer structures identified a single chain form of IFN-g. However this structure consisted of a linked version of the homo-dimeric chains, where the N- and C-termini had been fused together. This was carried out to overcome the highly flexible nature of the last C-terminal swapped helix, found in the swapped form, making high resolution structural determination of this protein difficult (Randal and Kossiakoff, 1998).

Mannose-specific agglutinin (snowdrop) (PDB code – 1msa)

The homodimeric structure of mannose specific snowdrop lectin (Hester *et al.*, 1995) is shown in Figure 6.4e. This retroviral inhibitor consists of three antiparallel four stranded beta sheets, forming a twelve stranded barrel structure, that forms a dimeric association with an identical domain through C-terminal strand exchange. The exchanged strand can be seen to form part of one four stranded sheet in the adjacent subunit. It is thought that these hybrid beta-sheets are the sites for high affinity mannose binding in the dimer interface (Hester *et al.*, 1995). A search for monomeric forms of the domain swapping protein revealed a two-domain lectin from bluebell, Figure 6.4f (Wright *et al.*, 2000), PDB code 1dlp, whose N-terminal domain shows significant homology to the snowdrop lectin, again with a similar twelve stranded beta barrel fold. Here the exchanged strand is not swapped, but connects to an eleven-residue linker that passes to the N-terminal strand of domain 2 (Wright *et al.*, 2000).

Figure 6.4 Structures of potential domain swapped proteins identified by the search algorithm.

a) The domain swapped structure of flavin mononucleotide (FMN) binding protein from *Methanobacterium Thermoautotrophicum* (Christendat *et al.*, 2000).

b) & c) The three-dimensional structures of 6-phosphogluconate dehydrogenase (6PGDH) (enzymes forming part of the pentose phosphate pathway) from the protozoan *Trypanosoma brucei* (B) (Phillips *et al.*, 1998) and sheep liver (C) (Somers *et al.*, 1992).

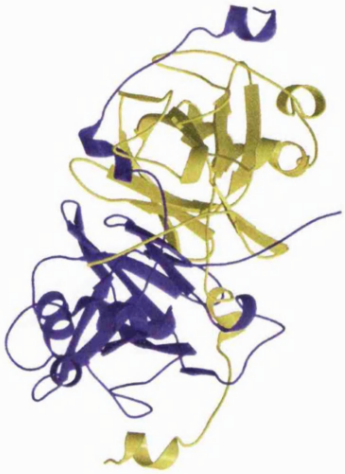
d) The homo-dimeric crystal structure of interferon-gamma (IFN-g) (Randal and Kossiakoff, 2000).

e) & f) The homodimeric structure of mannose specific snowdrop (E) (Hester *et al.*, 1995). A search for monomeric form revealed a two-domain lectin from bluebell (F) (Wright *et al.*, 2000), whose N-terminal domain shows significant homology to the snowdrop lectin with a similar twelve stranded beta barrel fold.

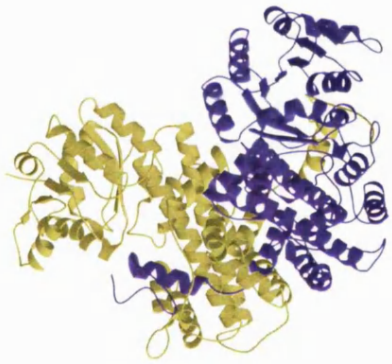
g) & h) The structure of the L1 metallo-beta-lactamase protein, from *Stenotrophomonas maltophilia* (G) (Ullah *et al.*, 1998). The structure of the metallo-beta-lactamase protein in *Bacteriodes fragilis* has also been determined (H) (Concha *et al.*, 1996). The N-terminal residues of the *B. fragilis* structure, appear to turn back into the main domain forming an additional strand with the main domain beta sheet, rather that extending from the main domain.

i) & j) The trimeric structure of the head-binding domain of P22 tailspike (I) (Steinbacher *et al.*, 1997), and the similar trimeric crystal structure of the feline immunodeficiency virus DUTP Pyrophosphatase (J) (Prasad *et al.*, 2000).

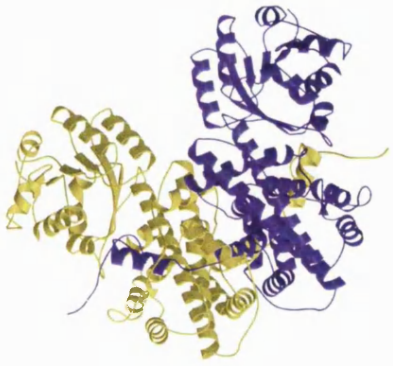
a)



b)



c)



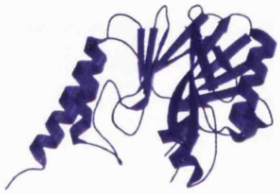
d)

e)

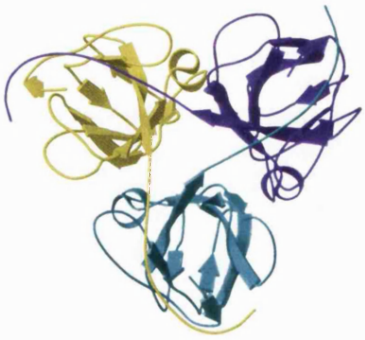
f)

g)

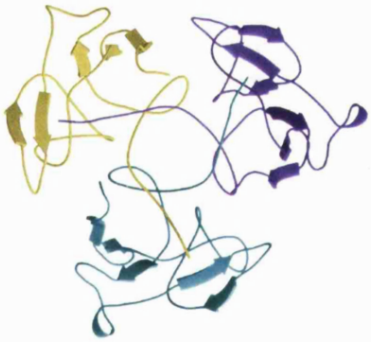
h)



i)



j)



L1 metallo-beta-lactamase protein (PDB code – 1sml)

The structure of the L1 metallo-beta-lactamase, a protein that is responsible for antibiotic resistance, from *Stenotrophomonas maltophilia* is shown in Figure 6.4g (PDB code 1sml). The L1 protein forms an active tetramer (shown as a dimer in Figure 6.4g) where the N-terminal residues of each chain make interactions with the opposing residues, formed by the side-chains of residues at the more distal end of the structure (Ullah *et al.*, 1998). The structure of the metallo-beta-lactamase protein in *Bacteriodes fragilis* (*B. fragilis*) has also been determined (Concha *et al.*, 1996), shown in Figure 6.4h. The monomeric *B. fragilis* structure lacks the extended N-terminal region. Other structural differences include a number of elongated helix and loop regions, and an additional strand in the N-terminal domain of the two-domain structure. However, although these metallo-beta-lactamase proteins from *S. maltophilia* and *B. fragilis* share little sequence similarity, the overall fold appears conserved (Ullah *et al.*, 1998). The N-terminal residues of the *B. fragilis* structure, rather that extending from the main domain, appear to turn back into the main domain forming an additional strand with the main domain beta sheet. As such, these protein structures may represent an example of *quasi* domain swapping.

The head-binding domain of P22 tailspike protein (PDB code – 1lkt) and Feline immunodeficiency virus DUTP Pyrophosphatase (PDB code – 1f7o)

The trimeric structure of the head-binding domain of P22 tailspike protein is shown in Figure 6.4i, where the N-terminal polypeptide chain can be seen to interact with the neighbouring subunit. Similarily, Figure 6.4j shows the trimeric crystal structure of the feline immunodeficiency virus DUTP Pyrophosphatase. Here it can be seen that the C-terminal region of the polypeptide chain crosses over and interacts with a neighbouring trimer subunit.

6.3.4 The secondary structure of swapped-domain linkers and swapped-domains

The secondary structure assignments given by DSSP for the swapped-domain linkers can be seen in Table 6.3a. These assignments were made for the protein chains in the swapped open conformation. The eight secondary structure states are shown together with the percentage of swapped-domain linker residues assigned to

Table 6.3 Secondary structure assignments of swapped-domain linkers

a) The frequency of secondary structures assignments in the swapped-domain linkers. These assignments were made for the protein chains in the swapped open conformation. The eight different secondary structure states defined by the DSSP algorithm are shown together with the percentage of swapped-domain linker residues assigned to each secondary structure state by DSSP. Also shown are the percentage frequencies of swapped-domain linker residues adopting coil, strand or helical conformations as defined by the simplified secondary structure assignment scheme (section 6.2.5).

b) The values for the *bona fide* swapped-domain linkers in their closed conformation.

c) The secondary structure frequencies for the *bona fide* swapped-domain linkers in their open conformations.

a)

Swapped-domain linkers (open conformation)		
Secondary structure state (DSSP)	DSSP assignment	Simplified scheme
	Percentage of residues	
C	43.0	81.0
S	11.6	0.0
T	9.2	0.0
E	14.4	10.9
B	1.8	0.0
H	14.8	8.1
G	5.3	0.0
I	0.0	0.0

b)

Bona fide swapped domain linkers (closed conformation)		
Secondary structure state (DSSP)	DSSP assignment	Simplified scheme
	Percentage of residues	
C	63.9	91.8
S	13.4	0.0
T	4.1	0.0
E	4.1	3.1
B	0.0	0.0
H	14.4	5.2
G	0.0	0.0
I	0.0	0.0

c)

Bona fide swapped domain linkers (open conformation)		
Secondary structure state (DSSP)	DSSP assignment	Simplified scheme
	Percentage of residues	
C	48.5	80.4
S	13.4	0.0
T	6.2	0.0
E	16.5	14.4
B	1.0	0.0
H	14.4	5.2
G	0.0	0.0
I	0.0	0.0

each state by DSSP. Also shown are the secondary structure assignments using the simplified secondary structure assignment scheme (section 6.2.5). It is evident from the DSSP eight state assignments that swapped-domain linker residues tend to be found in a coil conformation. Using the simplified scheme shows that nearly 20% contribute to helix or strand elements. Of the 47 swapped-domain hinge regions, 3 were found to be all helical, and 3 all strand.

The 14 *bona fide* domain-swapped structures were analysed to allow comparison of their swapped-domain linkers in the open and closed conformation. Table 6.3b shows the percentage secondary structure assignments for the linker residues in their closed conformation, whilst Table 6.3c shows the secondary structure frequencies for their open conformations. It can be seen from the simplified scheme that though the predominant structural class is coil, the percentage of strand linker residues increases from 3% (non-swapped conformations) to over 14% (swapped conformation). Visual inspection shows that this is mainly due to the linker forming beta sheet structures on interaction with the adjacent domain subunit, that were not present in the monomer structure.

The swapped-domains varied greatly in their secondary structure composition (Table 6.2). For example, in the swapped dimer of RNase A (Liu *et al*, 2002) a single (N-terminal) alpha-helix is swapped whilst in the Diphtheria toxin dimer (Bennet *et al*, 1994) the whole 148 residue C-terminal domain is swapped.

6.3.5 Length distribution of swapped-domain linkers

The length distribution of the swapped-domain linkers is shown in Figure 6.5. Overall the length of the linkers showed a large variation between 3 and 22 residues, although the longer linkers were less frequent. Most linkers were found to have lengths between 3 to 10 residues (87%), with the mode of the distribution being 4 residues. The mean linker length 6.8 residues.

6.3.6 Swapped-domain linker amino acid propensities

The amino acid propensities of the amino acids in the swapped-domain linkers compared to those found in proteins generally are shown in Table 6.4. Propensity values above 1.0 show the corresponding amino acid to be favoured,

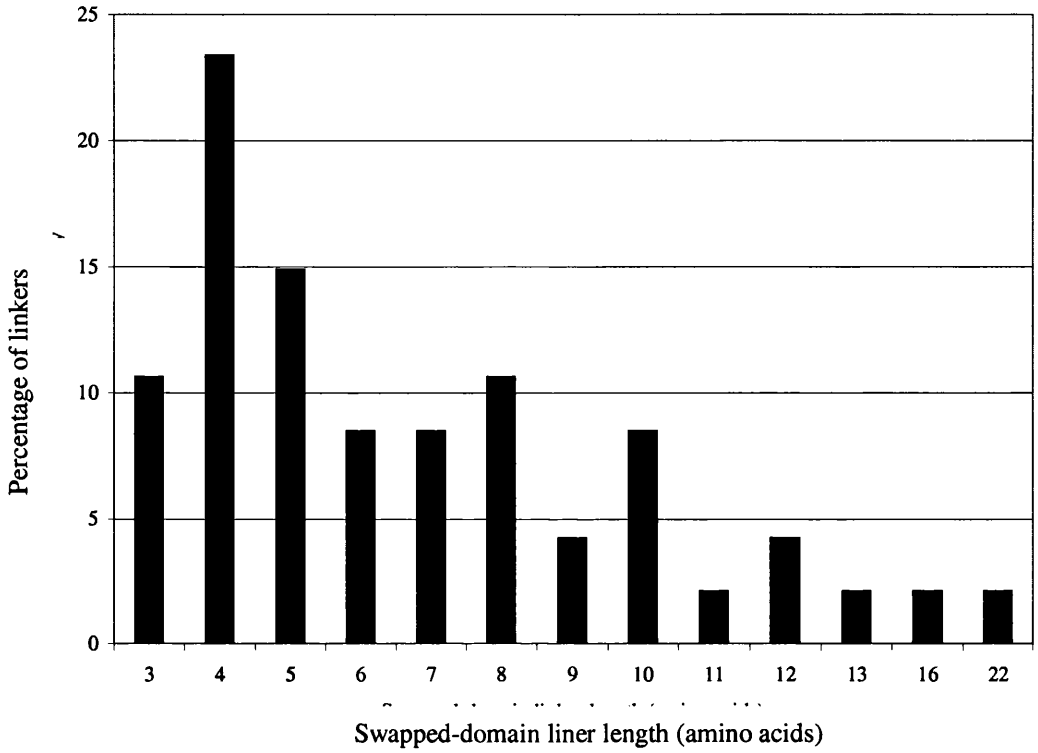


Figure 6. 5 Length distribution of swapped-domain linker peptides

Amino acid	Percentage		Propensity
	In Domain Swapped Linkers	In all Proteins	
PRO	8.03	5.04	1.79
LYS	9.12	5.64	1.62
HIS	4.38	2.44	1.59
SER	7.30	6.06	1.20
CYS	1.82	1.53	1.19
GLN	4.38	3.73	1.17
TRP	1.82	1.59	1.15
GLY	8.39	7.68	1.09
GLU	6.20	6.35	0.98
ASN	4.38	4.58	0.96
ASP	5.47	5.90	0.93
ARG	4.38	4.86	0.90
VAL	6.20	6.95	0.89
PHE	3.65	4.10	0.89
LEU	6.93	8.36	0.83
THR	4.74	5.74	0.83
ALA	6.57	8.00	0.82
TYR	2.19	3.71	0.59
MET	1.09	2.10	0.52
ILE	2.92	5.62	0.52
Charged	25.18	22.75	1.11
Polar	31.02	29.40	1.06
Hydrophobic*	35.77	42.81	0.84

