

The Characterisation and Prediction of Protein-Protein Interfaces

Thomas Kabir

Thesis submitted in fulfilment of the requirements for the degree
of Doctor of Philosophy from the University of London

Department of Biochemistry and Molecular Biology
University College London

September 2004

UMI Number: U602709

All rights reserved

INFORMATION TO ALL USERS

The quality of this reproduction is dependent upon the quality of the copy submitted.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if material had to be removed, a note will indicate the deletion.



UMI U602709

Published by ProQuest LLC 2014. Copyright in the Dissertation held by the Author.
Microform Edition © ProQuest LLC.

All rights reserved. This work is protected against
unauthorized copying under Title 17, United States Code.



ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

Abstract

Understanding how proteins interact with each other is of fundamental importance and is one of the most important goals of molecular biology. In order to study the characteristics of protein-protein interaction sites datasets of non-homologous protein-complexes have been compiled. These datasets include 142 obligate homo-complexes, 20 obligate hetero-complexes, 20 enzyme-inhibitor complexes, 15 antibody-antigen complexes, and 10 signaling complexes. Overall, the protein-protein interfaces of obligate complexes were found to be closely packed, relatively hydrophobic when compared to the entire protein exterior, planar, and enriched in residues such as tyrosine, phenylalanine, and isoleucine. In comparison to the protein-protein interfaces found within obligate protein-complexes the protein-protein interfaces of non-obligate protein-complexes were found to be on average much smaller in size and contain larger numbers of polar and charged residues. The bulk properties of the obligate and non-obligate protein-complexes are also discussed. A neural network was used together with the Patch Analysis method of Jones and Thornton (1997) to predict the location of the protein-protein interfaces in selected datasets of obligate homo and hetero-complexes. The Patch Analysis method is based upon defining a series of contiguous patches over the surface of a protein. The physical and chemical characteristics of each patch are encoded in the form of six parameters. One of these parameters is hydrophobicity. Another parameter that is used is accessible surface area (ASA). By comparing average values of these six parameters for the residues in a given surface patch with those covering known protein-protein interfaces the likelihood of a patch corresponding to a protein-protein interface can be assessed. Based upon the results for a dataset of 76 homo-dimers the use of a neural network enhances the accuracy of the original Patch Analysis method by some thirteen percent.

Acknowledgements

All generalisations are dangerous including this one! – Alexandre Dumas

I never thought that I would get this far. I only managed it because of the amount of help I received from some very kind and talented people. Firstly I thank Fr. Stephen Williams for all his support over the last few years. I also thank the superior and brethren of the Community of the Resurrection for housing me and being such good company over the last year. This thesis is dedicated jointly to the Community of the Resurrection and Fr. Stephen Williams.

I owe a lot to my supervisors Janet Thornton and Denise Gorse. Both have shown an astounding amount of patience with me that has been beyond the call of duty. I also thank the graduate tutor, Peter Swann, who get me out of quite a few scrapes. Amongst those I have worked with directly I thank Susan Jones, William Valdar, Hannes Ponstingl, Irene Nooren, and Adrian Shepherd.

A number of NHS People have been responsible for keeping me intact over the last four years. Amongst them I thank Jeff Halperin, Nadia Davies, David Webb, and Judy Stubbings. I wouldn't have made it through the last four years without them. In the Union I have benefited from the sound advice of Robyn Simms and Ian Moran.

The EBI and BSM people have been great fun. I particularly thank Ian Sillitoe, James Bray, Jennifer Dawe, Kevin Murray, Roman Laskowski, Craig Porter, Gail Bartlett, and Gabby 'moral support' Reeves. All of these people have not only been good company but have also provided me with a lot of help with my work.

Finally I thank my family for all the support. Did I really make it this far? Yes I did. Good. My apologies to anyone I have forgotten to thank here. I'm exhausted. I want to go home now.

Table of Contents

Abstract.....	2
Acknowledgements.....	3
Table of Contents.....	4
Chapter 1.....	7
1. Introduction.....	7
1.1 Introduction.....	7
1.2 Determination of Protein Structure.....	9
1.3 The Protein Structural Hierarchy.....	12
1.4 Protein Symmetry and Quaternary Structure.....	18
1.5 The Forces driving Protein-Protein Interactions.....	20
1.5.1 The Hydrophobic Effect.....	21
1.5.2 Non Covalent Interactions.....	21
1.6 Physical Characteristics of Protein-Protein Binding Sites.....	23
1.6.1 Size (Accessible Surface Area).....	23
1.6.2 Shape Complementarity.....	25
1.6.3 Packing at the Protein-Protein Interface.....	26
1.7 Classifying Protein-Protein Interactions.....	27
1.7.1 Permanent or Transitory.....	27
1.7.2 Homo or Hetero.....	27
1.7.3 Affinity (Strength of Interaction).....	28
1.7.4 Specificity.....	28
1.7.5 Biological relevance.....	29
1.8 Reasons for Oligomeric Proteins.....	29
1.8.1 Error Control.....	29
1.8.2 Coding Efficiency.....	30
1.8.3 Reduction of Surface Area.....	30
1.8.4 Stability.....	30
1.8.5 Regulation of Activity.....	31
1.8.6 Enzymatic Efficiency and Function.....	32
1.9 Rationale and Outline of Thesis.....	33
Chapter 2.....	35
2 Datasets of Multi-Subunit Protein Complexes.....	35
2.1 Introduction.....	35
2.2 Protein Databases.....	36
2.3 Assembly of Datasets of Multi-Subunit Complexes.....	37
2.4 The datasets.....	42
2.4.1 Datasets of Homo-Complexes.....	42
2.4.2 Datasets of Hetero-Complexes.....	46
2.5 Distribution of Multimeric States in Protein Databases.....	59
Chapter 3.....	61
3 Obligate Homo-Complexes.....	61
3.1 Introduction.....	61
3.2 Classification of Residues.....	63
3.3 Size (ASA) of Protein-Protein Interfaces.....	65

3.4	Symmetry.....	73
3.5	Amino Acid Composition.....	76
3.6	Hydrophobic Content	83
3.7	Hydrophobicity.....	85
3.8	Hydrogen Bonds.....	91
3.9	Salt Bridges.....	95
3.10	Secondary Structure.....	97
3.11	Packing at Subunit Interfaces	100
3.12	Planarity.....	103
3.13	Protrusion of Residues at Protein-Protein Interfaces.....	106
3.14	Flexibility of Interface Residues.....	108
3.15	Conclusions	111
Chapter 4.....		115
4	Hetero-Complexes.....	115
4.1	Introduction	115
4.2	Obligate Hetero-Multimers.....	121
4.2.1	Size (ASA) of Protein-Protein Interfaces	121
4.2.2	Planarity.....	123
4.2.3	Hydrogen Bonding	125
4.2.4	Amino Acid Composition.....	127
4.2.5	Secondary Structure Content	129
4.2.6	Subunit Shape	131
4.2.7	Hydrophobicity.....	133
4.2.8	Subunit Organisation	135
4.2.9	Subunit Genetic Structure.....	142
4.2.10	Subunit Homology.....	142
4.2.11	Subunit Assembly.....	147
4.3	Non Obligate Hetero-Multimers.....	148
4.3.1	Enzyme-Inhibitors	148
Enzyme interface.....		153
Inhibitor Interface.....		153
4.3.2	Antibody-Antigen Complexes.....	163
Antibody interface.....		164
Antigen Interface.....		164
4.3.3	Signalling Proteins.....	166
4.4	Comparison of Obligate Vs Non-Obligate Hetero-Complexes.....	167
4.5	Comparison of Homo Vs Hetero Obligate-Complexes.....	171
Chapter 5.....		175
5	Prediction of Protein-Protein Interaction Sites Using a Neural Network	175
5.1	Introduction	175
5.2	Patch Analysis	177
5.2.1	Definition of a Surface Patch.....	177
5.2.2	Definition of Patch Parameters.....	180
5.2.3	The Scoring Algorithm.....	182
5.3	Neural Networks.....	183
5.3.1	Feed Forward Neural Networks	184
5.3.2	Supervised Learning	187

5.4	Neural Network based Patch Analysis	188
5.4.1	Training and Testing the Neural Network.....	188
5.4.2	Evaluating the Results	192
5.5	Results	194
5.5.1	Homo-Complexes.....	196
5.5.2	Hetero-Complexes.....	200
5.6	Rationalising the Results	202
5.6.1	Understanding Incorrect Predictions	203
5.6.2	Assessing the Statistical Significance of the Results.....	209
5.7	Future Work.....	211
5.8	Conclusions	213
6	Conclusions.....	215
7	Appendix.....	220
8	References.....	224

Chapter 1

Introduction

1.1 Introduction

Proteins are nature's 'beasts of burden'. At the molecular level proteins have a part to play in carrying out virtually every biological process. These processes range from transporting oxygen around the blood stream to identifying and destroying foreign bodies. Proteins are structurally complicated and often form highly intricate complexes in order to carry out the manifold tasks that are required to sustain any living organism. A good example of this is ATP synthase (Stock et al., 1999). However, it is remarkable that such diversity of function is attained using proteins constructed from a rather limited repertoire of protein folds. It is estimated that there are only 1000 distinct protein folds which account for the enormous diversity in the structures and functions that proteins perform (Chothia, 1992). Proteins generally interact with other proteins or smaller ligands in order to carry out their function. An examination of the protein-protein interactions in yeast cells confirms this tendency showing that yeast proteins interact with 5 or more other proteins (Tucker et al., 2001). An analysis of the regions where proteins interact with other proteins or ligands is therefore the key to understanding how to interfere with the functions that they perform. The applications of understanding how proteins bind to each other are innumerable. For example once the active site of an enzyme has been characterised compounds can be designed to bind to and inhibit the enzyme. The most obvious application of such interfering with protein interactions is in the field of medicine. The elucidation of the structures of HIV protease and reverse transcriptase has led to the design of HAART (highly active anti-retroviral therapy) which is currently the most effective treatment of the HIV/AIDS epidemic. HAART is a therapy based upon inhibiting HIV protease and reverse transcriptase with synthetic compounds that were selected once the active sites of the two enzymes were characterised.

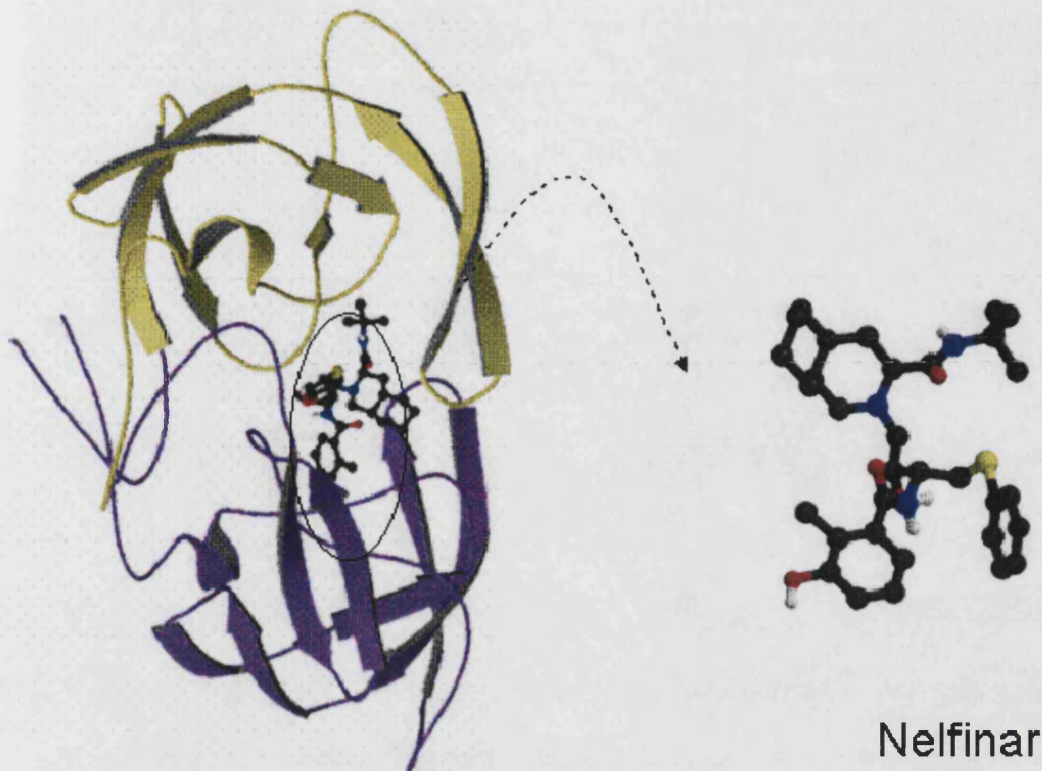


Figure 1.1: HIV protease in complex with nelfinavir (1ohr, Kaldar et al., 1997).

The structure of HIV protease in complex with the synthetic inhibitor nelfinavir is shown in figure 1.1 (Kaldor et al., 1997). The structures of other proteins that are known to be active in diseases such as BSE, diabetes, and malaria are also leading to the development of ways of treating these conditions (a more comprehensive list can be found in Stryer et al., 2002). A convergence of new experimental methods such as two-hybrid experiments, micro-arrays, and proteomic data together with the advent of structural genomics holds the key to determine the full set of protein-protein interactions that take place within an organism. With such knowledge comes the potential to understand nature at the molecular level as never before.

1.2 Determination of Protein Structure

X-ray crystallography and nuclear magnetic resonance (NMR) are the two major experimental techniques that have been used to determine the atomic structure of proteins. The first stage in determining the structure of a protein by x-ray crystallography is to isolate, purify and grow well ordered crystals of the protein. None of these steps is easy. Recent advances have included expressing proteins in synthetic bacteria in order to produce large quantities of the purified protein and automating crystallisation experiments using robots. Once crystals have been obtained the next stage is to bombard the crystal with an x-ray beam. The x-ray beam is then scattered by the atoms of the protein within the unit cell and interferes constructively and destructively according to Bragg's law ($2d \sin \theta = n\lambda$) to produce a diffraction pattern. The diffraction pattern provides information regarding the electron density of the atoms that scatters the x-ray beam. From this diffraction pattern the three-dimensional coordinates of the atoms that make up the protein can be computed.

The quality of a protein structure can in part be assessed using the resolution and the R factor of the structure (Branden & Tooze, 1998). The R factor is a measure of how well the electron density of the calculated protein structure matches up with experimental data. Typically, protein structures are reported to have an R factor of around 0.2 (Laskowski in Boune & Weissig, 2003). An R factor between 0.4-0.6 can be obtained from a completely disordered structure. The resolution of an x-ray structure is an indication of the quality of the electron-density map used to solve the structure and thus an indirect measure of the precision to which the three-dimensional co-ordinates have been determined. The resolution of a structure is expressed in Angstroms (\AA). At low resolutions of $>4\text{\AA}$ the shape of the protein and some α -helices can be resolved (figure 1.2(a)). At a resolution of 3\AA it is possible to determine the position of the polypeptide chain(s) of the protein and the positions of individual amino acids along the chain. At 2.5\AA the conformation of the amino acid side-chains can be readily observed (figure 1.2(b)). At high resolutions of $\sim 1.5\text{\AA}$ the positions of hydrogen atoms and solvent molecules can be seen (figure 1.2(c)).

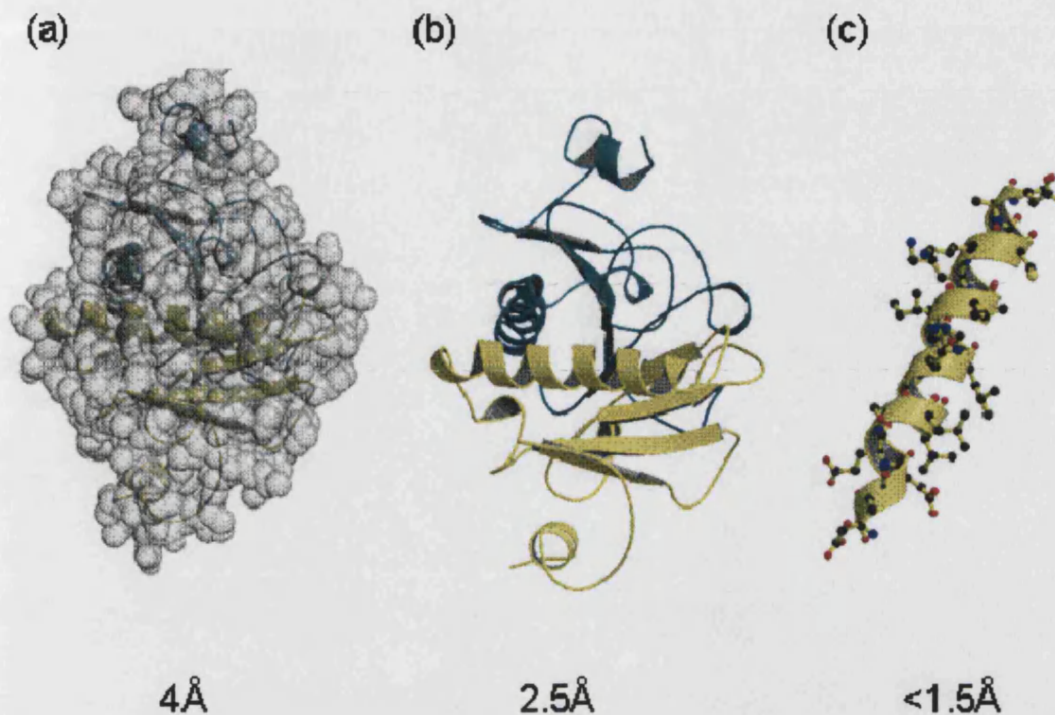


Figure 1.2: (a) at a resolution of 4Å the shape of a protein and some secondary structure motifs can be resolved. (b) At 2.5Å the positions of amino acid side chains are reasonably well defined. (c) At resolutions <1.5Å the positions of hydrogen atoms and solvent molecules can be observed.

The majority of the structures studied in later chapters have been solved to a resolution of 2.5Å or less. Ultra-high resolution structures (<1Å) can be obtained using either neutron diffraction or more commonly x-ray crystallography. At this level of resolution the positions of hydrogen atoms are clearly resolved and the dynamics of individual chemical bonds can be investigated. This can be useful in probing the catalytic mechanism of an enzyme. A disadvantage of x-ray crystallography is that some information regarding the gross dynamics of the protein is lost during the crystallisation process when proteins adopt a largely static configuration. The advantage of NMR over x-ray crystallography is that using NMR the dynamic nature of a protein structure can be revealed. This is especially useful in working out the exact mechanisms that a protein employs to carry out its function. In 2002, 15% of all protein structures in the PDB have been solved using NMR (Berman, 2002). Up to recently only rather small proteins have been accessible to analysis using NMR. However, more recently larger protein structures like GroEL have been successfully subjected to NMR (Fiaux et al., 2002).

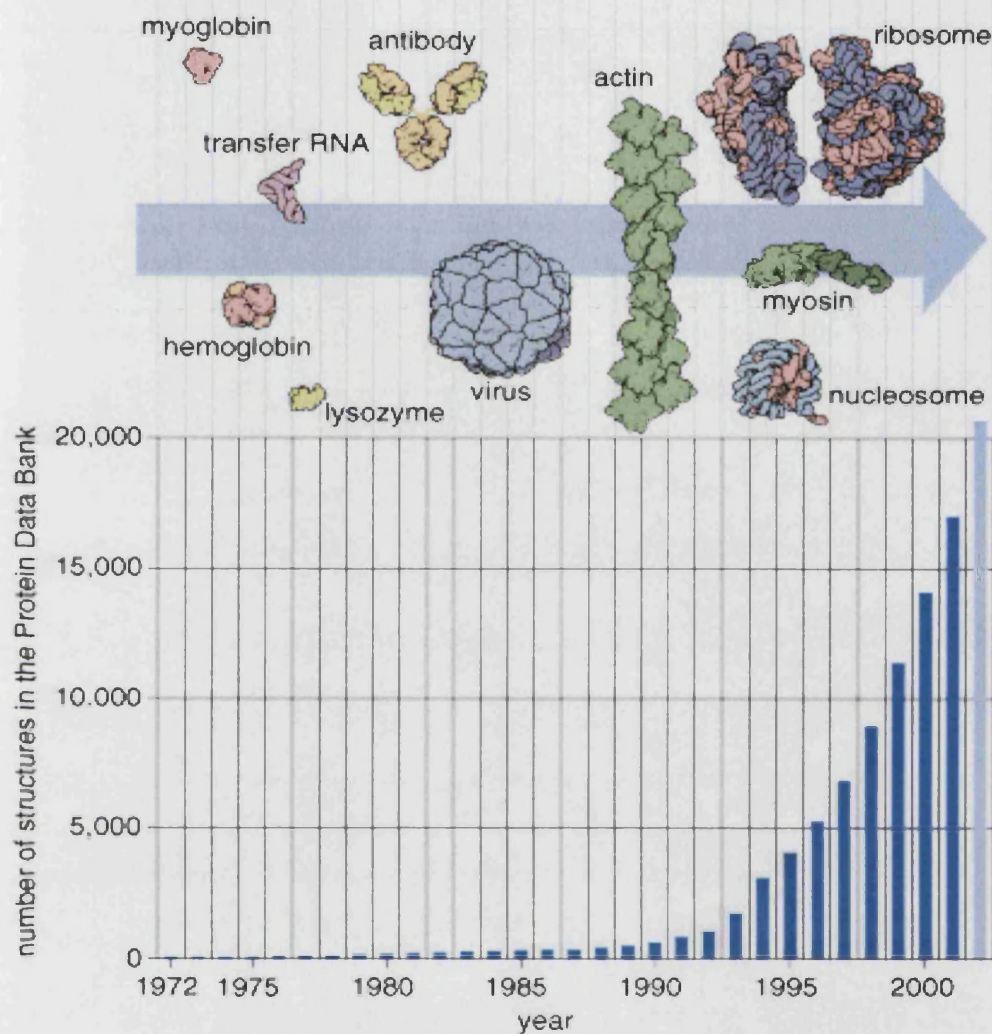


Figure 1.3: PDB content growth. The chart illustrates the rapid growth in the number of protein structures being deposited in the PDB. Over the course of last thirty years advances in technology and experimental techniques have allowed the structures of complex proteins such as the ribosome to be determined. This figure is reproduced from Berman et al., 2002.

The PDB (Berman et al., 2000) is available at <http://www.pdb.org> and contains the coordinates of more than 22,000 structures. The number of *unique* protein structures in the PDB is much smaller. For some proteins the PDB contains separate entries for the same structure determined to differing resolutions and the coordinates of many closely related structures such as trypsin complexes each with different point mutations in the active site of the enzyme. The rapid rise in the number of protein structures whose coordinates have been deposited in the PDB is shown in figure 1.3. The increasing rate at which protein structures are being solved is in part due to

technological advances and the increasing world-wide effort that is currently being invested in protein structure determination. It is worth noting that alongside the increase in the rate of protein structure determination has come a corresponding increase in the numbers of large and complex protein structures such as ATP synthase which have been deposited in the PDB (Stock et al., 1999). Structural genomics initiatives should increase the number of known protein structures greatly in the years to come (Brenner, 2001).

1.3 The Protein Structural Hierarchy

Proteins are polymers of the twenty naturally occurring amino acids joined together covalently via peptide bonds. Proteins do occasionally contain amino acids outside the normal twenty, which are formed by chemical modification of an amino acid after the polypeptide chain has been synthesised. The twenty amino acids all share a similar basic structure with differing side chains. The differing side chains confer distinct chemical characteristics on each of the amino acids. Broadly speaking the side chains are of a polar, charged, and non-polar (hydrophobic) character. The sole exception to this is glycine which has no side chain. A table illustrating the side chains of the twenty amino acids arranged according to their chemical character is shown in figure 1.4.

Although the side-chains of the 20 amino acids have been classified in figure 1.4 as having a hydrophobic, polar, or charged character, these three classifications are not mutually exclusive. For instance, amino acid side chain can have both hydrophobic and charged properties and many other classifications (other than charged, polar, or hydrophobic) are possible. Tryptophan has a polar character due in part to the nitrogen atom in one of its two carbon rings but is also highly hydrophobic as revealed by many experimentally derived hydrophobicity scales. The properties of amino acid side chains can also vary according to environmental conditions. As an example the side-chain of histidine can be charged or polar depending on the pH of the local environment. Some other characteristics that can be used to classify the side-chains of the twenty amino acids are given in the following paragraphs (labelled (a) to (f)). Many of the amino acid side chain properties noted in (a) to (f) are taken

from Cozzone, 2002 and are given here in a somewhat modified form. Given that the side chains of the twenty amino acids can simultaneously be described by more than one of the characteristics in (a) to (f) or as being charged, polar, or hydrophobic (or a combination of the three), show that there is no one satisfactory way to categorise amino acids in a way that fully reflects the complex properties of the amino acid side chains.

(a) Hydrophobicity. Several experimental and theoretical methods have been employed to quantify the hydrophobicity of the side-chain of each amino acid. Most experimentally derived hydrophobicity scales are derived from transferring amino acids from non-polar solvents such as octanol, cyclohexane, and linear alkanes, to water and measuring the free energy of transfer (Chan, 2002, Chan & Dill, 1997). Many experimental hydrophobicity scales have been reported over the last thirty years. As yet there “there has been a lack of quantitative agreement between hydrophobicity scales” (Chan, 2002). The reasons for the discrepancies between hydrophobicity scales are complex and are examined by Chan & Dill, 1997. The hydrophobicity scale that is used throughout this thesis to quantify the hydrophobicity of protein interiors, exteriors, and protein-protein interfaces is one proposed by Fauchere and Pliska, 1983, and is detailed in section 3.7 in chapter 3. Due to the differences between the various hydrophobicity scales the classification of any amino acid or any grouping of amino acids (such as protein-protein interfaces) as being ‘hydrophobic’ is somewhat subjective and is dependant on the exact way that the ‘hydrophobicity’ of amino acids has been measured. Indeed there are some discrepancies between the groupings of amino acids into charged, polar, or hydrophobic and the Fauchere and Pliska hydrophobicity scale shown in table 3.7 in chapter 3. For example according the hydrophobicity scale tryptophan is the most hydrophobic amino acid but it has been classified as being polar in figure 1.4. The explanation for this is that (as mentioned previously) tryptophan has both polar and charged characteristics and so either the ‘polar’ or ‘hydrophobic’ classification can be used. Other amino acids for which there is a disagreement between the classification scheme shown in figure 1.4 and the Fauchere and Pliska hydrophobicity scale are cysteine, and tyrosine. Neither of these amino acids fall easily into either the ‘polar’ classification. Since the Fauchere and Pliska scale is derived from experimental results it can be considered to be a more reliable indicator of the hydrophobicity for

any given set of residues rather than by numerically counting the fraction of residues that fall into the charged, polar, and hydrophobic classifications set out in figure 1.4.

(b) Cyclic or Aliphatic. A cyclic amino acid has a side-chain containing closed rings of carbon atoms. Proline, phenylalanine, tyrosine, tryptophan, and histidine are cyclic amino acids all other amino acids are aliphatic. An aliphatic amino acid is one whose side-chain contains carbon atoms in chains rather than in closed rings.

(c) Amphipathic. An amphipathic amino acid is one whose side-chain has both a polar and non-polar (or hydrophobic) character. As described previously tryptophan can be considered to be amphipathic having both polar and hydrophobic characteristics.

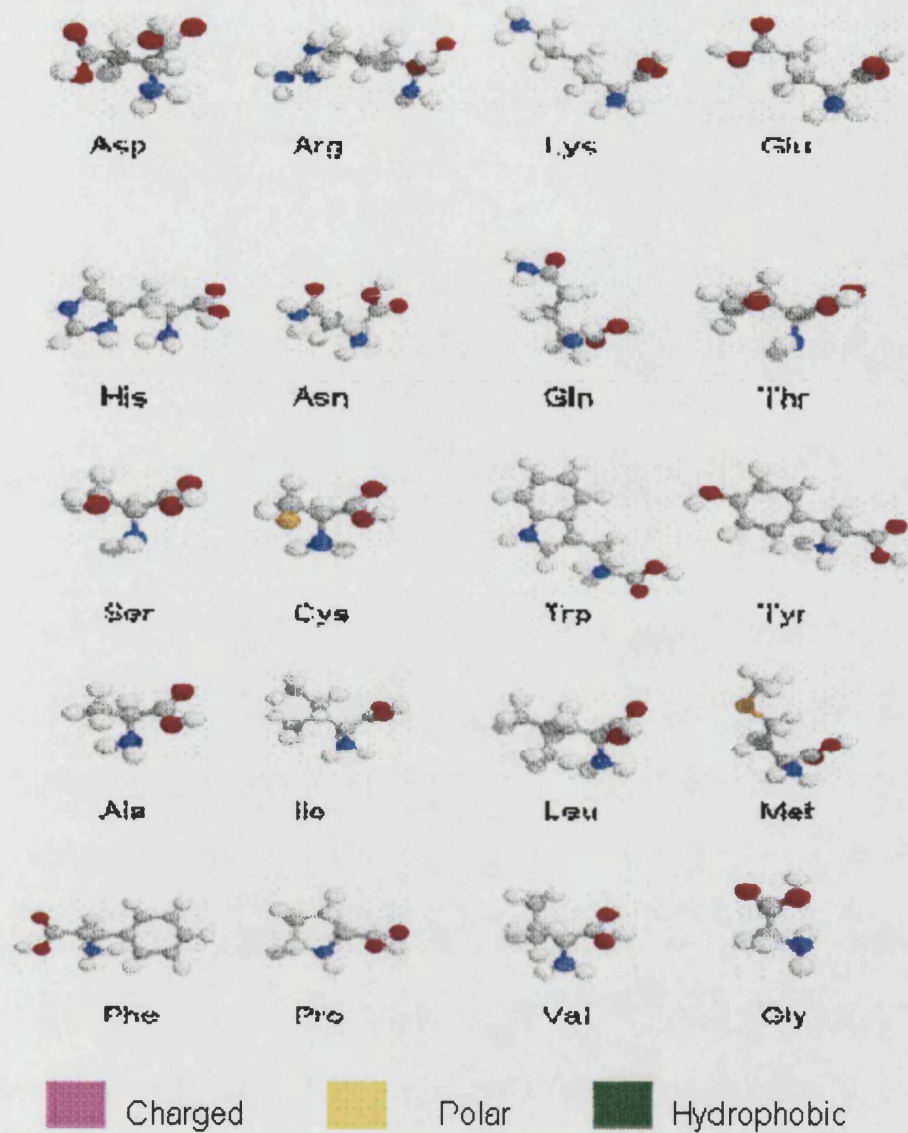
(d) Basic or acidic. The side-chains of lysine and arginine have a net positive charge at physiological pH and can be considered basic amino acids. Histidine can also be considered a basic amino acid under certain conditions. Acidic amino acids include aspartic acid and glutamic acid both having negatively charged groups in their side-chains at physiological pH. Asparagine and glutamine can also be considered to be acidic.

(e) Size. The physical size of the amino acid side-chains varies widely from glycine which has no side chain to alanine which has the smallest side-chain of any amino acid (113\AA^2 of solvent accessible surface area) to tryptophan which has the largest side-chain with an accessible surface area of 259\AA^2 .

(f) Sulphur-containing. The side-chains of both methionine and cysteine contain sulphur atoms. Methionine has a hydrophobic side-chain while cysteine has a polar character due to the SH group in its side-chain.

Protein structures can be described by a number of different levels, which are summarised in figure 1.5 and are detailed in the next section. The primary structure of a protein is simply the linear amino acid sequence of the polypeptide chain(s) that make up the protein figure 1.5(a). The secondary structure of a protein is the level at

which the amino acids form structural motifs such as α -helices and β -strands figure 1.5(b). The arrangement of secondary structure motifs into relatively compact three dimensional units is known as the tertiary structure of a protein figure 1.5(c). Many proteins are also organised into domains. The definition of what is a 'domain' is controversial. A structural domain is usually thought of as a region of the polypeptide chain that is folded into a compact and stable structure that may exist independently of the remainder of the protein.



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

Figure 1.4: The structures of the twenty naturally occurring amino acids. Each acid has also classified as being charged, polar, or hydrophobic in character.

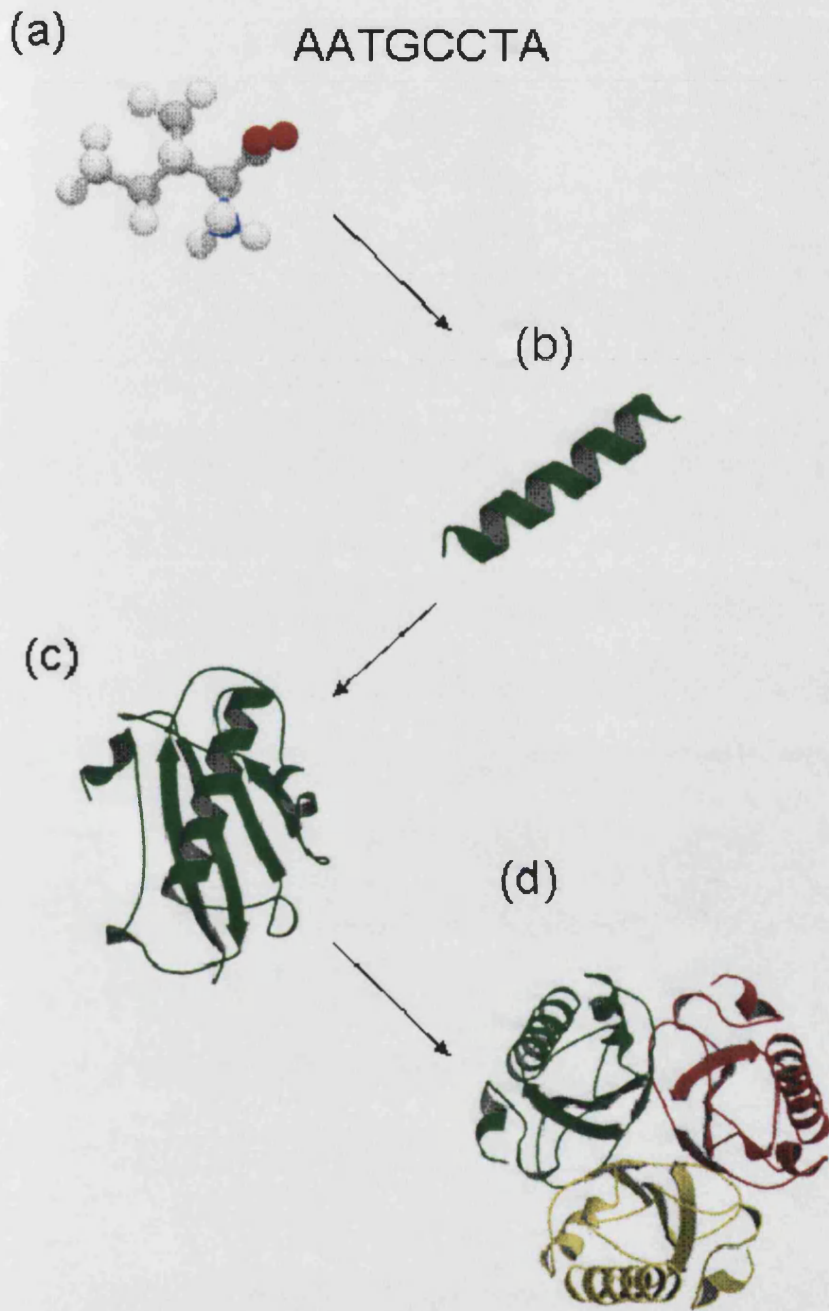


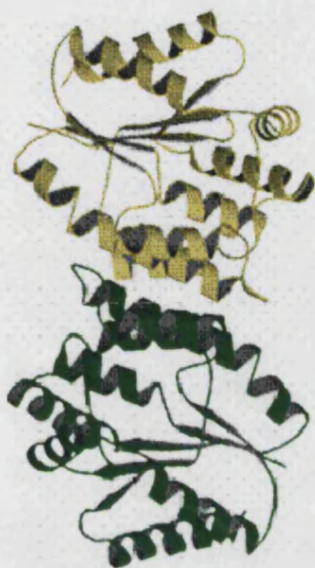
Figure 1.5: The protein structure hierarchy. (a) The linear amino acid sequence of a protein is denoted the primary structure of a protein. (b) The secondary structure of a protein is the level at which amino acids form regular structural motifs such as α -helices and β -sheets. (c) The arrangement of secondary motifs and loops into a relatively stable structure is known as the tertiary structure of a protein. The level at which individual polypeptide chains bind to one another to form a protein complex is referred to as a protein's quaternary structure

Domains are usually linked to each other via loops (regions of the polypeptide chain with no or little regular secondary structure). The presence of domains in a protein is usually related to its function. Phosphoglycerate kinase (3pgk) is a protein consisting of two domains (Watson et al., 1982). The active site of the enzyme is located in the region between the two domains. Movement of the domains relative to one another thus controls access to the active site and the catalytic activity of the enzyme. Proteins that are composed of only one polypeptide chain are known as monomeric proteins. Proteins composed of more than one polypeptide chain are known as multimeric proteins. An example of a multimeric protein is chorismate mutase (5csm) which is composed of three polypeptide chains as shown in figure 1.5(d) (Strater et al., 1997). The full arrangement of a number of polypeptide chains bound to each other to form a complex is termed the quaternary structure of a protein (see figure 1.5(d)). The level at which independent proteins associate with each other permanently or otherwise to form a protein complex is sometimes denoted the quaternary structure of a protein.