Table 6.4 Amino acid propensities in swapped-domain linkers

* excluding proline (see section 6.3.6)

whilst a propensity below 1.0 shows a residue to be disfavoured. The overall propensity value for the polar residues of 1.06 shows that there is a preference for this amino acid subset, whilst a propensity value of 1.11 for the charged residue subset shows these residues are favoured in swapped-domain linkers. In contrast, hydrophobic residues are less favoured, with an overall propensity value of 0.84 excluding proline (0.92 including proline). These observations are made more salient when considering the individual amino acid propensities. With the exception of proline and glycine, all the residues with a propensity above one are polar or charged. Those amino acids with a propensity close to one are similarly polar or charged. The disfavoured amino acids tend to be hydrophobic with threonine and tyrosine being exceptions. Tyrosine may be disfavoured as it is amongst the larger polar residues, having an aromatic side chain, though both tryptophan with a propensity of 1.15 and histidine with a propensity 1.59, also have aromatic side chains and are favoured. As a whole, the most favoured amino acid is proline, followed by lysine, histidine and serine, all of which have propensities above 1.2.

As described in section 6.3.4, the majority of swapped-domain linker residues are found in a coil conformation. To assess if there was similarity in the amino acid propensities of swapped-domain linkers and general coil residues (as calculated in Chapter 2 of this thesis), a ranked correlation between the two distributions was calculated. A correlation coefficient of r=0.12 was determined, showing no significant correlation between the two ranked distributions. The correlation for the two sets of data can be seen in Figure 6.6a. Amongst the amino acids that appear to differ most in their propensity are histidine and lysine, that are also amongst the most favoured swapped-linker residues whilst glycine is less favoured in swapped-domain linkers compared to non-linker coil.

The propensities obtained for the swapped-domain linkers in this chapter were compared to the propensities calculated for the domain-linkers in Chapter 2. A ranked correlation coefficient of r=0.47 was calculated between the distributions, that although weak, is significant. The comparison of the swapped-domain linker propensities and general domain linkers propensities can be seen in Figure 6.6b. Amongst the largest outliers are cysteine, tryptophan both of which are favoured in swapped-domain linkers (propensities of 1.19 and 1.15) but are disfavoured in domain-linkers – in fact tryptophan (propensity of 0.57) is the least favoured residue in domain-linkers. Lysine and histidine differ considerably – though lysine is

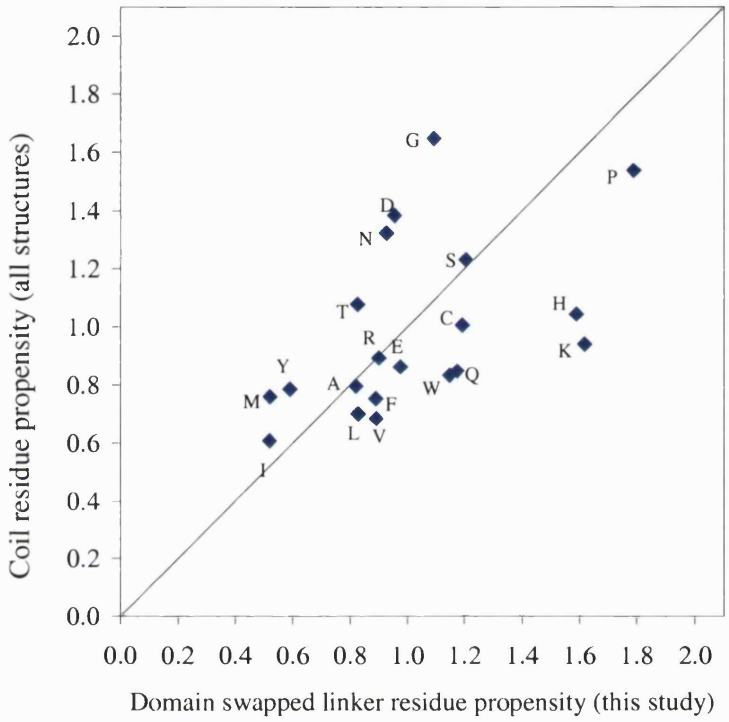
Figure 6.6 Comparison of swapped-domain linker amino acid propensities

Swapped-domain linker amino acid propensities calculated in this study were compared to:

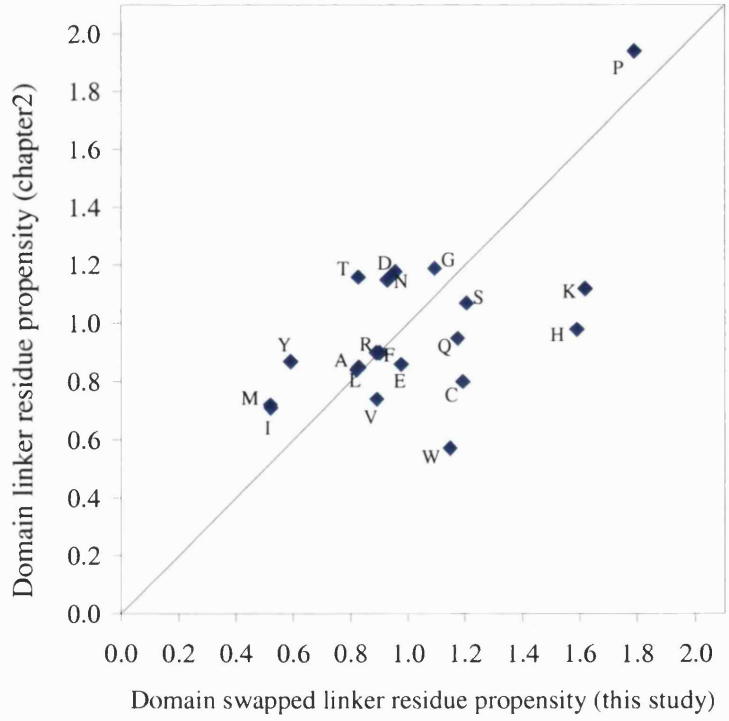
a) Amino acid propensities calculated for non-linker coil residues (in Chapter 2).

b) Domain linker propensities (calculated in Chapter 2).

a)


Coil residue propensity (all structures)
Domain swapped linker residue propensity (this study)
0.0 0.2 0.4 0.6 0.8 1.0 1.2 1.4 1.6 1.8 2.0
G
P
D
N
S
T
C
H
K
R
E
Y
A
Q
W
M
F
L
V
I

b)


Domain linker residue propensity (chapter2)
Domain swapped linker residue propensity (this study)
0.0 0.2 0.4 0.6 0.8 1.0 1.2 1.4 1.6 1.8 2.0
P
T
D
N
G
K
S
H
Y
R
F
Q
A
L
E
C
M
I
V
W

favoured (propensity of 1.12) and histidine is tolerated (propensity of 0.98) in domain linkers, these residues are ranked second and third in domain-swapped linkers, with propensities of 1.62 and 1.59 respectively.

Proline is the most favoured residue in domain-swapped linkers with a propensity of 1.79. This finding is similar to that obtained in Chapter 2, where proline was also the most frequent residue in domain-linkers. As discussed, this residue has a key role in conferring flexibility to the polypeptide chain, especially when paired with glycine. Interestingly, whereas glycine was the second most frequent residue in domain-linkers, it is the eighth most frequent in swapped-domain linkers, albeit still having a propensity above one.

6.3.7 C-alpha extension of swapped-domain linkers

The mean C-alpha extension (in Å) between swapped-domain linker residue pairs, shown in Table 6.5, was calculated by measuring the distance between the two terminal C-alpha atoms, and dividing by the linker length minus 1. The mean extension for all swapped-domain linkers in their domain-swapped open conformation was 2.51 Å (±1.13). This value is intermediate between the mean extension of general coil residues (2.84 Å (±0.53)) and the mean value of 2.21 Å (±0.88) for general domain linkers (both values were taken from the analysis in Chapter 2). The mean extension values for the *bona fide* linkers in their closed and open conformations, are also shown in Table 6.5. The mean extension of the swapped-domain linkers in their swapped conformation is greater than the mean extension value in their monomer conformation (2.37 Å (±1.17) versus 2.05 Å (±1.27)), demonstrating the more extended conformation achieved by the linkers in the swapped oligomer.

6.3.8 Solvent accessibility of swapped-domain linker residues

The extent to which swapped-domain linking residues are exposed to solvent in the swapped oligomer complex was calculated using relative solvent accessibility measurements (section 6.2.9). The percentage of swapped-domain linker residues defined as exposed (RSA >10%) is shown in Table 6.6. The equivalent values for non-linker helical, strand, coil and domain-linker residues taken from Chapter 2 of

	C-alpha extension (Å)	
	mean	s.d.
*Helical residues	1.54	0.07
*Strand residues	3.21	0.18
*Coil residues	2.84	0.53
*All domain-domain linker residues	2.21	0.88
Swapped-domain linker residues open monomer (all)	2.51	1.13
Swapped-domain linker residues (*bona fide* open monomer)	2.37	1.17
Swapped-domain linker residues (*bona fide* closed monomer)	2.05	1.23

Table 6.5 C-alpha extension

The mean C-alpha extent per residue pair is shown for non-linker helical, strand and coil residues, as well as residues found in domain linking peptides. Also shown is the mean C-alpha extent for all swapped-domain linker residues in the data set. Additionally the C-alpha extension for swapped-domain linker residues in the *bona fide* domain swapped structures are shown, calculated for their monomer (i.e. closed conformation) and swapped (i.e. open conformation) states.

*Values taken from Chapter 2 of this thesis.

	% total exposed residues	RSA	
		mean	s.d.
*Helical residues	56.4	23.3	24.0
*Strand residues	42.6	15.0	18.8
*Coil residues	72.3	33.6	27.8
*Domain linker residues	71.3	31.8	26.5
Swapped-domain linker residues	79.0	33.1	27.2

Table 6.6 Relative solvent exposure

The proportions of exposed residues for each structural state are shown as well as the mean relative accessible surface area and associated standard deviation for non-linker helical, strand and coil, domain-linker and swapped-domain linker residues.
*Values taken from Chapter 2 of this thesis.

this thesis are also shown in this table. Of these residue sub-sets, swapped-domain linker residues have the largest percentage (79%) of residues assigned as exposed in the protein oligomer, higher than both domain-linker residues (71% exposed) and non-linking coil residues, (72% exposed), and considerably higher than helical residues (56% exposed) and strand residues (42% exposed). Comparison of mean relative solvent accessibility again shows swapped-domain linker residues to have similar values to non-linker coil and domain-linker residues, all having mean relative solvent accessibility values in the low 30's (mean relative solvent accessibility 33.1, 33.6 and 31.8 respectively). The large standard deviation values demonstrate the high degree of variation of the accessibility values, even within the structural subsets, making any differences between the mean relative solvent accessibility values of the general coil, swapped-domain and domain-linking resides difficult to interpret with confidence.

6.3.9 Hydrogen bonding

The mean number of hydrogen bonds made between residues within the swapped-domain linkers was calculated for all swapped-domain linkers in the dataset (open conformation) (section 6.2.10). A mean value of 0.51 hydrogen bonds per swapped-domain linking residue was calculated. The same calculation for domain-linker residues (calculated in Chapter 2 of this thesis), gave 0.39 hydrogen bonds per residue, suggesting that swapped-domain linkers tend to make on average, more internal hydrogen bonds.

The swapped-domain linker is found in different conformations in the open and closed states (Figure 6.1). An analysis of the *bona fide* domain-swapped structures was also made to assess the difference in internal hydrogen-bonding (if any) within their swapped-domain linkers in the two conformations. Only the *bona fide* swapped-domain linkers were considered since by definition, only these proteins had their structures solved in both the open and closed conformations. The linkers of this sub-set in the swapped open state were found to have 0.58 hydrogen bonds per residue, whilst 0.91 hydrogen bonds per residue were found in the closed conformation. It appears the conformation that the linker chain attains in the closed state allows more potential for internal hydrogen bonding.

Finally, the difference in the number of hydrogen bonds made by residues in the swapped-domain, (within the swapped-domain and externally to other domains) was computed for the *bona fide* domain-swapped proteins, in the monomer and swapped structures (section 6.2.10), and is shown in Table 6.7. A similar number of hydrogen bonds between the two states would indicate little difference in the conformation of the swapped-domain in the open and closed conformation, as might be expected if *inter*chain interactions made between subunits were similar to the *intra*chain interactions made within subunits. In Table 6.7 it can be seen that for most *bona fide* swapped-domains, the number of hydrogen bonds made by the swapped-domain residues are similar in the two conformational states. The most marked difference was found between the monomer and swapped form of the Cyanovirin-N antiviral protein, PDB code 1iiy (swapped dimer) and PDB code 3ezm (monomer), each making 40 and 55 hydrogen bonds respectively. Analysis of the swapped-domain structure of Cyanovirin-N shows that though the swapped-domain is by no means the largest of the *bona fide* proteins, at 49 residues, it appears to intertwine with the adjacent dimer subunit, which may result in a degree of conformational change thereby creating more opportunities for hydrogen bonding in the open state than are available in the closed state.

The surface area of the swapped *bona fide* swapped-domains are shown in Table 6.7, calculated by subtracting the total surface area of the closed monomer structure from that of the open monomer (both single subunit). It can be seen that the sizes of interface areas can vary considerably from structure to structure in this subset of proteins, (from 1581 $Å^2$ to 4162 $Å^2$) and does not appear to correlate to the size of the swapped-domain.

6.4 Discussion

The work described in this Chapter has aimed to make a general analysis of domain-swapped protein structures, in particular the properties of the linker regions of peptide that join the swapped-domain to the main domain. The process of domain swapping may provide a method for the evolution of protein oligomers through the exchange of structures that form identical interactions in the swapped state as those that are made in the monomeric non-swapped state. This mechanism may have bypassed the need for the evolution of compatible interfaces between the

Monomer		Swapped oligomer		% change in	Surface area of
structure	No. H-bonds	structure	No. H-bonds	no. H-bonds	swapped interface Å^2
1brn	35	1yvs	34	-2.9	2825
4icb	38	1ht9	35	-8.6	1790
1hng	49	1cdc	52	5.8	4162
1iiy	40	3ezm	55	27.3	2931
1qlx	34	1i4m	36	5.6	1818
1hz5	20	1jml	24	16.7	1698
5rsa	15	1a2w	15	0.0	1666
5rsa	13	1f0v	14	7.1	1827
1qmp	14	1dz3	17	17.6	1900
1snc	28	1snd	27	-3.7	1581
1www	7	1wwa	8	12.5	1620
1mdt	136	1ddt	134	-1.5	4106

Table 6.7 Hydrogen bonding in, and interface size of *bona fide* domain-swapped structures

The difference in the number of hydrogen bonds in the swapped domains of *bona fide* domain swapped proteins was calculated in their monomer and swapped states as indicated in the table. The percentage change in number of hydrogen bonds is also given. Also shown is the surface area (Å^2) of the interface formed between the swapped domain and the main domain.

oligomerised structures. The energy barrier between the monomeric fold and the domain-swapped fold must be low enough to allow transition between the two conformations (Newcomer, 2002).

An initial search for potential domain-swapped structures in the PDB was made, identifying structures with highly exposed non-globular chain termini extending from the main domain. The degree of residue exposure to solvent was assessed by calculating the number of contacting residues within a 10 Å cut-off. The domain-swapped search algorithm gave eight additional domain-swapped structures for the analysis data set, although many more false positive structures were found with highly exposed termini regions. It appeared that many of the false positive 'swapped structures' could be assigned as artefacts of crystallisation. For example, termini-peptide of several structures were seen to 'invade' and interact with neighbouring protein subunits, as might be expected for swapped-domains. However, whilst exchanging identical units of chain, from the visual inspection it did not appear as though the interaction sites of these swapped units were equivalent to thise that would be made in the monomer. In other words the swapped-domain region would not be able to form the same interaction to its own main domain in the closed monomer form, and were therefore not true domain swaps.

Another potential source of false positives resulted from the common practice of dissecting a protein into its constituent domains to facilitate biochemical studies. The structure of the N-terminal domain of the nitrogen fixation protein, FIXL (Gong *et al.*, 1998), PDB code 1d06, is shown in Figure 6.7. This protein forms a five stranded beta barrel with a C-terminal protruding helix that would normally lead into a kinase domain that is not present in this structure. It can be seen that the C-terminal helix of each subunit exchanges between the domains, possibly the best conformation to shield hydrophobic residues within the helix from solvent. The residues in the helix are therefore 'exposed' in the monomeric subunit structure and as such this structure was identified by the domain swapping search algorithm. The occurrence of this oligomeric structure is a result of the truncation of the kinase domain and would therefore be unlikely to be found in nature.

The analysis data set used in this study contains two swapped protein structures, having only one domain of their two-domain closed monomer counterparts, (CD2 dimer of N-terminal binding domain and the Grb-SH2 domain

Figure 6.7 Structure of the homo-dimer formed by the N-terminal domain of the nitrogen fixation protein FIXL

The structure of the N-terminal domain of the nitrogen fixation protein FIXL (Gong *et al.*, 1998). The C-terminal helices exchange between the domains and are therefore highly extended and exposed in the monomeric subunit structure, and were therefore identified by the domain swapping search algorithm. However the 'domain swapping' in this structure appears to be a result of the common practice of dissecting a protein into its constituent domains to facilitate biochemical studies.

dimer). However, the domain swapping in both of these cases is clear, with considerable interchange of structural elements (see Table 6.2) whereas the swapping in FIXL does not seem to be so clear and so remains a false positive, the orientation of the C-terminal helices appearing to be more a factor of the domain truncation than an overly favourable interchange of structures.

The high concentrations of expressed protein together with the non physiological conditions (such as low pH) that are sometimes used to achieve crystallisation of protein is thought to be a factor attributable to several cases of domain swapping (Bennett *et al*, 1995; Liu and Eisenberg, 2002). Therefore a degree of reticence over the biological relevance of some swapped structures must be applied. It is important to keep this in mind when considering domain swapping as an evolved mechanism for oligomer formation and potential regulation of protein function *in vivo* rather than an artifact of analysis.

The search for closed monomer structures of the domain-swapped proteins found by the search method came up with two potential matches. First a match to bovine interferon-gamma was found, however, this was actually found to be a fused form of the swapped homodimer. Second, a search for a monomeric form of the homo-dimeric mannose specific snowdrop lectin identified a monomeric two-domain lectin from bluebell, whose N-terminal domain showed significant homology to the snowdrop lectin domain. Although both the homologous domains have a similar 12 stranded barrel fold, the swapped beta-strand of the snowdrop domain remains in the N-terminal domain of the bluebell lectin where it is connected to the domain-linking peptide spanning between the domains.

A general analysis of swapped-domain-linkers was made for all the structures in the data set. The properties of these swapped-domain linkers were also compared to the characteristics found for domain-linking peptides made in Chapter 2. Swapped-domain linkers can be seen to act in two ways; in the closed monomer form the swapped-domain linker region can appear as just another part of protein structure, for example an all coil linker may appear as general coil. However in the open monomer the swapped-domain linker is found in a different conformational state bridging the gap between the swapped-domain and main domain. The ability to change conformations is clearly an essential property of the swapped-domain linker

and such properties are also essential in many domain-linkers that are often responsible for facilitating large conformational changes between domains in multi-domain structures (Gerstein *et al.*, 1994).

The assessment of the secondary structures formed by the swapped-domain linkers showed them to be mainly coil which is similar to the findings for domain-linkers, and is not surprising as coil residues are amongst the most flexible regions in protein structure. Of the 47 swapped-domain linkers analysed, 6 were found to consist of only helical or only strand residues. It was also shown that the secondary structure states varied between the closed and open monomer conformations. Interactions made by the *bona fide* swapped-domain linkers in the swapped oligomer can lead to sheet formation, in place of coil found in the closed monomer. Such new interactions made by the linker in the swapped-dimer may favour oligomerisation helping to overcome the associated loss of entropy.

Though a smaller sample set than the domain-linkers, the distribution of swapped-domain linker lengths still showed considerable variation from 3 to 22 residues, the most common length being 4 residues compared to 8 for domain linkers. An average length of 6.8 residues was found, shorter than the average length of 9.8 residues found for domain-linkers (Chapter 2). The length of the swapped-domain linker has been shown to be an important factor for domain swapping, where the addition or deletion of linker residues affect or hinder domain swapping (Murray *et al.*, 1995; Albright *et al.*, 1996).

The average residue pair C-alpha extension showed the swapped-domain linkers to be more extended than domain-linkers, but not as extended as non-linker coil. Comparisons were also made between the closed and open monomer conformation of the *bona fide* swapped-domain linkers. Here the change in conformation between the two states was suggested by the larger mean residue pair C-alpha extension of 2.37 Å of the open monomer, compared to 2.05 Å of the closed monomer. This also suggests that the swapped-domain linker tend to achieve a less-strained conformation in the swapped state favouring oligomerisation.

Solvent accessibility characteristics (measured as relative solvent accessibility) of swapped-domain linker residues showed them to be on average as accessible as coil and domain-linker residues, with a mean relative solvent accessibility of 33.1% exposed surface area. Though some regions of swapped-domains become buried when 'invading' the neighbouring subunit, in most domain-

swapped oligomers the majority of residues are found on the surface of the protein complex.

The exposed nature of the swapped-linker residues is supported by observations of their amino acid propensities. Hydrophobic residues were not favoured (with the exception of proline), whilst charged and polar residues are tolerated or preferred. The amino acid propensities of domain-swapped linkers were compared to those found in non-linker coil where differences included an increased preference for histidine and lysine and a decreased preference for glycine in the swapped-domain linkers. Comparison of the domain-swapped linker residue propensities to those obtained for the domain-linkers in Chapter 2 again showed some of the biggest differences for histidine and lysine both of which are more favoured in the swapped-domain linkers. Interestingly tryptophan, the least favoured residue in domain-linkers, occurs much more frequently in swapped-domain linkers (propensity of 0.57 in domain-linkers compared to 1.15 in swapped-domain linkers). Histidine may be favoured because its aromatic side-chain can act as both a hydrogen-bond donor (the nitrogen bonded to a hydrogen) and a hydrogen-bond acceptor (via the other nitrogen in the aromatic ring). This versatility in hydrogen-bond formation may be a factor in this residue's high occurrence in swapped-domain linkers allowing the different conformations of the open and closed monomer state to be tolerated.

Proline was found to be the most favoured of the residues in the swapped-domain linkers, mirroring the findings for the domain-linkers (Chapter 2). As has been discussed in Chapter 2, a high frequency of proline residues will confer turn propensity to the polypeptide chain, facilitating changes in direction. The importance of proline has been demonstrated by the mutation of two proline residues in the linker region of the p13suc1 protein. Mutation of residues in the linker has also been shown to shift the equilibrium between the monomer or dimer state (Bergdoll *et al.*, (1997).

Analysis of hydrogen-bond formation internal to the swapped-domain linkers showed there were 0.51 hydrogen-bonds per residue (in the open monomer state), higher than the value of 0.39 calculated for the domain-linkers in Chapter 2. Comparison of the average number of internal linker hydrogen-bonds per residue pair for the *bona fide* swapped-domain linkers in the closed and open monomer state calculated that over a third as many hydrogen bonds are made on average in the

closed conformation (mean of 0.91 hydrogen-bonds per residue pair), compared to the open conformation (mean of 0.58 hydrogen-bonds per residue pair). This supports the C-alpha extension calculations and observations of the structures themselves, that the linkers are more extended in the open conformation, and therefore have less chance for formation of internal hydrogen-bonds. Measuring the change in number of hydrogen-bonds between the *bona fide* swapped-domains in the closed and open monomer conformations showed only small changes in hydrogen-bonds made. Cyanovirin-N antiviral protein was an exception possibly due to the degree of interchange of the swapped-domain, that may be the cause of some change in structure in comparison to the closed monomer form. A structural alignment of these proteins would be required to confirm this. The sizes of the swapped-domain interfaces were found to vary considerably, from just under 1600 $Å^2$ to nearly 4200 $Å^2$. However the size of the swapped-domain interface was unrelated to the size of the swapped-domain.

The analysis of the swapped-domain linkers has shown that they tend to be exposed, usually coil in structure with a preference for polar and charged residues. The linkers are capable of changes in their conformation that enables the formation of oligomers from monomer subunits by the exchange of structure. The high propensity for proline will enable large changes in direction of the polypeptide chain whilst the preference for residues such as histidine will enable a degree of versatility in the hydrogen bonds that can be formed within the linkers in their open and closed conformations. From this analysis it appears many of the properties of swapped-domain linkers are similar to domain linkers – the high proline content demonstrating the importance of this residue in sections of proteins that can mediate conformational changes.

The assembly of protein monomers to form domain-swapped oligomers must over-come the energy barrier associated with the loss of entropy of independent monomer subunits (Newcomer, 2002). Because the interaction interface of the swapped-domain is almost identical in the closed and open conformations, favourable changes in the swapped-linker region may be enough to allow domain swapping to occur. The changes in the properties of the swapped-domain linker, such as the increase in C-alpha extent, demonstrating a less constrained conformation, and decrease in internal hydrogen-bonding may play a part in this. Newly formed

interactions made by the swapped-domain linker at the interface were also shown by the increase in strand residues in the linkers of open monomer compared to their closed monomer counterparts. Such interactions may also be favourable enough to tilt the equilibrium towards a domain-swapped state (Newcomer, 2002).

Several experiments have effected domain swapping by engineering the swapped-domain linker. Shortening the linker has lead to the formation of domain-swapped dimers in a number of proteins, including staphylococcal nuclease where the deletion of residues caused a helix to extend from the monomer, and in turn swap with a neighbouring subunit (Green *et al*, 1995). The lengthening of swapped-domain linkers has been shown to convert swapped-domain structures to a monomer form, for example the DNA-binding protein Cro repressor protein (Albright *et al.* 1996). This protein forms a stable dimer in solution, Figure 6.8a shows the structure (Ohlendorf *et al.*, 1998). Each dimer subunit contains a helix-turn-helix motif characteristic of many prokaryotic DNA binding proteins. Interaction of the C-terminal strands of each subunit is essential for dimerisation and provide the correct distance and orientation between the helix-turn-helix motif to allow high affinity binding to the DNA double helix (Branden and Tooze, 1999). A monomeric form of the Cro protein was engineered (Albright *et al.* 1996), and its structure subsequently determined as shown in Figure 6.8b. This was achieved by inserting five amino acids into the wild type Cro sequence. This insertion forms a beta turn allowing the following beta strand to mimic the position of the 'swapped' beta strand in the wild-type Cro dimer, allowing the C-terminal residues to bind within the main domain hydrophobic core (Albright *et al*, 1996). The wild-type C-terminal strand is not long enough to allow for it to turn back and its C-terminal residues to bind to its own core resides. Thus the monomeric form would be highly unlikely to be seen *in vivo.* However the engineering of potential swapped-domain linker sites in monomer chains, as described for Cro, could mean, in principal, that any protein could be made to domain swap.