1.4 Protein Symmetry and Quaternary Structure

There are a number of descriptions of the way that protein chains interact within a complex. The first way of describing the interactions is to look at the spatial arrangement of protein subunits that make up the complex. This means describing the symmetry or asymmetry of the protein complex. The second approach is to look at the surfaces that make up a protein-protein interface. In complexes made up of identical protein subunits (homo-complexes) there are two classes of protein-protein interfaces. Protein interfaces within homo-complexes are either isologous or heterologous (Monod, Wyman, and Changeux 1965). An isologous interface is one made up from the same sets of residues from each of the interacting subunits. The interface between the two subunits that make up thymidylate kinase (5tmp) is isologous as in figure 1.6(a) (Lavie et al., 1998). These surfaces must be self-complementary. A heterologous interface is formed from two different sets of residues. The interfaces within chloramphenicol acetyltransferase (3cla) are all heterologous being formed from different surfaces as in figure 1.6(b) (Leslie, 1990).

(a)



(b)

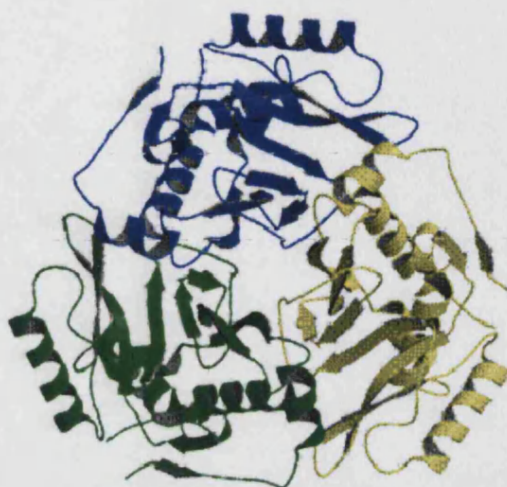


Figure 1.6: (a) The interface between the two subunits of thymidylate kinase is isologous being formed by the same sets of residues on either side of the interface (5tmp, Louie et al., 1998). (b) The chloramphenicol acetyltransferase trimer (3cla, Leslie et al., 1990). The subunit interfaces of 3cla are all heterologous being formed by different sets of residues from either subunit.

Isologous interactions give rise to protein complexes with a finite number of subunits. In contrast heterologous interactions between proteins can lead to protein subunits polymerising indefinitely. Examples of this are the polymerisation of tubulin heterodimers into microtubules and the formation of actin filaments. The fact that isologous interactions between proteins give rise to protein complexes of a finite size points to isologous rather than heterologous interactions being prevalent in homo-complexes. Monod et al., 1965, predicted this by considering the ways that protein complexes might have evolved. Using isologous interfaces alone Monod also forecast that dimers and tetramers would be the prevailing classes of protein complex. All examinations of the available protein structures have confirmed that this is indeed the case (Jones & Thornton, 1995, Goodsell & Olson, 2000). Cornish-Bowden & Koshland, 1970, however concluded on thermodynamic grounds that “there was no evidence for an advantage between isologous versus heterologous binding pairs in protein design”.

Several authors have commented on the symmetry of protein complexes and its consequences for protein function (Goodsell & Olson, 2000, Blundell, 1996 and

references therein). Homo-complexes are generally symmetric. The proteins in figures 1.6(a) and (b) illustrate this and possess two fold and three fold rotational symmetry respectively. Protein complexes composed out of non-identical protein subunits can be symmetric but are not necessarily so. From a geometric point of view there are a limited number of ways of packing together a number of identical proteins to form a 'closed' structure (as opposed to an 'open' chain-like structure). The symmetry of some protein complexes may simply be a reflection of this.

1.5 *The Forces driving Protein-Protein Interactions*

In thermodynamic terms the active three dimensional structure of a protein is only marginally more stable than the protein in its unfolded state. Correspondingly a protein complex is often not much more stable than any of its component proteins in their free states. The thermodynamic quantity that determines whether two proteins will bind to each other to form a complex in solution is the change in Gibbs free energy of the reaction ΔG . The change in the free energy of a reaction is given by equation 1.1.

$$\Delta G = \Delta H - T\Delta S \quad (1.1)$$

For a reaction to occur ΔG must be negative. The more negative the ΔG of a reaction the greater the stability of the resulting protein complex. ΔH is the enthalpy change of the reaction and T is the absolute temperature. ΔS is the change in entropy (or disorder) of the system. The physical origin of the ΔS term lies in the change in flexibility of the residues that make up each of the interacting proteins in their going from a free to a bound state and the release of ordered water molecules into bulk solvent. When two proteins associate the side chains of the amino acids at the protein-protein interface become less flexible. There is therefore a loss in entropy which opposes binding. A favourable ΔH results from the formation of non-covalent interactions such as inter-subunit hydrogen bonds, van der Waals interactions and salt bridges. These interactions are described in detail in sections 1.5.2.1 to 1.5.2.3. An advantageous $T\Delta S$ term arises primarily from the hydrophobic effect detailed in

section 1.5.1 and the increase in the entropy of solvent molecules that accompanies protein-complex formation.

1.5.1 The Hydrophobic Effect

The hydrophobic effect is the major force driving both protein folding and protein-protein interactions. The presence of a protein with exposed hydrophobic regions in aqueous solution disrupts the extensive network of hydrogen bonds that exist between water molecules. To minimise this disruption water molecules form relatively well ordered cages around the exposed non-polar amino acid side chains. The result of this is that the entropy of the water molecules decreases since some of them are in a more ordered configuration. This is thermodynamically unfavourable. The aggregation of hydrophobic surfaces in solution so as to minimise the disruption to the electrostatic interactions that occurs in a polar solvent is known as the hydrophobic effect. The strength of the hydrophobic effect is a matter of some debate. Chothia and Janin, 1975, have estimated that every Angstrom squared of buried surface area gives rise to 25 Calorie mol⁻¹ of free energy. However some estimates of this figure are closer to 50 Calorie mol⁻¹. Recent evidence points to the free energy being closer to the 25 Calorie mol⁻¹ mark (Raschke, 2001). The sheer magnitude of the free energy produced by the hydrophobic effect makes it a major force in driving protein-protein interactions. The hydrophobic effect is non-specific in nature and in vivo hydrophobic surfaces will associate indiscriminately.

1.5.2 Non Covalent Interactions

Non covalent interactions provide the remainder of the free energy required to form a protein complex. Non-covalent interactions also play a large part in ensuring that protein-protein interactions are specific in nature.

1.5.2.1 Hydrogen Bonds

A hydrogen bond is formed between a hydrogen atom bound to an electronegative atom and another electronegative atom. Hydrogen bonds are strongest when both the hydrogen bond donor and acceptor lie in a straight line. In general the greater the

angle (or distance) between the donor and acceptor the weaker the bond. The strength of a hydrogen bond is usually taken to be between 3-7 Kcal Mol⁻¹ depending on the geometry of the bond and dielectric constant of the local environment. A comprehensive review of hydrogen bond geometries and strengths has been compiled by Hubbard, 2001. The number of hydrogen bonds across a protein-protein interface is proportional to its size. There are roughly 0.88 inter-subunit hydrogen bonds for every 100Å² of buried surface area (Jones & Thornton, 1996). The burial of an unpaired hydrogen bond donor or acceptor is extremely unfavourable in thermodynamic terms and will destabilise the interface between two interacting proteins. The presence of hydrogen bond donors and acceptors at a protein interface (and the consequent need to make sure that they are paired up correctly) is an effective mechanism by which a protein will bind with another protein in a highly specific manner. In consequence while protein-protein interactions are primarily driven by the hydrophobic interaction hydrogen bonds largely confer specificity on such interactions (Fersht, 1987).

1.5.2.2 van der Waals Interactions

All atoms and molecules have fluctuating clouds of electrons about them. An asymmetric charge distribution about an atom will cause it to have a transient dipole. A transient dipole in any given molecule or atom induces opposite dipoles in its neighbours. The resulting attraction between two opposite dipoles is the physical basis for van der Waals interactions. van der Waals interactions are relatively weak and short range in nature with strengths of ~1 Kcal Mol⁻¹. As such van der Waals interactions can only contribute significantly to the free energy of binding when large numbers of residues are in close proximity to one another. This fact helps to explain why protein-protein interfaces must be complementary in shape to one another.

1.5.2.3 Salt Bridges and Other Electrostatic Interactions

A salt bridge is formed by the electrostatic attraction between oppositely charged groups. Salt bridges usually involve lysine and arginine residues and aspartic acid and

glutamic acid residues (Lakey & Gokce, 2001). Most proteins contain only one or two inter-subunit salt bridges (Xu et al., 1997). There is some evidence to suggest that proteins that exist in harsh environments such as the extracellular space or in extreme temperatures are somewhat enriched in salt bridges (Scandurra et al., 1998). The delocalised ring of π electrons about the aromatic groups of some amino acids can also interact strongly with positively charged groups (e.g. the $-\text{NH}_3^+$ group of lysine). The existence of groups capable of forming salt bridges and other varieties of electrostatic interactions contributes not only to the binding energy of a protein-protein interaction but perhaps to a greater extent its specificity. It is very important to note that the strength of any electrostatic interaction is proportional to $1/\epsilon$ where ϵ is the dielectric constant of the local environment. For water ϵ varies from 88 at 0 °C to 55 at 100 °C. The dielectric constant of water is usually taken to be around 80. In protein environments ϵ is thought to be between 1 and 4 (Krumrine et al in Bourne & Weissig, 2003). The strength of a salt bridge is therefore quite dependent on whether the interacting residues are exposed to solvent or not.

1.6 Physical Characteristics of Protein-Protein Binding Sites

In this section some of the physical characteristics of protein-protein binding are outlined. A detailed review of the chemical characteristics of protein-protein binding sites is given in chapter 3.

1.6.1 Size (Accessible Surface Area)

Accessible surface area (ASA) is used to define molecular surfaces and quantify the areas over which residues are accessible to solvent. As such ASA calculations are fundamental in the study of proteins and their interactions with other proteins and are used extensively in later chapters. The ASA of a residue (or indeed any other group of atoms) is defined as the area formed by rolling a spherical probe of radius R over the atoms of that residue whilst maintaining contact with the atoms of that residue and no other (Lee & Richards, 1971). The value of R is usually taken to be 1.4Å.

The comparison of the ASA of an individual residue compared with the ASA that the residue would have when part of a completely unfolded polypeptide chain is known

as its relative accessible surface area (rASA). The rASA of a residue can be used to classify it as being on the surface of a protein, in its interior, or in a protein-protein interface. The scheme used to classify residues as being part of a proteins interior, exterior, or as part of a protein interface on the basis of rASA measurements is detailed further in chapter 3. The definition of the residues as being interior, exterior, or interface permits the chemical characteristics of these subsets of residues to be elucidated.

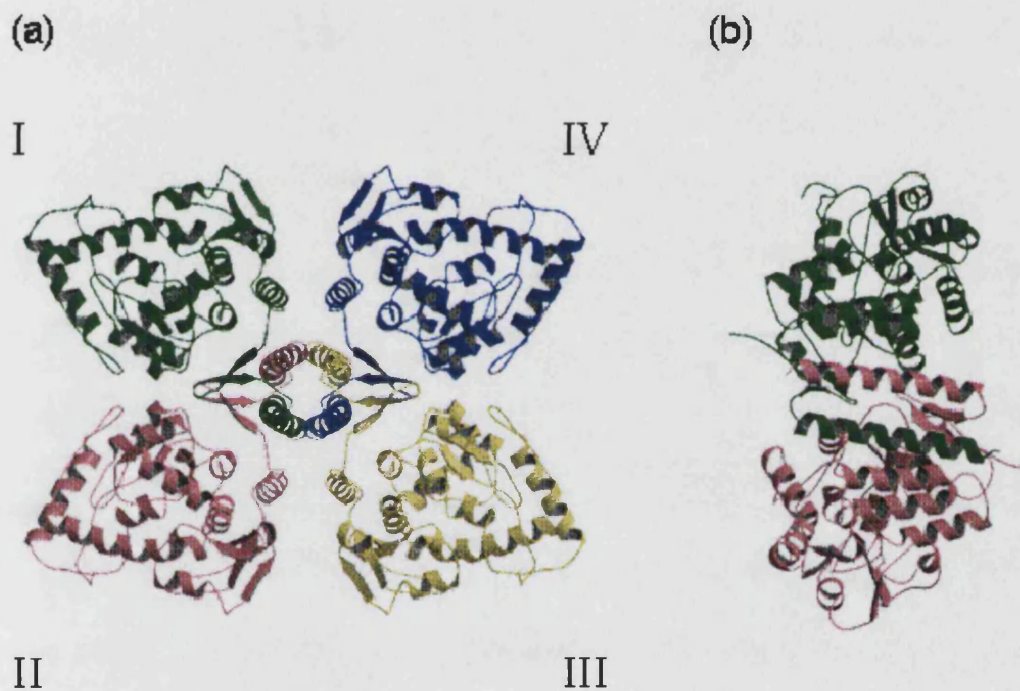


Figure 1.7: (a) The tyrosine hydroxylase tetramer (1toh, Goodwill et al., 1997). The area of contact between subunits I and II is 1650\AA^2 in size. In contrast the area of contact between the subunits labelled I and IV is only 540\AA^2 . 1toh can be considered to be a ‘dimer of dimers’, the dimer subunit being either subunits I and II in (b) or subunits III and IV.

The measurement of a protein’s ASA together with its molecular weight can also be used to find out how extended or compact is the overall shape adopted by a protein structure. The size (ASA) of protein-protein interfaces can yield useful information concerning the evolutionary history of proteins. For example many multimeric proteins can be considered as composites of lower multimers. Tyrosine hydroxylase in figure 1.7(a) is one such protein being a homo-tetramer comprised of a ‘dimer of dimers’ (Goodwill et al., 1997). The protein is involved in neurotransmitter biosynthesis and defective forms of the hydroxylase have been implicated as playing a

role in various psychiatric conditions such as bipolar affective disorder (Smyth et al., 1996). Subunits I and II in figure 1.7(b) are bound together into a dimer through an extensive interface $1,650\text{\AA}^2$ in size. In contrast subunits I and IV interact with each other through a tetramerization domain with an interface of only 540\AA^2 . Deletion of the tetramerization domain results in enzymatically active monomeric proteins (Vrana et al., 1994 & Lohse et al., 1993). This result together with the comparison contact areas between protein subunits may point to the enzyme having an evolutionary history in which the protein began as a monomer which then evolved to form a functional dimer and finally through association of dimers into a tetramer.

Another major application of ASA is looking at the size of protein-protein interfaces to distinguish between biological and non-biological contacts in protein crystals grown for X-ray crystallography. Proteins pack against each other in a number of different orientations in the crystalline state and discriminating between biologically relevant protein-protein interfaces and those which are artefacts of the crystallisation process is a requirement in determining the biologically relevant quaternary structure of a protein. The size of protein-protein interfaces is a powerful discriminator between biological and non-biological contacts within protein crystals. The Protein Quaternary Server (PQS, <http://pqs.ebi.ac.uk>) uses contact area as measured using ASA together with other physical quantities to determine the likely quaternary structure of a protein. The quaternary structures of 78% of a dataset of 76 homo-dimers are correctly predicted by the PQS (Postingl, personal communication, 2002). Using only the ASA of the protomers in the crystal, 76 homo-dimers and 96 monomers are predicted as being biologically relevant or not with a success rate of 84.6%. Considering the conservation of residues at the sequence level in addition to other methods this accuracy rate has further improved (Postingl, 2000, Valdar & Thornton, 2001).

1.6.2 Shape Complementarity

Two proteins may only associate in a specific manner if there is complementarity (both geometric and electrostatic) at the regions where they interact. The two protein surfaces that make up a protein-protein interface are therefore generally highly complementary in shape to each other. Several methods have been proposed to

quantify this. Lawrence et al., 1993, propose a shape correlation statistic S_c to quantify the geometric similarity between surfaces at the protein-protein interface. In this approach vectors normal to the surfaces of interacting proteins are defined. A comparison of the relative direction of these vectors then allows the shape complementarity of protein surfaces to be assessed. The value of the correlation statistic S_c lies between 0 and 1 with an S_c of 1 indicating perfect complementarity between protein surfaces. Antibodies have been observed to have a poorer surface complementarity with their respective antigens than other categories of protein interactions (Lawrence & Colman., 1993, Decanniere et al., 2001). This is unsurprising since antibodies evolve on a very rapid time scale to recognize their respective antigens relative to proteins involved in any other biological interaction.

1.6.3 Packing at the Protein-Protein Interface

Residues at the protein-protein interface are closely packed. They have to be. The interactions which maintain the three dimensional structure of a protein are the same as those which hold proteins together within a protein complex. These interactions such as hydrogen bonds and the hydrophobic effect are all short range in nature and can only become appreciable when large numbers of residues come into close proximity to each other. This necessitates that the residues at a protein-protein interface pack together in a similar way to those in the protein interior. The packing of residues at the protein-protein interface can be quantified using Voronoi polyhedra (Richards, 1974). Voronoi polyhedra have been used extensively to evaluate the packing of residues in proteins and a full account of this method is given in chapter 3. Conte et al., 1999, used the Voronoi method to calculate the packing of residues at the protein-protein interfaces of 75 protein-complexes. The protein-protein interface was indeed shown to be almost as closely packed as the protein interior in nearly every protein-complex. Furthermore, the volumes of amino acids in the protein interface were shown to be on average 5% smaller than the volumes occupied by amino acids in small molecule crystals, illustrating this fact.

1.7 *Classifying Protein-Protein Interactions*

Protein interactions can be described and classified in a number of different ways. One way of classifying protein interactions is to look at the biological function of the interacting proteins. Thus for example the interaction between elastase and elafin is designated as an enzyme-inhibitor interaction. Aside from looking at the functions of interacting proteins there are many other ways of describing the nature of a protein-protein interaction. Some of the different descriptors used in later chapters to characterise protein-protein interactions are set out in the following sections.

1.7.1 Permanent or Transitory

The interaction of two proteins can either have two outcomes. The first is a 'permanent' protein complex in which its constituent subunits are permanently bound to each other *in vivo*. The second outcome is a protein complex which is to some degree 'transitory' in which the interacting proteins are not permanently bound to each other and separate at some later stage. Citrate synthase (1csh) is a homo-dimer in which its subunits are permanently bound to each other. The complex between the human nerve growth factor and its receptor (1www) is an instance of a transient protein complex. The physiological conditions in which interacting proteins exist must be taken into account in deciding whether the resulting protein complex is either permanent or transitory.

1.7.2 Homo or Hetero

Protein interactions can be between identical proteins giving rise to a homo-complex or between different proteins resulting in a hetero-complex. An example of a homo-complex is chorismate mutase (2chs) as in figure 1.5(d) (Chook & Lipscomb, 1993). Haemoglobin is a hetero-complex composed of two non-identical subunits arranged in a $(\alpha\beta)_2$ organisation.

1.7.3 Affinity (Strength of Interaction)

The affinity with which two proteins interact is usually measured by determining the dissociation constant (K_d) of the complex. Proteins interact with a huge range of affinities ranging from proteins with a dissociation constant of 10^{-4} M^{-1} to 10^{-16} M^{-1} (Kleanthous, 2000). The dissociation constants of a number of enzyme-inhibitor complexes are examined in chapter 4. The inhibition of human angiogenin by the placental ribonuclease inhibitor (1a4y) is an exceptionally high affinity reaction with a dissociation constant of 10^{-16} M^{-1} . The inhibition of thrombin by hirudin is another high affinity reaction with a dissociation constant of 10^{-14} M^{-1} . In contrast red abalone lysin (2lyn) exists in a dynamic equilibrium between dimeric and monomeric forms (Kresge et al., 2000, Nooren & Thornton, 2003). The dissociation constant of the dimeric form of lysin is low (10^{-6} M^{-1}) illustrating the low stability of the complex.

1.7.4 Specificity

Protein interactions can be specific, non-specific or multi-specific in nature. Most protein-protein interactions are highly specific. The biological consequences of the unwanted aggregation of proteins are often severe. An example of this is the aggregation of amyloid fibers, which plays a part in the Alzheimer's disease (Auld et al., 2002).

The interaction of an antibody with its antigen is usually highly specific. An example of a specific antibody-antigen interaction is the interaction of the 13B5 antibody with the HIV p24 viral coat protein detailed in chapter 2.

Examples of protein-protein interactions that are non specific or only marginally specific are comparatively rare. The major histocompatibility complex typically binds to a large number of different peptide antigens and hence constitutes a class of low-specificity protein interactions.

The inhibition of various different serine proteases by the bovine pancreatic trypsin inhibitor (BPTI) is an instance of a multi specific protein-protein interaction.

1.7.5 Biological relevance

As discussed in section 1.5.1 the conditions under which proteins crystallise are very different to the ones that they are subject to in vivo. This can result in proteins participating in any number of contrived and artificial interactions with each other when growing crystals for x-ray crystallography. An example of a protein that crystallises in a form that is not biologically relevant is the cell cycle regulatory protein CksHs2 (Parge et al., 1993). The protein crystallises as a hexamer. Gel chromatography experiments however show that CksHs2 is a dimer in solution. Hexameric CksHs2 may be a low energy storage form of the protein but it is certainly not the biologically active one.

1.8 *Reasons for Oligomeric Proteins*

The reasons why proteins aggregate with one another to form oligomers are manifold. The most basic reason why certain proteins polymerise is simply the biological need for large structures. Prime examples of large protein structures are actin and collagen filaments. Such structures are polymers of identical subunits and are thus often symmetric. A class of large protein structures that must be comprised of a number of non-identical subunits are molecular motors such as ATP synthase and the ribosome. Other classes of large oligomer are numerous. Some of the possible advantages of a protein being an oligomer have been enumerated by other authors (D'Alessio 1999, Goodsell & Olson 2000) and are set out below. All of these advantages are equally applicable to proteins with internal structure such as multi-domain proteins.

1.8.1 Error Control

Errors inevitably occur during protein synthesis resulting in potentially defective proteins. The likelihood of a mistake occurring during protein synthesis increases with the length of the polypeptide chain being synthesised. Making a protein out of several identical or non-identical protein chains rather than one long polypeptide chain is therefore an important way of minimising the potentially disastrous effects resulting from errors that arise during the transcription and translation of the genetic code (Ibba & Soll, 1999). A related advantage of producing a protein composed of

several chains is related to the intrinsic instability of many protein subunits outside of a complex. If a protein is produced with a significant numbers of errors in it then those errors may prevent it from binding to other proteins to form a complex. Its lack of intrinsic stability means that the defective protein quickly denatures. It follows that producing a protein composed of multiple chains rather than one can potentially act to ensure that defective proteins are discarded.

1.8.2 Coding Efficiency

Expressing the same gene a number of times is a genetically compact way of coding for large structures composed of identical subunits (Crick & Watson, 1957). The need for coding efficiency is greatest in organisms with small genomes where genetic space is at a premium such as viruses. Viral capsids are an extreme result of this need often being composed of hundreds of identical subunits. For instance the bluetongue virus capsid (1bvp) is composed of 760 identical subunits. In contrast mammalian genomes such as those of *Homo sapiens* contain large amounts of non-coding DNA (Venter et al., 2001, Human Genome Consortium, 2001). There is therefore little need for genetic compactness in such organisms.

1.8.3 Reduction of Surface Area

By forming an oligomer a protein reduces the net amount of surface area (ASA) exposed to solvent and results in the formation of areas of contact between the proteins that make up the oligomer. Any reduction in the surface area of a protein exposed to solvent will reduce the disruption to the extensive network of hydrogen bonds that occurs within solvent. This is thermodynamically favourable and helps to improve the solubility of the protein (Goodsell & Olson, 2000).

1.8.4 Stability

Protein oligomers have the potential to be more stable than their constituent subunits. A basic reason for this is that proteins form multiple subunit interfaces. The need for stability is universal and enzymes are no exception. An example of a protein where this is the case is adenylate kinase (Vonnrhein et al., 1998). The enzyme is a monomer

in most organisms except in *Sulfolobus acidocaldarius* where it is trimeric. The bacteria *Sulfolobus acidocaldarius* grows around volcanic vents and around hot springs at temperatures in excess of 80°C. Each monomer within the trimer is stabilised by a large number of inter-subunit salt bridges and hydrogen bonds which collectively ensure that the protein can retain its native structure and enzymatic activity under the harsh conditions in which it is naturally found.

1.8.5 Regulation of Activity

Moving subunits relative to each other is a well known way of regulating the activity of a protein. There are broadly two main instances of this (although there are other examples).

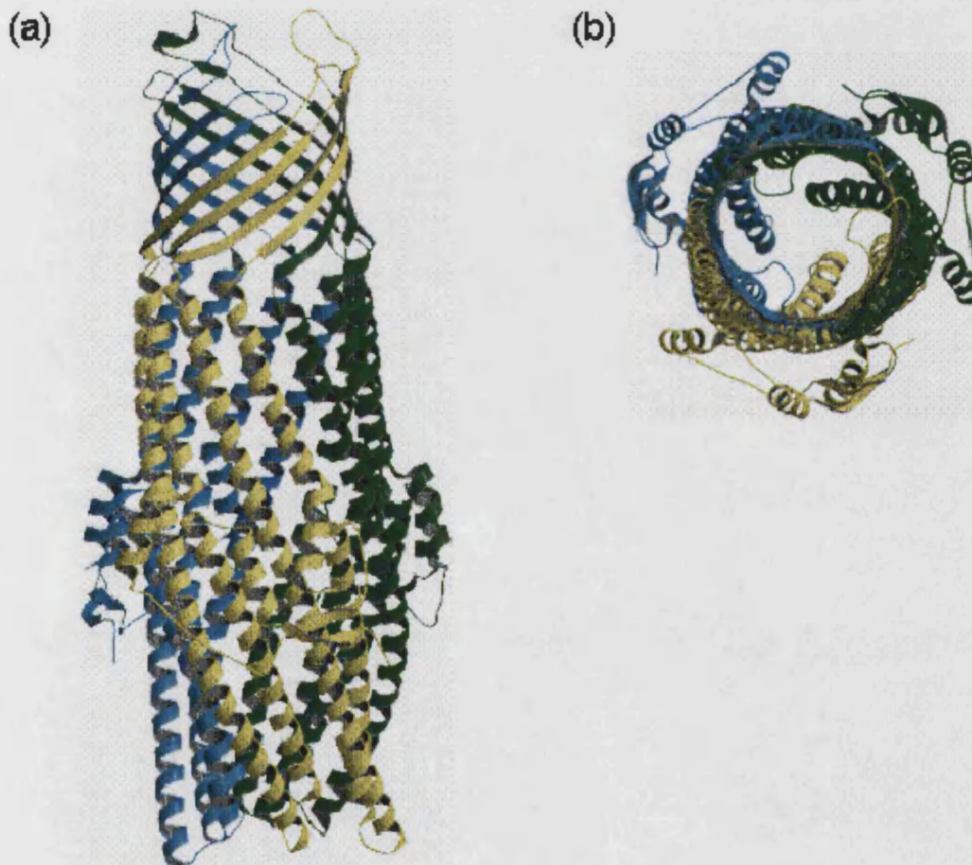


Figure 1.8: (a) The TolC integral outer membrane protein from *Escherichia coli* (Koronakis et al., 2000). (b) The coiled α -helices at the lower end of the protein form a three way valve which opens in response to TolC interacting with translocase (an inner membrane protein).

The first is where there are enzyme active sites or other binding pockets at a subunit interface. Movements between the subunits of such proteins will then affect the ability of the protein to interact with other proteins or catalyse a reaction. The second instance is where the protein's activity is regulated by means of allostery. In allosteric proteins the conformation of the protein changes between a biologically active and inactive state in response to ligand binding or covalent modification. There are many examples of proteins regulated by allosteric movements including haemoglobin, aspartate carbamoyltransferase (ATCase), and the bacterial membrane protein TolC. Three TolC subunits form a trimeric channel 140 Å² in length sealed at its end by a number of coiled helices as shown in figure 1.8(a) and (b). When TolC interacts with an inner membrane protein translocase the coiled helices at the end of protein open allowing the passage of a range of diverse proteins and other molecules through the channel (Koronakis, 2000).

1.8.6 Enzymatic Efficiency and Function

The catalytic efficiency of an enzyme can be enhanced if the enzyme is multimeric. The reasons for this are complex. The formation of a multimer 'hides' some of the surface area of each of its constituent subunits not involved in catalysis. Firstly a multimeric enzyme with several identical active sites has a larger cross sectional area than any of its component subunits. This increased cross sectional area can make a collision between a multimeric enzyme and substrate more likely than one between the substrate and the monomeric enzyme. Secondly it has been postulated that once a substrate collides with an enzyme it then performs a two-dimensional random walk along the enzyme's surface. The presence of multiple active sites and the reduced amount of surface area not involved in catalytic activity in a multimeric enzyme can also increase the efficiency with which a substrate diffuses to its active site compared with its monomeric counterpart.

Functional diversity can also be enhanced if an enzyme is multimeric. Some enzymes are composed of a number of subunits each of which catalyses a different reaction. This enables a number of chemical reactions to be catalysed within the same complex rather than by a series of sequential enzyme-substrate reactions which is

comparatively inefficient. One of the best characterised multi-enzyme complexes is tryptophan synthase which catalyses the last two stages in the biosynthesis of tryptophan (Miles, 2001).

1.9 Rationale and Outline of Thesis

The *general* principles governing protein interactions have been established in previous studies (Janin & Chothia, 1975, Jones & Thornton, 1996). However these and other studies have noted the diverse nature of protein-protein interfaces and underlined the fact that they are often not very much different from any other part of the protein exterior. In short it is still not clear what makes a binding site a binding site. Any significant improvements in designing ways of predicting the location of protein-protein interaction sites are likely to be dependent on better defining the characteristics of a binding site. The ever increasing number of protein structures makes the characterisation of protein-protein binding sites possible to an extent that has not been feasible in the past. As well as being of fundamental importance in its own right doing this may provide further insights in designing new or improving on existing methods to locate protein binding sites from structure alone.

In the introduction to this thesis the enormous diversity of proteins and their interaction with other proteins has been outlined. In chapter 2 the procedure that was used to compile each dataset of proteins studied in later chapters is outlined. The contents of each dataset are tabulated and presented. Chapter 3 is concerned with the characterisation of the datasets of obligate homo-proteins. The properties of the protein-protein interfaces in these datasets are characterized in terms of their size, planarity, and hydrophobicity. The number of hydrogen bonds and salt bridges across the protein-protein interface is also detailed and related to the size of the protein-protein interface. The bulk properties of each protein such as its shape and the relationship between the ASA of a protein and its molecular weight are given. In chapter 4 the obligate hetero-protein interfaces are also characterized using the same physical descriptors as in chapter 3. The properties of proteins in the datasets that describe non-obligate interactions (enzyme-inhibitor and antibody-antigen complexes together with proteins involved in signal transduction) are also presented in chapter 4.

The culmination of the work presented in chapters 2-4 is the comparison of the protein-protein interfaces in proteins representing the different modes of protein-protein interaction. A comparison is made between the properties of obligate and non-obligate protein complexes. A corresponding comparison is then made between proteins composed of identical (homo) and non-identical (hetero) subunits. In chapter 5 a neural network is used to improve on an existing method known as Patch Analysis (Jones & Thornton, 1997a). The neural network based Patch Analysis method is used to predict the location of the protein-protein interfaces of the protein-complexes in five of the datasets of proteins studied in chapters 3 and 4. The conclusions of all the work in this thesis are made in chapter 6.

One aspect of protein-protein interfaces that has not been addressed directly in this thesis is whether residues at a protein-protein interface are more conserved at the sequence level than the remainder of the proteins surface. Whether or not interface residues are indeed conserved at the sequence level has been addressed by many authors over the last few years and is currently the subject of much research and is thus not addressed substantially in this thesis (Lichtarge & Sowa, 2002, Valdar & Thornton, 2001, Bartlett et al., 2002, Nooren et al., 2003). There is evidence to suggest that residues at protein-protein interfaces are indeed often conserved. Valdar and Thornton (2001) analysed the protein-protein interfaces of six families of homo-dimers and concluded that the protein-protein interfaces of these proteins are indeed more significantly conserved than the remainder of the proteins exterior. Bartlett et al., 2002, found that residue within the active sites of enzymes are also conserved. Nooren et al., 2003, looked at the conservation of residues at the protein-protein interfaces of transient protein-complexes. It was concluded that that the “interface residues of the weak transient homo-dimers are generally more conserved than surface residues” and that “protein families that include members with different oligomeric states or structures are identified, and found to exhibit a lower sequence conservation at the interface” (Nooren et al., 2003). If it is that residues at protein-protein interfaces are significantly conserved can this information be used to help predict the location of protein-protein interfaces? This question is briefly considered in section 5.7 of chapter 5 where conservation score data is used together with other parameters to locate the protein-protein interfaces of 53 homo-dimers.

Chapter 2

Datasets of Multi-Subunit Protein Complexes

2.1 Introduction

In this chapter the procedure that was used to compile non-redundant and non-homologous datasets of proteins separated out by some property is described. The construction of such datasets is a necessary pre-requisite to studying and characterizing different types of protein-protein interactions. At a basic level the interaction of proteins with each other can be considered to result in permanent (obligate) and non-permanent (non-obligate) complexes. Datasets of obligate proteins composed of identical (homo) and non-identical (hetero) subunits have been assembled according to their multimeric state. The PDB is sparsely annotated regarding the biologically relevant multimeric state of protein structures. In view of this an entry is only included in a dataset if experimental evidence confirming the multimeric state of the protein in solution can be found in the literature or other protein databases. The datasets of obligate homo-complexes consist of 76 homo-dimers, 26 trimers, 31 tetramers, and 9 hexamers. The corresponding datasets of obligate hetero-proteins consist of 10 hetero-dimers, 7 tetramers, and 3 hexamers. Biological processes involving non-permanent protein complexes are extremely numerous and diverse in nature ranging from protein synthesis to apoptosis. In view of this, datasets representing three major categories of protein interactions resulting in non-permanent protein complexes have been compiled. These datasets include 20 enzyme-inhibitor complexes, 15 antibody-antigen complexes, and 10 proteins involved in signal transduction. All datasets are non-homologous to ensure that they are as representative as possible. Every protein within each dataset of obligate homo-proteins shares a sequence identity of less than 25% with any other protein within the dataset. With the exception of the antibody-antigen dataset each protein within every

other dataset contain one or more chains that have a sequence identity of less than 25% with a chain belonging to any other protein in the dataset.

2.2 Protein Databases

Three major databases are used to construct the datasets and ensure that they are non-redundant and non-homologous. These databases are the PDB, Swiss-Prot, and the FSSP. The co-ordinates of protein structures are obtained from the PDB (Berman et al., 2000). The PDB is the global archive of protein structures in the world which in July 2003 contained the co-ordinates of some 17,000-protein structures solved by x-ray crystallography. The level of annotation in the PDB is highly variable. Some protein structures are clearly labelled with their relevant multimeric state while most structures are not. Additionally the non-uniformity in the PDB file format can make the extraction of such information where it does exist difficult. The PDB data uniformity project has gone some way to addressing this problem but was only just underway when many of the datasets detailed in this chapter were being compiled (Westbrook et al., 2002). This makes the extraction of protein structures of a given multimeric state from the PDB alone impossible.

As a result a second protein database, Swiss-Prot (Bairoch & Apweiler, 2000) was used in assembling the datasets. Swiss-Prot is a manually curated database of protein sequences together with related structural data (where available) and is to be found at <http://www.expasy.ch/sprot>. Every Swiss-Prot entry is also extensively annotated with data relating to the biological function of the protein. The co-ordinates of the biologically relevant multimeric state of the protein (once it is known) are obtained from the PQS (Protein Quaternary Structure, Kim & Thornton, 1999).

The FSSP database (Holm & Sander, 1996) is used in order to make sure that the datasets are non-homologous. FSSP stands for Fold classification based on Structure-Structure alignment of Proteins. All protein chains > 30 residues in length from each structure in the PDB are included in the FSSP database. In the FSSP every protein chain is classified into around 600 non-homologous sequence families. From each sequence family a representative protein chain is taken to form a 'representative set'. The families are non-homologous in that every chain within each sequence family

shares a sequence identity of < 25% with any chain in any of the other sequence families. Protein chains within each sequence family are homologous to each other having a sequence identity of >25% with each other. All-against-all structural comparisons using the Dali server (Dietmann & Holm, 2001) are carried out on the representative set to describe the structural similarities between proteins. A fold index is attached to each protein chain. The first number gives the fold class for the protein chain. Subsequent digits indicate the number of standard deviations of the structural similarity compared to the database average. As an example the prolactin-human growth hormone complex is composed of two chains with FSSP indices of 1.16.1.5.1.1 and 238.1.5.6.1.1 respectively. The first numbers 1 and 238 indicate that the two chains belong to a different fold class and are thus non-homologous to each other. In this way it is possible to screen datasets for homology. The various databases are all organized differently with data being presented in a number of differing formats. In order to address this problem the Sequence Retrieval System (SRS, <http://srs.ebi.ac.uk>) at the EBI is used. Using SRS it is possible to search a database, extract information in other databases that may relate to the search query, and present the results in a user-defined format.

2.3 Assembly of Datasets of Multi-Subunit Complexes

In this section the procedures used to assemble the various datasets of proteins is detailed. As an example the steps that were taken to compile the dataset of homo-trimers is shown in a flow chart in figure 2.1.

- (a) The first stage in assembling a non-redundant and non-homologous dataset of homo-trimers is to search for protein structures annotated as being trimers in the PDB and Swiss-Prot. The PDB is searched for trimers using the search terms 'trimer' and 'homo-trimer' (both with wildcards added). The Swiss-Prot database is then searched using the same search terms.
- (b) The result is lists of PDB and Swiss-Prot entries each containing some reference to the entry being that of a homo-trimeric protein.

- (c) Each Swiss-Prot entry in the list is then labeled up with the PDB code of the PDB entry that it refers to. The PDB and Swiss-Prot derived lists are then merged to produce a unique list of PDB codes.
- (d) Every entry in the list is then further labeled with the FSSP code of each chain within the protein and the resolution to which the structure is determined. The final list is then sorted according to the FSSP code of the protein chains. The highest resolution structure from each FSSP fold family in the list is then selected as a representative and is provisionally deposited in the dataset. Only proteins whose structures have been determined by x-ray crystallography to a resolution $\leq 3 \text{ \AA}$ are included in the dataset.
- (e) The result is lists of PDB and Swiss-Prot entries each containing some reference to the entry being that of a homo-trimeric protein.
- (f) Each Swiss-Prot entry in the list is then labelled up with the PDB code of the PDB entry that it refers to. The PDB and Swiss-Prot derived lists are then merged to produce a unique list of PDB codes.

The final stage is to manually check the literature and the Swiss-Prot entry of each protein in the dataset to confirm that the predominant biologically relevant state of the protein in solution is trimeric. This is difficult. Often the biologically relevant quaternary structure of a protein is simply not known or there is contradictory experimental evidence.

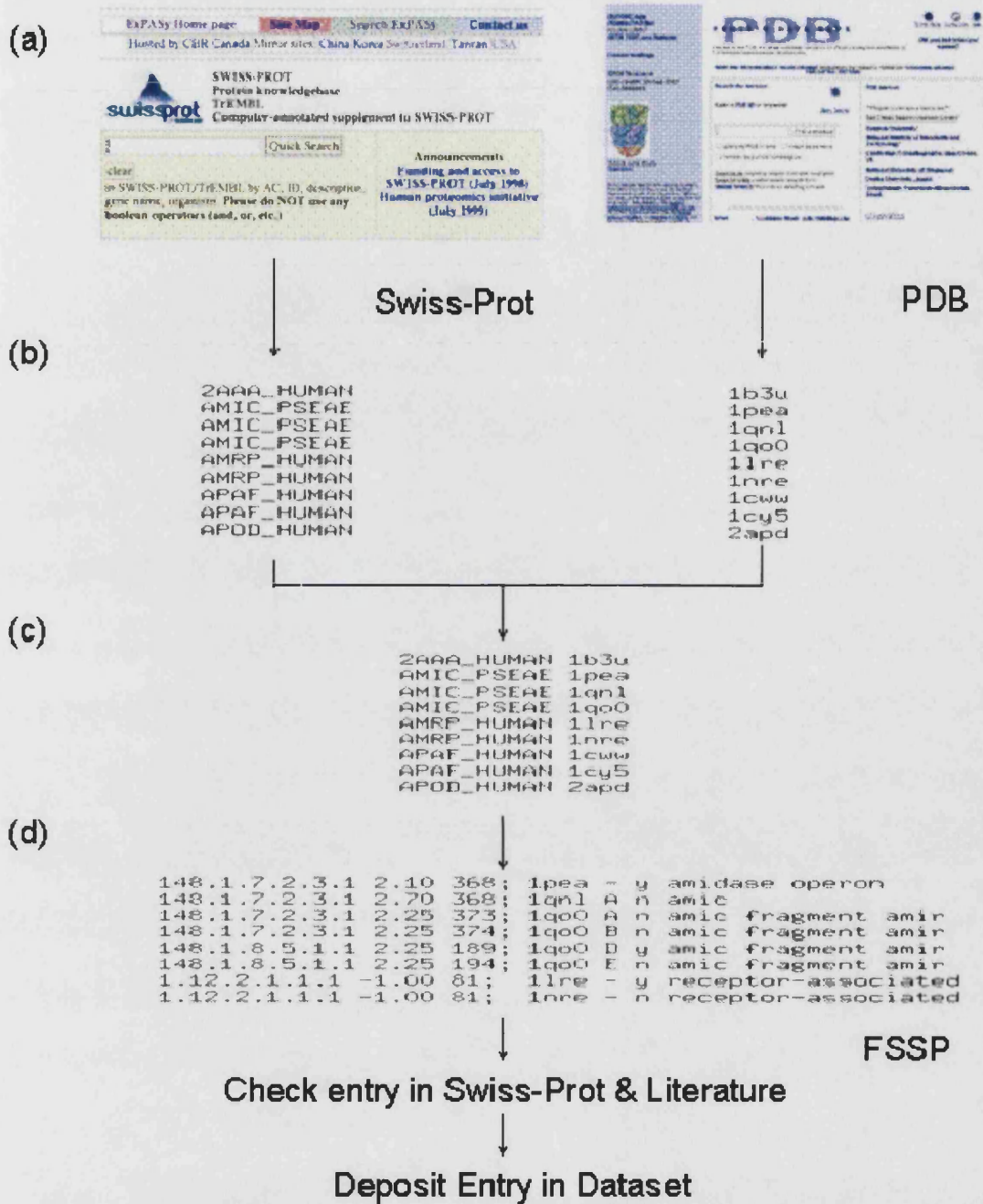


Figure 2.1: Compiling the datasets. (a) The PDB and Swiss-Prot are searched for proteins annotated with the required multimeric state or other property. The resulting lists of entries from the PDB and Swiss-Prot shown in (b) are then merged to produce the list of non-redundant protein structures depicted in (c). Each protein structure is then labelled up with its FSSP code and the resolution that it has been solved to. The highest resolution structure from each FSSP fold family is then checked using Swiss-Prot and any relevant literature to confirm the multimeric state of the protein and that all protein-protein interfaces are complete. The structure is then finally deposited in the dataset.

Usually what experimental evidence there is relating to the quaternary structure of a protein is in the form of ultracentrifugation, light scattering, gel-electrophoresis, and SDS-page experiments amongst others. An entry is only included if such evidence can be found to confirm the multimeric state of the protein in solution.

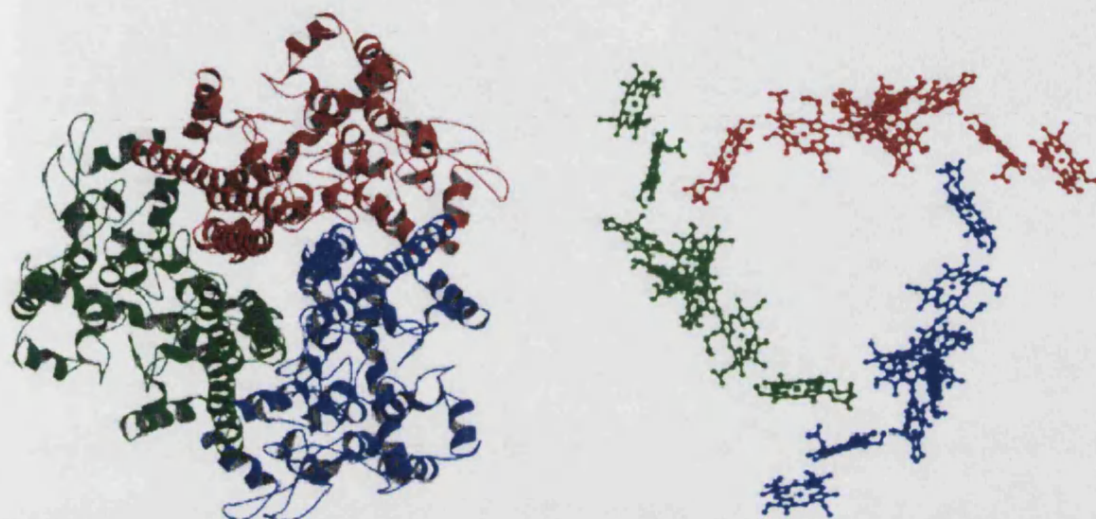


Figure 2.2: Hydroxylamine oxidoreductase (1fgj, Igarashi et al., 1997). (a) The functional homo-trimer. (b) The 24 heme groups lining the interior of the trimer.

There are a very small number of exceptions to this. These are invariably proteins whose ability to carry out a particular function is dependent on existing in a particular multimeric state. As an example hydroxylamine oxidoreductase (1fgj) shown in figure 2.2(a) is a trimeric protein but no experimental evidence could be found in the literature to confirm this. However trimerization is necessary for the catalytic activity of the enzyme. Twenty-four heme groups (eight from each subunit) line the basin of the protein forming an electron transport chain as shown in figure 2.2(b). In addition the substrate binding sites of the enzyme are thought to be located at the subunit interfaces (Igarashi et al., 1997). In light of this and the large size of the subunit interfaces (5300\AA^2) one can be reasonably certain that the functional form of the protein is a trimer. Finally each entry in the dataset is then checked in the PDB and Swiss-Prot to ensure that it is complete. If the entry is not complete but all inter-subunit binding regions are still represented in its PDB entry then the protein is retained in the dataset, otherwise the protein is discarded. In addition non-soluble proteins (primarily membrane proteins) were also excluded. The co-ordinates of the trimeric form of each protein in the dataset are then obtained from the PQS.

In general the above procedure was repeated in constructing each of the other datasets using different search terms to query the PDB and Swiss-Prot. The dataset of homodimers as well as a dataset of 95 monomers used later in chapter 4 were originally compiled by Hannes Postingl at the EBI (Postingl et al., 2000). Every entry in the datasets of obligate and non-obligate hetero-proteins contains only one chain that is non-homologous with all other protein chains in the dataset. For instance in the dataset of enzyme-inhibitor complexes a number of the enzymes are homologs being serine proteases but the inhibitors are non-homologous with each other. Some of the entries in the dataset of signaling proteins are membrane associated. All of the receptor complexes in this dataset are fragments with only the ligand binding regions of the complexes being represented in the PDB entry.

A different procedure had to be employed to compile the dataset of antibody-complexes. All antibodies are members of the immunoglobulin (Ig) super-family and are thus homologous to each other. As well as this most antibody structures in the PDB are fragments with the only the Fab or F_v antigen binding regions being represented. A list of all antibody structures in the PDB is maintained at the University of Reading and is available at <http://www.bioinf.org.uk/abs>. This list was screened by hand in conjunction with the PDB to find all antibody-antigen complexes for which the antigen is a protein. Each entry in the resulting list of complexes is then examined to determine that the antigen-antibody interface is predominantly complete. If the interface is complete then the entry is included in the dataset. The PDB was also searched to see if any structures of the free antibodies exist. Out of the 15 antibody-antigen complexes most of the antibodies have been solved in their free state and can be found in the PDB.

2.4 The datasets

A summary of the contents of all datasets of proteins studied in this work is shown in the table 2.1. Tables showing the contents of each individual dataset are given in the next section. References for individual protein structures are contained in the PDB.

	Obligate	Non-Obligate
Homo-Complexes		
	76 Dimers 36 Trimers 31 Tetramers 9 Hexamers	
Hetero-Complexes		
	10 Dimers 7 Tetramers 3 Hexamers	20 Enzyme-Inhibitors 10 Signaling Proteins 15 Antibody-Antigens

Table 2.1: Summary of the datasets of obligate and non-obligate multimers set out in sections 2.4.1 to 2.4.2.

2.4.1 Datasets of Homo-Complexes

PDB Code	Protein	Source	Resolution (Å)
1a3c	Pyrimidine Operon Regulatory Protein	<i>Bacillus Subtilis</i>	1.60
1ad3	Aldehyde Dehydrogenase	<i>Rattus Norvegicus</i>	2.60
1af5	I-CreI Endonuclease	<i>Chlamydomonas Reinhardtii</i>	3.00
1afw	3-ketoacyl-CoA thiolase	<i>Saccharomyces Cerevisiae</i>	1.80
1ajs	Aspartate Aminotransferase	<i>Sus Scrofa</i>	1.60
1alk	Alkaline phosphatase	<i>Escherichia Coli</i>	2.00
1alo	Aldehyde oxidoreductase	<i>Desulfovibrio gigas</i>	2.00
1amk	Triose Phosphate Isomerase	<i>Leishmania Mexicana</i>	1.83
1aom	Nitrite Reductase	<i>Thiosphaera Pantotropha</i>	1.80
1aor	Aldehyde Ferredoxin Oxidoreductase	<i>Pyrococcus Furiosus</i>	2.30
1aq6	L-2-Haloacid Dehalogenase	<i>Xanthobacter Autotrophicus</i>	1.95
1aou	Carboxylesterase	<i>Pseudomonas Fluorescens</i>	1.80
1bam	Restriction endonuclease bamHI	<i>Bacillus Amyloliquefaciens</i>	1.95
1bif	6-Phosphofructo-2-Kinase/Fructose-2,6-Bisphosphatase	<i>Rattus Norvegicus</i>	2.00
1bsr	Bovine Seminal Ribonuclease	<i>Bovine (Bos Taurus)</i>	1.90
1buo	Promyelocytic Leukemia Zinc Finger Protein Plzf	<i>Homo Sapiens</i>	1.90
1cg2	Carboxypeptidase G2	<i>Pseudomonas Sp.</i>	2.50
1chm	Creatine Amidinohydrolase	<i>Pseudomonas Putida</i>	1.90

1cmb	Met Apo-Repressor DNA Binding Protein	<i>Escherichia Coli</i>	1.80
1cp2	Nitrogenase Iron Protein	<i>Clostridium Pasteurianum</i>	1.93
1csh	Citrate Synthase	<i>Gallus Gallus</i>	1.60
1ctt	Cytidine Deaminase	<i>Escherichia Coli</i>	2.20
1czj	Octaheme Cytochrome c3	<i>Desulfomicrobium Baculatum</i>	2.16
1daa	D-Amino Acid Aminotransferase	<i>Bacillus (Thermophilic)</i>	1.94
1fip	FIS DNA Binding Protein	<i>Escherichia Coli</i>	1.90
1fro	Glyoxalase I	<i>Homo Sapiens</i>	2.20
1gvp	Gene V DNA Binding Protein	<i>Escherichia Coli</i>	1.60
1hjr	Ruvc Resolvase	<i>Escherichia Coli</i>	2.50
1hss	Alpha-Amylase Inhibitor	<i>Triticum Aestivum</i>	2.06
1icw	Interleukin-8 Mutant	<i>Homo Sapiens</i>	2.01
1imb	Inositol Monophosphatase	<i>Homo Sapiens</i>	2.20
1isa	Iron(II) Superoxide Dismutase	<i>Escherichia Coli</i>	1.80
1iso	Isocitrate Dehydrogenase	<i>Escherichia Coli</i>	1.90
1jhg	Trp Repressor Mutant V58I	<i>Escherichia Coli (synthetic)</i>	1.30
1jsg	Oncogene Product p14TCL1	<i>Homo Sapiens</i>	2.50
1kba	Kappa-Bungarotoxin	<i>Bungarus Multicinctus</i>	2.30
1kpf	Protein Kinase C Interacting Protein	<i>Homo Sapiens</i>	1.50
1lyn	Abalone Sperm Lysin	<i>Haliotis Rufescens</i>	2.75
1mj1	Methionine Repressor Protein Metj	<i>Escherichia Coli</i>	2.10
1mka	Beta -Hydroxydecanoyl Thiol Ester	<i>Escherichia Coli</i>	2.00
1moq	Glucosamine 6-Phosphate Synthase	<i>Escherichia Coli</i>	1.57
1nox	NADH Oxidase	<i>Thermus Aquaticus</i>	1.59
1nsy	NAD Synthetase	<i>Bacillus Subtilis</i>	2.00
1oac	Copper Amine Oxidase	<i>Escherichia Coli</i>	2.00
1opy	Ksi, 3-Ketosteroid Isomerase;	<i>Pseudomonas Putida</i>	1.90
1otp	Thymidine Phosphorylase	<i>Escherichia Coli</i>	2.80
1pgt	Glutathione S-Transferase	<i>Homo Sapiens</i>	1.80
1pre	Proaerolysin	<i>Aeromonas Hydrophila</i>	2.80
1puc	Yeast Cell-Cycle Control Protein, p13suc1	<i>Schizosaccharomyces Spombe (Synthetic Construct)</i>	1.95
1rfb	Recombinant Bovine Interferon Gamma	<i>Bovine (Bos Taurus)</i>	3.00
1rpo	Rop (Cole1 Repressor Of Primer) Mutant	<i>Escherichia Coli</i>	1.40
1ses	Seryl-tRNA Synthetase	<i>Thermus Thermophilus</i>	2.50
1slt	S-lectin	<i>Bovine (Bos Taurus)</i>	1.90
1smn	Serratia Endonuclease	<i>Serratia Marcescens</i>	2.10
1smt	Transcriptional Repressor Smtb	<i>Synechococcus</i>	2.20
1sox	Sulfite Oxidase	<i>Gallus Gallus</i>	1.90
1tox	Diphtheria Toxin	<i>Candida Albicans</i>	2.30
1trk	Transketolase	<i>Saccharomyces cerevisiae</i>	2.00
1tys	Thymidylate Synthase Mutant	<i>Escherichia Coli</i>	1.80
1uby	Farnesyl Pyrophosphate Synthetase	<i>Gallus Gallus</i>	2.40
1utg	Oxidized Uteroglobin	<i>Oryctolagus Cuniculus</i>	1.34
1wgj	Inorganic Pyrophosphatase	<i>Saccharomyces Cerevisiae</i>	2.00
1xso	Superoxide Dismutase	<i>Xenopus Laevis</i>	1.50
2ccy	Ferricytochrome C' (C Prime)	<i>Rhodospirillum Molischianum</i>	1.67
2ilk	Interleukin-10	<i>Homo Sapiens</i>	1.60
2rsp	Aspartic Protease	<i>Rous Sarcoma Retrovirus</i>	2.00
2tct	Tetracycline Repressor	<i>Escherichia Coli</i>	2.10
2tgi	Transforming Growth Factor- Beta Two (TGF-B2)	<i>Homo Sapiens</i>	1.80
3grs	Glutathione Reductase	<i>Homo Sapiens</i>	1.54
3pgh	Cyclooxygenase-2	<i>Mus Musculus</i>	3.00
3sdh	Hemoglobin I	<i>Scapharca Inaequalvis</i>	1.40

3ssi	Subtilisin Inhibitor	<i>Streptomyces Albogriseolus</i>	2.30
4kbp	Purple Acid Phosphatase	<i>Phaseolus Vulgaris</i>	2.70
5csm	Chorismate Mutase	<i>Saccharomyces cerevisiae</i> (Synthetic Construct)	2.00
5tmp	Thymidylate Kinase	<i>Escherichia Coli</i>	1.98
9wga	Agglutinin Isolectin	<i>Triticum Vulgaris</i>	1.80

Table 2.2: Dataset of 76 non-homologous homo-dimers.

PDB Code	Protein	Source	Resolution (Å)
1aa0	Fibritin mutant	<i>Coliphage T4</i>	2.20
1b77	Rb69, Sliding clamp protein	<i>Bacteriophage Rb6</i>	2.10
1bro	Bromoperoxidase a2	<i>Streptomyces Lividans</i>	2.05
1bvp	Bluetongue viral coat protein	<i>Bluetongue Virus</i>	2.60
1ca4	Tnf receptor associated factor 2	<i>Homo Sapiens</i>	2.20
1cb0	Mta phosphorylase	<i>Escherichia Coli</i>	1.70
1cbu	Cobu	<i>Salmonella Typhimurium</i>	2.30
1ce0	Gnc4 leucine zipper model	<i>HIV type 1 virus</i>	2.40
1cjd	Bacteriophage prd1 coat protein,	<i>Bacteriophage Prd1</i>	1.85
1dpt	D-dopachrome tautomerase	<i>Homo Sapiens</i>	1.54
1dun	Eiav dutpase	<i>Equine Infectious Anemia Virus</i>	1.90
1e2a	Enzyme iia from lactococcus lactis	<i>Lactococcus Lacti</i>	2.30
1fgj	Hydroxylamine oxidoreductase	<i>Nitrosomonas Europaea</i>	2.80
1nif	Nitrite reductase	<i>Achromobacter Cycloclastes</i>	1.60
1nks	Adenylate kinase	<i>Escherichia Coli</i>	2.57
1ppr	Peridinin-chlorophyll protein	<i>Amphidinium Carterae</i>	2.00
1qex	Bacteriophage t4 tail constriction	<i>Bacteriophage T4</i>	2.30
1qlm	Methenyltetrahydromethanopterin Cyclohydrolase	<i>Methanopyrus Kandleri</i>	2.00
1rla	Rat liver arginase	<i>Rattus Norvegicus</i>	2.10
2chs	Chorismate mutase	<i>Bacillus Subtilis</i>	1.90
2pii	Pii, glb product	<i>Escherichia Coli</i>	1.90
2std	Scytalone dehydratase	<i>Magnaporthe Grisea</i>	2.10
3cla	Chloramphenicol acetyltransferase	<i>Escherichia Coli</i>	1.75
3csu	Aspartate transcarbamoylase	<i>Escherichia Coli</i>	1.88
3tdt	Tetrahydrodipicolinate n Succinyltransferase	<i>Mycobacterium Bovis</i>	2.00
4bcl	Bacteriochlorophyll a protein	<i>Prosthecochloris Aestuarii</i>	1.90

Table 2.3: Dataset of 26 non-homologous homo-trimers.

PDB Code	Protein	Source	Resolution (Å)
1a0l	Human beta-tryptase	<i>Homo Sapiens</i>	3.00
1a2z	Pyrrolidone Carboxyl Peptidase	<i>Thermococcus Litorali</i>	1.73
1a4e	Catalase A	<i>Saccharomyces Cerevisiae</i>	2.40
1ado	Fructose 1,6-Bisphosphate Aldolase	<i>Oryctolagus Cuniculus</i>	1.90
1az9	Aminopeptidase	<i>Escherichia Coli</i>	2.00
1b25	Formaldehyde Ferredoxin Oxidoreductase	<i>Pyrococcus Furiosus</i>	1.85
1bfd	Benzoylformate Decarboxylase	<i>Pseudomonas Putida</i>	1.60
1bsm	Superoxide Dismutase	<i>Propionibacterium Shermanii</i>	1.35
1buc	Butyryl-CoA dehydrogenase	<i>Megasphaera Elsdenii</i>	2.50
1bvq	4-Hydroxybenzoyl Coa Thioesterase	<i>Pseudomonas Sp. Strain C</i>	2.00
1csl	Cystathionine -Gamma- Synthase (Cgs)	<i>Escherichia Coli</i>	1.50
1cuk	DNA recombination protein RuvA	<i>Escherichia Coli</i>	1.90
1dco	DcoH	<i>Rattus Norvegicus</i>	2.30
1eta	Transthyretin Variant	<i>Homo Sapiens</i>	1.70
1euh	Nadp Dependent Aldehyde Dehydrogenase	<i>Streptococcus Mutan</i>	1.82
1ftr	Tetrahydromethanopterin Formyltransferase	<i>Methanopyrus Kandleri</i>	1.70
1gp1	Glutathione Peroxidase	<i>Bos Taurus</i>	2.00
1gsh	Glutathione Synthetase	<i>Escherichia Coli</i>	2.00
1ith	Hemoglobin	<i>Urechis Caupo</i>	2.50
1mpy	Catechol 2,3-Dioxygenase (Metapyrocatechase)	<i>Pseudomonas Putida</i>	2.80
1mxb	S-Adenosylmethionine Synthetase	<i>Escherichia Coli</i>	2.80
1nhk	Nucleoside Diphosphate Kinase	<i>Myxococcus Xanthus</i>	1.90
1nhp	NADH peroxidase mutant	<i>Enterococcus Faecalis</i>	2.00
1sml	L1 Metallo-Beta-Lactamase	<i>Stenotrophomonas Maltophilia</i>	1.70
1toh	Tyrosine Hydroxylase	<i>Rattus Norvegicus</i>	2.30
1uox	Urate Oxidase	<i>Aspergillus Flavus</i>	2.00
1xva	Methyltransferase	<i>Escherichia Coli</i>	2.20
2fua	L-Fuculose 1-Phosphate Aldolase	<i>Escherichia Coli</i>	1.92
2izg	Streptavidin-Biotin	<i>Streptomyces Avidinii</i>	1.36
4pga	Glutaminase- Asparaginase	<i>Pseudomonas Sp. 7A</i>	1.70
5pgm	Phosphoglycerate Mutase	<i>Phosphoglycerate Mutase</i>	2.12

Table 2.4: Dataset of 31 non-homologous homo-tetramers.

PDB Code	Protein	Source	Resolution (Å)
1a3g	Branched-Chain Amino Acid Aminotransferase	<i>Escherichia Coli</i>	2.50
1bgv	Glutamate Dehydrogenase	<i>Clostridium Symbiosum</i>	1.90
1dci	Dienoyl-Coa Isomerase	<i>Rattus Norvegicus</i>	1.50
1dxe	2-Dehydro-3-Deoxy-Galactarate Aldolase	<i>Escherichia Coli</i>	1.80
1lcp	Bovine Lens Leucine Aminopeptidase	<i>Bos Taurus</i>	1.65
1ndc	Nucleoside diphosphate kinase	<i>Dictyostelium discoideum</i>	2.00
2cev	Arginase	<i>Bacillus Caldevelox</i>	2.15
2eip	Inorganic Pyrophosphatase	<i>Escherichia Coli</i>	2.20
3gcb	Mutant Yeast Bleomycin Hydrolase	<i>Saccharomyces Cerevisiae</i>	1.87

Table 2.5: Dataset of 9 non-homologous homo-hexamers.

2.4.2 Datasets of Hetero-Complexes

2.4.2.1 Datasets of Obligate Hetero-Proteins

PDB Code	Protein	Source	Resolution (Å)
1ajq	Penicillin Amidohydrolase	<i>Escherichia Coli</i>	2.05
1ft1	Protein Farnesyltransferase	<i>Rattus Norvegicus</i>	2.25
1h2a	Hydrogenase	<i>Desulfovibrio Vulgaris</i>	1.80
1hcn	Chorionic gonadotropin	<i>Homo Sapiens</i>	2.60
1hfe	Fe-Only Hydrogenase	<i>Desulfovibrio Desulfuricans</i>	1.60
1ixx	Coagulation factors ix/x-binding protein (ix/x-bp)	<i>Trimeresurus Flavoviridis</i>	2.50
1luc	Bacterial Luciferase	<i>Vibrio Harveyi</i>	1.50
1req	Methylmalonyl-Coa Mutase	<i>Propionibacterium Freudenreichii Subsp. Shermanii</i>	2.00
2frv	Oxidized Form Of Ni-Fe Hydrogenase	<i>Desulfovibrio Gigas</i>	2.54
4mon	Orthorhombic Monellin	<i>Dioscoreophyllum Cumminisii Diels</i>	2.30

Table 2.6: Dataset of 10 non-homologous obligate hetero-dimers.

PDB Code	Protein	Source	Resolution (Å)
1apy	Aspartylglucosaminidase	<i>Homo Sapiens</i>	2.00
1b7y	Phenylalanyl tRNA Synthetase Complexed With Phenylalaninyl- Adenylate	<i>Thermus Aquaticus</i>	2.50
1bou	4.5-Dioxygenase, Ligab	<i>Pseudomonas Paucimobilis</i>	2.20
1ccw	Glutamate Mutase	<i>Clostridium Cochlearium</i>	1.60
1qdl	Anthranilate Synthase	<i>Sulfolobus Solfataricus</i>	2.50
1qsh	Hemoglobin	<i>Homo Sapiens</i>	1.70
2scu	Succinyl-Coa Synthetase	<i>Escherichia Coli</i>	2.30

Table 2.7: Dataset of 7 non-homologous obligate hetero-tetramers.

PDB Code	Protein	Source	Resolution (Å)
1mro	Methyl-Coenzyme M Reductase	<i>Methanobacterium Thermoautotrophicum</i>	1.16
1tii	Escherichia Coli Heat Labile Enterotoxin Type IIb	<i>Escherichia Coli</i>	2.25
1eg9	Naphthalene 1,2-dioxygenase	<i>Pseudomonas Putida</i>	1.60

Table 2.8: Dataset of 3 non-homologous obligate hetero-hexamers.

The datasets of obligate hetero-proteins consist of a number of protein subunits that are permanently bound to one another. In general such complexes are biologically inactive when dissociated into their constituent subunits. Most of the proteins in these datasets are enzymes. A notable exception to this is hemoglobin (1qsh) being involved in oxygen storage and transport. Obligate hetero-complexes are relatively scarce there being only twenty in total in the current datasets compared with 142 for the homo-complexes.

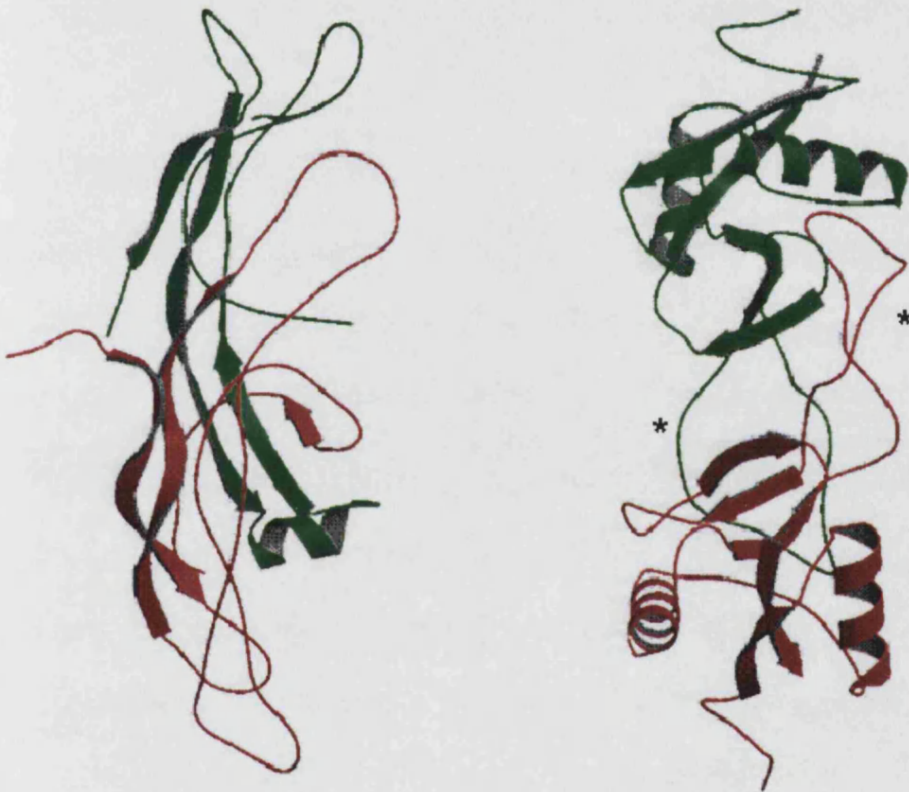


Figure 2.3: (a) The human chorionic gonadotropin dimer (1hcn, Wu et al., 1994). The coagulation factor IX/X-bp binding protein (1ixx, Mizuno et al., 1997). The domain swapped loops of 1ixx are marked *.