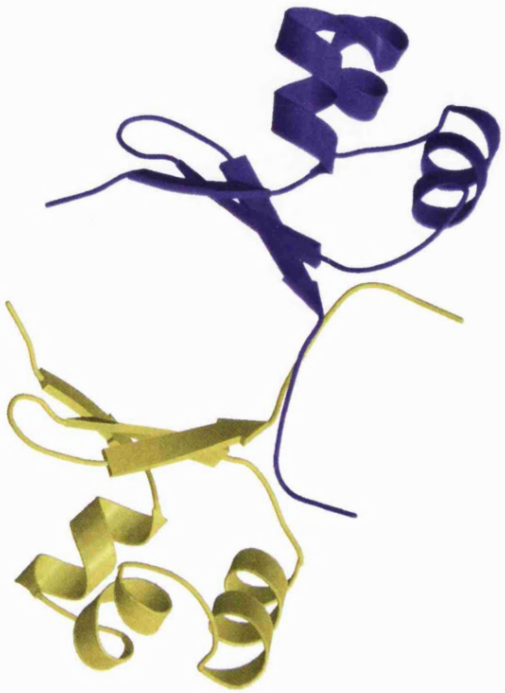
The research for domain-swapped proteins in the PDB gave two examples of putative domain-swapped viral proteins as shown in Figures 6.4i and j. It can be seen that the trimeric structure of the head-binding domain of P22 tailspike and the trimeric crystal structure of the feline immunodeficiency virus DUTP Pyrophosphatase both show exchanged regions of chain. A further example of a

Figure 6.8 Structures of homo-dimeric and engineered monomeric bacteriphage lambda Cro

a) Wild-type Cro forms a stable dimer in solution (Ohlendorf *et al.*, 1998).

b) A monomeric form of the Cro protein has been engineered and its structure subsequently determined (Albright *et al.*, 1996). This was achieved by inserting five amino acids into the wild type Cro sequence.

a)



b)

candidate for domain swapping used in this analysis is the simian virus 40 (SV40) coat protein oligomer (Stehle *et al*, 1996), that functions as a viral coat protein. The SV40 coat protein C-terminal arms, made up of a helix and strands 'invade' a neighbouring subunit, forming part of the jelly-roll domain of the invaded structure. The neighbouring subunits are viral coat building blocks tied together by the C-terminal arms rather than being cemented together across preformed complementary interfaces (Branden and Tooze, 1999). It is interesting to note that the exchange of peptide between viral proteins has been observed in a number of cases - polypeptide arms extending over or under domains of neighbouring subunits, intertwining with others, are a remarkable feature of virus structures (Steinbacher *et al.*, 1997). Interestingly none of these three viral domain-swapped proteins have a know closed monomer structure. Though these viral oligomers clearly have swapped-domains, it important to consider if closed monomer structures would exist *in vivo*, for example in the case of SV40 capsid protein. If the closed monomer occurred, would it form a biologically relevant protein? It may be possible that these proteins initially fold as monomers and then swap domains, or it is also possible that oligomer formation occurs as they fold.

Perhaps for such cases where it is unclear as to whether a closed monomeric form would exist in nature, the domain swapping definition should be modified slightly to highlight structures that swap domains but the swapped interface may not be replicating a intra-domain interaction present in a closed monomer structure as one may not exist.

Many avenues of future work exist in the field of domain-swapped proteins. An analysis of the conservation of linker residues should be made, especially focusing on the linkers of *quasi* domain-swapped proteins. Differences in the linker residues of these homologous closed and open monomers must play some part in allowing one form of the homologue to swap, and the other to remain as a monomer. It would also be of interest to consider the main domain interface that is formed between the swapped oligomers. Though the subunits are identical it would be of interest to see if the surface residue propensities at the interface site have evolved to form favourable residue interactions, as is the case for non domain-swapped homo-oligomers.

Further work into the prediction of domain-swapped structures is also worthwhile. The search method here is limited in only identifying non-globular

domain-swapped structures. Identifying the exposed swapped-domain linkers of the open monomers may allow globular domain-swapped structures to be identified. Identifying loops that may be capable of acting as a hinge loop has been addressed by calculation loop contact distance values from structure co-ordinates, outlined in the paper by Linhananta *et al.*, (2002). Combining these methods may provide a useful prediction method.

Chapter 7

Final discussion

The identification of the domain content and corresponding boundaries within a protein sequence is an important first step in many areas of molecular biology. This thesis has described two new methods for domain prediction; A method to predict domains and their boundaries from protein sequence using predicted secondary structure has been described (DomSSEA) and a method that delineates protein domains by post-processing PSI-BLAST sequence alignments has also been developed (DPS). Both methods have also been combined, with DPS acting as a pre-filter stage to DomSSEA.

The DPS (Domains Parsed by Sequence) method was developed to use PSI-BLAST to identify homologues to a query sequence, and then delineate domains from the N- and C-termini of locally aligned sequence fragments. In cases where a domain boundary is found, DPS predicts the target sequence as multi-domain. Just over 50% of the multi-domain sequences containing continuous domains were correctly predicted by DPS, with a selectivity of 79%. In turn, nearly 56% of the continuous domain boundaries *predicted* by DPS were correct within ±20 residues of a true domain boundary (with a corresponding sensitivity of 30%). Clearly not all domain boundaries predicted by DPS coincided with the domain boundaries given in the CATH domain database. A closer examination of the false-positive boundary predictions and how they might relate to corresponding regions of protein structure would be worthwhile. Also, although the DPS method was designed to be a simple approach to domain assignment from sequence, boundary prediction might be improved by the creation of a more accurate multiple-alignment of the PSI-BLAST (Altschul *et al.*, 1997) hits to the query sequence before the domain identification is made.

The results for prediction of discontinuous domain boundaries using DPS were disappointing, but perhaps to be expected, as this method was not designed with the specific intention of recognising such domains. Since most protein structure prediction methods are aimed at the assignment of continuous domains it was felt that it was important to accurately assign continuous domain boundaries. Therefore, although discontinuous domain assignment was considered, the benchmarking of DomSSEA and DPS focussed on the assignment of continuous domains. Because of this, for both methods described (DPS and DomSSEA), the accuracy of discontinuous domain assignment was below that for continuous domain assignment

mainly due to the inherent complexity of recognising discontinuous domain fragments.

Many discontinuous domains are thought to have arisen from domain insertion (Russell, 1994), and are therefore found as two or more individual segments along the amino acid sequence. Such discontinuous domain fragments might be identified by searching for segments of chain that are separated in sequence, but that align to the same sequences, whilst the intervening residues between them do not. A similar procedure has been carried out by George and Heringa (2002). However, the identification of discontinuous segments by this method does not appear to improve corresponding boundary prediction, since the discontinuous domain boundary prediction by George and Heringa (2002) does not appear to be substantially better than that calculated for DPS in this study.

The DomSSEA method is based upon the automatic analysis of the predicted secondary structure of a query sequence. A human sequence analyst will usually resort to this method in an attempt to parse a protein into domains when homology-based approaches have been unsuccessful. DomSSEA acts as a simple fold recognition algorithm based upon the mapping of predicted secondary structures to observed secondary structure patterns of proteins of known 3-D structure. DomSSEA was able to successfully predict domain content of a protein with the prediction of continuous domain boundaries achieving a sensitivity of 31%, with a corresponding selectivity of 32%. DomSSEA was able to predict both the correct domain content *and* domain boundaries for 25% of the multi-domain test set (±20 residues). Although this method is not 100% accurate it could act as a rapid pre-filtering stage in automatic genome annotation and threading methods where domain boundaries cannot be located purely from comparative sequence analysis. A further benefit of DomSSEA is the fact that a number of predictions can be give, from which domain assignments can be chosen and tested.

Combining DPS and DomSSEA gave an approach in which domain assignment by sequence comparison could first be made by DPS leaving the more difficult assignment cases to be predicted by DomSSEA. This combined approach to domain assignment by DPS and DomSSEA gave correct domain content predictions for 73% of the single and multi-domain test set, with 76% of the multi-domain chains

correctly predicted to contain more than one domain. Furthermore, correct assignments were made for over 55% of the continuous domain boundaries (with a selectivity of 45%) using both methods (± 20 residues).

A number of additional methods and improvements could be made for a combined domain prediction algorithm. For instance, the inclusion of a pre-stage to DPS in which the query sequence is searched against a domain database such as Pfam would be useful, in order to detect more obvious matches. The inclusion of transmembrane prediction (Jones, 1994) and the detection of internal sequence repeats which may correspond to domains (Heringa, 1998) could be used to improve the domain prediction process. Using the length and composition of domain linkers, either applied in hidden Markov models or used to train a neural network in order to predict domain boundaries might also be worthwhile. The analysis of domain linkers showed that they tend to be mainly coil, favouring proline residues, although all-helical and all-strand linkers do also exist. Even if direct prediction of linkers using such characteristics did not achieve a high level of accuracy, they may still act as a useful post-processing stage for domain boundaries predicted by DPS or DomSSEA. This could include the analysis of predicted domain boundaries and their degree of agreement with the characteristics observed for domain linkers. Post-processing stages for predicted domains could also be implemented. For example, a strong relationship was found between sequence length and surface area. The surface area of a putative protein domain could therefore be predicted and compared to the observed distribution of the expected surface area for the given chain length.

All the domain definitions used to test the methods used in this study were based on those given in the CATH domain database. CATH domains are identified using a number of algorithms, and in cases of disagreement, manual domain boundary assignments are made (Orengo *et al.*, 1997). As such, benchmarking domain prediction methods on CATH is not reliant on a single assignment procedure providing the 'true' domain boundaries. However there are disagreements in domain assignments between classification databases such as SCOP and CATH (Hadley and Jones, 1999). The construction of a test data set based on chains which are given similar domain assignments by both of these databases might be a better approach.

Domain assignments can differ because methods of assignment can be different according to their overall definition of a structural domain. In SCOP,

domain recurrence is important because a domain will only be assigned if it is known to exist as an individual unit (Murzin *et al.*, 1995). This means that some proteins in SCOP could be further subdivided into smaller, more compact domains, as might be the case when using CATH. Assigning domains on the criteria of compactness however may be problematic since defining the level of compactness used may be difficult. Domains can be defined as either units of compact structure, or as units of sequence homology. Both are acceptable definitions although the two may result in different domain assignments (Marchler-Bauer *et al.*, 2002). There are several levels of structural independence, where a domain may fold independently, whilst still containing smaller subdomains. In addition, small units of structure can correspond to autonomous folding units, which do not necessarily correspond to domains. Though Wetlaufer (1973) described domains as independent folding units this definition may not agree with all domains assigned as independent structural units.

The assignment of domains will never be perfect because ultimately it will always be a subjective process, and consequently, there is no universally applied domain definition. This is a problem for domain prediction methods, especially as they become more accurate. For example when comparing predictions that have been made using different criteria for a domain, or results that have been benchmarked on different domain classifications, differences will always be present.

The delineation of domain boundaries is a difficult problem because it relates to protein folding and understanding the principals by which amino acid sequence confers protein structure. The protein folding problem is still a key issue to be resolved and the accurate prediction of protein structure will be an essential tool in understanding structure and function and is of great importance in this post-genomic age. The formation of a hydrophobic core plays a key role in the folding of protein domains. It was shown that the design of simple prediction methods based upon the percentage or distribution of hydrophobic residues within proteins does not appear to be a feasible method. There appeared to be no obvious separation in the percentage of hydrophobic residues in single and multi-domain proteins, whilst the distribution of hydrophobic residues in sequence appears to be random. The use of factors such as the percentage of exposed residues and some measure of hydrophobicity may however have some use in domain prediction, although not in isolation. The approach used for DomSSEA was to try to identify the more conserved protein

secondary structure corresponding to protein structural folds which form the hydrophobic structural core of protein domains.

In this study it has been shown that domain linkers may play an important role in the folding of multi-domain proteins. Unstructured linkers were found to have a high frequency of proline residues, isolating the linker from the adjacent. The presence of the *cis*-proline residues in domain linkers is known to slow protein folding because the *cis-trans* isomerisation of proline peptides is intrinsically slow. It would be interesting to determine whether multi-domain proteins, which undergo post-translational folding in the cytosol, are more likely to have proline rich linker regions.

The rate of domain folding may also be essential in the efficient formation of multi-domain structures. If one domain of a two domain protein folds quickly and the other more slowly, the likelihood of the two structures misfolding by aggregation may be reduced. Work by Plaxco and colleagues (1998) found that the folding rate of single domain chains correlates well with the mean sequence separation between residues within a given cut-off in 3D structure. Furthermore, proteins with a large number of local contacts fold faster than proteins with more long-range non-local contacts. The analysis and possible prediction of contact order and folding rate in multi-domain proteins may be of some use in domain analysis and prediction.

Protein domain-swapping provides a mechanism for oligomer formation from identical protein subunits. In this study it has been shown that swapped domains can differ considerably in size and secondary structure content. The analysis of swapped-domain linkers has shown that, like domain-linkers, these swapped-domain linkers have a preference for proline residues. Again, proline residues will confer both independence and some turn propensity to a swapped-domain linker. The ability to form oligomers by swapping identical structures, and thus replicating *intra*-domain interfaces with *inter*-domain interfaces is a fascinating mechanism for protein complex formation. Structural studies have suggested that domain swapping may occur through the partial unfolding of the monomer to an intact core structure, leaving the terminal domains free to move and swap (Bennet *et al.*, 1995; Liu *et al.*, 2001). Understanding the mechanism and cellular conditions required for the unfolding and refolding of these proteins will be of great interest, especially in light of the possible association of domain-swapping and amyloid diseases. The search

method implemented in Chapter 6 was limited in that it only identified non-globular domain-swapped structures. The identification of swapped-domain linkers in their monomer form may allow globular domain-swapped structures to be identified. Further work regarding the prediction of domain-swapped structures is clearly needed.

Chapter 8

References

Ahmad, K.F., Engel, C.K. and Prive, G.G. (1998). Crystal structure of the BTB domain from PLZF. *Proc. Natl. Acad. Sci. USA.* **95**, 12123-12128.

Albright, R.A., Mossing, M.C. and Matthews, B.W. (1996). High-resolution of an engineered Cro monomer shows changes in conformation relative to native dimer. *Biochemistry*, **35**, 735-742.

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

Altschul, S.F. and Koonin, E.V. (1998). Iterated profile searches with PSI-BLAST – a tool for discovery in protein databases. *Proteins Struc. Funct. Genet.* **23**, 444-447.

Anfinsen, B.C., Haber, E., Sela, M. and White, F.H. (1961). The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. *Biochemistry*, **47**, 1309-1314.

Anfinsen, C.B. (1973). Principles that govern the folding of protein chains. *Science*, **181**, 223-230.

Apic, G., Gough, J. and Teichmann,S.A. (2001). An insight into domain combinations. *Bioinformatics*, **17,** (Suppl.), 83-89.

Apweiler, R., Attwood, T.K., Bairoch, A., Bateman, A., Birney, E., Biswas, M., Bucher, P., Cerutti, L., Corpet, F., Croning, M.D., Durbin, R., Falquet, L., Fleischmann, W., Gouzy, J., Hermjakob, H., Hulo, N., Jonassen, I., Kahn, D., Kanapin, A., Karavidopoulou, Y., Lopez, R., Marx, B., Mulder, N.J., Oinn, T.M., Pagni, M., Servant, F., Sigrist, C.J. and Zdobnov, E.M. (2000). InterPro-an integrated documentation resource for protein families, domains and functional sites. *Bioinformatics*, **16**, 1145-1150.

Arai, R., Ueda, H., Kitayama, A., Kamiya, N. and Nagamune, T. (2001). Design of the linkers which effectively separate domains of a bifunctional fusion protein. *Protein Eng.* **8**, 529-532.

Argos, P. (1990). An investigation of oligopeptides linking domains inprotein tertiary structures and possible candidates for general gene fusion. *J. Mol. Biol.* **211**, 943-958.

Arnes, J,G., Augustine, J.G., Moras, D. and Franklyn, C.S. (1997). The first step of aminoacylation at the atomic level in histidyl-tRNA synthetase. *Proc. Nat. Acad. Sci. USA*, **97**, 7144-7154.

Arold, S.T., Hoellerer, M.E. and Noble, E.M. (2002). The structural basis of localisation and signaling by the focal adhesion targeting domain. *Structure*, **10** 319-327.

Arvai, A.S., Bourne, Y., Hickey, M.J. and Tainer, J.A. (1995). Crystal structure of the human cell cycle protein CksHs1: single domain fold with similarity to kinase N-lobe domain. *J. Mol. Biol.* **249**, 835-842.

Attwood, T.K. (2002). The PRINTS database: a resource for identification of protein families. *Brief. Bioinform*, **3**, 252-263.

Aurora, R., Creamer, T.P., Srinivasan, R. and Rose, G.D. (1997). Local interactions in protein folding: lessons from the alpha-helix. *J. Biol. Chem.* **272**, 1413-1416.

Bairoch, A. and Apweiler, R.(2000). The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucl. Acids Res.* **28**, 45-48.

Banner, D.W., Bloomer, A.C., Petsko, G.A., Phillips, D.C., Pogson, C.I., Wilson, I.A., Corran, P.H., Furth, A.J., Milman, J.D., Offord, R.E., Priddle, J.D. and Waley, S.G. (1975). Structure of chicken muscle triose phosphate isomerase determined crystallographically at 2.5 Angstrom resolution using amino acid sequence data. *Nature*, **255**, 609-614.

Bateman, A., Birney, E., Durbin, R., Eddy, S.E., Lowe, K.L. and Sonnhammer, E.L. (2000). The Pfam protein families database. *Nucl. Acids Res.* **28**, 263-266.

Bax, B., Lapatto, R., Nalini, V., Driessen, H., Lindley, P.F., Mahadevan, D., Blundell, T.L. and Slingsby, C. *Nature.* **347**, 776-780.

Bennett, M. J. and Eisenberg, D. (1994). Refined structure of monomeric diphtheria toxin at 2.3 Å resolution. *Protein. Sci.* **3**, 1464-1475.

Bennett, M. J., Choe, S. and Eisenberg, D. (1994). Refined structure of dimeric diphtheria toxin at 2.0 Å resolution. *Protein. Sci.* **3**, 1444-1463.

Bennett, M.J., Schlunegger, M.P. and Eisenberg, D. (1995). 3D domain swapping: a mechanism for oligomer assembly. *Protein Sci.* **4**, 2455-2468.

Bennett, W.S. and Huber, R. (1984). Structural and functional aspects of domain motions in proteins. *C R C Crit. Rev. Biochem.* **15,** 291-384.

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

Bewley, C.A., Gustafson, K.R., Boyd, M.R., Covell, D.G., Bax, A., Clore, G.M.and Gronenborn, A.M. (1998). Solution structure of cyanovirin-N, a potent HIV-inactivating protein. *Nat. Struct. Biol.* **5,** 571-578.

Bianchet, M.A., Bains, G., Pelosi, P., Pevsner, J., Snyder, S. H., Monaco, H.L. and Amzel, L.M. (1996). The three-dimensional structure of bovine odorant binding protein and its mechanism of odor recognition. *Nat. Struct. Biol.* **3**, 934-939.

Birney, E., Clamp, M., Kraspcyk, A., Slater, G., Hubbard, T., Curwen, V., Stabenau, A., Stupka, E., Huiniecki, L. and Potter, S. (2001). Ensemble: a multi-genome computational platform. *Am. J. Hum. Genet.*, **69**, 219.

Blake, C.C.F., Koenig, D.F., Mair, G.A., North, A.C.T., Phillips, D.C. and Sarma, V.R. (1965). Structure of hen egg-white lysozyme. *Nature*, **206**, 757-761.

Blake, C.C.F. (1978). Do genes-in-pieces imply proteins-in-pieces? *Nature*, **273**, 267.

Blundell, T.L. and Mizuguchi, K. (2000). Structural genomics: an overview. *Prog. Biophys. Mol. Biol.* **73**, 289-295.

Bocskei, Z., Groom, C.R., Flower, D.R., Wright, C.E., Phillips, S.E., Cavaggioni, A., Findlay, J.B. and North, A.C. (1992). Pheromone binding to two rodent urinary proteins revealed by X-ray crystallography. *Nature*, **360**, 186-188.

Bode, W., Engh, R., Musil, D., Thiele, U., Huber, R., Karshikov, A., Brzin, J., Kos, J. and Turk, V. (1988). The 2.0 Å X-ray crystal structure of chicken egg white cystatin and its possible mode of interaction with cysteine proteinases. *EMBO J.* **7**, 2593-2599.

Bork, P. (1991). Shuffled domains in extracellular proteins. *FEBS Lett.* **286**, 47-54.

Bourne, Y., Arvai, A.S., Bernstein, S.L., Watson, M.H., Reed, S.I., Endicott, J.E., Noble, M.E., Johnson, L.N. and Tainer, J.A. (1995). Crystal structure of the cell cycle-regulatory protein suc1 reveals a beta-hinge conformational switch. *Proc. Natl. Acad. Sci.* **92**, 10232-10236.

Branden, C.I. and Tooze, J. (1999). *Introduction to protein structure*. Garland, New York, Second edition.

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

Buckle, A.M. and Fersht, A.R. (1994). Subsite binding in an RNase: structure of a barnase-tetranucleotide complex at 1.76 Å resolution. *Biochemistry*, **33**, 1644-1653.

Busetta, B. and Barrans, Y. (1984). The prediction of protein domains. *Biochim. Biophys. Acta.* **790**, 117-124.

Cambell, I.D. and Baron, M. (1991). The structure and function of protein modules. *Philos. Trans. R. Soc. Lond. (Biol.)* **332**, 165-170.

Campbell, I.D. and Downing, A.K. (1994). Building protein structure and function from modular units. *Trends. Biotechnol.* **12**, 168-172.

Carugo, O. (2001). Detection of breaking points in helices linking separate domains. *Proteins Struct. Funct. Genet.* **42**, 390-398.

Chattopadhyaya, R., Meador, W.E. and Quiocho, F.A. (1992). Calmodulin structure refined at 1.7 Å resolution. *J. Mol. Biol.* **228**, 1177-1192.

Chothia, C. (1975) Structural invariants in protein folding. *Nature.* **254**, 304-308.

Christendat, D., Yee, A., Dharamsi, A., Kluger, Y., Savchenko, A., Cort, J.R., Booth, V., Mackereth, C.D., Saridakis, V., Ekiel, I., Kozlov, G., Maxwell, K.L., Wu, N., Mcintosh, L.P., Gehring, K., Kennedy, M.A., Davidson, A.R., Pai, E.F., Gerstein, M., Edwards, A.M. and Arrowsmith, C.H. (2000). Structural Proteomics of an Archaeon. *Nat. Struct. Biol.* **7**, 903-909.

Concha, N.O., Rasmussen, B.A., Bush, K. and Herzberg, O. (1996). Crystal structure of the wide-spectrum binuclear zinc beta-lactamase from *Bacteriodes fragilis. Structure*, **4**, 823-836.

Copley, R.R. and Bork, P. (2000). Homology among (beta/alpha)(8) barrels: implications for the evolution of metabolic pathways. *J. Mol. Biol.* **303**, 627-641.

Corpet F., Gouzy J. and Kahn D. (1998). The ProDom database of protein domain families. *Nucl. Acids Res.* **26**, 323-326.

Creighton, T.E. (1992). *Proteins: Structures and molecular properties.* Freeman, New York, Second edition.

Creighton, T.E. (1988). Disulphide bonds and protein stability. *BioEssays*, **8**, 57-63.

Crestfield, A.M., Stein, W.H. and Moore, S. (1962). On the aggregation of bovine pancreatic ribonuclease. *Arch. Biochem. Biophys.* **1**, 217-222.

Crippen, G.M. (1978). The tree structural organisation of proteins. *J. Mol. Biol.* **126**, 315-332.

Das, S. and Smith, T.F. (2000). Identifying nature's protein Lego set. *Adv. Protein Chem.* **54**, 159-183.

Davidson, J.N., Chen, K.C., Jamison, R.S., Musmanno, L.A. and Kern, C.B. (1993). The evolutionary history of the first three enzymes in pyrimidine biosynthesis. *Bioessays*, **15**, 157-164.

Dayhoff, M.O., Schwartz, R.M. and Orcutt, B.C. (1978). Atlas of protein sequences and structure. **5**, 345-352. Nat. Biomed. Res. Found., Washington DC.

de Souza, S.J., Long, M., Klein, R.J., Roy, S., Lin, S. and Gilbert, W. (1998). Toward a resolution of the introns early/late debate: only phase zero introns are correlated with the structure of ancient proteins. *Proc. Natl. Acad. Sci. USA.* **95**, 5094-5099.

Dean, M. and Allikmets, R. (1995). Evolution of ATP-binding cassette transporter genes. *Curr. Opin. Genet. Dev.* **5**, 779-785.

Diederichs, K., Jacques, S., Boone, T. and Karplus, P.A. (1991). Low-resolution structure of recombinant human granulocyte-macrophage colony stimulating factor. *J. Mol. Biol.* **221**, 55-60.

Dill, K.A. (1985). Theory of folding and stability in globular proteins. *Biochemistry,* **24**, 1501-1509.

Dill, K.A. and Chan, H.S. (1997). From Levinthal to pathways to funnels. *Nat. Struc. Biol.* **4**, 10-19.

Dobson, C.M. and Hore, P.J. (1998). Kinetic studies of protein folding using NMR spectroscopy. *Nat. Struct. Biol.* **5**, 504-507.

Dobson, C.M. and Karplus, M. (1999). The fundamentals of protein folding: bringing together theory and experiment. *Curr. Opin. Struct. Biol.* **9**, 92-101.

Donaldson, L. and Capone, J.P. (1992). Purification and characterization of the carboxyl-terminal transactivation domain of Vmw65 from herpes simplex virus type 1. *J. Biol. Chem.* **267**, 1411-1414.

Doolittle, W.F. (1978). Genes in pieces: were they ever together? *Nature*, **272**, 581-582.

Dumas, P., Bergdoll, M., Cagnon, C. and Masson, J.M. (1994). Crystal structure and site directed mutatgenesis of bleomycin resistance protein and their significance for drug sequestering. *EMBO J.* **13**, 2483-2493.

Eisenberg, D., Marcotte, E.M., Xenarios, I. and Yeates, T.O. (2000). Protein function in the post-genomic era. *Nature*, **405**, 823-826.

Ellis, R.J. and Hartl, F.U. (1999). Principles of protein folding in the cellular environment. *Curr. Opin. Struct. Biol.* **9**, 102-110.