Human chorionic gonadotropin (hCG) is the hormone that effects the secretion of the pregnancy-sustaining hormone, progesterone. Due to the central role played by hCG in human pregnancy there is considerable interest in developing antagonists of the hormone to act as contraceptives and agonists to promote fertility. The hormone itself is a hetero-dimer of subunits rich in disulphide bonds and $K_d \approx 10^{-7}$ M in vivo (Forastieri, 1982). A diagram of the dimer is shown in figure 2.3(a). The interface between the subunits is extensive burying a total of 3860\AA^2 (Wu et al., 1994). The

protein adopts a very open structure with each subunit of the dimer having only a very small identifiable hydrophobic core. Blood coagulation factor Ix/X-bp is also a hetero-dimer of two homologous disulphide linked subunits. The protein itself is an anticoagulant found in the venom of the habu snake. The interface between the two subunits is highly unusual being formed by loops having no regular secondary structure joined together by a disulphide bond. The loops that make up the protein interface are thought to be the result of domain swapping (Mizuno, 1997). The locations of the domain swapped loops are indicated with an * in figure 2.3(b). The Ix/X-bp protein is currently the only known instance of domain-swapping occurring in the middle of the complex rather than at the N or C terminus (Liu & Eisenberg, 2002).



Figure 2.4: Heat labile enterotoxin LT-IIb (1tii, Van der Akker et al., 1996). The large α subunit situated on top of the pentameric ring of β subunits is proteolytically cleaved at the point marked with an arrow to produce the mature heptamer.

Heat-labile enterotoxin LT-IIb has been included in the dataset of hetero-hexamers even though technically it is a hetero-heptamer (Akker et al., 1996). Heat-labile enterotoxin LT-IIb is synthesized as a $\alpha\beta_5$ hexamer. The large α subunit that sits on top of the pentameric ring of β subunits is then proteolytically cleaved to produce the structure shown in figure 2.4. Since only eight residues remote from any protein-protein interface are cleaved from the α subunit it was deemed appropriate to treat the protein as a hetero-hexamer.

2.4.2.2 Datasets of Non-Obligate Hetero-Proteins

PDB Code	Protein Complex	Source	Resolution (Å)
1a4y	Ribonuclease Inhibitor-Angiogenin Complex	<i>Homo sapiens</i> Expressed in: <i>Escherichia Coli</i>	2.00
1acb	Alpha -Chymotrypsin-Eglin c complex	<i>Oxen (bos taurus)</i> and leech (<i>hirudo medicinalis</i>)	2.00
1avw	Porcine Pancreatic Trypsin/ Soybean Trypsin Inhibitor	<i>Sus scrofa</i> and Soybean	1.75
1ldt	Porcine Trypsin/Leech-Derived Trypsin Inhibitor	<i>Sus scrofa</i> and <i>Hirudo medicinalis</i> Expressed in: <i>Saccharomyces Cerevisiae</i> .	1.90
1slu	Rat Anionic Trypsin/A86H-Ecotin	<i>Escherichia Coli</i> and <i>Rattus Norvegicus</i>	1.80
1tab	Trypsin/Bowman-Birk inhibitor	<i>Bos Taurus</i> and <i>Phaseolus Angularis</i>	2.30
3bth	Trypsin/BPTI complex	<i>Bos taurus</i> . Expressed in: <i>Escherichia Coli</i>	1.75
2sic	Subtilisin BPN/subtilisin inhibitor	<i>Bacillus Amyloliquefaciens</i> and <i>Streptomyces Albogriseolus</i>	1.80
1brs	Barnase-Barstar Complex	<i>Bacillus Amyloliquefaciens</i> Expressed in: <i>Escherichia Coli</i>	2.00
1clv	Yellow Meal Worm Alpha Amylase/Amylase Inhibitor	<i>Tenebrio molitor</i>	2.00
1dp5	Aspartic Proteinase A/IA3 mutant inhibitor	<i>Saccharomyces Cerevisiae</i> Expressed in: <i>Escherichia Coli</i> .	2.20
1dtd	Human Carboxypeptidase A2 (LCpa2)/ Leech Carboxypeptidase inhibitor	<i>Homo sapiens</i> and <i>Hirudo Medicinalis</i>	1.65
1fle	Elafin/Elastase complex	<i>Sus scrofa</i> and <i>Homo Sapiens</i>	1.90
1hia	Kallikrein Complexed With Hirustasin	<i>Homo sapiens</i> and <i>Hirudo Medicinalis</i>	2.40
1smp	<i>Serratia Marcescens</i> Metallo-Protease/ <i>Erwinia Chrysanthemi</i> Inhibitor	<i>Serratia Marcescens</i> and <i>Erwinia Chrysanthemi</i> . Expressed in: <i>Escherichia Coli</i> .	2.30

41stf	Papain/Stefin	<i>Carica Papaya and Homo Sapiens Expressed in: Escherichia Coli</i>	2.37
1ugh	Human Uracil-DNA Glycosylase/Protein Mimic Inhibitor	<i>Homo sapiens and Bacteriophage pbs2 Expressed in: Escherichia Coli</i>	1.90
1viw	Alpha-Amylase/Bean Phaseolus Vulgaris Inhibitor	<i>Tenebrio Molitor and Phaseolus Vulgaris</i>	3.00
4htc	Hirudin-Thrombin Complex	<i>Homo sapiens and Hirudo medicinalis Variant 2</i>	2.30
4sgb	Streptomyces griseus proteinase B/Potato Inhibitor PC1-1	<i>Streptomyces Griseus and Solanum Tuberosum</i>	3.20

Table 2.9: Dataset of 20 non-homologous enzyme-inhibitor complexes.

Most of the enzyme-inhibitor complexes belong to the serine protease super-family, which includes the digestive enzymes, chymotrypsin, trypsin, thrombin and elastase amongst others. There are a large number of naturally occurring inhibitors of serine proteases of varying sizes and shapes that can be broadly grouped into about 18 families on the basis of sequence identity (Laskowski, 1980). Serine proteases share a common catalytic mechanism with similarly constituted active sites. Consequently, the same inhibitor can bind to and inhibit several different serine proteases. As an example BPTI (bovine pancreatic trypsin inhibitor) inhibits both human and bovine trypsin, chymotrypsin, and various kallikreins.

Diagrams of two of the trypsin-inhibitor complexes in the dataset are shown in figures 2.5(a)(i) and (ii). Both soybean trypsin inhibitor (STI) and BPTI bind to and inhibit trypsin through a convex binding loop which is highly complementary in shape to the active site of the enzyme (Apostoluk, 1998). It is interesting to note that the actual conformation of the protease binding loop is similar to a number of inhibitors from different families (a good example of convergent evolution). In both complexes the actual area of contact between BPTI and STI and trypsin is quite small. In spite of this the interaction between trypsin and its cognate inhibitors is very strong with disassociation constants ranging from 10^{-9} to 10^{-14} M (Sweet, 1974). There are only a small number of enzyme complexes in the dataset that do not contain a serine protease component. These include human uracil DNA glycosylase in complex with a mimic protein inhibitor (1ugh) and an angiogenin-inhibitor complex (1a4y).

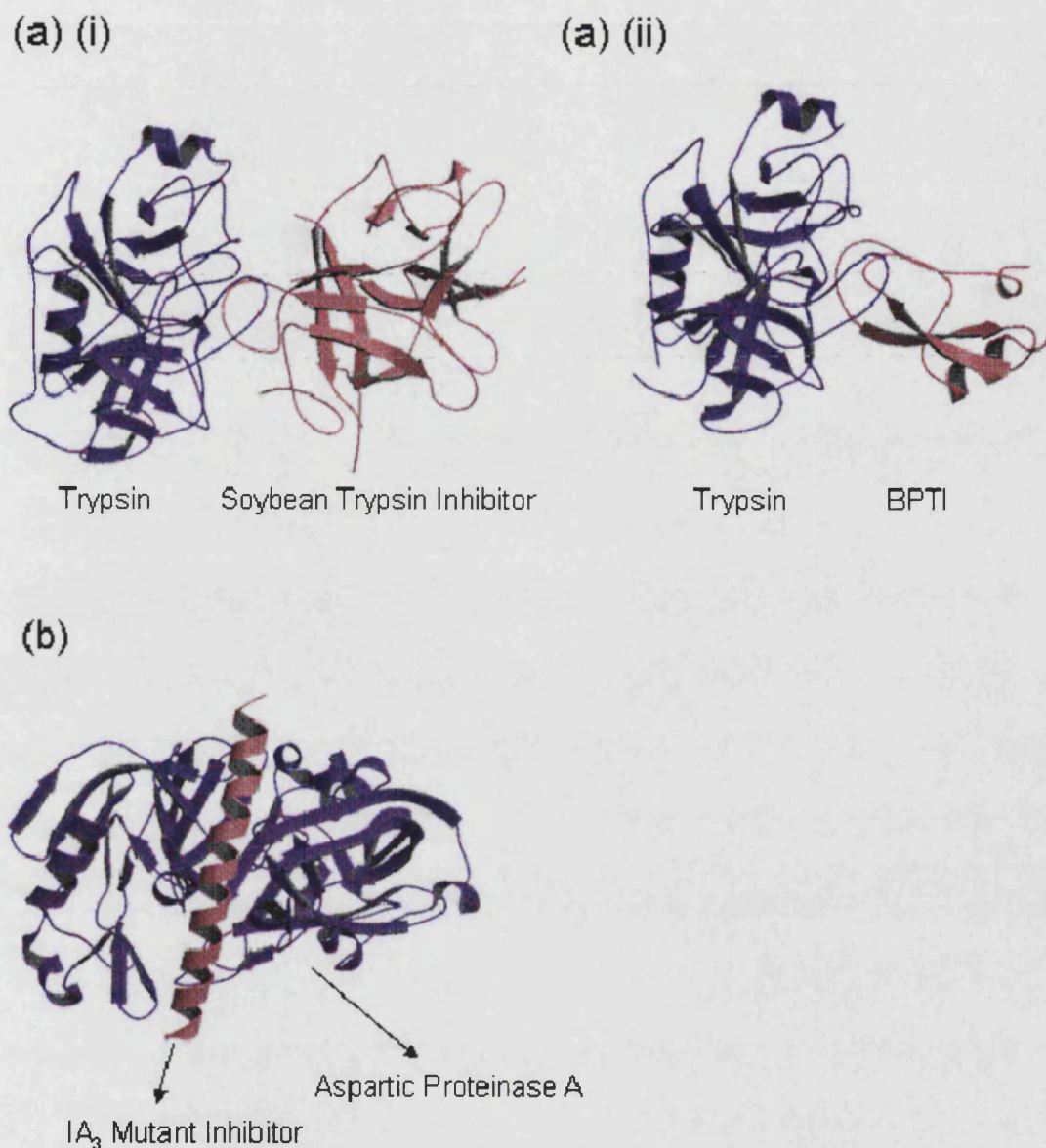


Figure 2.5: (a) (i) Trypsin in complex with the soybean trypsin inhibitor (1avw, Song & Suh, 1998) and BPTI (a) (ii) (3bth, Helland et al., 1999). Trypsin is a multi-specific enzyme capable of being selectively inhibited by a number of different inhibitors. (b) Aspartic proteinase A in complex with the IA₃ mutant inhibitor (1dp5, Li et al., 2000)

An intriguing entry in the dataset is aspartate protease A from *Saccharomyces cerevisiae* coupled with a highly specific and potent 8kDa inhibitor called IA₃ (1dp5). A diagram of the protease-inhibitor complex is given in figure 2.5(b). The inhibitor is highly unusual in that it has no detectable secondary structure in solution (Phylip et al., 2001). Upon binding to the protease IA₃ adopts a near perfect alpha helical

conformation. The mechanism by which IA₃ adopts the alpha helical conformation is not fully understood.

PDB Code	Protein Complex	Source	Resolution (Å)
2trc	Gβγ / Phosducin Complex	<i>Bos Taurus</i>	2.40
1buh	Cdk2 / Ckshs1 Complex	<i>Homo sapiens</i>	2.60
1g3n	Cdk6/ Cyclin K /P18(Ink4C)	<i>Homo sapiens</i>	2.90
1jsu	Cdk2/ Cyclin A / p27(Kip)	<i>Homo sapiens</i>	2.30
1tx4	Rho A/ P50-Rhogap	<i>Homo sapiens</i>	1.65
1ds6	Ras Toxin Substrate 2 / Rho GDP-Dissociation Inhibitor 2	<i>Homo sapiens</i>	2.35
1e0o	Fibroblast Growth Factor 2/Receptor Complex	<i>Homo sapiens</i>	2.80
1www	Nerve Growth Factor / Domain 5 of the Trka Receptor	<i>Homo sapiens</i>	2.20
1bp3	Human Growth Hormone / Prolactin Receptor (Extracellular Domain)	<i>Homo sapiens expressed in: Escherichia coli</i>	2.90
1flt	Vascular Endothelial Growth Factor (Receptor Binding Domain) / Fms-Like Tyrosine Kinase 1 (Extracellular Igg Like Domain)	<i>Homo sapiens</i>	1.70

Table 2.10: Dataset of 10 non-homologous signaling protein-complexes.

The dataset of signaling proteins can be broadly considered to fall into two classes of proteins. The first class consists of the large GTP-binding proteins (or G proteins). There are two major categories of G proteins, the first being the hetero-trimeric G proteins and the second consisting of the monomeric proteins of the Ras family. The majority of G proteins are hetero-trimers. G proteins oscillate between their active GTP bound form and inactive GDP bound state. The binding of GTP to G proteins induces conformational changes in the molecule, resulting in its interaction with other proteins (frequently receptors), and the subsequent transmission of signals from one place to another (Blundell, 2000). In the dataset there is one hetero-trimeric protein 2trc which is active in the rod-cell visual transduction system. A diagram of the βγ-phosducin complex in two different orientations is shown in figure 2.6. Phosducin acts as an inhibitor by binding to the βγ subunits in place of the GTP binding α subunit. The phosducin bound βγ protein is also rendered incapable of interacting with the cell-membrane receptor further inhibiting the signal transduction process.

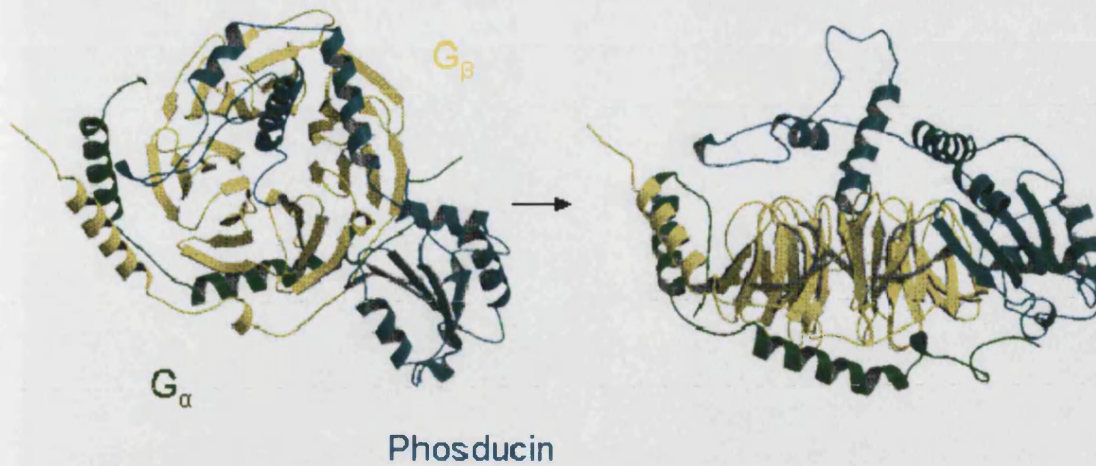


Figure 2.6: The transducin $\beta\gamma$ -phosducin complex shown in two different orientations (2trc, Gaudet et al., 1996).

Phosducin is composed of two non-interacting domains that bury a total of 2300 \AA^2 upon formation of the hetero-trimer and induce conformational changes in the blades of the propeller β subunit. There are two proteins in the dataset that belong to the Ras superfamily. These are the Rho A / P50-Rhogap complex (1tx4) shown in figure 2.7(a) and the Ras toxin Substrate 2 / Rho GDP-Dissociation Inhibitor 2 structure (1ds6).

Cyclin dependent kinases (cdk's) are a class of protein kinases that regulate the progression of a cell through the different phases of the cell cycle. Cdk's themselves have no catalytic activity and only become active when bound to proteins known as cyclins. Binding to an inhibitor can subsequently render the cdk/cyclin complexes inactive. There are two entries in the dataset that represents full cdk-cyclin-inhibitor complexes (PDB codes 1g3n, and 1jsu).

A diagram of the cdk2/Cyclin A/p27(Kip) complex is shown in figure 2.7(b). Only the binding domain of the p27 inhibitor is represented in the PDB structure. p27 actually functions as an inhibitor by covering the catalytic cleft of the cdk subunit and inducing conformational changes in the cdk-cyclin structure. The protein-protein interfaces within this protein are very large. The binding domain of the p27 inhibitor adopts an extended structure with no hydrophobic core. This results in the inhibitor burying a total of 5750 \AA^2 upon binding to the cyclin A-cdk2 protein (Russo et al.,

1996). The interface between cdk2 and cyclin A is also large being 3500\AA^2 in size (Jeffrey, 1995).

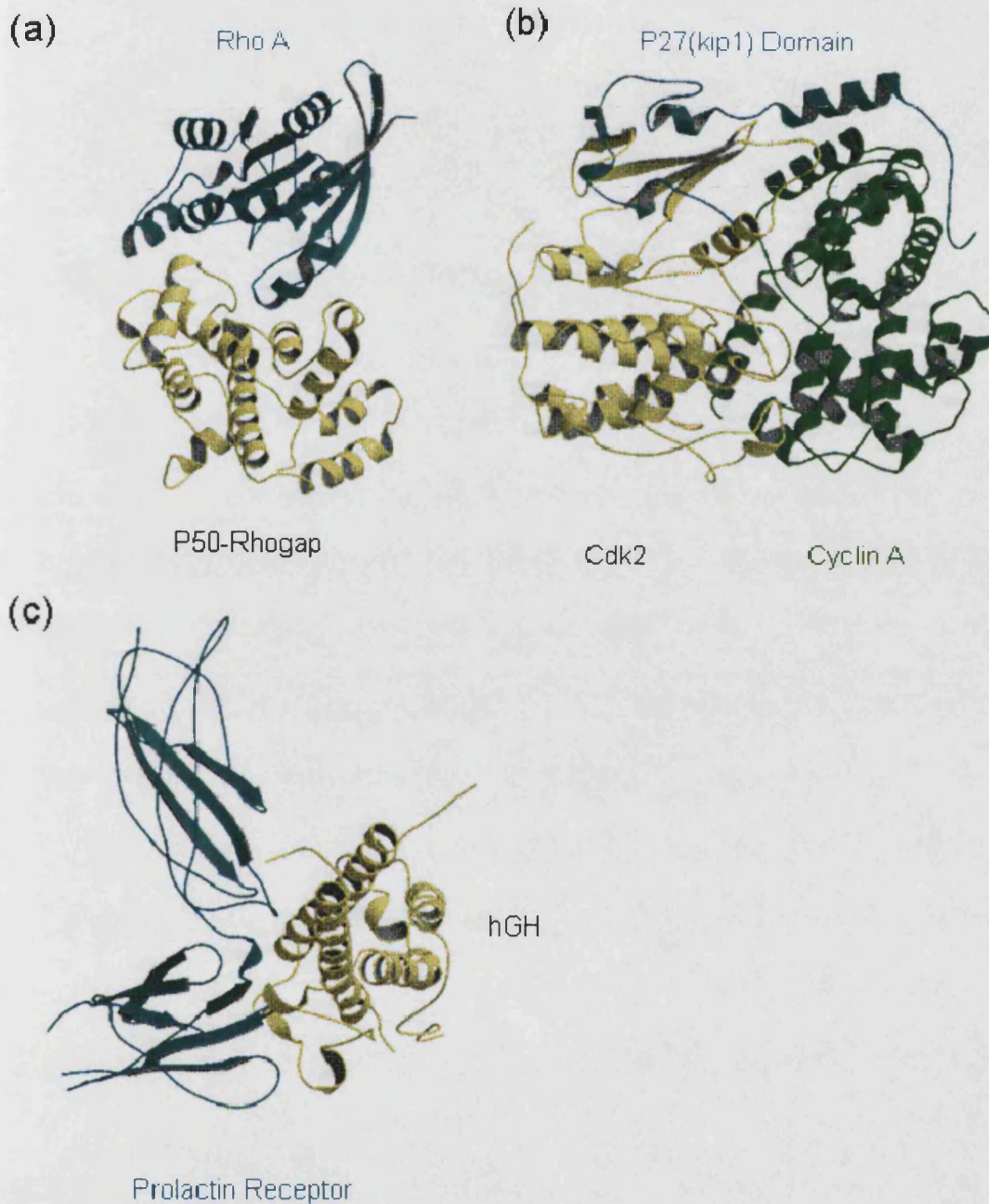


Figure 2.7: (a) The rho A/p50 rhogap complex (1tx4, Rittinger et al., 1997). (b) The complex between cdk2, cyclin A, and p27(kip 1) (1jsu. Russo et al., 1996). (c) The human growth hormone (hGH) in complex with the prolactin receptor (1bp3, Somers et al., 1994).

Four proteins 1ev2, 1www, 1flt and 1bp3 in the dataset are growth factor-receptor complexes with only the extra-cellular ligand-binding regions of the receptor being

represented in the crystal structures. Growth factors themselves are a class of signaling proteins, which stimulate cell growth and proliferation. The human growth hormone-prolactin-receptor complex (1bp3) is shown in figure 2.7(c). The prolactin receptor regulates milk production in mammals and is activated when bound to the growth hormone. “The receptor is composed of two domains with the hormone binding site being formed by loops somewhat like the antigen binding site of an antibody (Somers et al., 1994). The hormone-binding loop of prolactin forms a strong binding site for a zinc atom that links both the hormone and receptor. The presence of zinc at the binding site increases the affinity of the hormone for the receptor in vitro by a factor of 10,000” (Branden & Tooze, 1998).

PDB Code	Antibody	Antigen	Resolution (Å)
1jrh	A6 Fab	<i>Interferon γ Receptor (N terminal Domain)</i>	2.80
1fdl	Anti-Lysozyme Antibody D1.3 Fab	<i>Hen Egg-White Lysozyme</i>	2.50
1nfd	Anti-TCR Fab	<i>N15 Alpha-Beta T-Cell Receptor (a heterodimer)</i>	2.80
1mlc	D44.1 Fab	<i>Hen Egg-White Lysozyme</i>	2.10
1kb5	Desire 1 Fab	<i>KB5-C20 T-Cell Antigen Receptor Fv</i>	2.50
1dqj	HYHEL-63 Fab	<i>Hen Egg-White Lysozyme</i>	2.00
1qfu	IgG 1 Kappa Antibody Fab	<i>Hemagglutinin</i>	2.80
1nsn	IgG 1 Kappa Antibody Fab	<i>Staphylococcal nuclease ribonucleate</i>	2.90
1ahw	Immunoglobulin 5g9 Fab	<i>Thromoplastin Coagulation Factor III (Extracellular Domain)</i>	3.00
2jel	Jel42 Fab	<i>Histidine Containing Protein (HPR)</i>	2.50
1e6j	Monoclonal Antibody 13B5 Fab	<i>HIV Viral Capsid Protein (P24)</i>	3.00
1egj	Monoclonal Antibody BION-1 Fab	<i>Cytokine Receptor Beta Cahin Precursor (Domain 4)</i>	2.80
1nca	N9 Neuraminidase-NC41 Fab	<i>Influenza Virus A neuraminidase</i>	2.50
1g9m	Neutralizing Antibody 17B Fab	<i>HIV Envelope Protein Gp120</i>	2.20
1fns	NMC-4 (IGG1) Fab	<i>Von Willebrand Factor</i>	2.00

Table 2.11: Dataset of 15 antibody-antigen complexes.

In total there are 15 entries in the dataset of antibody complexes. All of the entries in the dataset are fragments consisting of the antigen binding Fab antibody fragments

bound to their respective antigens. Three proteins in the dataset are antibody-lysozyme complexes. Lysozyme is a relatively small and stable enzyme of approximately 130 amino acids. The stability of the enzyme makes it an ideal subject for mutagenesis experiments to probe the relationship between protein structure and function. The enzyme has been so extensively studied that there are currently more lysozyme structures in the PDB than any other single protein structure. There were 836 lysozyme structures in the PDB July 2003. The total PDB contains ~21,800 structures.

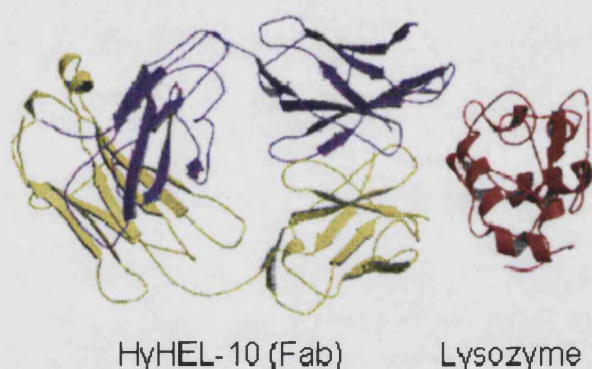


Figure 2.8: (a) The HyHEL-10 antibody bound to lysozyme (3hfm, Padlan et al., 1989). The interface between the antibody and lysozyme is so closely packed that there are no cavities large enough to accommodate a water molecule.

The HyHEL-10 Fab lysozyme complex (3hfm) is shown in figure 2.8. Lysozyme is shown in red with the heavy and light chains of the fab fragment in yellow and purple. The actual surface area of lysozyme that is in contact with the antibody is quite large being 774\AA^2 in size. The interaction between the antibody and antigen is very strong, the association coefficient being $1.5 \times 10^9 \text{ M}^{-1}$ (Padlan et al., 1989). The areas of lysozyme and HyHEL-10 that are in contact with each other are so complementary in shape that that there are no cavities at the antibody-antigen interface large enough to accommodate a water molecule. A large number of the residues of the antibody that make contact with lysozyme are aromatic further enhancing the hydrophobic nature of the antibody-antigen interface. Originally eight lysozyme-antibody structures were included in the dataset (1fdl, 1jhl, 1mlc, 1fbi, 1bql, 1dqj, and 3hfm). To avoid the overall interface characteristics of the antigen interfaces being biased by the large

number of lysozyme structures only entries of lysozyme bound to antibodies in substantially different orientations were selected. The three lysozyme-antibody structures that were finally selected for inclusion in the dataset are 1mlc, 1dqj, and 1fdl. Figure 2.9 shows the lysozyme component of 1mlc, 1dqj, and 1fdl bound to their respective antibodies in different positions.

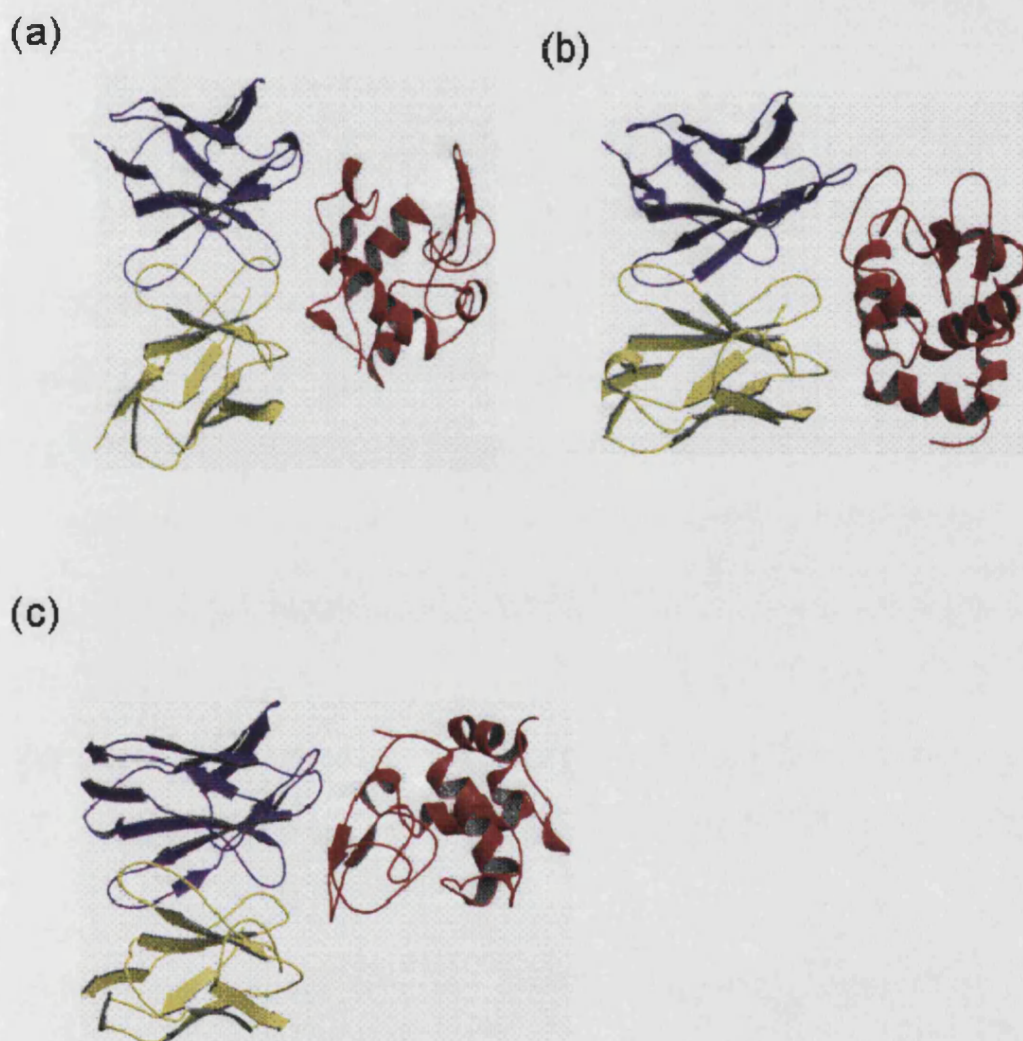


Figure 2.9: Three antibody-hen egg-white lysozyme complexes. The light and heavy chains of the antibody are coloured yellow and purple respectively. The antigen is coloured in red. In each of the three complexes lysozyme is bound to its antibody in a different orientation. (a) 1fdl, Fischmann et al., 1991. (b) 1dqj, Li et al., 2000. (c) 1mlc, Braden et al., 1994.

Antibody-viral complexes can be considered to comprise a second class of antibody interaction. Four proteins (1qfu, 1e6j, 1nca, and 1g9m) in the dataset are structures of antibodies in complex with viral proteins. Two of these proteins are antibody-HIV complexes (1e6j, 1g9m). 1e6j is a complex between a 13B5-Fab antibody and the p24 HIV-1 capsid protein. The detection of antibodies produced in response to p24 is commonly used as a diagnostic for HIV infection (Janvier et al., 1993). The association rate for the interaction between the 13B5-Fab antibody and p24 is quite low being $3.5 \times 10^5 \text{ M}^{-1}\text{s}^{-1}$ (Monaco-Malbet et al., 2000).

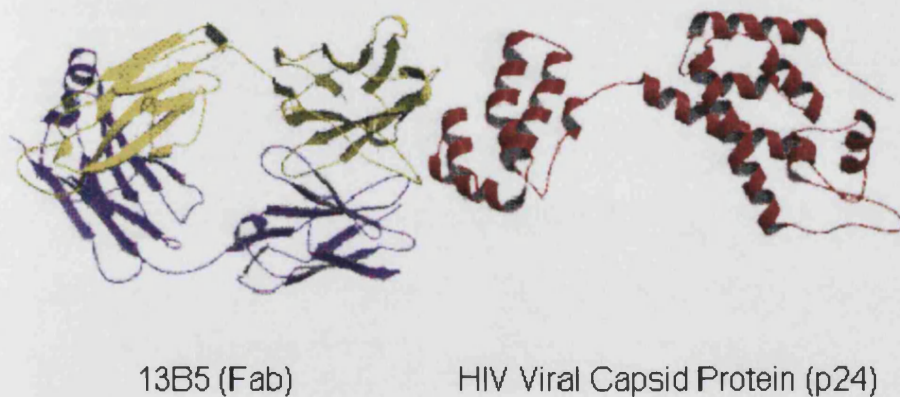


Figure 2.10: The complex between the HIV viral capsid protein (p24) and its antibody (1e6j, Monaco-Malbet et al., 2000).

The low affinity of the interaction is reflected by the small contact area of 609\AA^2 between the antibody and the p24 protein (Monaco-Malbet et al., 2000). The small contact area given the large size of the antigen can be explained by the fact that the p24 protein adopts a quite extended structure. This can be seen in figure 2.10. The 13B5 antibody binds asymmetrically to p24 in the sense that the heavy chain of the antibody makes up 82% of the contact area with p24. In common with the HyHEL-10 lysozyme complex many of the residues that make contact with p24 are aromatic most of them being tyrosines. The remaining structures in the dataset are antibody-receptor complexes (1nfd, 1jrh, 1kb5).

2.5 Distribution of Multimeric States in Protein Databases

Ascertaining the true distribution of multimer types found in nature has always been a great challenge. In 1975 Darnell & Klotz surveyed all available structures of homo-complexes and concluded that dimers and tetramers were the two most common multimeric states. Trimers, pentamers, and other multimers containing an odd number of subunits are observed but much less frequently than complexes containing an even number of subunits. Jones & Thornton confirmed the findings of Darnell & Klotz by examining the contents of July 1993 release of the PDB. The results of this investigation are shown in fig 2.11.