Enright, A.J., Ouzounis, C.A. (2000). GeneRAGE: a robust algorithm for sequence clustering and domain detection. *Bioinformatics*, **16**, 451-457.

Fischer, D., Elofsonnn, A., Rychlewski, L., Pazos, F., Valencia, A., Rost, B., Ortiz, A.R. and Dunbrack, R.L. (2001) CAFASP2: The second critical assessment of fully automated structure prediction methods. *Proteins Struc. Funct. Genet.* **45(S5)**, 171-183.

Fisher, H.F. (1965). An upper limit to the amount of hydration of a protein molecule. A corollary to the limiting law of protein structure. *Biochim. Biophys. Acta.* **109**, 544-550.

Frydman, J., Erdjument-Bromage, H., Tempst, P. and Hartl, F.U. (1999). Co-translational domain folding as the structural basis for the rapid *de novo* folding of firefly luciferase. *Nat Struct Biol.* **6**, 697-705.

George, R.A. and Heringa, J. (2002a). SnapDRAGON: a method to delineate protein structural domains from sequence data. *J. Mol. Biol.* **316**, 839-851.

George, R.A, and Heringa, J. (2002b). Protein domain identification and improved sequence similarity searching using PSI-BLAST. *Proteins Struct. Funct. Genet.* **48**, 672-681.

Gerstein, M., Lesk, A.M., Baker, E.N., Anderson, V., Norris G. and Chothia C. (1993). Domain Closure in Lactoferrin: Two Hinges produce a See-saw Motion between Alternative Close-Packed Interfaces. *J. Mol. Biol.* **234**, 357-372.

Gerstein, M., Lesk, A.M. and Chothia, C. (1994). Structural mechanisms for domain movements in proteins. *Biochemistry*, **33**, 6739-6749.

Gerstein, M. (1997). A structural census of genomes: comparing bacterial, eukaryotic and archaeal genomes in terms of protein structure. *J. Mol. Biol.* **274**, 562-576.

Gerstein, M. and Krebs, W. (1998). A database of macromolecular motions. *Nucl. Acids Res.* **26**, 4280-4260.

Ghelis, C. and Yon, J.M. (1979). Conformational coupling between structural units. A decisive step in the functional structure formation. *C R Seances Acad. Sci. D,* **289**, 197-199.

Gilbert, W. and Glynias, M. (1993). On the ancient nature of introns. *Gene*, **135**, 137-144.

Gokhale, R,S and Khosla, C. (2000). Role of linkers in communication between protein modules. *Curr. Opin. Chem. Biol.* **4**, 22-27.

Gong, W., Hao, B., Mansy, S.S., Gonzalez, G., Gilles-Gonzalez, M.A. and Chan, M.K. (1998). Structure of a biological oxygen senseor: A new mechanism for heme-driven signal transduction. *Proc. Natl. Acad. Sci.* **95**, 15177-15182.

Gouzy, J., Corpet, F. and Kahn, D. (1999). Whole genome protein domain analysis using a new method for domain clustering. *Comput. Chem.* **23**, 333-340.

Gracy, J. and Argos, P. (1998). DOMO: a new database of aligned protein domains. *Trends Biochem. Sci.* **23**, 495-497.

Green, S.M., Gittis, A.G., Meeker, A.K. and Lattman, E.E. (1995). One-step evolution of a dimer from a monomeric protein. *Nat. Struct. Biol.* **2**, 746-751.

Hadden, J.M., Convery, M.A., Declais, A.C., Lilley, D.M.J. and Phillips, S.E.V. (2001). Crystal structure of holliday junction resolving enzyme T7 endonuclease I at 2.1 Å. *Nat. Struc. Biol.* **8**, 62-67.

Hadley, C. and Jones, D.T. (1999). A systematic comparison of protein structure classifications: SCOP, CATH and FSSP. *Structure Fold. Des.* **7**, 1099-1112.

Hakansson, M., Svensson, A.L., Fast, J. and Linse, S. (2001). An Extended Hydrophobic Core Induces EF-Hand Swapping *Protein. Sci.* **10**, 927-933.

He, M.M., Clugston, S.L., Honek, J.F. and Matthews, B.W. (2000). Determination of the Structure of Escherichia Coli Glyoxalase I Suggests a Structural Basis for Differential Metal Activation. *Biochemistry,* **39**, 8719-8721.

Hegvold, A.B. and Gabrielsen, O.S. (1996). The importance of the linker connecting the repeats of the c-Myb oncoprotein may be due to a positioning function. *Nucl. Acids Res.* **24**, 3990-3995.

Henikoff, J.G., Greene, E.A., Pietrokovski, S. and Henikoff, S. (2000). Increased coverage of protein families with the blocks database servers. *Nucl. Acids Res.* **28**, 228-230.

Henikoff S. and Henikoff J.G. (1992). Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. USA*, **89**, 10915-10919.

Henikoff, S., Greene, E.A., Pietrokovski, S., Bork, P., Attwood, T.K. and Hood, L. (1997). Gene families: the taxonomy of protein paralogs and chimeras. *Science*, **278**, 609-614.

Heringa, J. and Taylor, W. (1997). Three-dimensional domain duplication, swapping and stealing. *Curr. Opin. Struct. Biol.* **7**, 416-421.

Heringa, J. (1998) Detection of internal repeats: how common are they? *Curr. Opin. Struct. Biol.* **8**, 338-345.

Hester, G., Kaku, H., Goldstein, I.J. and Wright, C.S. (1995). Structure of mannose-specific snowdrop (*Galanthus nivalis*) lectin is representative of a new plant lectin family. *Nat. Struct. Biol.* **6**, 472-479.

Holm,L. and Sander,C. (1994). Parser for protein folding units. *Proteins Struct. Funct. Genet.* **19**. 256-268

Holm, L. and Sander, C. (1997). Dali/ FSSP classification of three-dimensional protein folds. *Nucl. Acids Res.* **25**, 231-234.

Holm, L. and Sander, S. (1998). Removing near-neighbour redundancy from large protein sequence collections. *Bioinformatics.* **14**, 423-429.

Hubbard, S.J. and Argos, P.A. (1996). Functional role for protein cavities in domain:domain motions. *J. Mol. Biol.* **261**, 289-300.

Hubbard, S.J. (1998). The structural aspects of limited proteolysis of native proteins. *Biochim. Biophys. Acta.* **1382**, 191-206.

Huynen, M., Doerks, T., Eisenhaber, F., Orengo, C., Sunyaev, S., Yuan, Y. and Bork, P. (1998). Homology-based fold predictions for Mycoplasma genitalium proteins. *J. Mol. Biol.* **280**, 323-326.

Ikura, M., Clore, G.M., Gronenborn, A.M., Zhu, G., Klee, C.B. and Bax, A. (1992). Solution structure of a calmodulin-target peptide complex bu multidimensional NMR. *Science*, **256**, 632-638.

Islam, S.A., Luo, J. and Sternberg, M.J.E. (1995). Identification and analysis of domains in proteins. *Protein Eng.* **8**, 513-525.

Janin, J. and Wodak. S.J. (1983). Structural domains in proteins and their role in the dynamics of protein function. *Prog. Biophys. Mol. Biol.* **42**, 21-78.

Janowski, R., Kozak, M., Jankowska, E., Grzonka, Z., Grubb, A., Abrahamson, M., Jaskolski, M. (2001). Human Cystatin C, an Amyloidogenic Protein, Dimerizes Through Three-Dimensional Domain Swapping. *Nat. Struct. Biol.* **8**, 316-320.

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

Jones, D.T., Taylor, W.R. and Thornton, J.M. (1994). A model recognition approach to the prediction of all-helical membrane protein structure and topology. *Biochemistry*, **33**, 3038-3049.

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

Jones, D.T. and Hadley, C. (2000). Threading methods for protein structure prediction. In Higgins, D and Taylor, W (eds), Bioinformatics, Sequence, structure and databanks. Oxford University Press, Oxford. pp. 1-13.

Jones, E.Y., Davis, S J., Williams, A.F., Harlos, K. and Stuart, D.I. (1992). Crystal structure at 2.8 A resolution of a soluble form of the cell adhesion molecule CD2. *Nature*, **360**, 232-239.

Jones, S., Stewart, M., Michie, A., Swindells, M., Orengo, C. and Thornton, J.M. (1998). Domain assignment for protein structure using a consensus approach: Characterisation and analysis. *Protein Sci.* **7**, 233-242.

Jones, S., Marin, A. and Thornton, J.M. (2000). Protein domain interfaces: characterisation and comparison with oligomeric protein interfaces. *Protein Eng.* **13**, 77-82.

Kabsch, W. and Sander, C. (1983). Dictionary of protein secondary structure; pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577-2637.

Kartha, G., Bello, J. and Harker, D. (1967). Tertiary structure of ribonuclease. *Nature*, **213**, 862-865.

Kauzmann, W. (1959). Some factors in the interpretation of protein denaturation. *Adv. Prot. Chem.* **14**, 1-63.

Khazanovich, N., Bateman, K., Chernaia, M., Michalak, M. and James, M. (1996). Crystal structure of the yeast cell-cycle control protein, p13suc1, in a strand-exchanged dimer. *Structure*, **4**, 299-309.

Kikuchi,T., Némethy,G. and Scheraga,H.A. (1988). Prediction of the location of structural domains in globular proteins. *J. Protein Chem.* **7**, 427-471.

Kishan, K.V., Scita, G., Wong, W.T., Di Fiore, P.P. and Newcomer, M.E. (1997). The SH3 domain of Eps8 exists as a novel intertwined dimer. *Nat. Struct. Biol.* **4**, 739-743.

Knaus, K.J., Morillas, M., Swietnicki, W., Malone, M., Surewicz, W.K. and Yee, V.C. (2001). Crystal Structure of the Human Prion Protein Reveals a Mechanism for Oligomerization. *Nat. Struct. Biol.* **8**, 770-774.

Koretke, K.K., Russell, R.B. and Lupas, A.N. (2002). Fold recognition from sequence comparisons. *Proteins. Struct. Funct. Genet.* (Suppl) **5**, 68-75.

Koshland, D.E. (1958). Application of a theory of enzyme specificity to protein synthesis. *Proc. Natl. Acad. Sci. USA*, **44**, 98-104.

Kraulis, P.J. (1991). MolScript; a program to produce both detailed and schematic plots of protein structures. *J. Appl. Crystallogr.* **24**, 946-950.
Kriventseva, E.V., Biswas, M. and Apweiler, R. (2001). Clustering and analysis of protein families. *Curr. Opin. Struct. Biol.* **11**, 334-339.

Kuhlman, B, O'Neill, J.W., Kim, D.E., Zhang, K.Y.J. and Baker, D. (2001). Conversion of monomeric protein L to an obligate dimer by computational protein design. *Proc. Nat. Acad. Sci.* ***98***, 10687-10691.

Kuroda, Y., Tani, K., Matsuo, Y. and Yokoyama, S. (2000). Automated search of natively folded protein fragments for high-throughput structure determination in structural genomics. *Protein Sci.* **9**, 2313-2321.

Labeit, S. and Kolmerer, B. (1995). Titins: giant proteins in charge of muscle ultrastructure and elasticity. *Science*, **270**, 293-296.

Langmuir, I. (1938). Protein denaturation. *Cold Spring Harb. Symp. Quant. Biol.* **6**, 159.

Larsen, T.M., Laughlin, L.T., Holden, H.M., Rayment, I. and Reed, G.H. (1994). Structure of rabbit muscle pyruvate kinase complexed with Mn2+, K+, and pyruvate. *Biochemistry*, **33**, 6301-6309.

Larsson, G., Svensson, L. A. and Nyman, P. O. (1996). Crystal structure of the Escherichia coli dUTPase in complex with a substrate analogue (dUDP). *Nat. Struct. Biol.* **3**, 532-538.

Lesk, A.M., Conte, L.L. and Hubbard, T.J.P. (2002). *Proteins. Struct. Funct. Genet.* (Suppl) **5**, 98-118.

Levinthal, C. (1968). Are there pathways for protein folding? *J. Chem. Phys.* **65**, 44-45.

Levitt, M. and Chothia, C. (1976). Structural patterns in globular proteins. *Nature.* **261**, 552-558.

Lewis, R.J., Brannigan, J.A., Muchova, K., Barak, I., Wilkinson, A.J. (1999). Phosphorylated aspartate in the structure of a response regulator protein. *J. Mol. Biol.* **295**, 9-15.

Lewis, R.J., Muchova, K., Brannigan, J.A., Barak, I., Leonard, G. and Wilkinson, A.J. (2002). Domain-swapping in the sporulation response regulator Spo0A. *J. Mol. Biol.* **297**, 757-770.

Linhananta, A., Zhou, H. and Zhou, Y. (2002). The dual role of a loop with low loop contact distance in folding and domain swapping. *Protein Sci.* **11**, 1695-1701.

Liu, Y., Hart, P.J., Schlunegger, M.P. and Eisenberg, D. (1998). The crystal structure of a 3D domain-swapped dimer of RNase A at a 2.1-A resolution. *Proc. Natl. Acad. Sci.* **95**, 3437-3442.

Liu, Y. and Eisenberg, D. (2002). 3D domain swapping: As domains continue to swap. *Protein Sci.* **11**, 1285-1299.

Liu, Y.S., Gotte, G., Libonati, M. and Eisenberg, D.S. (2001). A Domain-Swapped Rnase a Dimer with Implications for Amyloid Formation. *Nat. Struct. Biol.* **8**, 211-214.

Loll, P.J. and Lattman, E.E. (1989). The crystal structure of the ternary complex of staphylococcal nuclease, Ca2+, and the inhibitor pdTp, refined at 1.65 Å. *Proteins. Struct. Funct. Genet.* **5**, 183-201.

Luo, Y., Bertero, M.G., Frey, E.A., Pfuetzner, R.A., Wenk, M.R., Creagh, L., Marcus, S.L., Lim, D., Sicheri, F., Kay, C., Haynes, C., Finlay, B.B. and Strynadka, N.C.J. (2001). Structural and Biochemical Characterization of the Type III Secretion Chaperones Cest and Sige. *Nat. Struct. Biol.* **8**, 1031-1036.

Lupas, A.N., Ponting, C.P. and Russell, R.B. (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.

Maignan, S., Guillotea, J.P., Fromage, N., Arnoux, B., Becquart, J. and Ducruix, A. (1995). Crystal structure of mammalian Grb2 adaptor. *Science*, **268**, 291-293.
Malby, R.L., McCoy, A.J., Kortt, A.A., Hudson, P.J. and Colman, P.M. (1998). Three-dimensional structures of single-chain Fv-neuraminidase complexes. *J. Mol. Biol.* **279**, 901-910.

Marchler-Bauer, A., Panchenko, A.R., Ariel, N. and Bryant, S.H. (2002). Comparison of sequence and structure alignments for protein domains. *Proteins Struc. Funct. Genet.* **15**, 439-446.

Marcotte, E.M., Pellegrini, M., Ng, H.L., Rice, D.W., Yeates, T.O. and Eisenberg, D. (1999). Detecting protein function and protein-protein interactions from genome sequences. *Science*, **285**, 751-753.

Marsden, R.L., McGuffin, L.J. & Jones, D.T. 2002. Rapid protein domain assignment from amino acid sequence using predicted secondary structure. *Protein Sci.***11**, 2814-2824.

Mattison, K., Oropeza, R. and Kenny, L. (2002). The linker region plays an important role in the interdomain communication of the response regulatoe OmpR. *J. Bio. Chem.* **277**, 32714-32721.

McDonald, I.K. and Thornton, J.M. (1994). Satisfying hydrogen bonding potential in proteins. *J. Mol. Biol.* **238**, 777-793.

McGuffin,L.J., Bryson,K. and Jones,D.T. (2001). What are the baselines for protein fold recognition? *Bioinformatics*, **17,** 63-72.

McGuffin,L.J. and Jones,D.T. (2002). Targeting novel folds for structural genomics. *Proteins Struct. Funct. Genet.* **48,** 44-52.

Meador, W.E., Means, A.R. and Quiocho, F.A. (1992). Target enzyme recognition by calmodulin: 2.4Å structure of a calmodulin-peptide complex. *Science*, **257**, 1251 1255.

Merckel, M.C., Fabrichniy, I.P., Salminen, A., Kalkkinen, N., Baykov, A.A., Lahti, R. and Goldman, A. (2001). Crystal structure of *Streptococcus mutans* pyrophosphatase: A new fold for an old mechanism. *Structure*, **9**, 289-297.

Merritt, E.A. and Bacon, D.J. (1997). Raster3D: Photorealistic molecular graphics. *Methods Enzymol.* **277**, 505-524.

Milburn, M.V., Hassell, A.M., Lambert, M.H., Jordan, S.R., Proudfoot, A.E., Graber, P. and Wells, T. N. (1993). A novel dimer configuration revealed by the crystal structure at 2.4 A resolution of human interleukin-5. *Nature*, **363**, 172-176.

Miller, S., Janin, J., Lesk. A.M. and Chothia, C. (1987). Interior and surface of monomeric proteins. *J. Mol. Biol.* **196**, 641-656.

Mizuno, H., Fujimoto, Z., Koizuni, M., Kano, H., Atoda, H. and Morita, T. (1997). Structure of coagulation factors IX/X-binding protein, a heterodimer of C-type lectin domains. *Nat Struct. Biol.* **4**, 438-441.

Montelione, G.T. and Anderson, S. (1999). Structural genomics: keystone for a Human Proteome Project. *Nat. Struct. Biol.* **6**, 11-12.

Moore, J.D. and Endow, S.A. (1996). Kinesin proteins: a phylum of motors for microtubule-bases motility. *Bioessays*, **18**, 207-262.

Muller, A., MacCallum, R.M. and Sternberg, M.J. (1999). Benchmarking PSI-BLAST in genome annotation. *J. Mol. Biol.* **293**, 1257-1271.

Murray, A.J., Lewis, S.J., Barclay, A.N. and Brady, R.L. (1995). One sequence, two folds: a metastable structure of CD2. *Proc. Natl. Acad. Sci.* **92**, 7337-7341.

Murvai, J., Vlahovicek, K., Barta, E. and Pongor, S. (2001). The SBASE protein domain library, release 8.0: a collection of annotated protein sequence segments. *Nucl. Acids Res.* **29**, 58-60.

Murzin, A.G., Brenner, S.E., Hubbard , T. and Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247**, 536-540.

Musacchio, A., Saraste, M. and Wilmanns, M. (1994). High-resolution crystal structures of tyrosine kinase SH3 domains complexed with proline-rich peptides. *Nat. Struct. Biol.* **1**, 546-551.

Nagano, N., Orengo, C.A. and Thornton, J.M. (2002). One fold with many functions: the evolutionary relationships between TIM barrel families based on their sequences, structures and functions. *J. Mol. Biol.* **321**, 741-765.

Najmudin, S., Nalini, V., Driessen, H.P.C., Slingsby, C., Blundell, T.L., Moss, D.S. and Lindley, P.F. (1993). Structure of the bovine eye lens protein gamma-b(gamma-ii)-crystallin at 1.47 angstrom. *Acta. Crystallogr. D. Biol. Crystallogr.* **49**, 223-231.

Nalini, V., Bax, B., Driessen, H., Moss, D.S., Lindley, P.F. and Slingsby, C. (1994). Close packing of an oligomeric eye lens beta-crystallin induces loss of symmetry and ordering of sequence extensions. *J. Mol. Biol.* **236**, 1250-1258.

Needleman, S.B and Wunsch, C.D. (1970). A general method applicable to the search for similarities in amino acid sequence of two proteins. *J. Mol. Biol.* **48**, 443-453.

Netzer, W.J. and Hartl, F.U. (1997). Recombination of protein domains facilitated by co-translational folding in eukaryotes. *Nature*, **388**, 343-349.

Netzer, W.J. and Hartl, F.U. (1998). Protein folding in the cytosol: chaperonin-dependent and -independent mechanisms. *Trends. Biochem. Sci.* **23**, 68-73.
Newcomer, M.E. (2002). Protein folding and three-dimensional domain swapping: a strained relationship? **12**, 48-53.

Newcomer, M.E. (2002) Protein folding and three-dimensional domain swapping – a strained relationship? *Curr. Opin. Struct. Biol.* **12**, 48-53.

Nixon, A.E., Warren, M.S. and Benkov, S.J. (1997). Assembly of an active enzyme by the linkage of two protein modules. *Proc. Natl. Acad. Sci. USA*, **94**, 1069-1073.

Ogihara, N.L., Ghirlanda, G., Bryson, J.W., Gingery, M., DeGrado, W.F. and Eisenberg, D. (2001). Design of three-dimensional domain-swapped dimers and fibrous oligomers. *Proc. Natl. Acad. Sci.* **98**, 1404-1409.

O'Hare, P. and Williams, G. (1992). Structural studies of the acidic transactivation domain of the Vmw65 protein of herpes simplex virus using NMR. *Biochemistry*, **31**, 4150-4156.

Ohlendorf, D.H., Tronrud, D.E. and Matthews, B.W. (1998). Tefined structure of cro repressor protein from bacteriophage lambda suggests both flexibility and plasticity. *J. Mol. Biol.* **280**, 129-136.

O'Neill, J.W., Kim, D.E., Baker, D. and Zhang, K.Y.J. (2001). Structures of the B1 domain of protein L from Peptostreptococcus magnus with a tyrosine to tryptophan substitution. *Acta. Crystallogr.* **57**, 480-487.

Orengo, C.A., Jones, D.T. and Thornton, J.M. (1994). Protein superfamilies and domain superfolds. *Nature*, **372**, 631-634.

Orengo, C.A., Michie, A.D., Jones, S., Jones, D.T., Swindells, M.B. and Thornton, J.M. (1997). CATH- a hierarchic classification of protein domain structures. *Structure*, **5**, 1093-1108.

Ostermeier, M. and Benkovic, S.J. (2000). Evolution of protein function by domain swapping. *Adv. Protein Chem.* **55**, 29-77.

Parge, H.E., Arvai, A.S., Murtari, D.J., Reed, S.I. and Tainer, J.A. (1993). Human CksHs2 atomic structure: a role for its hexameric assembly in cell cycle control. *Science*, **262**, 387-395.

Park, J, Karplus, K., Barrett, C., Hughey, R., Haussler, D., Hubbard, T. and Chothia, C. (1998). Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. *J. Mol. Biol.* **284,** 1201-1210.

Park, J. and Teichmann, S.A. (1998). DIVCLUS: an automatic method in the GEANFAMMER package that finds homologous domains in single- and multi-domain proteins. *Bioinformatics*, **14**, 144-150.
Parthasarathy, S. and Murthy, M.R. Analysis of temperature factor distribution in high-resolution protein structures. *Protein Sci.* **6**, 2561-2567.

Patthy, L. (1999). Genome evolution and the evolution of exon-shuffling - a review. *Gene*, **238**, 103-114.

Pearson, W.R. and Lipman, D. J. (1988). Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. USA,* **85**, 2444-2448.

Perisic, O., Webb, P.A., Holliger, P., Winter, G. and Williams, R.L. (1994). Crystal structure of a diabody, a bivalent antibody fragment. *Structure*, **2**, 1217-1226.

Phillips, C., Dohnalek, J., Gover, S., Barrett, M. and Margaret, A.J. (1998). A 2.8 Å resolution structure of 6-phosphogluconate dehydrogenase from the protozoan parasite *Trypanosoma brucei*: Comparison with the sheep enzyme accounts for differences in activity with coenzyme and substrate analogues. **282**, 667-681.

Plaxco, K.W., Simons, K.T. and Baker, D. (1998). Contact order, transition state placement and the refolding rates of single domain proteins. *J. Mol. Biol.* **10**, 985-994.

Plewniak, F., Thompson, J.D. and Poch, O. (2000). Ballast: blast post-processing based on locally conserved segments. *Bioinformatics*, **16**, 750-759.

Pollastri, G., Balidi, P., Fraselli, P. and Casadio, R. (2002). Prediction of coordination number and relative solvent accessibility in proteins. *Proteins Struc. Funct. Genet.* **47**, 142-153.

Ponting, C.P. and Russell, R.R. (2002). The natural history of protein domains. *Annu. Rev. Biophys. Biomol. Struct.* **31**, 45-71.

Prasad, G.S., Stura, E.A., Elder, J.H. and Stout, C.D. (2000). Structures of Feline Immunodeficiency Virus Dutp Pyrophosphatase and its Nucleotide Complexes in Three Crystal Forms. *Acta. Crystallogr.* **56**, 1100-1109.

Przytycka,T., Aurora,R. and Rose,G. (1999). A protein taxonomy based on secondary structure. *Nat. Struc. Biol.* **6,** 672-682.

Raaijmakers, H., Vix, O., Toro, I., Golz, S., Kemper, B. and Suck, D. (1999). X-Ray Structure of T4 Endonuclease Vii: A DNA Junction Resolvase with a Novel Fold and Unusual Domain Swapped Dimer Architecture. *EMBO J.* **18**, 1447-1458.

Randal, M. and Kossiakoff, A.A. (1998). Crystallisation and preliminary X-ray analysis of a 1:1 complex between a designed monomeric interferon-gamma and its soluble receptor. *Protein Sci.* **7**, 1057-1060.

Randal, M. and Kossiakoff, A.A. (2000). The 2.0 Å Crystal Structure of Bovine Interferon-Gamma; Assessment of Structural Differences between Species *Acta. Crystallogr.* **56**, 14-24.

Rashin, A. (1985). Location of domains in globular proteins. *Methods Enzymol.* **115**, 420-440.

Remington, S., Weigand, G. and Huber, R. (1982). Crystallographic refinement and atomic models of two different forms of citrate synthase at 2.7 and 1.7 Å resoulution. *J. Mol. Biol.* **158**, 158-163.

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

Ridderstrom, M., Cameron, A.D., Jones, T.A. and Mannervik, B. (1998). Involvement of an active-site Zn2+ ligand in the catalytic mechanism of human glyoxalase I. *J. Biol. Chem.* **273**, 21623-21628.

Rigden, D.J. (2002). Use of covariance analysis for the prediction of structural domain boundaries from multiple protein sequence alignments. *Protein Eng.* **15**, 65-77.

Rose, G.D. (1979). Hierarchic organisation of domains in globular proteins. *J. Mol. Biol.* **234**, 447-470.

Rose, G.D., Geselowitz, A.R., Lesser, G.J., Lee, R.H. and Zehfus, M. H. (1985). Hydrophobicity of amino acid residues in globular proteins. *Science.* **229**, 834-838.

Rost, B. (1999a). Twilight zone of protein sequence alignments. *Protein Eng.* **12**, 85-94.

Rost, B. (1999b). Short yeast orfs: Expressed protein or not? CUBIC preprint: CUBIC, Colombia University, Department of Biochemistry and Molecular Biophysics. http://cubic.bioc.columbia.edu/papers/1999_globe.

Rost, B. and Sander, C. (1993). Prediction of protein secondary structure at better than 70% accuracy. *J. Mol. Biol.* **232** 584-599.

Rubin, G.M., Yandell, M.D., Wortman, J.R., Gabor Miklos, G.L., Nelson, C.R., Hariharan, I.K., Fortini, M.E., Li, P.W., Apweiler, R. Fleischmann, W., *et al.*, (2000). Comparative genomics of the eukaryotes. *Science*, **287**, 2204-2215.