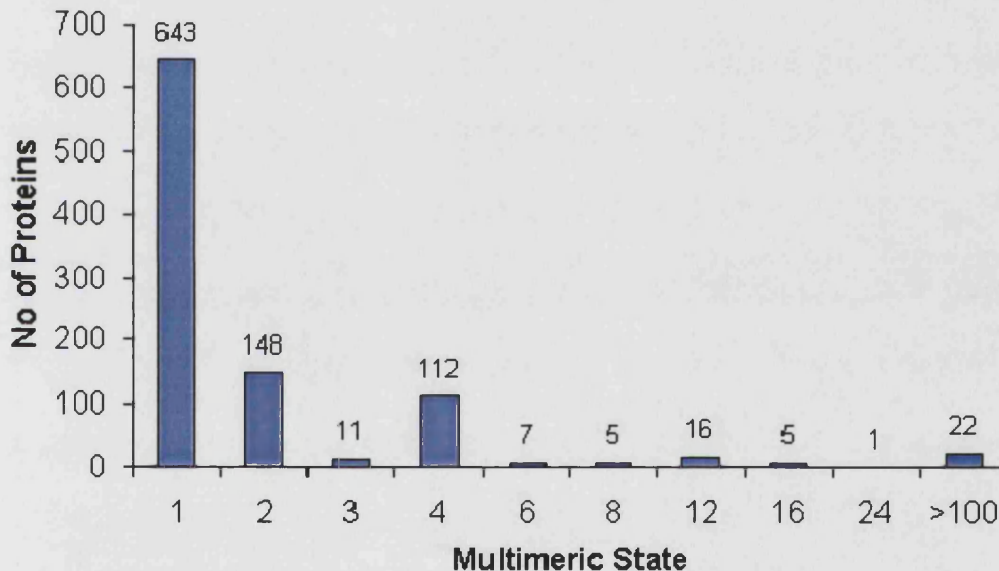


Figure 2.11: Multimeric states of proteins in the July 1993 release of the PDB. 1 = monomer, 2 = dimer, etc. Adapted from Jones & Thornton, 1996.

Dimers and tetramers do indeed appear to be the most common multimeric states adopted by proteins. Trimers seem to be the only class of multimer containing an odd number of subunits that is to be found in appreciable numbers. Hetero-complexes are less frequently observed than homo-complexes. It does need to be pointed out that different proteins will be present in different quantities in the cell. Thus just because it seems that there are more examples of different proteins that are trimeric rather than hexameric does not mean that there are numerically more trimers than hexamers in the cell.

	E-coli ^(a)		PQS ^(b)		^(c)	
	Homo	Hetero	Homo	Hetero	Homo	Hetero
Monomers	72	-	6619	-	95	-
Dimers	115	27	6160	1745	76	10
Trimers	15	5	726	768	26	0
Tetramers	62	16	1685	1004	31	7
Pentamers	1	1	159	80	0	0
Hexamers	20	1	491	335	9	3

Table 2.12: Numbers of monomers and multimers up to hexamers in (a) *Escherichia coli* (Goodsell & Olson, 2000). (b) PQS holdings of obligate and non-obligate multimers in July 2003. (c) datasets of obligate multimers detailed in this chapter.

The contents of the PQS in July 2003 are shown in table 2.12 and broadly point towards the distribution of multimers in the PDB not having changed very much over the last ten years. The most conspicuous difference is that homo-dimers are almost as common as monomers. This may verify the suspicion expressed by Jones & Thornton that the PDB at the time contained a disproportionate number of small monomeric proteins due in part to the relative ease with which such proteins can be crystallised. An analysis of all the Swiss-Prot entries for proteins from *Escherichia coli* actually indicates that monomers are slightly less common than homo-dimers (Goodsell & Olson, 2000). The relative sizes of the obligate homo-dimers, trimers, tetramers, and hexamers correspond roughly to what one would expect from a comprehensive analysis of Swiss-Prot. Querying the ‘subunit’ field in the ‘comment’ subentry of Swiss-Prot entries for the respective multimer class yielded fractions of 5:8:1:4:1 for monomers up to hexamers. The relative sizes of the datasets of the homo-complexes used in this thesis are described by the ratios 5:8:2:4:1 indicating a substantial difference only for trimers which are over-represented by a factor of two (Postingl, Kabir, and Thornton, 2003).

Chapter 3

Obligate Homo-Complexes

3.1 Introduction

The interfaces between proteins are complex environments. Because of the enormous functional and physical diversity of proteins there is no universal set of characteristics that definitively distinguishes a binding site from any other part of the surface of a protein (Ringe, 1995). From a thermodynamic point of view an assembly of proteins is only marginally more stable than its constituent monomers. The free energy that makes the difference is principally provided by the hydrophobic effect (the burial of hydrophobic regions on the surface of the protein during the binding process). Point-point interactions such as hydrogen bonds and salt bridges provide the remainder of the free energy required to form a protein complex. The hydrophobic effect, the predominant driving force behind protein folding and oligomerisation, is non-specific; *in vivo* hydrophobic surfaces tend to aggregate indiscriminately so as to minimise contact with solvent and disruption to the extensive network of hydrogen bonds that surround them. It is the point-point interactions, requiring reasonably good complementarity between the interacting surfaces, which allow specific binding interactions to occur. This was noted by Chothia & Janin (1975) who concluded that: “Hydrophobicity is the major factor stabilising protein-protein association, while complementarity plays a selective role in deciding which proteins may associate”. As such patches of hydrophobic residues enclosed by a few polar or charged residues may provide the strongest single indicator of a protein-protein interaction site. It has been observed (Larsen et al., 1998) that protein-protein interfaces may resemble a cross section through a protein, consisting of a hydrophobic core surrounded by a few polar residues that take part in the formation of point-point interactions such as

hydrogen bonds and salt bridges and help to exclude water from the hydrophobic core.

An example of a protein where the interface region can be clearly seen to have a well-defined hydrophobic core surrounded by a few polar residues is the trimeric bacteriophage rb69 sliding clamp protein (Shamoo, 1999). A representation of the protein surface coloured up by electrostatic potential rendered by GRASP (Nicholls, Sharp, and Honig, 1991) is shown in figure 3.1.

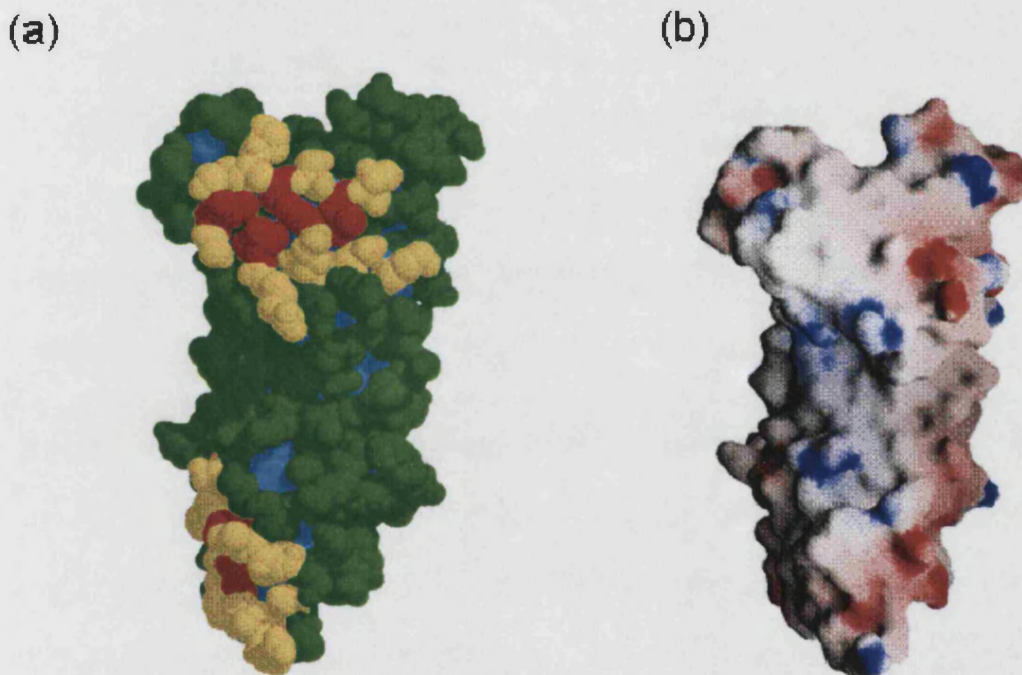


Figure 3.1: The bacteriophage rb69 sliding clamp monomer (1b77, Shamoo, 1999). (a) Residues that make up protein-protein interfaces are coloured yellow and red. (b) A representation of the sliding clamp monomer generated using GRASP (Nicholls et al,1991). Regions of negative and positive electrostatic potential are coloured in red and blue respectively. The protein-protein interfaces of 1b77 appear to be composed of a central hydrophobic core surrounded by polar and charged residues.

The issue is somewhat confused however by the fact that many proteins also have a number of small hydrophobic patches surrounded by polar residues scattered over the entire extent of the protein-protein interface, as shown in figure 3.16(b) (Chakrabarti & Janin, 2002). In addition it should be noted that protein structures are still evolving. Consequently at any one time it is only possible to directly observe a

'snapshot' through the process of evolution. Some protein families are clearly further along in the evolutionary process than others and their binding sites may be expected to be more finely tuned to assist in the binding process; this may further complicate the problem of determining what characterises a protein-protein interface region. Nevertheless the lesson to be drawn from looking at the protein-protein interface shown in figure 3.1 is that protein-protein interaction sites do have some characteristic attributes and that it is worthwhile determining what these characteristics are for reasons that have already been discussed in chapter 1.

In this chapter non-homologous datasets of obligate protein-complexes are assembled according to their multimeric state. The protein-protein interfaces within these complexes are then characterised in terms of their physical and chemical characteristics for the most part using the same descriptors used by Jones, 1995. The protein-protein interfaces of homo-dimeric proteins (Jones & Thornton, 1996) and tetrameric proteins (Miller, 1989) have been previously studied in this way, but to our knowledge no systematic investigation has been made of the protein-protein interfaces of trimeric and hexameric proteins since the study carried out by (Janin, Chothia, and Miller, 1988) at which time relatively few crystal structures were known. The characterisation of the protein-protein interfaces of proteins of differing multimeric states may provide some indication as to how and why individual proteins aggregate into the wide variety of complexes and multimeric states that we observe. The rapid growth in the number of crystal structures means that the characteristics of interface regions can be elucidated in a more comprehensive way than has been previously possible.

3.2 *Classification of Residues*

The residues of each protein are split into three classes: interior, surface, and interface. The classification of residues as belonging to the interior or exterior of a protein was based on the relative accessible surface area (rASA) of each residue of the protomer within the multimer. rASA is defined as the ratio between the ASA that an amino acid has in its current position, and the ASA it would have when adopting its

most solvent exposed sidechain conformation within a fully extended polypeptide. A rASA of 0% means that the residue is completely inaccessible to solvent while a rASA of 100% means that the residue is fully exposed to solvent.

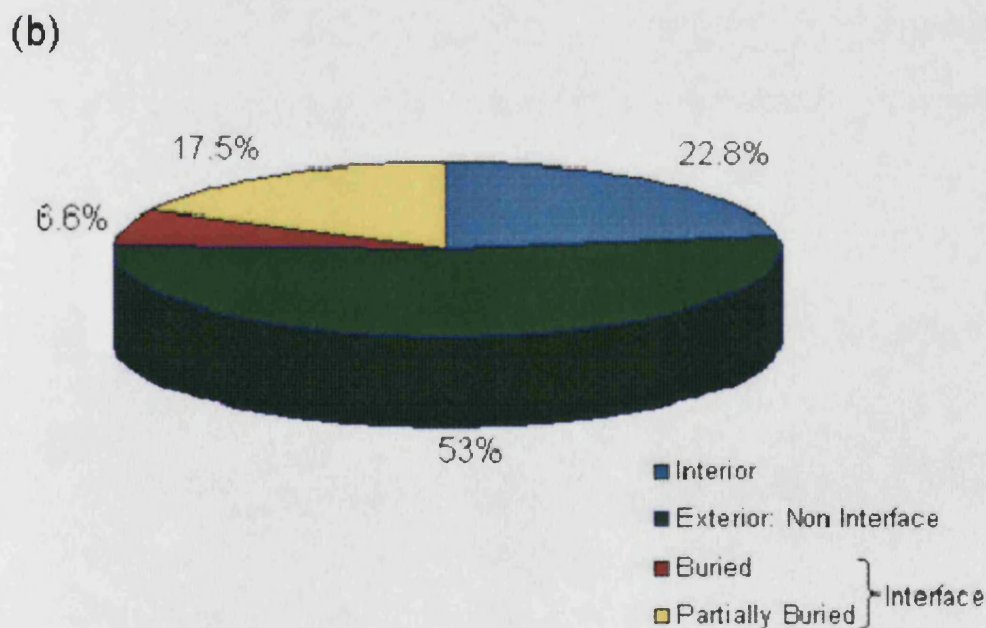
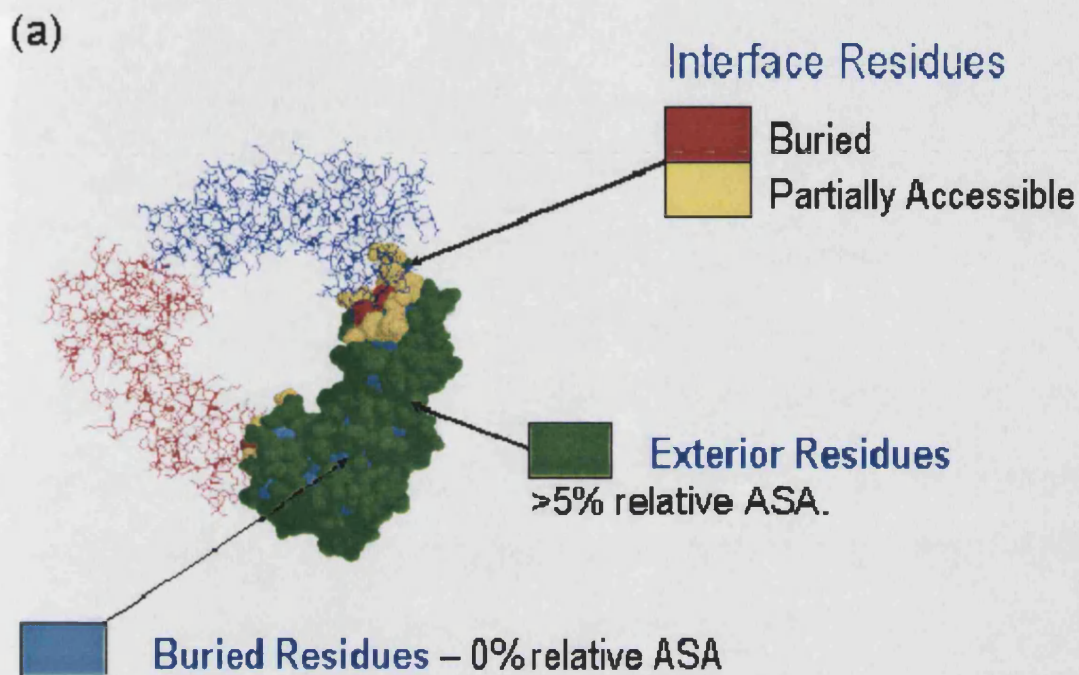


Figure 3.2: (a) The scheme used to classify residues as being part of the protein interior, exterior, and interface regions. 'Interface' residues are sub-divided as being part of a 'buried' or 'partially accessible zone'. (b) A pie chart showing the average composition of the 142 homo-complexes.

A residue is classified as being part of the interior class if it has a rASA < 5%. Conversely a residue is classified as being part of the protein exterior if it has a rASA > 5%. The 5% cut-off was originally devised and implemented by Miller et al., 1987.

The classification of a residue as belonging to a protein-protein interface is based on its accessible surface area (ASA) in isolation from a complex compared with its ASA in complex. A residue is classified as being 'interface' if its ASA when the protein is part of a complex is >1Å² lower than the ASA of the residue when the protein is in isolation from the rest of the complex. Interface residues were further classified as being part of a 'buried' or a 'partially accessible' zone. 'Buried' residues are those residues in the protein-protein interface that are completely inaccessible to solvent. 'Partially accessible' residues are those residues in the interface that have a non-zero ASA.

A diagram showing the bacteriophage rb69 sliding clamp protein (Shamoo, et al., 1999) with residues coloured up according to the classification scheme that has been outlined in this section is shown in figure 3.2(a). The average percentages of a protein that are interior, exterior and interface is shown in figure 3.2(b)

3.3 Size (ASA) of Protein-Protein Interfaces

The accessible surface area (ASA) of each protein was calculated using NACCESS (Hubbard, 1990), an implementation of the algorithm developed by Lee & Richards (1971), with the probe radius set to 1.4Å; hetero groups and water molecules were ignored when calculating ASA for the purposes of this study. The total buried ASA of each protomer within a multimer is here defined by

$$\text{Buried ASA} = \frac{n \sum \text{ASA}_{\text{monomer}} - \sum \text{ASA}_{\text{multimer}}}{n} \quad (1)$$

where n is the number of subunits in the complex.

The amount of ASA buried for the protomers of the dimers, trimers, tetramers, and hexamers can be seen in table 3.1. Trimeric, tetrameric and hexameric proteins tend to bury a greater fraction of their total ASA on formation of a multimer than dimers.

	Dimers	Trimers	Tetramers	Hexamers
Monomer Weight (Da)				
Mean	29000	26200	33000	35100
Min	6950	4500	12000	16800
Max	96900	61600	68800	53800
SD	19153	12700	14525	13797
Buried ASA (\AA^2)				
Mean	1890	2520	3090	3650
Min	540	880	950	1980
Max	7150	5390	10040	5740
SD	1170	1266	1789	1358
(%) ASA Buried				
Mean	15.9	22.5	22.5	25.9
Min	4.2	6.2	9.9	17.1
Max	31.3	40.3	40.1	37.1
SD	6.92	9.54	7.01	6..21

Table 3.1: A table showing basic statistics for the total amount of ASA buried in protein-protein interfaces per protomer within the datasets of homo-dimers, trimers, tetramers, and hexamers. Statistics are also given for the molecular weight of the protomers from the different datasets of homo-complexes.

For each protein within each dataset the total interface ASA of one protomer within the multimer has been plotted against its molecular weight and is shown in figure 3.3. The increasing gradients of the lines of best fit for each dataset in figure 3.3 illustrate quite well that the higher the multimer the greater the fraction of surface area of the protomer that is buried in protein-protein interfaces. The buried ASA figures for the different types of multimer do vary from those found in the past. From table 3.1 dimers and trimers on average bury 1890 and 2520 \AA^2 of ASA respectively compared with the figure of 1685 \AA^2 for a dataset of 32 homo-dimers compiled by Jones & Thornton, 1996.

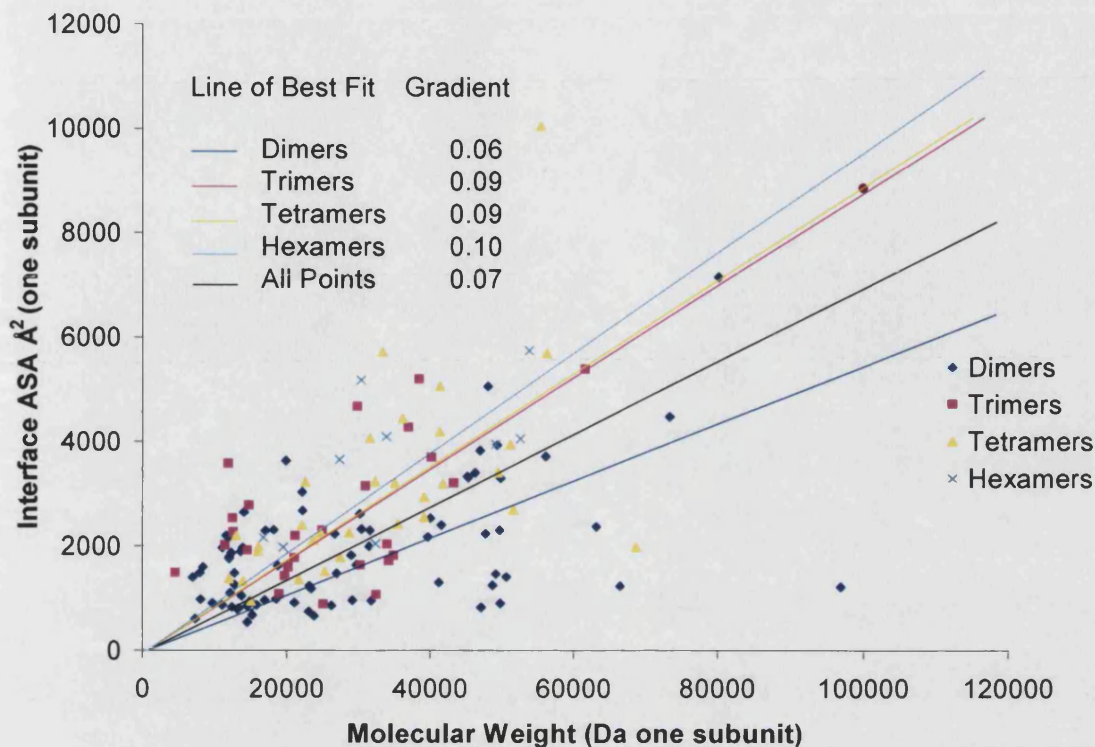


Figure 3.3: A plot of the molecular weight for each protein subunit together with its total interface ASA. The correlation coefficient (r) for all points in this plot is 0.49.

Millet et al., 1987, using a dataset of nine homo and hetero tetramers found that the subunits within tetramers each on average bury 4185\AA^2 of ASA in protein-protein interfaces compared with the value of 3090\AA^2 found here. The differences found here demonstrate the value of recalculating such quantities as buried ASA based upon larger datasets of proteins. The amount of ASA buried in each category of multimer is of course dependent on the ways that a number of individual subunits can pack together to form a 'closed' complex in which the areas of contact between subunits is maximised.

It is tempting to think that large contact areas between individual subunits equate with the stability of the complex, in theory the larger the total area of contact between subunits, the greater the number of point-point interactions like salt bridges and hydrogen bonds between subunits and the greater the stability of the complex. However in sections 3.8 and 3.9 it is shown that although the number of hydrogen bonds does scale with the size of the interface, the number of salt bridges remains essentially constant. It is probable therefore that the dominant reason for higher

multimers burying proportionally more subunit surface within interfaces, as seen in table 3.1, relates to their need to maintain an optimal surface-to-volume balance in spite of very different protomer shapes and multimer packing arrangements.

The number of ways that two, three, four, or six identical subunits can pack together to form 'closed' arrangements is strictly limited. Trimers exclusively form triangular structures with a three fold rotational symmetry axis as illustrated by chloramphenicol acetyltransferase (3cla) in figure 3.5(b). Homo-tetramers most often (but not always) adopt a square like configuration an example of which is human beta tryptase (1a0l) in figure 3.5(c). By treating protomers as if they were spheres Teller, 1976, estimated that each subunit that forms a square-like tetramer buries 43% of its total ASA in protein-protein interfaces. The figure for a protomer forming a dimer is predicted to be 14%. The differences between these predicted values and the actual values in table 3.1 can be attributed to the variety of shapes adopted by the protomers.

For higher multimers (trimers, tetramers, hexamers) one may ask if the multiple interface regions are playing equivalent roles, or whether a multimer is more tightly bound at some interfaces than others. This question is strongly related to the types of symmetries possible within the higher multimer types, and is of interest because of the light it may throw on the evolutionary histories (and present functions) of multi-subunit complexes. This question is investigated in table 3.2 and figures 3.5 and 6.

Table 3.2 lists the mean sizes of the interfaces a protomer makes with its neighbours in the various multimer types. In this table interface 1 (I_1) is the largest protein-protein interface, interface 2 (I_2) is the second largest and so on. The numbers of proteins with the given interface are shown in bold. The percentage of the total interface ASA buried that each individual interface represents is given in brackets next to the absolute size of the interface concerned.

	Dimers	Trimers	Tetramers	Hexamers
Interface 1 (I_1)	(76)	(26)	(31)	(9)
Mean (\AA^2)	1890 (100%)	1310 (51%)	1790 (59%)	1643 (42%)
Min	540	511	510	687
Max	7150	2625	4537	3886
SD	1170	633	940	1037
Interface 2 (I_2)	-	(26)	(31)	(9)
Mean (\AA^2)		1252 (49%)	999 (33%)	935 (27%)
Min		455	397	603
Max		2593	4083	1903
SD		639	740	401
Interface 3 (I_3)	-	-	(22)	(9)
Mean (\AA^2)			448 (11%)	730 (22%)
Min			65	427
Max			1939	1043
SD			441	223
Interface 4 (I_4)	-	-	-	(7)
Mean (\AA^2)				456 (11%)
Min				167
Max				617
SD				151

Table 3.2: The mean sizes of the interfaces a protomer makes with its neighbours in homo-dimers, trimers, tetramers, and hexamers. Interface 1 (I_1) is the largest protein-protein interface, interface 2 (I_2) is the second largest and so on. The numbers of proteins with the given interface are shown in bold. The percentage of the total interface ASA buried that each individual interface represents is given in brackets next to the absolute size of the interface concerned.

Figure 3.4 gives a plot of I_1 against I_2 for each protein in each dataset. Both table 3.2 and figure 3.4 display a clearly distinct behaviour for members of the trimer dataset, whose two interface regions appear to be equivalent, resulting in highly symmetrical structures as in the example of figure 3.5(b), chloramphenicol acetyltransferase, (Leslie, 1990).

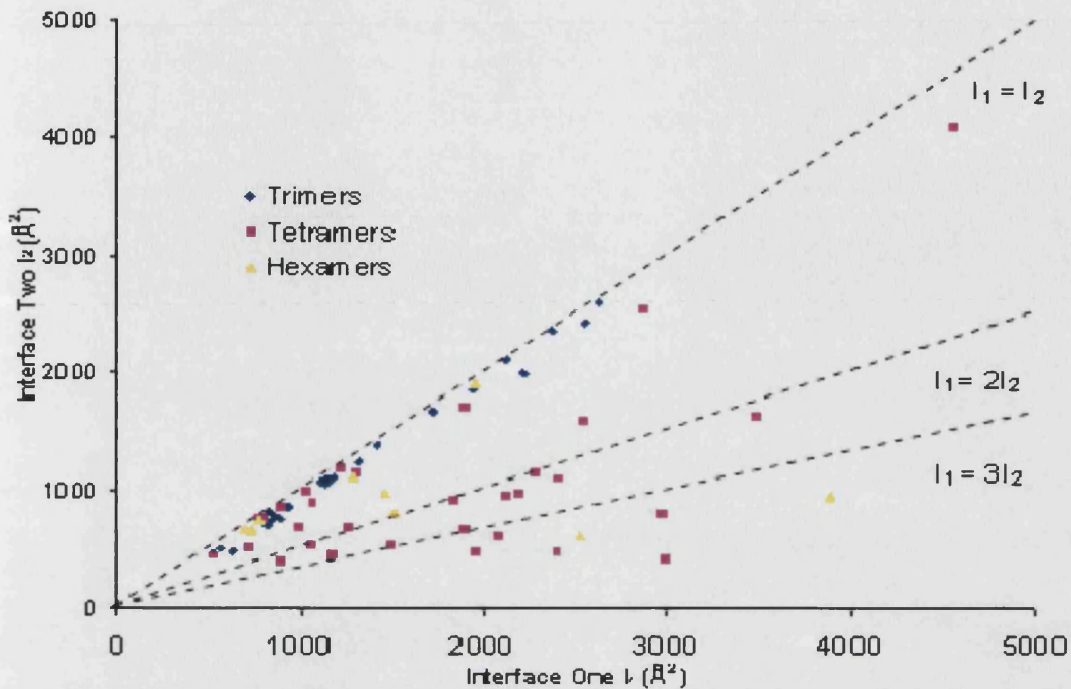
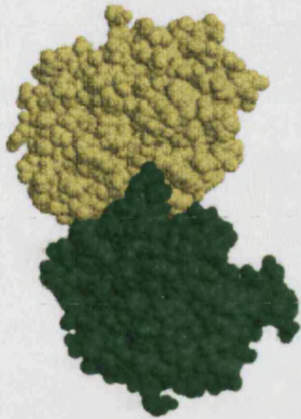


Figure 3.4: A plot of the ASA of the largest protein-protein interface (I_1) for each trimer, tetramer, and hexamer against the ASA of its second largest interface (I_2).

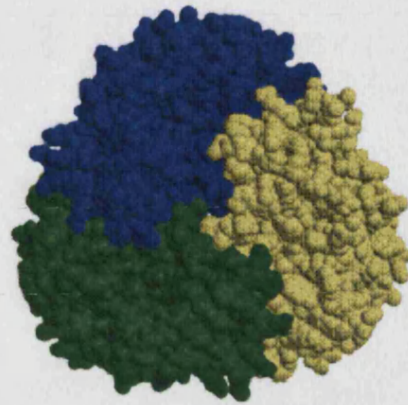
Tetramers conversely typically have asymmetric interfaces. Each protomer has one large and one small interface. Approximately ninety per cent of the total buried ASA for a tetramer is contained within the two largest protein-protein interfaces. This reinforces the idea that many tetramers can be considered as a ‘dimer of dimers’, as in the example of figure 3.5(c) (human beta tryptase), which has a large intra-dimeric interface and a smaller, weaker inter-dimeric one. These smaller interfaces are often used for channelling substrates between the active sites of the tetramer (Miller, 1989). The subunits that make up a tetramer can also associate more intimately so that each subunit contacts every other subunit in the tetramer, resulting in three protein-protein interfaces per protomer. Twenty two out of the thirty one tetramers are of this type, a typical example being catalase (Mate et al., 1999), shown in figure 3.5(d). From table 3.1 the protomers that form trimers and tetramers both typically bury around 23% of their ASA in protein-protein interfaces. This is due to the observation that the protomers of trimers and tetramers usually form two protein-protein interfaces with each other protomer within the complex. Each protomer in eight of the nine hexamers forms four protein-protein interfaces with the other protomers in the complex.

(a)



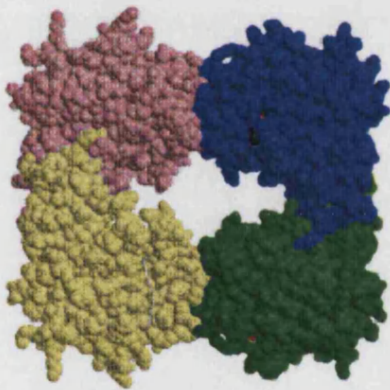
1amk

(b)



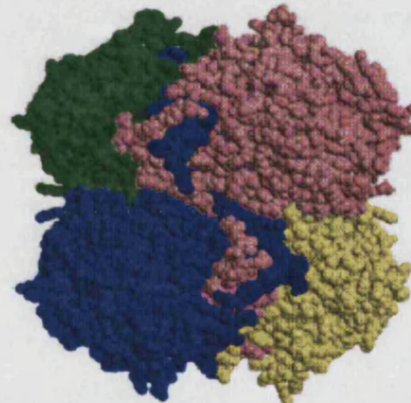
3cla

(c)



1a0l

(d)



1a4e

Figure 3.5: (a) Triose phosphate isomerase (1amk, Williams et al., 1999). (b) The chloramphenicol acetyltransferase trimer (3cla, Leslie, 1990). (c) The human beta tryptase tetramer (1a0l, Pereira et al., 1998). (d) Catalase (Mate et al., 1999).

The two largest interfaces of half of the proteins in the dataset of hexamers are of roughly the same size, which would suggest a ‘dimer of trimers’ arrangement, with the two trimers bound back to back; other hexamers studied here take the form of a ‘trimer of dimers’, with the three dimers associated around a threefold axis. Branched-chain amino acid aminotransferase (Okada et al., 1997) is a hexamer which

can be considered as a dimer of trimers. Another example of a hexamer that can be considered to be 'dimer of trimers' is inorganic pyrophosphatase (2eip) shown in figure 3.6 (Kankare et al., 1996). Hexamers can also be formed from the association of six equivalent monomers to form a ring-like structure, although this class of hexamer is not found in the dataset of hexamers that is studied here. Some of these ring-like hexamers are to some degree transient, breaking up into their constituent subunits at some point during their time in the cell prior to degradation. .

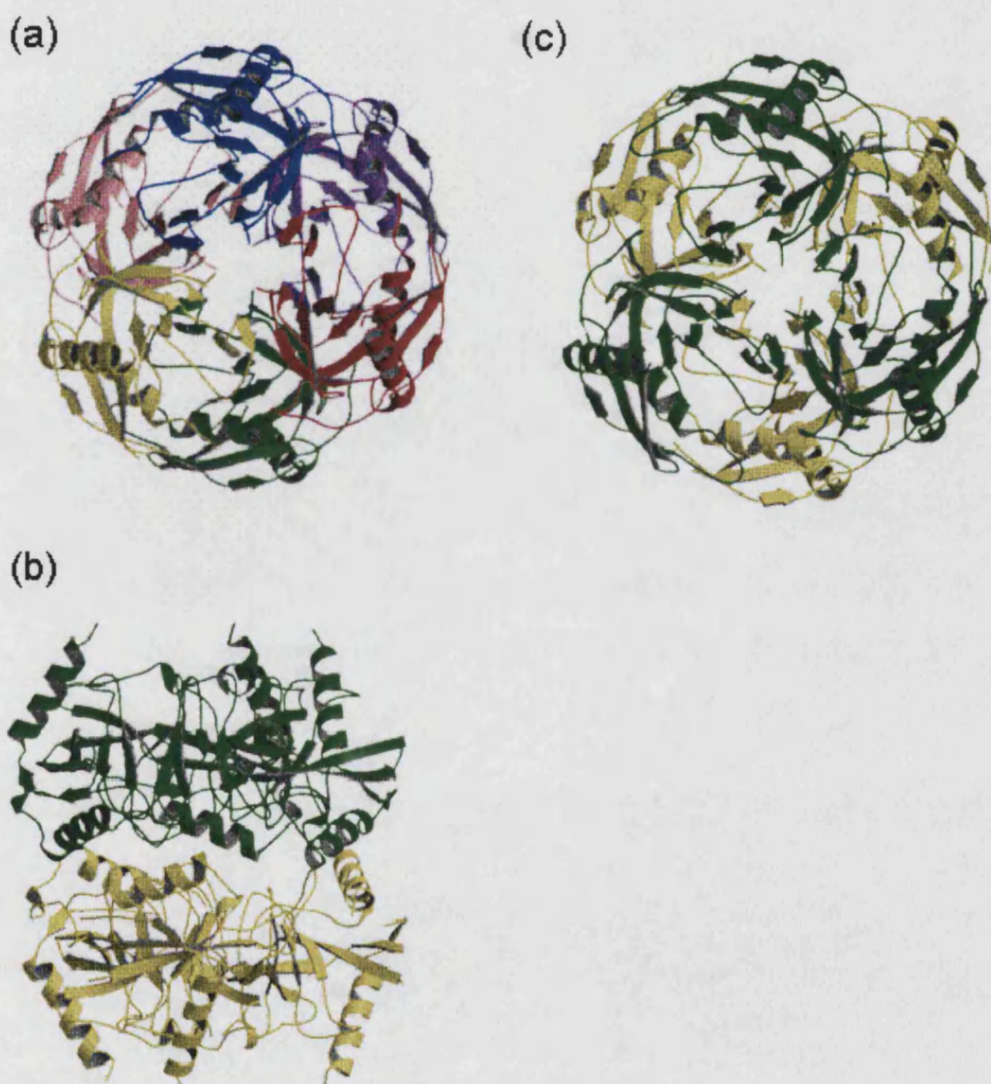


Figure 3.6: The inorganic pyrophosphatase hexamer (2eip, Kankare et al., 1996). (a) The hexamer with each subunit coloured up in a different colour. The hexamer can be considered to be a 'dimer of trimers'. In (b) and (c) the hexamer is shown in two different orientations with the two trimers that make up the full complex coloured in green and yellow.