Russell, R.B. (1994). Domain insertion. *Protein Eng.* **7**, 1407-1410

Russell, R.B., Copley, R.R. and Barton, G.J. (1996). Protein fold recognition by mapping predicted secondary structures. *J. Mol. Biol.* **259**, 349-365.

Russell, R.B. and Ponting, C.P. (1998). Protein fold irregularities that hinder sequence analysis. *Curr. Opin. Struct. Biol.* **8**, 364-371.

Salamov, A.A., Suwa, M., Orengo, C.A. and Swindells, M.B. (1999). Genome analysis: Assigning protein coding regions to three-dimensional structures. *Protein Sci.* **8**, 771-777.

Sander, C. and Schneider, R. (1991). Database of homology derived protein structures and the structural meaning of sequence alignment. *Proteins Struct. Funct. Genet.* **9**, 56-68.

Sauer, J., Christensen, T., Frandsen, T.P., Mirgorodskaya, E., McGuire, K.A., Driguez, H., Roepstorff, P., Sigurskjold, B.W. and Svensson, B. (2001). Stability and function of interdomain linker varients of glucoamylase 1 from *Aspergillus niger. Biochemistry,* **40**, 9336-9346.

Sayle, R.A. and Milner-White, E.J. (1995). Rasmol; biomolecular graphics for all. *Trends Biochem. Sci.* **20**. 374.

Schiering, N., Casale, E., Caccia, P., Giordano, P. and Battistini, C. (2000). Dimer Formation Through Domain Swapping in the Crystal Structure of the Grb2-Sh2 Ac-Pyvnv Complex. *Biochemistry*, **39**, 13376-13382.

Schultz, J., Copley, R., Doerks, T., Pomting, C.P. and Bork, P. (2000). SMART: a web based tool for the study of genetically mobile domains. *Nucl. Acids Res.* **28,** 231-234.

Senda, T., Shimazu, T., Matsuda, S., Kawano, G., Shimizu, H., Nakamura, K.T. and Mitsui, Y. (1992). Three-dimensional crystal structure of recombinant murine interferon-beta. *EMBO J.* **11**, 165-182.

Servant, F., Bru, C., Carrere, S., Courcelle, E., Gouzy, J., Peyruc, D. and Kahn, D. (2002). ProDom: automated clustering of homologous domains. *Brief. Bioinform.* **3**. 246-251.

Siddiqui, A.S. and Barton, G.J. (1995). Continuous and discontinuous domains: an algorithm for the automatic generation of reliable protein domain definitions. *Protein Sci.* **4**, 872-884.

Sigrist, C.J., Cerutti, L., Hulo, N., Gattiker, A., Falquet, L., Pagni, M., Bairoch, A. and Bucher, P. (2002). PROSITE: a documented database using patterns and profiles as motif descriptors. *Brief. Bioinform.***3**, 265-274.

Singleton, M.R., Sawaya, M.R., Ellenberger, T. and Wigley, D.B. (2000). Crystal Structure of T7 Gene 4 Ring Helicase Indicates a Mechanism for Sequential Hydrolysis of Nucleotides *Cell*, **101**, 589-600.

Skolnick, J., Fetrow, J.S. and Kolinski, A. (2000). Structural genomics and its importance for gene function analysis. *Nat. Biotechnol.* **18**, 283-287.

Smith, T.F. and Waterman, M.S. (1981). Identification of common molecular subsequences. *J. Mol. Biol.* **147**, 195-197.

Somers, D.O., Medd, S.M., Walker, J.E. and Adams, M.J. (1992). Sheep 6-phosphogluconate dehydrogenase, revised protein sequence based upon the sequences of cDNA clones obtained with the polymerase chain reaction. *J. Biochem.* **288**, 1061-1067.

Sonnhammer, E.L. and Kahn, D. (1994). Modular arrangement of proteins as inferred from analysis of homology. *Protein Sci.* **3**, 482-492.

Sorensen, B.R., Eppel, J.T. and Shea, M.A. (2002). An interdomain linker increases the thermostability and decreases the calcium affinity of the calmodulin N-domain. *Biochemistry*, **41**, 15-20.

Sowdhamini, R. and Blundell, T. (1995). An automatic method involving cluster analysis of secondary structures for the identification of domains in proteins. *Protein Sci.* **4**, 506-520.

Spitzfaden, C., Grant, R.P., Mardon, H.J. and Campbell, I.D. (1997). Module-module interactions in th ecell binding region of fibronectin: stability, flexibility and specificity. *J. Mol. Biol.* **265**, 565-579.

Stehle, T., Gamblin, S.J., Yan, Y. and Harrison, S.C. (1996). The structure of simian virus 40 refined at 3.1 Å resolution. *Structure*, **4**, 165-182.

Steinbacher, S., Miller, S., Baxa, U., Budisa, N., Weintraub, A., Seckler, R. and Huber, R. (1997). Phage P22 tailspike protein: crystal structure of the head-binding domain at 2.3 Å, fully refined structure of the endorhamnosidase at 1.56 Å resolution, and the molecular basis of O-antigen recognition and cleavage. *J. Mol. Biol.* **267**, 865-880.

Story, R.M., Weber, I.T. and Steitz, T.A. (1992). The structure of the E. coli. recA protein monomer and polymer. *Nature*, **355**, 318-325.

Strop, P., Smith, K.S., Iverson, T.M., Ferry, J.G. and Rees, D.C. (2001). Crystal Structure of the 'Cab' Type Beta Class Carbonic Anhydrase from the Archaeon Methanobacterium Thermoautotrophicum. *J. Biol. Chem.* **276**, 10299-10305.

Svensson, L.A., Thulin, E. and Forsen, S. (1992). Proline cis-trans isomers in calbindin D9k observed by X-ray crystallography. *J. Mol. Biol.* **223**, 601-606.

Swindells, M.B. (1995). A procedure for detecting structural domains in proteins. *Protein Sci.* **4**, 103-112.

Tatusov, R.L., Natale, D.A., Garkavtsev, I.V., Tatusova, T.A., Shankavaram, U.T., Rao, B.S., Kiryutin, B., Galperin, M.Y., Fedorova, N.D. and Koonin, E.V. (2001). The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucl. Acids Res.* **29**, 22-28.

Taylor, W. R. (1999). Protein structural domain identification. *Protein Eng.* **12**, 203-216.

Teichmann, S.A., Park, J. and Chothia, C. (1998). Structural assignments to the Mycoplasma genitalium proteins show extensive gene duplications and domain rearrangements. *Proc. Natl. Acad. Sci. USA.* **95**, 14658-14663.

Thanaraj, T.A. and Argos, P. (1996). Ribosome-mediated translational pause and protein domain organization. *Protein Sci.* **5**, 1594-1612.

Thompson, J.D., Higgins, D.G. and Gibson, T.J. (1994). CLUSTAL-W Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucl. Acids Res.* **22**, 4673-4680.

Traut, T.W. (1988). Do exons code for structural or functional units in proteins? *Proc. Natl. Acad. Sci. USA.* **85**, 2944-2948.

Ullah, J.H., Walsh, T.R., Taylor, I.A., Emery, D.C., Verma, C.S., Gamblin, S.J. and Spencer, J. (1998). The crystal structure of the L1 metalo-beta-lactamase from *Strenotrophomonas maltophilia* at 1.7 Å resolution. *J. Mol. Biol.* **284**, 125-136.

Ultsch, M.H., Wiesmann, C., Simmons, L.C., Henrich, J., Yang, M., Reilly, D., Bass, S.H. and De Vos, A.M. (1999). Crystal Structure of the neurotrophin-binding domain of Trka, Trkb and Trkc. *J. Mol. Biol.* **290**, 149-159.

Varenne, S., Buc, J., Lloubes, R. and Lazdunski, C. (1984). Translation is a non-uniform process. Effect of tRNA availability on the rate of elongation of nascent polypeptide chains. *J. Mol. Biol.* **180**, 549-576.

Vijayalakshmi, J., Mukhergee, M. K., Graumann, J., Jakob, U. and Saper, M. A. The 2.2 A Crystal Structure of Hsp33. (2001). A Heat Shock Protein with Redox-Regulated Chaperone Activity. *Structure*, **9**, 367-375.

Vonderviszt, F. and Simon. I. (1986). A possible way for prediction of domain boundaries in globular proteins from amino acid sequence. *Biochem. Biophys. Res. Commun.* **139**, 11-17.

Waksman, G., Krishna, T.S.R., Williams, C.H. and Kuriyan, J. (1994). Crystal structure of *Escherichia coli* thioredoxin reductase refined at 2 Å resolution: Implications for a large conformational change during catalysis. *J. Mol. Biol.* **236**, 800-816.

Wang, J.H., Yan, Y., Garrett, T.P.J., Liu, J.H., Rodgers, D.W., Garlick, R.L., Tarr, G.E., Husain, Y., Reinherz, E.L. and Harrison, S.C. (1990). Atomic structure of a fragment of human CD4 conatining 2 immunoglobulin-like domains. *Nature*, **348**, 411-418.

Watson, J.D. and Crick, F.H.C. (1953). A structure of deoxyribose nucleic acid. *Nature*, **171**, 737-738.

Waugh, D.F. (1954). Protein-protein interactions. *Adv. Prot. Chem.* **9**, 325-437.

Weis, W.I., Kahn, R., Fourme, R., Drickamer, K. and Hendrickson, W. A. (1991). Structure of the calcium-dependent lectin domain from a rat mannose-binding protein determined by MAD phasing. *Science*, **254**, 1608-1615.

Wernisch, L., Hunting, M. and Wodak, S.J. (1999). Identification of structural domains in proteins by a graph heuristic. *Proteins. Struct. Funct. Genet.* **35**, 338-352.

Wetlaufer, D.B. (1973). Nucleation, rapid folding and globular intrachain regions in proteins. *Proc. Natl. Acad, Sci. USA*, **70**, 697-701.

Wheelan, S.J, Marchler-Bauer, A. and Bryant, S.H. (2000). Domain size distributions can predict domain boundaries. *Bioinformatics*, **16,** 613-618.

White S.H. and Jacobs, R.E. (1990). Statistical distribution of hydrophobic residues along the length of protein chains: Implications for protein folding and evolution. *J. Biophys.* **57**, 911-921.

White, S.H. (1992). Amino acid preferences of small proteins. *J. Mol. Biol.* **227**, 991-995.

Wiesmann, C., Ultsch, M.H., Bass, S.H. and De Vos, A.M. (1999). Crystal structure of nerve growth factor in complex with the ligand-binding domain of the Trka receptor receptor. *Nature*, **401**, 184-188.

Wlodawer, A., Bott, R. and Sjolin, L. 1982. The refined crystal structure of ribonuclease A at 2.0 Å resolution. *J. Biol. Chem.* **257**. 1325-1332.

Wodak, S.J. and Janin, J. (1981). Location of structural domains in proteins. *Biochemistry*, **20**, 6544-6552.

Wooton, JC. (1994). Non-globular domains in protein sequences: automated segmentation using complexity measures. *Comput. Chem.* **18**, 269-285.

Wright, L.M., Reynolds, C.D., Rizkallah, P.J., Allen, A.K., Els. J.M., Damme, V., Donovan, M.J. and Peumans W.J. (2000). Structural characterisation of the native fetuin-binding protein *Scilla campanulata* agglutinin: a novel two-domain lectin. *FEBS letters.* **468**, 19-22.

Wright, P.E. and Dyson, H.J. (1999). Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *J. Mol Biol.* **293**, 321-331.

Wu, C.H., Huang, H., Arminski, L., Castro-Alvear, J., Chen, Y., Hu, Z., Ledley, R.S., C. K., Hans-Werner Mewes, L., Orcutt, B.C., Suzek, B.E., Vinayaka, A., Yeh, L.L., Zhang, J. and Barker. W.C. (2002). The Protein Information Resource: an integrated public resource of functional annotation of proteins. *Nucl. Acids Res.* **30**, 35-37.

Wu, H., Myszka, D.G., Tendian, S.W., Brouillette, C.G., Sweet, R.W., Chaiken, I.M. and Hendrickson, W.A. (1996). Kinetic and structural analysis of mutant CD4 receptors that are defective in HIV gp120 binding. *Proc. Natl. Acad. Sci.* **93**, 15030-15035.

Xu, D. and Nussinov, R. (1997). Favourable domain size in proteins. *Folding and Design*, **3**, 11-17.

Yang, F., Bewley, C.A., Louis, J.M., Gustafson, K.R., Boyd, M.R., Gronenborn, A.M., Clore, G.M. and Wlodawer, A. (1999). Crystal structure of cyanovirin-N, a potent HIV-inactivating protein, shows unexpected domain swapping. *J. Mol. Biol.* **288**, 403-412.

Zahn, R., Liu, A., Luhrs, T., Riek, R., Von Schroetter, C., Garcia, F.L., Billeter, M., Calzolai, L., Wider, G. and Wuthrich, K. (2000). NMR solution structure of the human prion protein. *Proc. Nat. Acad. Sci.* **97**, 145-150.

Zdanov, A., Schalk-Hihi, C., Gustchina, A., Tsang, M., Weatherbee, J. and Wlodawer, A. (1995). Crystal structure of interleukin-10 reveals the functional dimer with an unexpected topological similarity to interferon gamma. *Structure*, **3**, 591-601.

Zegers, I., Deswarte, J. and Wyns, L. (1999). Trimeric domain-swapped barnase. *Proc. Natl. Acad. Sci.* **96**, 818-822.

Zehfus, M.H. and Rose, G.D. (1986). Compact units in proteins. *Biochemistry,* **25**, 5759-5765.

Zemela, A., Venclovas, C., Fidelis, K. and Rost, B. (1999). A modified definition of Sov, a segment-based measure for protein secondary structure prediction assessment. *Proteins Struct. Funct. Genet.* **34,** 220-223.