In the above section dimers and hexamers have been considered to be composites of lower multimers based on ASA measurements of the individual protein-protein interfaces that exist within each complex. However it is important to note that doing this is somewhat risky. More evidence such as the conservation of residues at each protein-protein interface is needed to unequivocally say that any given multimer is a composite structure. In contrast to the findings of Jones & Thornton, 1997 there is only a rather weak correlation between the total number of residues in a protomer and the total number of residues in all its protein-protein interfaces within the multimer. The Pearson correlation coefficient is 0.55 for the line of best fit that was fitted to all datasets. There is a stronger correlation in the dataset of hexamers (0.71) while the correlation coefficient for the dataset of trimers the coefficient is 0.60.

3.4 Symmetry

In homo-complexes symmetry is the rule rather than the exception. There are asymmetric homo-complexes such as hexokinase which is a dimer but such complexes are rare. Since the work of Monod et al., 1965 this observation has been a source of fascination for many authors. A number of reasons have been proposed to explain the abundance of proteins that possess various kinds of symmetries (single rotational axes such as 2-folds in most homo-dimers, or combinations of intersecting axes, such as 222 in most homo-tetramers, or 32 in many homo-hexamers) but as yet no satisfactory explanation has emerged. There are already several comprehensive reviews of protein symmetry and function and these provide a much more complete treatment of the subject than is possible here (Blundell & Srinivan, 1996, Goodsell & Olson, 2000, Kumar et al., 2000). Proteins are chiral objects being composed of chiral amino acids. This together with the fact that from a geometric point of view there are a limited number of ways that a number of identical proteins can pack together to form a closed and reasonably compact complex is a good starting point in explaining why homo-complexes so often observed to possess various kinds of symmetry. Homo-dimers such as triose phosphate isomerase in figure 3.5(a) usually possess a single two fold rotational axis of symmetry (point group C_2). The interfaces between the two subunits of homo-dimers are always isologous. Trimeric proteins have a three

fold axis of rotational symmetry (point group C_3) as illustrated by chloramphenicol acetyltransferase in figure 3.5(b). The protein-protein interfaces of trimeric proteins are exclusively heterologous

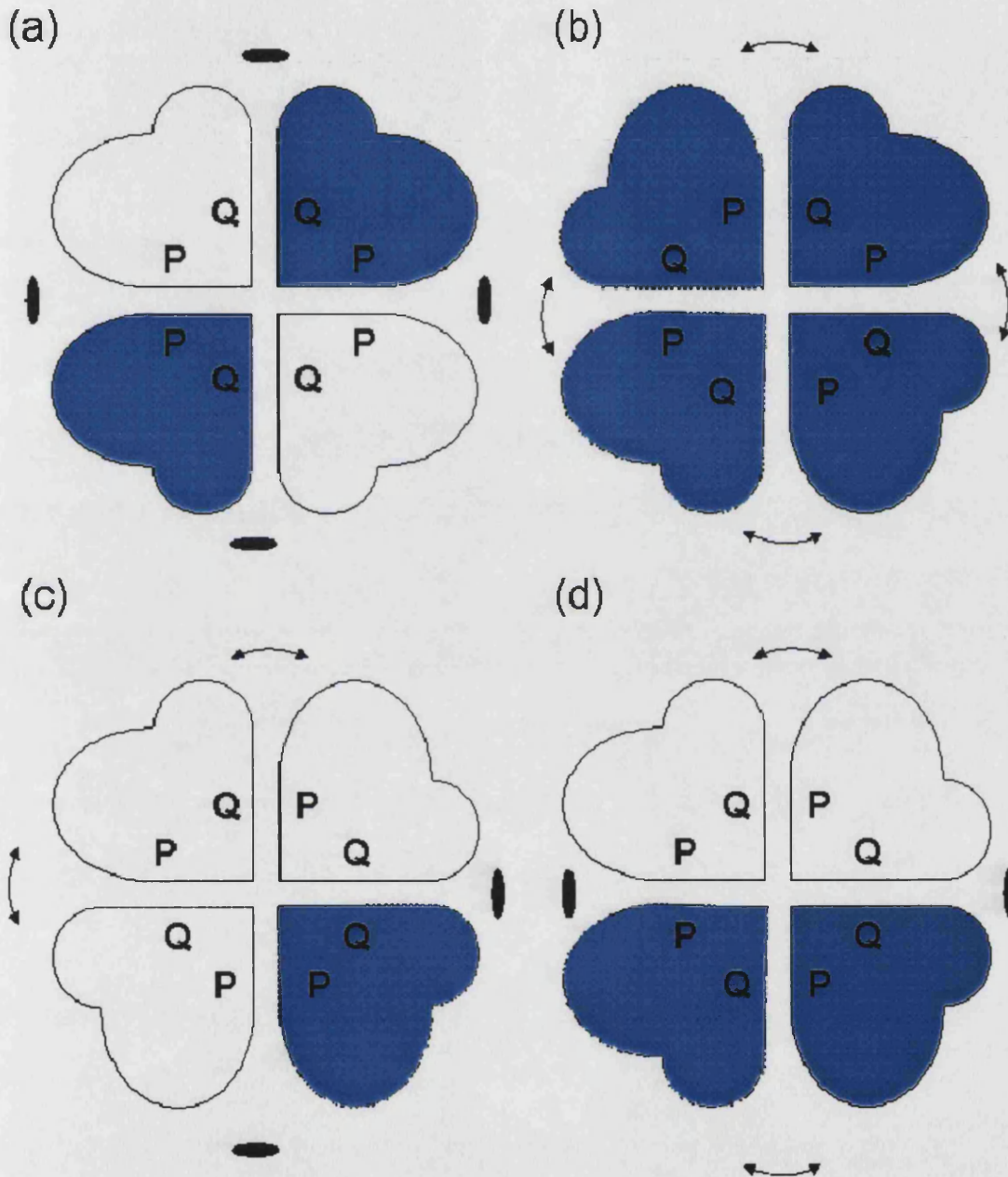


Figure 3.7: The four possible ways that four subunits can pack together to form a planar tetramer. The two different binding surfaces are labelled as P and Q. Subunits coloured in blue are rotated by 180° about a two fold axis in the plane compared with unshaded subunits. The two symmetrical arrangements shown in (a) are predicted to be more stable than the two asymmetrical arrangements in (b) (Cornish-Bowden et al., 1971). This figure is adapted from figure 3, Goodsell & Olson, 2000.

Tetrameric proteins can possess translational as well as rotational axes of symmetry. Most homo-tetramers have three mutually perpendicular two-fold axes of rotational symmetry described by the point group D_2 or in crystallographic notation by the symmetry group 222.

Tetramers are of special interest because they are the simplest case of protein-complexes in which both isologous and heterologous protein-protein interfaces may be found. Cornish-Bowden & Koshland, 1971, considered the number of different ways that four identical subunits could pack together to form a 'closed' and planar homo-tetramer. Four different tetramers are possible given that the complex must be both planar and closed and these are shown in figure 3.7.

Each subunit is considered to have two separate binding surfaces marked P and Q. Thus PP and QQ interfaces are isologous whereas PQ interfaces are heterologous. Binding energies are then assigned to the PP, QQ, and PQ interfaces and the stability of the overall complex is then assessed by simply summing the energies of all the different interactions within the tetramer. By systematically varying the binding energies of the three different interfaces Cornish-Bowden and Koshland showed that the two symmetric arrangements in figures 3.7(a) and (b) are strongly favoured over the two asymmetric tetramers in figures 3.7(c) and (d). This is true even when the differences between the binding energies of the PP, QQ, and PQ interfaces are quite minor.

The arrangements of subunits within each of the 31 homo-tetramers was analysed to test the hypothesis that the two symmetric arrangements with either all isologous or all heterologous interfaces are strongly preferred over the asymmetric arrangements. Twenty nine out of the thirty one tetramers adopt the arrangement in figure 3.7(a) or other non-planer arrangements in which the complex has exclusively isologous interfaces. The remaining two complexes, 1cuk, and 2fua in figure 3.19(b) adopt the arrangement shown in figure 3.7(b) with both complexes having cyclical symmetries described by the point group C_4 and protein-protein interfaces that are exclusively heterologous. Whether this means that isologous protein-protein interfaces are stronger than heterologous interfaces or that the exclusive use of either of these types of interface in a complex equates with stability compared to complexes in which both

isologous and heterologous are found is controversial and very much open to question. As pointed out in chapter 1 the exclusive use of isologous interfaces leads to closed structures thereby avoiding further polymerisation. This has to be an important consideration in understanding why isologous interfaces are so prevalent in homo-tetramers (and indeed in the other homo-complexes as shown in section 3.12)

Homo-hexamers most commonly exist in two forms. Hexamers can form cyclical structures in which subunits are related by a six fold rotational axis with C_6 point symmetry. The insulin hexamer is one such protein with this kind of symmetry. The alternative to this is an arrangement of subunits around a three fold axis of rotational symmetry with point symmetry D_3 . All of the proteins in the hexamer dataset are arranged in complexes with D_3 symmetry.

3.5 Amino Acid Composition

In relation to surface properties like amino acid composition there are two central questions. Firstly, can binding regions be differentiated from non-binding regions? Secondly, if this is possible, can the oligomerisation state (and the geometry of the multimer) be deduced from the number, size and position of an individual protein's predicted interface regions? It will be seen here and in later sections of this thesis that the former can to some degree be answered in the affirmative for certain of these surface properties, but that the results presented here indicate that the latter is not presently within the reach of our predictive methodologies.

	Amino Acid
Hydrophobic	Ala, Gly, Ile, Leu, Met, Phe, Pro, and Val
Polar	Asn, Cys, Gln, His, Ser, Thr, Trp, and Tyr
Charged	Asp, Arg, Lys, and Glu

Table 3.3: The scheme used to classify residues as being hydrophobic, polar, or charged.

The frequency distribution of each of the twenty amino acids was analysed for each dataset of proteins. In common with (Jones & Thornton, 1996), non-polar amino acids

are taken to be Ala, Gly, Ile, Leu, Met, Phe, Pro, and Val. Polar amino acids are taken to be Asn, Cys, Gln, His, Ser, Thr, Trp, and Tyr. Asp, Arg, Lys, and Glu are taken to be charged amino acids. This classification scheme is summarised in table 3.3.

In assessing the differences between binding and non-binding regions a logical starting point is to compare the surfaces of monomers with the surfaces of multimers.

The percentage frequency occurrence for each subset of residues (interior, interface, and exterior) in terms of its composition by polar, non-polar, and charged amino acids is given in table 3.4. From table 3.4 the surfaces of the 92 monomers and 142 homo-complexes are very similar. The monomer dataset was compiled using the same protocol used to construct the datasets of homo-complexes and is a kind gift from Hannes Postingl at the EBI, UK.

	Non-Polar (%)	Polar (%)	Charged (%)
Monomers			
Interior	73.4	22.7	3.9
Exterior	38.9	31.3	29.8
All Homo-Complexes			
Interior	72.8	22.9	4.3
Exterior	41.8	29.1	29.1
Interface	44.8	30.6	24.6

Table 3.4: The average percentages of residues that are non-polar (hydrophobic), polar, and charged in the dataset of 92 monomers and of all the datasets of homo-complexes combined. The exteriors of the monomers and the homo-complexes appear to have a similar amino acid composition.

The average percentages of charged and polar residues on the surfaces of the monomers and the homo-complexes are the same to within two percent. The surfaces of the monomers typically contain about three percent fewer hydrophobic residues than the homo-complexes but this difference is still rather insignificant.

Figure 3.8 shows the differences in the averaged amino acid composition of the twenty amino acids between the monomers and homo-complexes. From figure 3.8

the exterior of the homo-complexes contain very slightly higher numbers of hydrophobic residues such as leucine, alanine, and isoleucine. On the other hand, the exterior of monomers are faintly enriched in charged or polar residues like lysine, asparagine, and glutamine. But in no case are the differences for any of the amino acids in figure 3.8 much more than one percent.

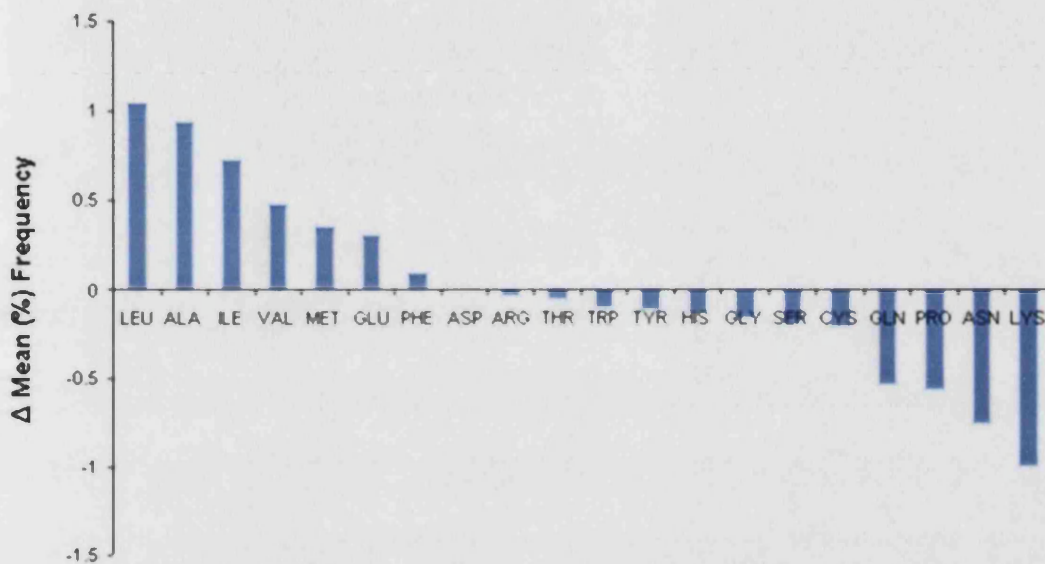


Figure 3.8: A chart showing the differences in the mean amino acid composition of the exteriors of all the 142 obligate homo-complexes together and the 92 monomers. A Δ mean (%) frequency >0 indicates that the exteriors of the homo-complexes are enriched in the residue compared with the monomers.

The interior of each class of multimer (and the monomers) are unsurprisingly dominated by hydrophobic residues. Hydrophobic residues on average make up about 73% of all interior residues while charged residues comprise just 4%.

The exteriors of the homo-dimers, trimers, tetramers, and hexamers are all made up out of quite comparable fractions of hydrophobic, polar, and charged residues. Hydrophobic residues typically make up ~40% of the exterior while polar and charged residues each comprise around 30% of the total number of exterior residues. Of course these are average values and there are proteins whose interfaces have very different amino acids compositions to those suggested by the data in table 3.4

The data in table 3.4 suggests that the surface of a multimeric protein and its protein-protein interface(s) have a broadly similar polar, non-polar and polar composition. This is unremarkable since interface residues are after all a subset of a protein's exterior.

Figure 3.10 shows the mean amino acid composition of the interior, exterior, and interface residues of all the homo-complexes together. From this figure it can be seen that on the level of the amino acids the exterior and interface are well correlated. However the protein interface appears to be slightly enriched, relative to the protein exterior when taken as a whole, in bulky hydrophobic residues and in certain aromatic residues such as tyrosine and phenylalanine. This makes sense. The greater the amount of ASA that is buried when several protomers bind to form a complex, the larger the amount of free energy produced helping to make the whole binding process thermodynamically favourable.

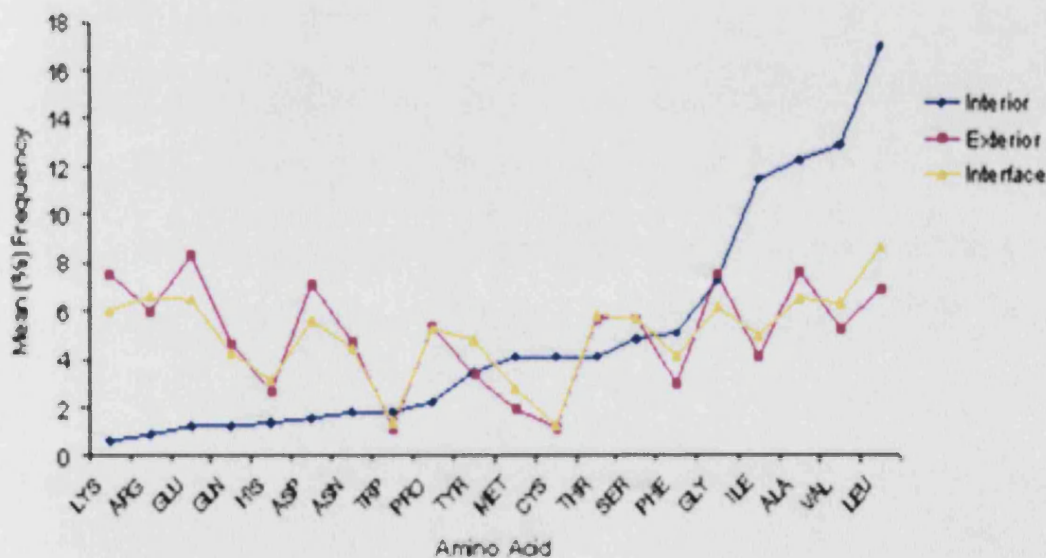


Figure 3.10: The mean percent frequencies of amino acids in the interior, exterior, and interface regions of all 142 homo-complexes. The amino acids are ordered according to the increasing % frequency in the interior region.

In addition aromatic residues have few rotatable bonds, so there is little loss of entropy upon binding. Aromatic residues such as tyrosine are also able to take part in the formation of a number of inter-subunit hydrogen bonds and cation- π interactions

that help to stabilise the subunits within the multimer. Histidine appears to be slightly more prevalent in the protein interface than in any other part of the protein. A possible explanation of this is that many of the proteins in the datasets that are studied here are enzymes in which active sites utilising histidine are often located at or near a protein interface.

To return to the question of whether dimers, trimers, tetramers or hexamers are distinguishable from each other using averaged amino acid composition data alone the answer seems to be no. The amino acid composition of the protein interior, exterior and interface regions does not vary significantly across the four classes of multimer. On the other hand protein-protein interaction sites seem to have an amino acid composition somewhere between that of the protein interior and exterior, as was found by Argos, 1998. Interface regions contain slightly larger fractions of hydrophobic residues than the entire exterior of a protein so it may be possible to distinguish at some level between binding and non binding regions.

To further investigate the nature of residues that are found within protein-protein interfaces the interface propensity of each amino acid has been calculated for all four classes of multimer. The residue interface propensity for any given amino acid is defined as:

$$\text{Residue Interface Propensity} = \frac{\% \text{ Frequency Protein Interface}}{\% \text{ Frequency Protein Surface}} \quad (2)$$

The interface propensity provides a quantitative measure of how likely a given residue is to be found in a protein-protein interface, with a propensity > 1 indicating that the residue is more likely to be found in an interface than on the surface of the protein.

A chart separately illustrating the residue interface propensities of the dimer, trimer, tetramer, and hexamer datasets is presented in figure 3.11; these values were calculated on the basis of 'binning' all interface residues (irrespective of whether or not the protomer has more than one interface within the multimer), effectively treating the protein as if it had only one protein-protein interface.

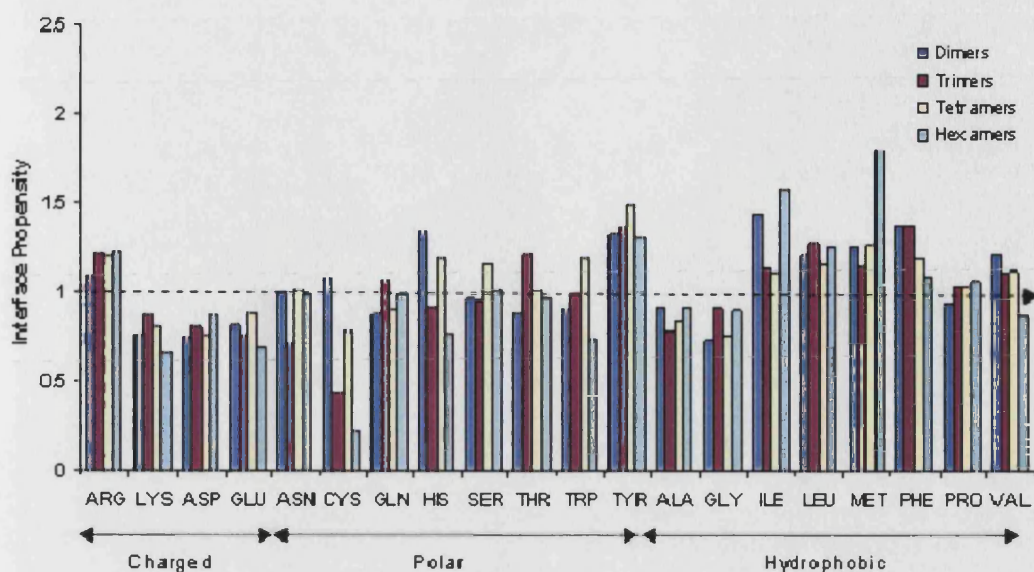


Figure 3.11: Mean residue interface propensities for the datasets of homo-dimers, homo-trimers, homo-tetramers, and homo-hexamers.

For comparison, averaged residue propensities are given in table 3.5 for the complete multimer dataset (dimers, trimers, tetramers and hexamers). Looking at figure 3.11 the interface propensities for dimers, trimers, tetramers, and hexamers do vary but in no systematic way. This suggests that distinguishing between dimers, trimers, tetramers, and hexamers would be very difficult using amino acid composition data alone. In all four classes of multimers hydrophobic residues have large interface propensities.

Looking at the interface propensities of all 142 homo-complexes taken together the two residues with the highest interface propensity are tyrosine (1.38) and phenylalanine (1.32). As previously mentioned the principle reasons why these residues are particularly common at protein-protein interfaces is due to their physical bulk leading to large amounts of free energy on binding and the ability of both residues to mediate various kinds of electrostatic interactions between the two sides of a protein-protein interface. Hydrophobic residues like isoleucine, methionine, and leucine are also preferred at the interface judging from their interface propensities.

Arginine has an interface propensity of 1.15 and is the only charged residue that protein-protein interaction sites are enriched in.

Amino Acid	Residue Interface Propensity
ARG	1.15
LYS	0.79
ASP	0.77
GLU	0.81
ASN	0.95
GLN	0.93
SER	1.00
GLY	0.78
HIS	1.19
THR	0.98
ALA	0.89
PRO	0.98
TYR	1.38
VAL	1.16
MET	1.28
CYS	0.83
LEU	1.21
PHE	1.32
ILE	1.32
TRP	0.98

Table 3.5: The residue interface propensities for all 142 homo-complexes. The higher the interface propensity the more likely the residue is to be found in a protein-protein interface. Residues placed in order of increasing hydrophobicity (see table 3.7)

The role of arginine residues at protein-protein interfaces has been established by other authors. Due to its positively charged side-chain arginine is frequently observed to ligate the aromatic rings of tyrosine, tryptophan, and phenylalanine creating interactions with quite substantial binding energies (Glasser et al., 2001). For this reason, its physical size and hydrogen bonding capability arginine has been experimentally observed to be located in ‘hot spots’ of binding energy in protein-protein interfaces along with tyrosine and phenylalanine (Bogan & Thorn, 1998). These ‘hot spots’ of binding energy have are further discussed in chapter 4. Along with the results of Ofra & Rost, 2003, our results point to tryptophan not being particularly prevalent at the protein-protein interfaces of obligate homo-complexes. The reasons for this are complicated and may have something to do with the difficulty

of accommodating such a large amino acid in the tightly packed protein-protein interfaces of many homo-complexes.

3.6 Hydrophobic Content

As already discussed in section 3.3 proteins that bind to form different classes of multimer must bury varying fractions of their surface area in protein-protein interfaces. It is therefore worth seeing if multimers are distinguishable from each other in terms of their total hydrophobic content. Initially, as in the previous section, classifying residues simply as either hydrophobic (Ala, Gly, Ile, Leu, Met, Phe, Pro, Val) or not, the percentage of such hydrophobic residues in the protomer of each protein was calculated and is plotted against the number of residues in the multimer in figure 3.12. The plot reveals that proteins belonging to each dataset appear to have similar distributions of hydrophobic residues, with the mean percentage of hydrophobic residues in proteins belonging to each dataset summarised in table 3.6.

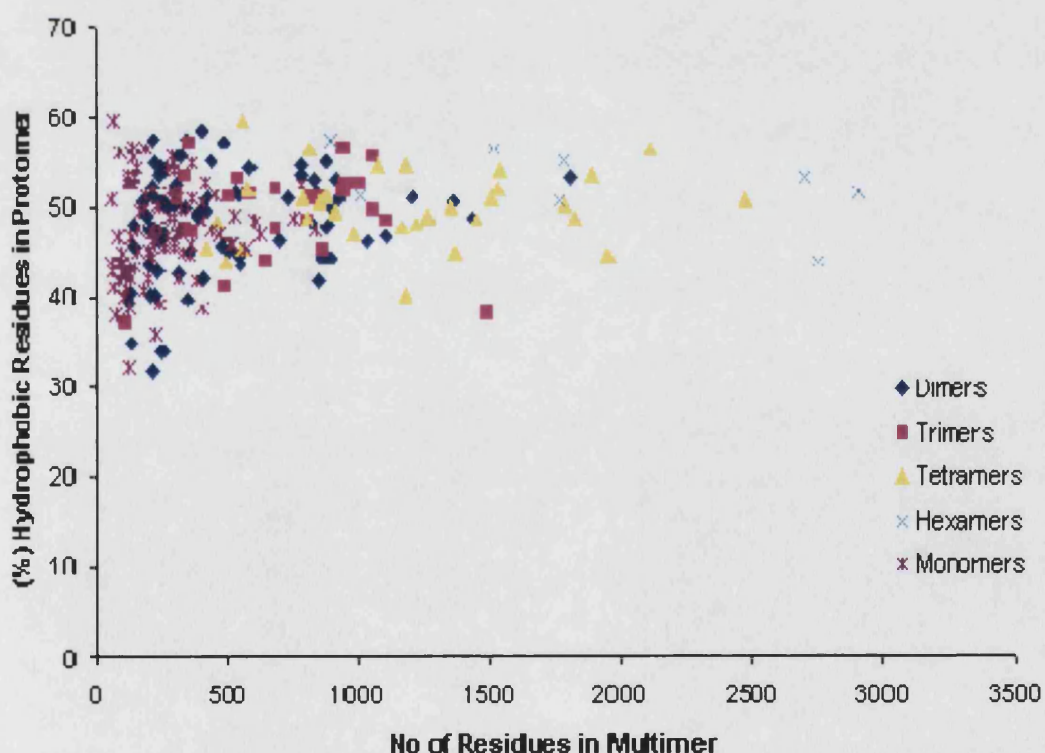


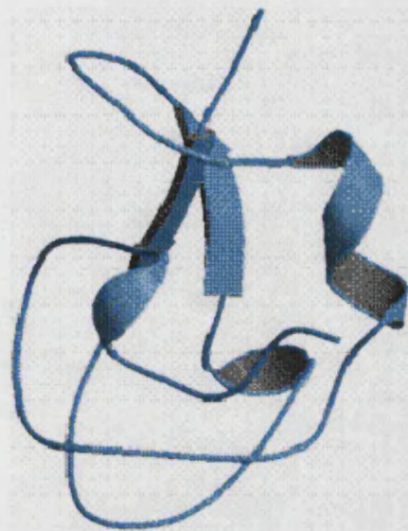
Figure 3.12: A plot of the number of residues in each multimer against the percentage of residues that are hydrophobic in one protomer (subunit) of the full multimer.

	Mean (%) Hydrophobic Residues in Entire Protomer	Min	Max	SD
Monomers	47.4	32.3	59.4	5.1
Dimers	48.3	31.8	58.5	5.6
Trimers	49.4	37.1	57.3	5.0
Tetramers	49.9	40.3	59.6	4.4
Hexamers	52.0	43.9	57.3	4.2

Table 3.6: The mean hydrophobic content of the 92 monomers, and of the protomers from the datasets of homo-dimers, trimers, tetramers, and hexamers.

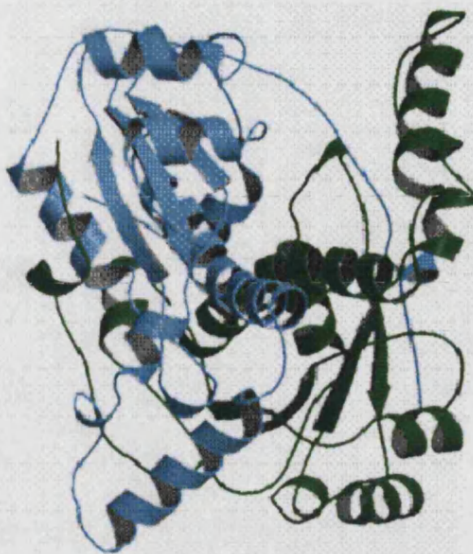
The higher the multimer, the greater the percentage of hydrophobic residues that make up the protein. However this trend on closer inspection is very weak. The differing standard deviation values on these mean percentages imply that this trend is not statistically significant. It must be concluded that the proteins from differing datasets nevertheless contain a very similar percentage of hydrophobic residues. This is despite a dimer monomer burying on average 16% of its ASA on binding compared with 26% for hexamers. However, one needs also to consider what is buried in the monomer interiors; from the point of view of thermodynamic stability it does not matter whether hydrophobic monomers are buried in the interior or interface, so long as they are in the main removed from solvent. It is interesting to see from the data in table 3.6 and figure 3.12 that there is no protein that has a hydrophobic content >60%. It is worthwhile asking why this is. The type III antifreeze protein (1ops) is a monomer and has the highest hydrophobic content of any protein considered in this thesis at 59.4%. The protein exists in sub-zero temperatures lowering the blood freezing point by absorbing ice and inhibiting its growth (Yang et al., 1998). NADH oxidase (1nox) is a dimer with a hydrophobic content of 58.5%. This protein is found in an extreme thermophile, *Thermus thermophilus*, and must withstand temperatures approaching boiling point (Hecht, 1995). Diagrams of both 1ops and 1nox are shown in figure 3.13. Both proteins have rather compact structures and may represent near optimal packing arrangements of hydrophobic residues in order to provide stability in extreme conditions. It is also possible that solubility becomes an issue in proteins with a hydrophobic content approaching the 60% threshold. When a plot is made of the number of residues in the protomer against the percentage of exterior residues that are hydrophobic a similar distribution is observed to that seen in figure 3.12.

(a)



1ops

(b)



1nox

Figure 3.13: (a) The type III antifreeze protein (1ops, Yang et al., 1998) and NADH oxidase (Hecht et al., 1996). Both of these proteins have a high hydrophobic content and function in extreme conditions.

No protein has an exterior with a hydrophobic content exceeding 55%. Homotetrameric haemoglobin (1lth) from *Urechis caupo* has the most hydrophobic protomer exterior with a hydrophobic content close to this threshold. The hydrophobic content of datasets of obligate hetero-complexes is considered in the next chapter.

3.7 Hydrophobicity

Hydrophobicity has been considered implicitly in the previous section dealing with amino acid content, but only at a coarse level, regarding individual amino acids as either hydrophobic or not. Numerical hydrophobicity scales such as the Fauchere and Pliska reference scale used later in this section allow a more subtle probing of the role of hydrophobicity at protein-protein interfaces. Moreover, hydrophobicity deserves special attention because of its central role in protein folding. It has been shown that every angstrom squared of buried ASA gives rise to about 25 cal mol^{-1} of free energy (Janin & Chothia, 1975) Since the largest single contribution to the free energy of

binding is from the hydrophobic effect it can be considered the major driving force in protein binding. But given the results regarding residue interface propensity, it cannot be expected that hydrophobicity will provide an unambiguous signal of a protein-protein interface region.

Looking more specifically at the role of hydrophobicity within protein-protein interfaces in relation to other regions of the protomers, numerical values for the hydrophobicity of the interior, exterior, and interface residues for each dataset were assigned using an experimentally determined hydrophobicity scale proposed by Fauchere & Pliska (1983). A hydrophobicity coefficient for each of the twenty amino acids is based upon measurements of the solubility of analogues for each acid in water/octane mixtures. The hydrophobicity value assigned to each amino acid is shown in table 3.7. The lower the hydrophobicity value the more hydrophilic the residue. Although there are many hydrophobicity scales in use the Fauchere & Pliska scale has been widely used by other authors and it is used here to allow for comparison with the work of Jones & Thornton, 1996.

The hydrophobicity of any given set of residues is simply a sum taken over the total number of each amino acid multiplied by its corresponding hydrophobicity value shown in table 7.

$$H_{total} = \sum_{A=1}^{20} N_A V_A \quad (3)$$

The total hydrophobicity of a set of residues is then given by equation 3. The mean hydrophobicity of the set of residues is then H_{total} divide by the total number of residues in the sample and is given by equation 4.

In equation 3 and 4: $\langle H_{total} \rangle$ is the mean hydrophobicity, N_A is the number of amino acids of type A , V_A is the hydrophobicity of the amino acid of type A , and N_{total} is the total number of amino acids in the sample over which the average is being taken.

Amino Acid	Hydrophobicity V_A
ARG	-1.01
LYS	-0.99
ASP	-0.77
GLU	-0.64
ASN	-0.60
GLN	-0.22
SER	-0.04
GLY	0.00
HIS	0.10
THR	0.26
ALA	0.31
PRO	0.72
TYR	0.96
VAL	1.22
MET	1.23
CYS	1.54
LEU	1.70
PHE	1.79
ILE	1.80
TRP	2.25

$$\langle H_{total} \rangle = \frac{\sum_{A=1}^{20} N_A V_A}{N_{total}} \quad (4)$$

Table 3.7: The Fauchere & Pliska hydrophobicity scale. Residues are placed in order of increasing hydrophobicity. Arginine is the most hydrophilic residue with tryptophan being the most hydrophobic.

A chart illustrating the average hydrophobicity of interior, interface and exposed residues in each dataset using this scale is given in figure 3.14. As can be seen in figure 3.14, the hydrophobicity of a protein-protein interface is intermediate between that of the interior and exterior of the protein. There is little variation in hydrophobicity across any of the four classes of multimers. This supports the notion that protein-binding sites are formed by patches of residues that are relatively hydrophobic in character on the surface of a protein. These findings are the same as Chothia & Janin, 1975, and Korn & Burnett, 1991. In short, dimers, trimers, tetramers, and hexamers are essentially indistinguishable from each other using hydrophobicity considerations alone. In order to further investigate the distribution of hydrophobicity across protein-protein binding sites, the residues that form the protein interface were sub-divided into a ‘buried’ zone (being completely inaccessible to solvent) and a ‘partially accessible’ zone as described in section 3.2.

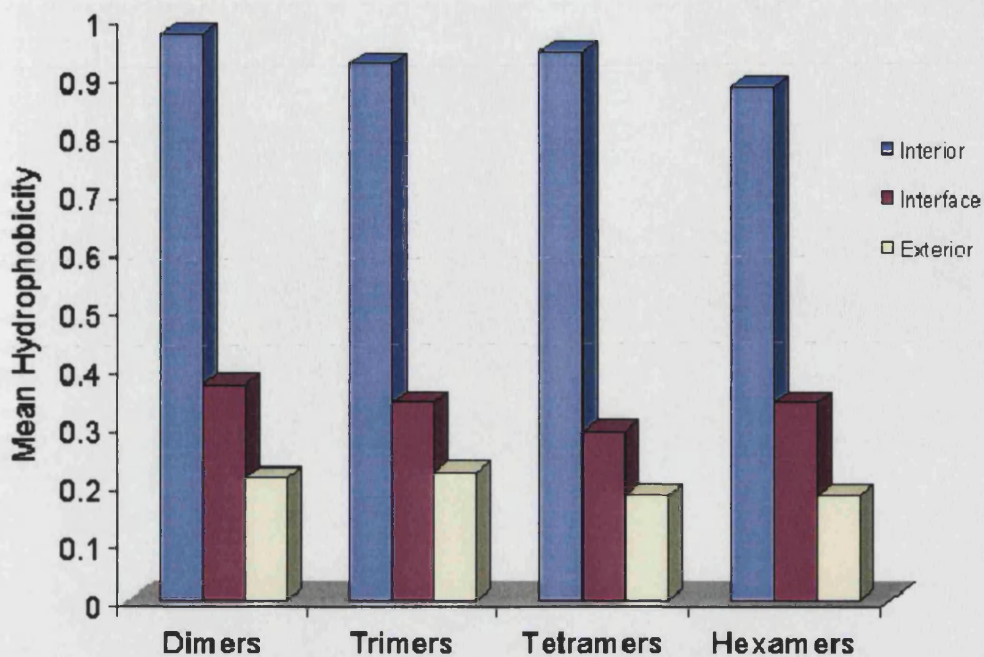


Figure 3.14: Mean hydrophobicities of the interior, interface, and exterior regions of the datasets of homo-dimers, trimers, tetramers, and hexamers. In each case the interface is intermediate in hydrophobicity between the protein interior and exterior.

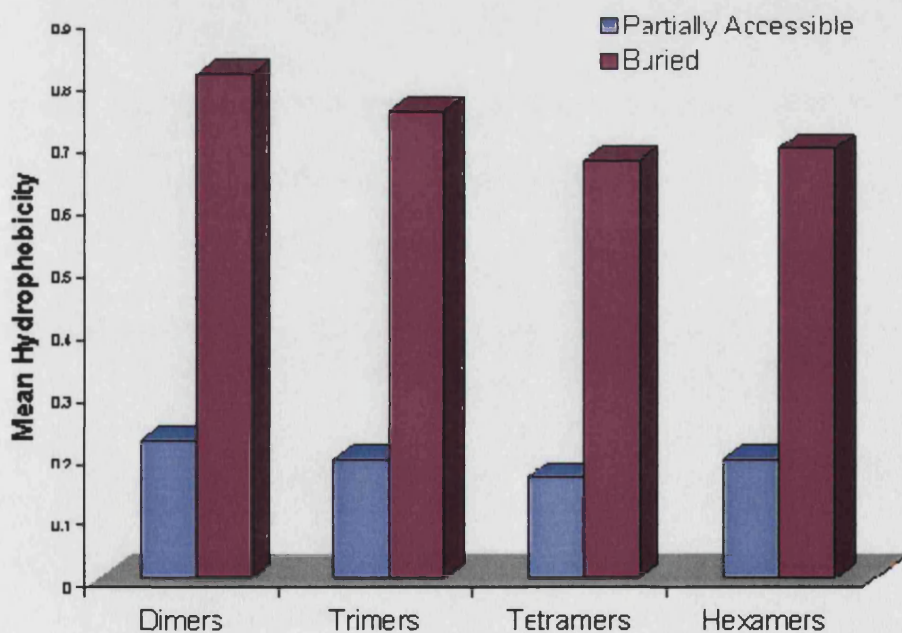


Figure 3.15: Mean hydrophobicities of the buried and partially buried regions of the datasets of homo-dimers, homo-trimers, homo-tetramers, and homo-hexamers.

The average hydrophobicity of all the residues that comprise each of the two subsets of residues was then calculated and is illustrated in figure 3.15. For all four datasets the residues that form part of the protein-protein interface but which are partially exposed to solvent are more polar in character than those that are completely buried. Since the partially buried zone usually surrounds those residues which are completely buried this would lead to a picture of a protein-binding site as being composed of a single hydrophobic core surrounded by an outer region of somewhat more polar residues. If this were in fact the case, protein-protein interfaces would be relatively easy to distinguish – but in practice they are not. This apparent contradiction can be explained by noting that protein-protein interfaces are often composed of a number of hydrophobic patches with a small number of polar or charged residues scattered across the entire extent of the interface. These polar or charged residues mediate salt bridges and hydrogen bonds at critical points that are required for the complex's stability.

Enolase is one such protein whose protein-protein interface is composed of a number of clusters of hydrophobic residues with no identifiable hydrophobic core (see figure 3.16(b)). Despite this the interface of enolase is still broadly hydrophobic. On the other hand, proteins whose interfaces do seem to have an identifiable hydrophobic 'core' include the rb69 sliding clamp protein in figure 3.1 and the Bence Jones Protein in figure 3.16(a).

Larsen et al (1998) conducted a visual survey of 136 homodimeric proteins in which only about a third of the proteins appeared to have protein-protein binding sites conforming to the idealised picture of a single hydrophobic core surrounded by a few polar residues shown in figure 3.16(a). The rest of the 136-protein dataset did indeed have less obviously distinguishable protein interfaces with hydrophobic patches, water molecules, and polar residues scattered across the entire interface like that of enolase in figure 3.16(b).

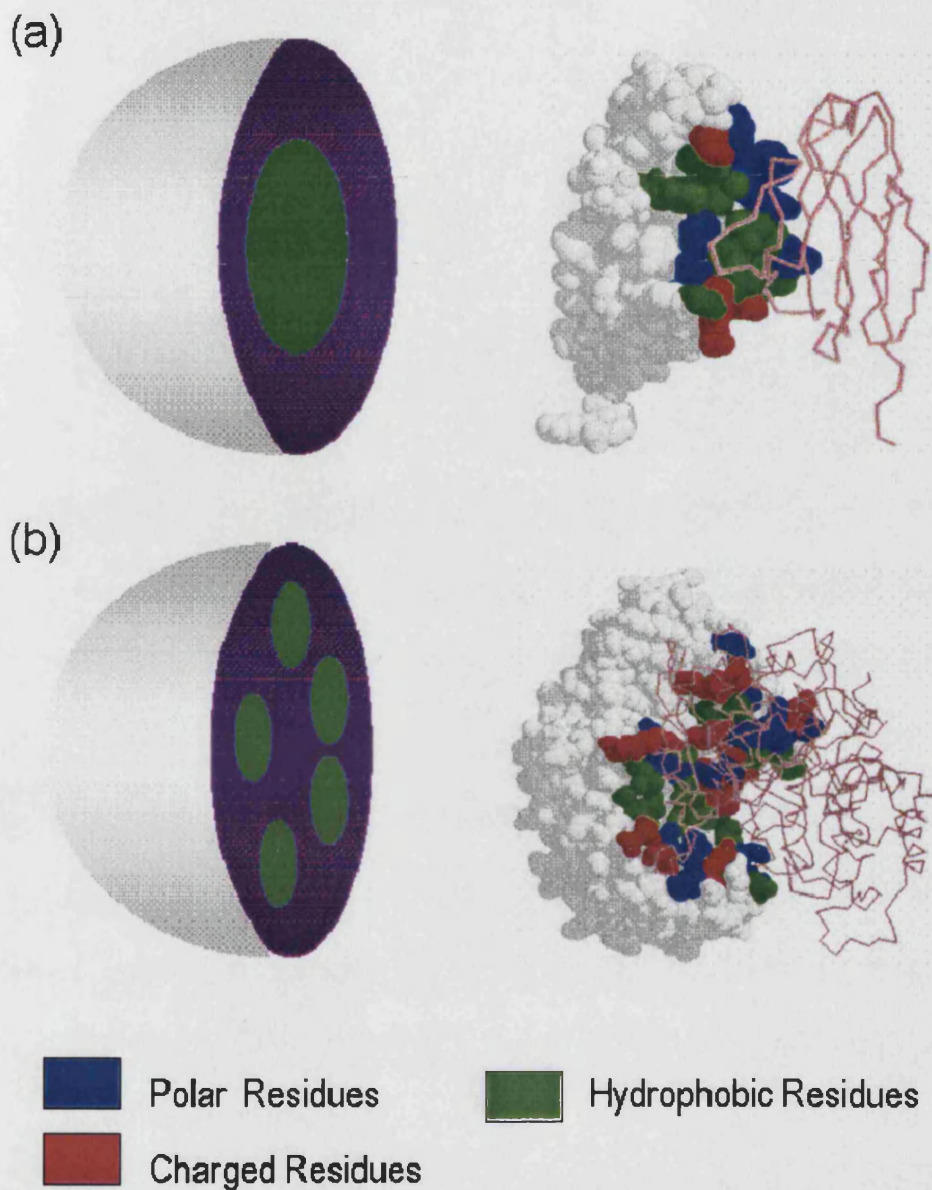


Figure 3.16: (a) The Bence Jones protein has a protein-protein interface with a single identifiable hydrophobic core surrounded by a few polar and charged residues. (2rhe, Furey et al., 1983). (b) Enolase (4enl, Lebiada et al., 1989) is a protein with clusters of hydrophobic residues scattered across the entire extent of its dimer interface with no identifiable hydrophobic core. These examples were taken from a survey carried out by Larsen et al., 1998 and is available at www.scrips.edu/pub/goodsell/interface.

3.8 Hydrogen Bonds

The nature of hydrogen bonds has been briefly explored in section 1.5.2.1 in chapter one. The correct formation of hydrogen bonds between interacting proteins is crucial to guaranteeing specificity in protein-protein interactions (Chothia & Janin, 1975, Fersht, 1987). It is therefore prudent to look at the numbers and types of hydrogen bonds at the protein-protein interfaces of the homo-complexes.

The number of inter-subunit hydrogen bonds was calculated using the program HBPLUS (McDonald & Thornton, 1994). The criterion used for defining acceptable hydrogen bond geometry is identical to that described in (Jones & Thornton, 1995), namely:

$$\begin{array}{ll} \text{D-A distance} < 3.9 \text{ \AA} & \text{DH-A angle} > 90^\circ \\ \text{H-A distance} < 2.5 \text{ \AA} & \text{HA-AA angle} > 90^\circ \\ & \text{DA-AA angle} > 90^\circ \end{array}$$

Where D is the hydrogen bond donor, A is the hydrogen bond acceptor, H is the hydrogen atom, and AA is the atom attached to the hydrogen bond acceptor. A diagram illustrating the above criteria is shown in figure 3.17(a). A more comprehensive account of the geometry of hydrogen bonds can be found in McDonald & Thornton, 1994. During protein folding both the backbone of the folding polypeptide chain and amino acid side chains are free to move in all directions in order to produce a compact globular structure with hydrogen bonds with optimal geometries. However when proteins form a complex the side chains of the amino acids between the interacting surfaces are constrained in the ways that they can move since they are bound to a relatively immobile polypeptide back-bone. This often leads to the formation of hydrogen bonds at protein interfaces with non-optimal geometries (and lengths) at the extremes of the criteria set out at the beginning of this section. Hydrogen bonds are strongest when the hydrogen bond donor and acceptor lay in a straight line. As discussed in section 1.5.2.1 in chapter one, hydrogen bonds play an important role in conferring directionality and specificity on protein-protein interactions of all kinds (Hubbard, 2001). An additional role hydrogen bonds have is in protein secondary structure. The two major regular secondary structure motifs α -helices and β -sheets are both maintained through an intricate network of hydrogen

bonds between the C=O and N-H groups that are to be found on the polypeptide backbone. A diagram of an α -helix is shown in figure A4(b). The α -helix is formed by a single polypeptide chain twisted about itself to form a “right-handed rod-like structure” (Hubbard, 2001). The structure is stabilised through hydrogen bonds between C=O of every amino acid and the N-H group of the amino acid to be found four amino acids way in the amino acid sequence as shown in figure A4(b). There are 3.6 residues per turn in an α -helix and the overall helix has a dipole moment (Braden & Tooze, 1998). A diagram of several β -strands hydrogen bonded to each other to form β -sheets is shown in figure A4(a). A β -strand is a single polypeptide chain of five to ten residues in a highly extended conformation. Individual β -strands are unstable in isolation of each other. As a result the C=O and N-H groups of adjacent β -strands hydrogen bond to each other to form β -sheets. Beta sheets can be parallel or anti-parallel depending on the relative directions of the hydrogen bonded β -strands (as shown in figure A4(a)).

A regression line was fitted to a plot of the total interface ASA per subunit against the numbers of inter-subunit hydrogen bonds for the dimer, trimer, tetramer, and hexamer datasets together shown in figure 3.17(b).

There is a clear linear relationship (with Pearson correlation coefficient across all datasets of 0.88) between the numbers of inter-subunit hydrogen bonds that a protein forms on becoming a multimer and its total interface ASA. This is true regardless of the multimeric state of the protein. As mentioned in chapter 4 water molecules are known to mediate hydrogen bonds at protein-protein interfaces (Conte et al., 1999).

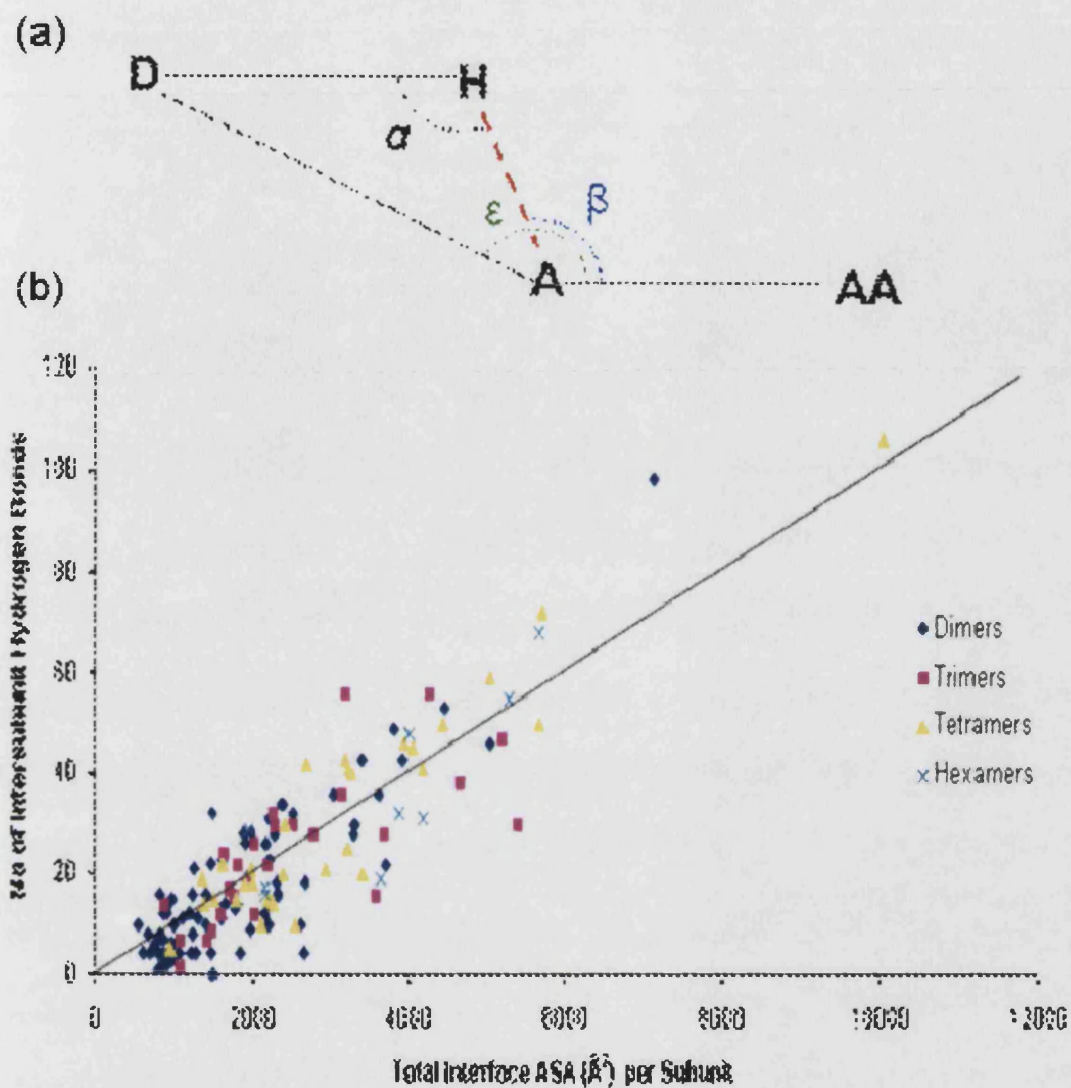


Figure 3.17: (a) A diagram illustrating the criteria used to determine acceptable hydrogen bond geometries. D denotes the hydrogen bond donor, H the hydrogen atom, A is the hydrogen bond acceptor, and AA is the atom bound to the hydrogen bond acceptor. The D-A distance should be less than 3.9\AA and the H-A distance should be less than 2.5\AA . The angles denoted α (DH-A), β (HA-AA), and ϵ (DA-AA) should all be $>90^\circ$. The actual hydrogen bond is shown as a dashed red line. (b) A plot of the number of inter-subunit hydrogen bonds for each protein subunit against its total interface ASA. The gradient of the line of best fit in the plot is 0.01. The correlation coefficient for the points in the plot is 0.88.

The actual numbers of hydrogen bonds across protein-protein interfaces is almost certainly higher than the data in figure 3.17(b) suggests. The mean numbers of inter-subunit hydrogen bonds per 100\AA^2 are tabulated in table 3.8.

	No of Hydrogen Bonds Per 100 Å ² Buried ASA	Min	Max	SD
Dimers	0.91	0.00	2.15	0.47
Trimers	0.97	0.19	1.75	0.38
Tetramers	0.97	0.39	1.56	0.30
Hexamers	0.95	0.52	1.49	0.30
All	0.94	0.00	2.15	0.41

Table 3.8: The mean numbers of inter-subunit hydrogen bonds for every 100Å² of buried ASA in the datasets of homo-dimers, trimers, tetramers, and hexamers. In each dataset there is ~1 inter-subunit hydrogen bond for every 100Å² of buried ASA.

Trimers, tetramers and hexamers form a similar number of inter-subunit hydrogen bonds (around one) for every 100Å² of buried ASA. Homo-dimers form slightly fewer hydrogen bonds per 100Å² of interface ASA than do the other classes of multimers, but the differences are still quite slight and easily accounted for given the differing sizes of the datasets.

	Type of Hydrogen Bond					
	MM	(%)	SS	(%)	SM	(%)
Dimers	4.4	24.3	6.6	36.6	7.0	39.2
Trimers	6.2	26.0	8.8	36.8	8.9	37.2
Tetramers	8.5	26.2	11.9	36.8	12.0	37.0
Hexamers	3.8	11.6	12.4	38.2	16.3	50.2
All	5.6	23.3	8.5	38.5	9.0	37.5

Table 3.9: Inter-subunit hydrogen bonds analysed by type. The average numbers and percentage frequencies of inter-subunit hydrogen bonds between main-chain groups (MM), side-chain groups (SS), and between main-chain and side-chain groups (SM) are given for each dataset of homo-complex.

Inter-subunit hydrogen bonds were further studied by analysing them on the basis of being between main chain groups (MM), side chains (SS), or between main chain and side chain groups (MS). The numbers and percentages of each of these three types of hydrogen bonds are given in table 3.9.

The percentages of MM, SS, and SM hydrogen bonds for all the homo-complexes in table 9 are very similar to the findings of Jones & Thornton, 1995, and Conte et al., 1999. As protein interfaces are primarily composed of interacting side-chains from

different subunits it is to be expected that the majority of inter-subunit hydrogen bonds involve amino acid side chains rather than main chain groups.

3.9 Salt Bridges

A salt bridge is formed by the electrostatic attraction between two oppositely charged groups. A salt bridge is taken to exist if any two oppositely charged groups are within 4Å of each other (Barlow & Thornton, 1983). As with hydrogen bonds, salt bridges between interacting subunits may potentially stabilise an otherwise weak interaction and additionally by their specific positioning help guide the interacting protein surfaces into the correct relative position and orientation. However in contrast to the results seen in the case of hydrogen bonds (presented in figure 3.17(b)), there is no correlation between the total interface ASA of a protein and the number of inter-subunit salt bridges in any of the datasets, as is shown in figure 3.18.

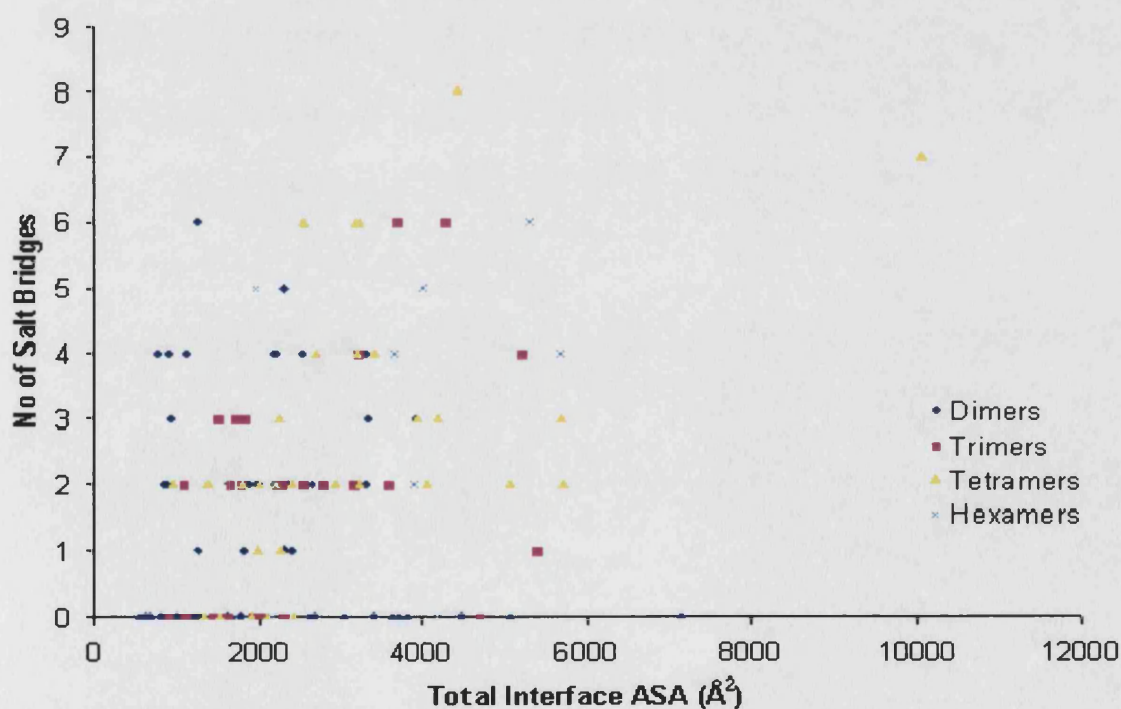


Figure 3.18: The number of inter-subunit salt bridges for each protein subunit plotted against its total interface ASA (Å²). There appears to be no relationship between the total interface ASA of a protein and the number of inter-subunit salt bridges.

These results are in agreement with work carried out by Xu et al., 1997 who found 623 salt bridges across 319 protein-protein interfaces, and average of 2 salt bridges per interface, concluding that “there appears to be no positive correlation between the size of the interface and the number of salt bridges across it”.

The number of inter-subunit salt bridges varies widely within each dataset, but is on average one for dimers, two for trimers, and three for both tetramers and hexamers. But these averages have no real meaning and it seems that the number of salt bridges a protein has is related primarily to the environment in which a protein exists and its function, rather than to the size or number of protein-protein interfaces that it may happen to have. It may be possible that salt bridges in general contribute more to the overall stability of the interface rather than providing the basis of any specificity since there are typically so few of them at a protein-protein interface. Salt bridges evidently do have a role in maintaining the structural integrity of proteins that function in extreme or adverse environments such as thermophiles (Kumar et al., 2000).

An example of a hyperthermophilic protein enriched in inter-subunit salt bridges is aldehyde ferredoxin oxidoreductase (PDB code 1aor). Aldehyde ferredoxin oxidoreductase is a dimeric protein with six inter-subunit salt bridges and an interface ASA of 1232\AA^2 (Chan et al., 1995). Other thermophilic proteins in the datasets that are used here such as adenylate kinase (1nks) from *Sulfolobus acidocaldarius* expressed in *Escherichia coli* are similarly enriched with inter-subunit salt bridges and hydrogen bonds. Salt bridges may also help to stabilise the active sites (if any) and protein-protein interfaces of this class of protein although it should be noted that other interactions such as hydrogen bonding would probably also play a role in stabilising the multimer.

3.10 Secondary Structure

Like amino acid content and hydrophobicity, secondary structure is another surface characteristic that could potentially be used to distinguish interface regions of isolated protomers from the rest of their surfaces. The secondary structure motifs considered here are those defined by Kabsch & Sander (1983), namely helix, strand, turn and coil. A table of the mean secondary structure content of each of the datasets in terms of these four classes of secondary structure is shown in table 3.10. As can be seen from table 3.10 our results indicate that the secondary structure content of a protein-protein interface is not really any different from that of the protein exterior when taken as a whole, in agreement with previous work (Argos, 1988, Jones & Thornton, 1995).

Dataset	Mean (%) Frequency Secondary Structure States			
	Helix	Strand	Turn	Coil
Dimers				
Interior	43.09	34.23	9.30	13.38
Interface	38.21	17.94	21.73	22.12
Exterior	37.00	16.50	23.82	22.69
Trimers				
Interior	31.0	33.8	13.2	21.9
Interface	34.5	26.1	20.8	18.8
Exterior	37.8	23.2	20.0	19.0
Tetramers				
Interior	42.53	34.17	8.60	14.70
Interface	33.40	21.03	25.21	20.36
Exterior	36.07	17.92	24.90	21.11
Hexamers				
Interior	45.25	34.42	7.82	12.52
Interface	44.53	12.65	25.30	17.52
Exterior	40.92	12.46	25.76	20.86

Table 3.10: The average percentages of interior, exterior, and interface residues that are helix, strand, turn, and coil for each dataset of homo-complex.

Helices are the most prevalent secondary structure state to be found in any part of a protein (including the protein-protein interface). Although some interfaces contain predominantly only one motif most are mixed.

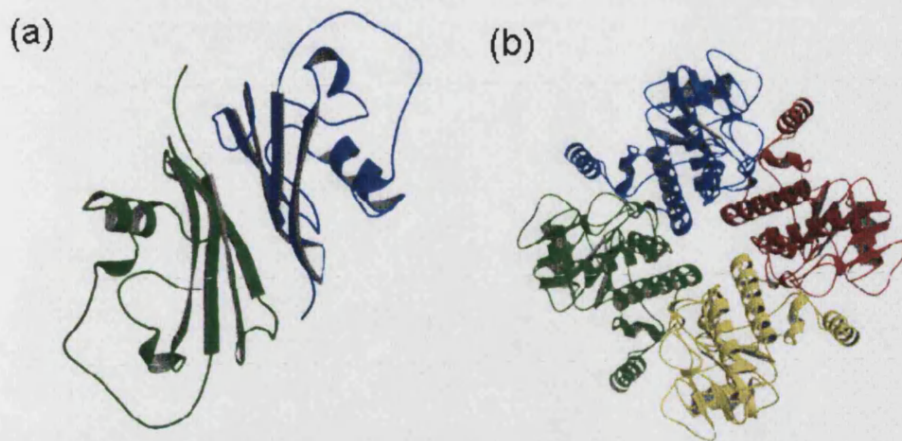


Figure 3.19: (a) The subtilisin inhibitor (3ssi, Mitsui et al., 1979) and the L-fucose 1-phosphate aldolase tetramer (2fua, Dreyer & Schultz, 1996). Proteins with interfaces composed mainly of beta sheets and alpha helices respectively.

An example of a dimeric protein whose interface contains mainly beta strands is the serine protease inhibitor from *Streptomyces albogriseolus* (3ssi) shown in figure 3.19(a). In contrast l-fucose-1-phosphate aldolase from *Escherichia coli* (2fua) in figure 3.19(b) is a tetramer which has protein-protein interfaces that consist mainly of alpha helices. The secondary structure content of the protein-protein interfaces of proteins in each dataset of proteins has been classified as being mainly alpha, mainly beta, alpha/beta, or coil.

Classification	Definition
Alpha	Alpha > 20% and Beta <20%
Beta	Alpha < 20% and Beta >20%
Alpha/Beta	Alpha > 20% and Beta >20%
Coil	Alpha ≤ 20% and Beta ≤20%

Table 3.11: The scheme use to classify protein-protein interfaces as having a mainly alpha, beta, or alpha/beta secondary structure content.

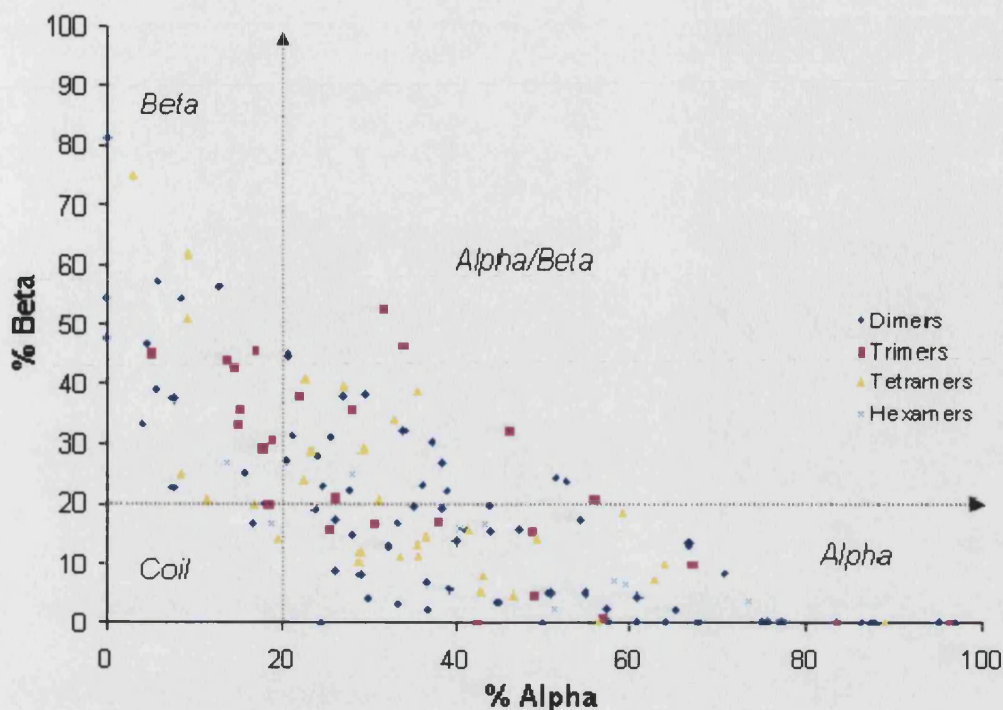


Figure 3.20: Secondary structure classification of protein-protein interface residues. The 20% cut-offs used to mark the boundaries between the alpha, alpha/beta and coil classifications are marked with dotted lines

The criteria employed for this classification scheme were identical to those used previously (Jones & Thornton, 1996) and are shown in table 3.11. A plot of the percentage alpha against percentage beta secondary structure content for every protein within each dataset is given in figure 3.20. The plot reveals that the vast majority of the 142 proteins have protein interfaces that contain alpha, beta, or both alpha and beta secondary structure elements. Only five proteins have protein interfaces that are classified as being coil, with few recognisable secondary structure elements. It follows from this that proteins usually interact with each other through 'ordered' interfaces, that is using interfaces comprising of a number of regular secondary structure motifs held together in reasonably well defined three dimensional arrangements. There are of course famous examples of proteins that bind using loops, an example being the serine protease inhibitors. But even in these proteins the loops are held in quite rigid conformations by means of a complex network of hydrogen bonding. Although there are more protein-protein interfaces which fall into the mainly alpha classification than any other there are still a significant number of protein-protein interfaces that fall into both the alpha/beta and beta classifications. It appears

therefore that a wide variety of secondary structural motifs can be packed efficiently at protein-protein interfaces, possibly (as discussed in sections 3.8 and 3.9 above) with the help of hydrogen bonds and salt bridges where the fit would otherwise be poor.

3.11 *Packing at Subunit Interfaces*

The association of a number of subunits into a stable complex requires that the surfaces at which they bind to each other are reasonably complementary in both physical shape and electrostatics. A number of ways in which the physical complementarity of interacting surfaces can be quantified have been suggested. Lawrence et al., 1993 uses a shape correlation statistic S_c to quantify complementarity (see section 1.6.2 in chapter 1). This and other studies have shown that interacting proteins generally exhibit good surface complementarity. If two interacting surfaces are complementary to one another in shape it follows that the residues at a protein-protein interfaces are quite closely packed. Therefore by quantifying the packing of residues at a protein-protein interface, an indirect measure of shape complementarity is also obtained.

This complementarity acts as an effective filter as to which proteins may associate, since the surfaces of interacting proteins must be sufficiently complementary in shape so as to allow the formation of hydrogen bonds and other interactions that drive the formation of a stable multimer. It should be noted however that methodologies which might attempt to detect protein-protein binding sites by looking for regions of geometric complementarity between protomers would be computationally extremely costly and are unlikely to lead to a predictive algorithm for interface detection. This is especially true given that the complementarity between interacting proteins is often hidden prior to the formation of the multimer, due to the interacting proteins undergoing conformational changes during binding or by the existence of

extensive networks of ordered water molecules between interacting subunits which mediate inter-subunit hydrogen bonds and thus improve surface complementarity.

In order to quantify how closely packed the protein-protein interface is the Voronoi polyhedra method is used (Richards, 1974). In this technique vectors are first constructed from the atom (or group) being considered to all neighbouring atoms (or groups). Along each of these vectors at a distance related to the Van der Waals radius of the atoms a perpendicular plane is constructed, the Voronoi polyhedron being formed by the intersection of these planes. The volume of the polyhedron is then inversely proportional to the packing efficiency of the central atom (Chothia 1975). Tsai et al., 1999, have calculated a set of standard Van der Waals radii and Voronoi volumes for every atomic group that is found in each of the twenty amino acids using large databases of organic compounds and high resolution protein crystal structures. By taking the ratio V/V_0 of the standard residue volume V_0 and calculated Voronoi volume V , it is possible to quantify how closely packed a given atomic group is. Using this method the V/V_0 packing ratios for residues in the protein-protein interfaces of each of the 142 homo-complexes has been calculated.

The Voronoi volumes and packing ratios were calculated using code that is available from: <http://bioinfo.mbb.yale.edu/geometry>. The simplified Richards B method was used in the calculation of the Voronoi volumes. The Voronoi volume of an atom or group of atoms is particularly sensitive to the distribution of atoms around the point being considered since it is these points that define the planes that intersect to form the Voronoi polyhedron itself. For this reason Voronoi volumes are only calculated for residues that are completely buried within the protein-protein interface. As a consequence only a rather small number of the residues that make up a protein-protein interface are considered when evaluating the packing efficiency of the interface as a whole. This is the most significant drawback of using Voronoi polyhedra.

The average packing ratio of all of the protein-protein interfaces for each protein was calculated using the coordinates of any water molecules and other hetero groups that may be present in the PDB entry for the protein concerned. This ensures that the

contribution to the overall packing density at subunit interfaces that the solvent makes is taken into account. This is important as water molecules at protein interfaces are often highly conserved (Janin, 1999) and can often play a considerable role in improving complementarity between the interacting surfaces at interfaces.

	% Atoms Buried	Mean Packing Ratio(V/V_0)	Min	Max	SD
Dimers	27.4	1.04	0.81	1.62	0.09
Trimers	25.8	1.03	1.00	1.11	0.03
Tetramers	25.6	1.02	0.70	1.14	0.09
Hexamers	31.0	1.02	0.96	1.07	0.04
All	27.8	1.03	0.70	1.62	0.08

Table 3.12: Mean packing ratios (V/V_0) for the datasets of homo-dimers, trimers, tetramers, and hexamers. For all datasets the mean V/V_0 ratio is close to one indicating that residues at protein-protein interfaces are almost as closely packed as residues in the protein interior.

The mean packing ratios for each dataset are shown in table 3.12, with a ratio < 1 indicating that the interface is more closely packed than the interior of a protein and a ratio > 1 that it is less so. These results suggest that the protein-protein interface is almost as closely packed (on average only $\sim 3\%$ less) as the protein interior. The mean packing ratio of 75 protein complexes (both homo and hetero) was similarly obtained by Conte & Chothia (1999) to be $1.01(\pm 0.06)$. Taking into account the standard deviation values for the mean packing ratios shown in table 3.12, these results also show that there is no statistically significant difference between how closely packed the protein-protein interfaces are for the different classes of multimer. A histogram of the V/V_0 packing ratios for the protein-protein interfaces of the homo-complexes is shown in figure 3.21. From figure 3.21 the distribution of packing ratios is narrow and is centred on 1.03 as suggested by the averages in table 12. But the fact that there is any distribution at all shows that interface residues are not all equally well packed.

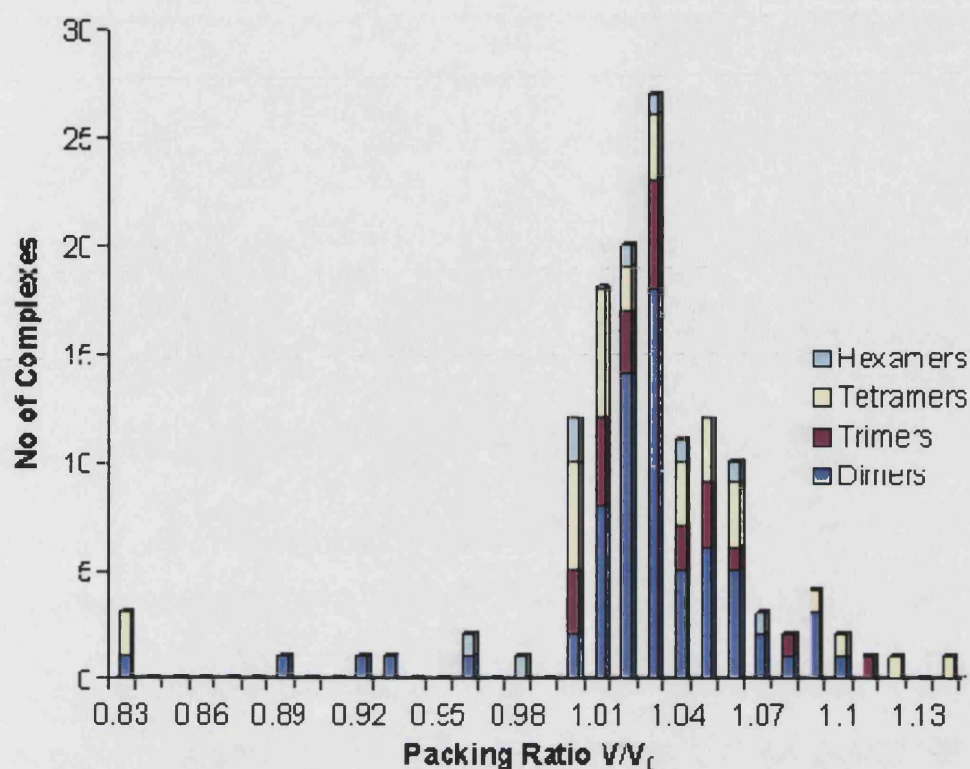


Figure 3.21: A histogram of the V/V_0 packing ratios for all ‘interface’ residues from the datasets of homo-complexes. The histogram shows that residues at protein-protein interfaces are almost as closely packed as those in the protein interior.

3.12 Planarity

The planarity of each protein-protein interface within the homo-complexes was calculated by fitting a least square plane through all interface atoms for each protein-protein interface. This was done using a program developed by Roman Laskowski. The root mean square deviation of atoms for each interface individually was then summed and averaged to give mean planarity values. A plot of each interface’s ASA against its planarity is given in figure 3.22. There is a good linear correlation between the physical size of a protein-protein interface and how planar or ‘flat’ or ‘interlocked’ it is, with small protein interfaces being comparatively flat with respect to larger ones. The Pearson correlation coefficient for all interfaces in this plot is 0.84 indicating the strength of this relationship. It should be strongly noted that the linear

variation of planarity with interface size (ASA) is a property of any well-behaved surface. For any well-behaved surface, the larger the surface being sampled the less planar it will be. To determine if protein-protein interfaces are genuinely more (or less) planar than the remainder of a protein surface the planarity of non-interface regions of protein-exterior would have to be compared with the planarity of protein-protein interfaces of roughly the same size. This is a matter for further study and is not the primary purpose of this section. The purpose of this section is twofold. The first aim is to establish the exact relationship between the planarity of the protein-protein interfaces and their physical size for each class of homo-complex (and in chapter 4, obligate and non-obligate hetero-multimers). The second aim is to *compare* the relationships between interface size and planarity for each class of homo-complex.

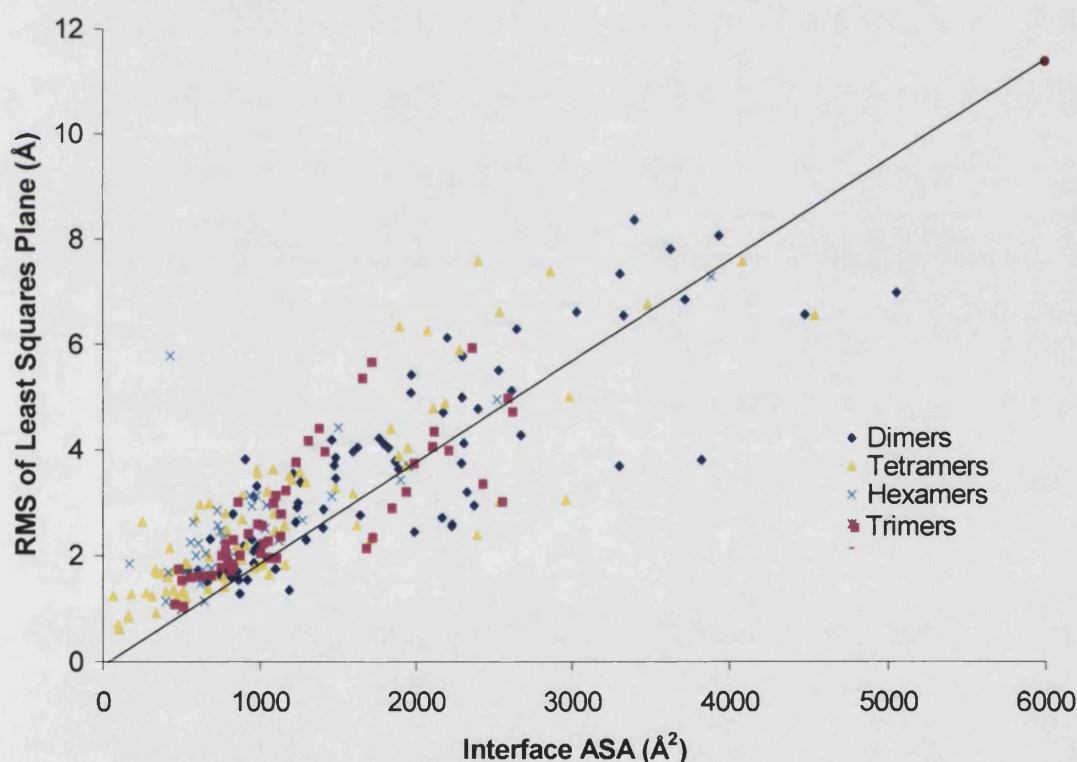


Figure 3.22: The planarity of each protein-protein interface plotted together with its size in \AA^2 . There is a linear relationship between the size of a protein interface and how planar it is. The gradient of the line of best fit in the plot is 0.0019.

In common with the work of Jones, 1996, it seems that protein-protein interfaces are in general quite planar, with 78% of all interfaces in the four datasets having a

planarity less than 4Å, suggesting that even quite large interfaces are reasonably flat. The planarity of a protein's interfaces has implications for its biological function and is further discussed in chapter 4. Separate values of the averaged planarities of the protein-protein interfaces for each class of multimer are given in table 3.13, ranging from 2.81Å for tetramers to 3.76Å for dimers. To give an idea of what a planar interface actually looks like, the sliding clamp protein in figure 3.2(a) has protein-protein interfaces that each has a planarity of 1.06Å. In contrast the protein-protein interfaces of catalase in figure 3.5(d) are some of the least planar to of all the 142 homo-complexes. The least planar interface in the catalase tetramer has a planarity of 7.57Å. Looking at figure 3.22 and table 3.13 it is apparent that the size of a protein-protein interface is the principle determinant of how planar it is. Another way of putting this is that the planarity of a protein-protein interface is a function of how large it is rather than what class of multimer it is to be found in.

	Mean Interface Size (Å ²)	Mean Planarity (Å)	Correlation coefficient
Dimers	1890	3.76	0.83
Trimers	1280	2.88	0.79
Tetramers	1160	2.81	0.85
Hexamers	970	2.67	0.76
All Datasets	1410	3.13	0.84

Table 3.13: Mean planarities of the protein-protein interfaces of homo-dimers, trimers, tetramers, and hexamers.

An additional factor that that may affect the planarity of a protein-protein interface is if the interface is isologous or heterologous. (An isologous interface is one that is formed by two sets of equivalent residues. In contrast, heterologous interfaces are formed by two different sets of interacting residues as described in chapter 1). Most interfaces within the 142 homo-complexes are isologous. For instance all the protein-protein interfaces of the 76 homo-dimers are isologous. Table 3.14 shows the mean interface size and planarity of the isologous and heterologous interfaces. The correlation between interface size and planarity is much weaker for heterologous interfaces than for isologous interfaces. This weaker correlation could be a reflection of the different ways in which heterologous interfaces may have evolved compared with isologous interfaces.

	Mean Interface Size (Å ²)	Mean Planarity (Å)	Correlation coefficient
Isologous	1490	3.22	0.86
Heterologous	1170	2.82	0.69

Table 3.14: Mean planarities of all the isologous and heterologous protein-protein interfaces within the datasets of obligate homo-complexes.

For example two ways that heterologous interface can arise is via domain swapping or through proteolytic cleavage of a single polypeptide chain. Basically heterologous interfaces are formed from two separate surfaces evolving to recognise each other and can be formed in a number of different ways. Isologous interfaces are formed from two identical proteins binding to each other using the same set of residues. Identical proteins already have largely inbuilt shape complementarity with each other. Only rather minor mutations in the gene coding for the protein would be needed to produce a protein-protein interface.

3.13 Protrusion of Residues at Protein-Protein Interfaces

Although it has been established in the preceding section that protein-protein interfaces are generally quite flat, this does not preclude there also being a number of residues that protrude from either side of a protein-protein interface interlocking with each other, anchoring the proteins together in a complex.

The degree to which the residues at protein-protein interfaces protrude from each subunit has been assessed by calculating the mean relative accessible surface area (rASA) of the twenty amino acids. As a means of comparison the rASAs of exterior residues were also calculated. The ratio between the rASA of an amino acid at the interface and on the protein exterior than gives an idea about how ‘protruding’ is the conformation taken up by a particular amino acid at a protein-protein interface. The results of these calculations are presented in table 3.15. Certain amino acids in interface regions, in particular hydrophobic and/or aromatic residues such as isoleucine, phenylalanine and tryptophan (in which the results of section 3.5 showed protein-protein interfaces to be relatively enriched compared to the protomer surface

as a whole) clearly have a significantly greater rASA (thus being more protrusive) than those generally found at the surface.

Amino Acid	Mean Relative Accessible Surface Area (%)		I/E
	Exterior (E)	Interface (I)	
ARG	40.95	48.84	1.19
LYS	50.25	55.44	1.10
ASP	48.72	54.62	1.12
GLU	49.29	51.81	1.05
ASN	45.06	49.39	1.10
CYS	25.57	37.38	1.46
GLN	45.74	50.76	1.11
HIS	36.02	43.06	1.20
SER	42.55	46.36	1.09
THR	38.74	44.32	1.14
TRP	27.13	37.85	1.40
TYR	31.1	41.3	1.33
ALA	39.41	44.5	1.13
GLY	44.83	50.31	1.12
ILE	28.01	38.49	1.37
LEU	30.02	40.79	1.36
MET	35.43	43.13	1.22
PHE	29.3	40.96	1.40
PRO	46.96	53.03	1.13
VAL	31.16	40.37	1.30

Table 3.15: The relative accessible surface areas of residues at the protein-protein interfaces (I) and exteriors (E) of all 142 homo-complexes. The ratio of I/E then provides a measure as to how ‘protruding’ a residue is at a protein-protein interface.

The reason why aromatic residues protrude at the interface in the way suggested by the data in table 3.15 is in part related to the way in which delocalised rings of π electrons interact with each other. The rings of aromatic amino acids do not generally stack directly on top of each other due to electrostatic repulsion between the delocalised π electrons. Because of this pairs of aromatic amino acids are most often observed to adopt edge to face or offset stacked orientations (McGaughey et al., 1998). Both of these preferred orientations require that the side chains of aromatic amino acids extend significantly into the space between interacting subunits. This is

best illustrated by figure 3.23(a) which shows the interface between abalone sperm lysin (1lyn).

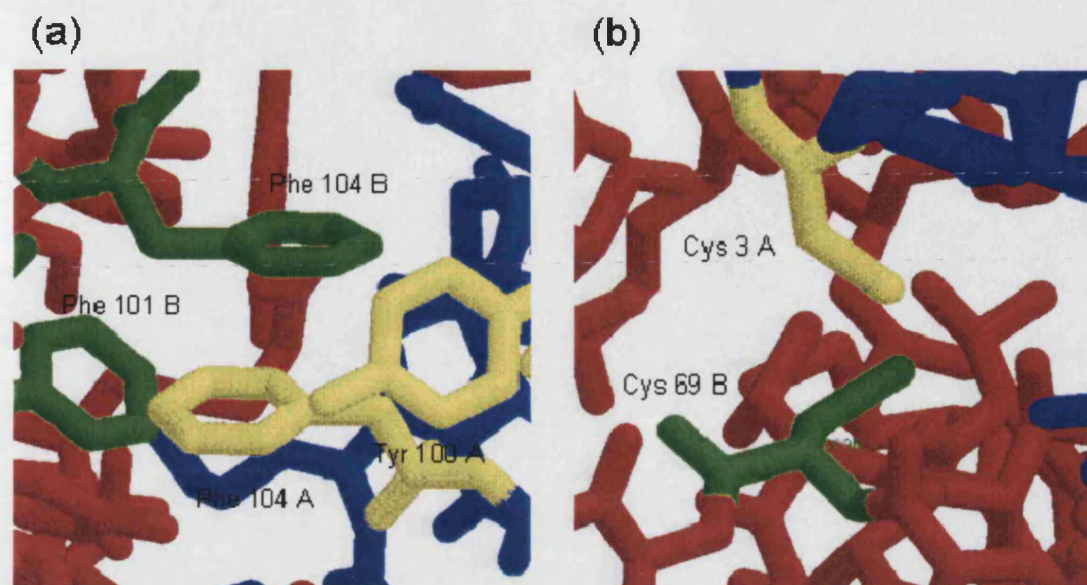


Figure 3.23: The dimer interface of (a) red abalone sperm lysin (1lyn, Kresge et al 2000) and (b) oxidized uteroglobin (Morize et al., 1987). In both (a) and (b) selected residues from both subunits are coloured in yellow and green.

From figure 3.23(a) the aromatic rings of phenylalanine residues from both subunits adopt an offset stacked packing arrangement. The data in table 3.15 also shows that the side chains of cysteine residues protrude strongly at protein-protein interfaces. There is a simple explanation for this. The few cysteine residues that are at protein-protein interfaces almost exclusively form disulphide bridges. A diagram of the dimer interface of oxidized uteroglobin (1utg) shows two cysteine residues from either side of the interface each extend quite deeply into the space between subunits in order to form the disulphide bridge between them (see figure 3.23(b)).

3.14 Flexibility of Interface Residues

In order to investigate the flexibility of interface residues, atomic temperature factors for the interior, exterior, and interface subsets for each protein in each dataset were analysed. The atomic temperature factor B is a measure of how flexible a given residue (or group of atoms) are in the crystal, and is proportional to the mean square

displacement of the given residue or group of atoms from its mean position. The B value of an atom is given by equation 5. In this expression x is difference in the position of the atom from its mean position.

$$B = 8\pi^2 \langle x^2 \rangle \quad (5)$$

A histogram of the B-values for all the residues that comprise the interior, exterior, and interface regions for all 142 homo-complexes together is shown in figure 3.24.

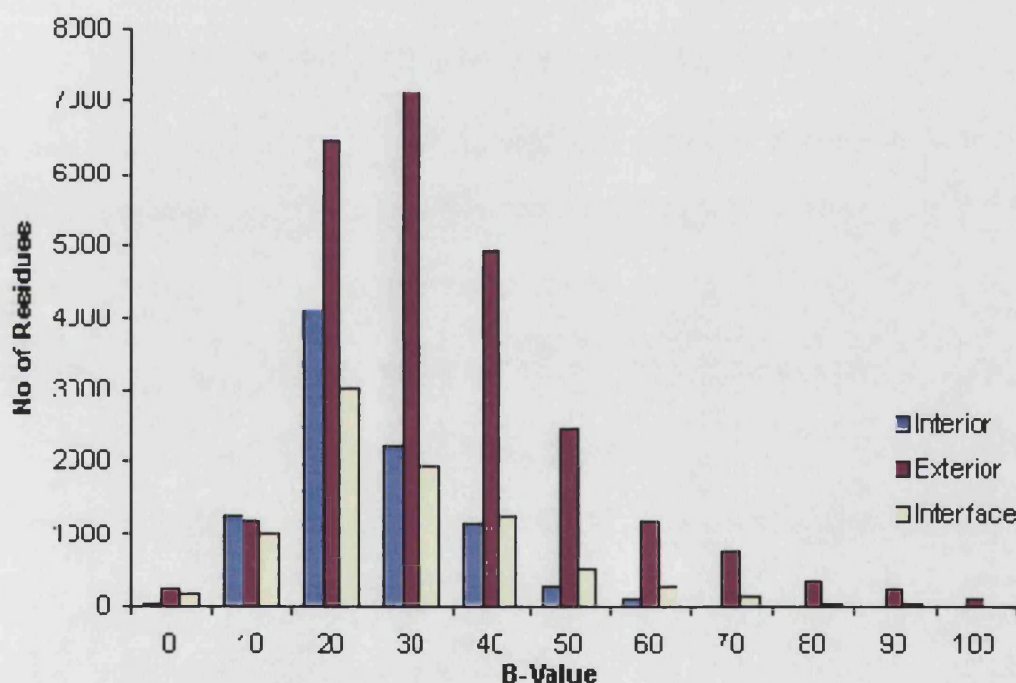


Figure 3.24: Histogram of atomic B-values for the interiors, exteriors, and the interface regions of all 142 homo-complexes.

The histogram shows that the residues that make up protein-protein interfaces, though on average more flexible those found in the interior of a protein, are in fact less flexible than those found on the surface of a protein when taken as a whole. The trend for interface residues to be less flexible than other surface residues is very strong with only 9 out of the total 142 protein complexes studied here not conforming to this trend. That interface residues appear to be less flexible than the protein exterior is

understandable. As discussed in the previous section residues from both sides of a protein-protein interface interdigitate with each other to form an interface stable enough to hold two proteins together in a stable complex. The result is that interface residues are usually less mobile than those on the protein exterior. On the other hand protein-protein (and indeed domain-domain) interfaces need some degree of flexibility. This could be so that allosteric conformational changes can take place. Many enzymes have active sites that are at or near subunit interfaces. Conformational changes at interfaces could therefore be required to assist in substrate binding and catalysis. One of the 9 proteins in which the interface residues are more flexible than the protein exterior is carboxylesterase from *Pseudomonas florescens* (1auo).

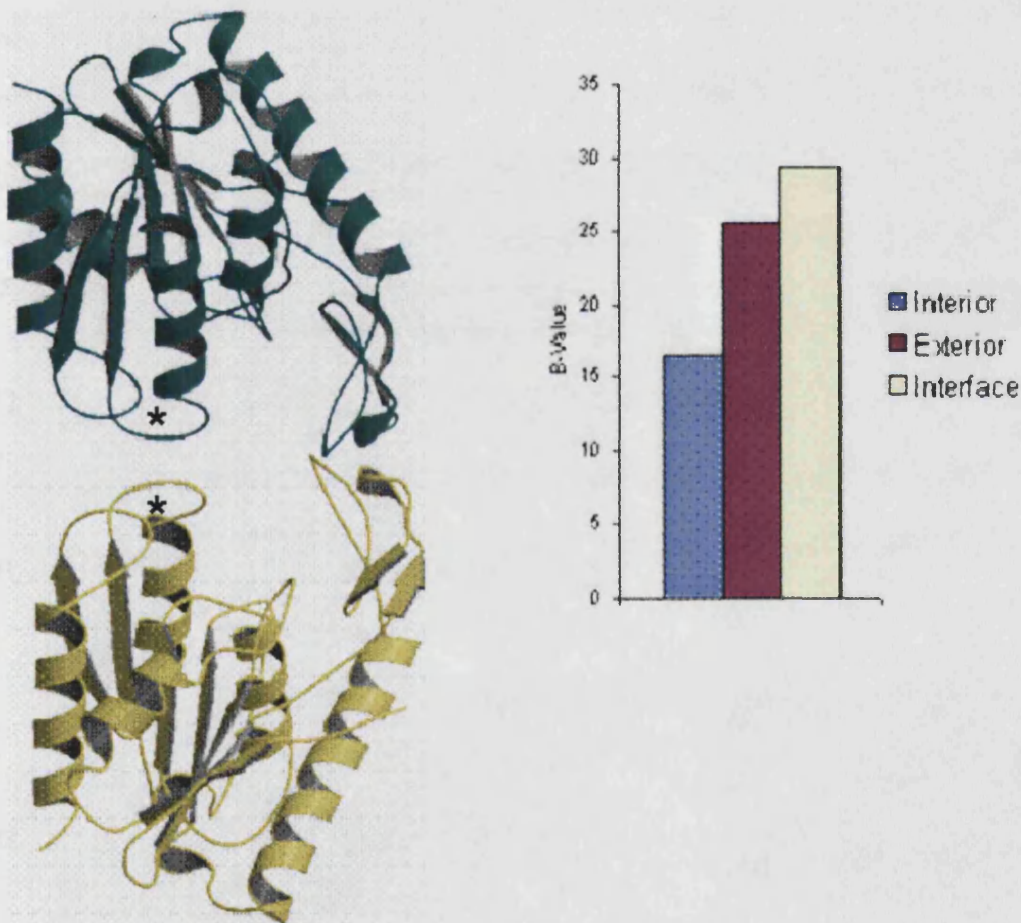


Figure 3.25: The carboxylesterase dimer (1auo, Kim et al., 1997). The two active sites of the enzyme are marked *. Averaged B-values of the interior, exterior, and interface regions of 1auo reveal that residues at the dimer interface are particularly flexible. This may help to explain the enzymes broad range of substrate specificity.

From the averaged B-values of 1 auo the residues at the dimer interface appear to be more flexible than those found on the protein exterior. The two active sites of the enzyme are at the dimer interface and are indicated by stars in figure 3.25. The carboxylesterase exhibits quite a broad range of substrate specificity (Kim et al., 1997). The flexibility of the residues at the interface may provide the active sites of the enzyme with the necessary capacity for structural deformity that is needed to bind a range of different substrates.

In the combined dataset of 142 homo-complexes there is no correlation between the size of a protein-protein interface and its average B value (data not shown). The rASA and B-values of residues at the protein interface taken together point to a picture in which interface residues are appreciably less flexible than the exterior and in some cases considerably more protrusive than those typically found on the exterior of a protein.

3.15 Conclusions

The principle focus of this chapter has been the physical and chemical nature of the protein-protein interfaces in homo-dimers, trimers, tetramers and hexamers. Overall the conclusions that can be drawn from the work presented in this chapter are broadly in line with the work carried out by other authors.

With regard to protein-protein interfaces in the general sense there seems to be an identifiable set of characteristics that sets these regions apart from non binding regions on a proteins exterior. The characteristic features of the protein-protein interfaces found within the 142 obligate homo-complexes are:

- Relatively hydrophobic. The hydrophobicity of interface regions using the Fauchere and Pliska scale is intermediate between that of the protein interior and exterior.

- Enriched in particular in bulky hydrophobic and/or aromatic residues such as tyrosine and phenylalanine, though also having a number of charged and polar and residues within them.
- Overall some 45% of the residues in protein-protein interfaces are hydrophobic, 25% are charged, and 31% are polar. The amino acid composition of the protein-protein interfaces is more closely related to the protein exterior than its interior.
- Less flexible than exterior residues as measured by atomic temperature (B-value) factors.
- Relatively planar, with an average RMS of least squares plane of less than 4Å for each dataset, though also having protrusive regions that may serve to enhance geometric complementarity, which as with the selectively positioned hydrogen bonds and salt bridges may play a crucial role in correctly orienting the interacting protomers during complexation.
- Conserved at the sequence level. Valdar et al., 2001 analysed the protein-protein interfaces of six protein families from the homo-dimer dataset and concluded that residues at protein-protein interfaces are more highly conserved than the protein exterior.

It is important to remember that the set of characteristics given above applies to obligate homo-complexes. The nature of the protein-protein interfaces of obligate hetero-complexes and non-obligate protein complexes is investigated in chapter 4. Protein-protein interfaces also appear in general to be extremely well packed, almost as closely packed (within 5%) as the protein interior; this indicates that there indeed must be a high degree of surface complementarity, even though this is hard to see from inspection of isolated protomers. One conspicuous feature of the homo-complexes is symmetry. All of the homo-complexes exhibit various kinds of symmetry (rotational, translational, or both). In tetramers symmetrical arrangements in which the protein-protein interfaces are all isologous are strongly favoured.

Whether or not this means that isologous interfaces represent a more stable binding arrangement than heterologous interfaces is a matter for further investigation.

One would ideally hope to characterise protein-protein interfaces to such a degree that these regions could be distinguished from the rest of the surface of an isolated protomer. Additionally it would be useful to be able to determine from the type and distribution of any predicted interface regions the likely oligomerisation state of the multimer. From the results presented this does not seem to be possible at present since there are no clear features that allow one to distinguish a dimer from a trimer interface, for example. The protein-protein interfaces found within dimers, trimers, tetramers, and hexamers all normally contain similar fractions of charged, polar, and hydrophobic amino acids. In addition the amino acid compositions of the protein-protein interfaces found within the four classes of multimer differ as shown in figure 3.9 but these differences do not follow any discernable pattern and are quite probably artefacts of the datasets of proteins used here. The major difference in the protein-protein interfaces found in the different multimers is size. As can be seen in tables 3.1 and 3.2 dimers have a protein-protein interface that is on average 1890\AA^2 in size while the figure for trimers is $\sim 1280\text{\AA}^2$.

One problem is that our current methods of analysis are based in the main on taking averages of physicochemical and geometric properties over surface regions; since proteins can recognize each other in vivo, in principle we should be able to predict interfaces with a reasonable degree of reliability, but at the current scale of analysis, looking at bulk surface properties may simply be too coarse. In addition there is the complicating fact that there is sometimes a considerable amount of surface adjustment during the formation of a multimer, so that some aspects of surface complementarity may be effectively hidden when considering the free protomers.

Work has been carried out to see to what degree interface regions can be predicted from locally-averaged values of properties such as hydrophobicity and planarity, using neural network techniques. This work builds on the non-adaptive 'patch analysis' methods of Jones & Thornton, 1997, and is presented in chapter 5. As well as using physicochemical and structural information originally used in the patch analysis method the utility of conservation scores is considered (Valdar & Thornton,

2001). Conservation scores give a quantitative measure as to how well conserved any given residue is from a sequence point of view and implicitly reflects the structural and functional importance of a given residue or residue group.

Chapter 4

Hetero-Complexes

4.1 Introduction

The tremendous variety of interactions that proteins are involved in gives rise to a many different categories of protein-complex. In this chapter proteins representing some of the more prevalent categories of hetero protein complex are characterised. This is done for the most part by looking at the size and chemical composition of the protein-protein interfaces of these hetero protein complexes in a similar way to the homo-complexes studied in chapter 3. In the previous chapter complexes made up of identical protein subunits permanently bound to each other (homo-complexes) have been characterised. The main subjects of this chapter are permanent and transient complexes of non-identical protein subunits (hetero-complexes). The different datasets of hetero-complexes studied in this chapter are set out below.

(a) Permanent or obligate hetero-complexes

10 hetero-dimers, 7 hetero-tetramers, and 3 hetero-hexamers

(b) Transitory or non-obligate hetero-complexes

20 enzyme-inhibitors, 15 antibody-antigens, and 10 signalling complexes

The contents of the above datasets can be found in section 2.4 in chapter 2. In this chapter a comparison is made between the proteins that form homo and hetero-complexes, and between obligate and non-obligate protein complexes. One aim of this is to construct a generalised set of characteristics describing the protein-protein interfaces within each of the datasets. The results of this should provide a useful reference for researchers looking to investigate the underlying mechanisms by which

proteins associate to form some of the different categories of protein complexes. Although this has been done before, since 1999 the number of protein structures in the PDB has doubled (see figure 1.3). The work presented in this chapter therefore provides a necessary update to previous work that has been carried out in this area (Conte, 1999, Jones & Thornton, 1996, and Argos, 1988). A second aim is to use these characteristics to see if it is possible to discriminate between the different classes of proteins studied here using structural data alone. This is done using a neural network in chapter 5.

Tables 4.1-4.4 give a summary of the structural characteristics of each dataset of proteins used in this thesis with the sole exception of the dataset of monomers. The data contained in tables 4.1-4.4 is referred to throughout the chapter.

Key for Tables 4.1-4.4

*^a The Δ ASA is the total accessible surface area of a subunit buried in protein-protein when part of a complex interfaces calculated using NACCESS (see section 3.2).

*^b The planarity of protein-protein interfaces is calculated using an algorithm from the SURFNET suite of programs (Laskowski, 1995).

*^c The residues that are classified as hydrophobic, polar, or charged are shown in table 3.3 in chapter 3.

*^d The number of hydrogen bonds for every 100\AA^2 of buried ASA was calculated using HBPLUS (McDonald & Thornton, 1994).

*^e The mean hydrophobicity of the interior, exterior, and interface regions was calculated using the Fauchere & Pliska hydrophobicity scale, 1983 (see section 3.7 in chapter 3).

Characteristic: Obligate Homo-Datasets					
	Dimers	Trimers	Tetramers	Hexamers	All
Size of Dataset	76	26	31	9	142
Mean Δ ASA (\AA^2) ^{*a}					
Min	540	880	950	1980	540
Max	7150	5390	10040	5740	10040
Mean	1890	2520	3090	3650	2380
SD	1170	1270	1790	1360	1470
% ASA buried per protomer					
Min	4.2	6.2	9.9	17.1	4.2
Max	31.3	40.3	40.1	37.1	40.3
Mean	15.9	22.5	22.5	25.9	19.2
SD	6.9	9.5	7.0	6.2	8.2
Planarity of all interfaces within multimer (\AA) ^{*b}					
Min	1.3	1.0	0.6	1.0	0
Max	8.4	5.9	7.6	7.3	8.4
Mean	3.8	2.9	2.8	2.7	3.1
SD	1.9	1.3	1.7	1.3	1.7
Interface averaged amino acid composition (%) ^{*c}					
Charged	23.9	24.6	25.0	25.3	24.6
Polar	30.1	30.8	32.3	25.9	30.6
Hydrophobic	45.8	44.6	41.8	48.8	44.8
No of hydrogen bonds per 100 (\AA^2) Δ ASA ^{*d}					
Min	0.00	0.19	0.39	0.52	0
Max	2.15	1.75	1.56	1.49	2.15
Mean	0.91	0.97	0.97	0.95	0.94
SD	0.47	0.38	0.30	0.30	0.41
Average Hydrophobicity (Fauchere & Pliska scale) ^{*e}					
Interior	0.97	0.92	0.94	0.88	0.95
Interface	0.37	0.34	0.29	0.34	0.34
Exterior	0.21	0.22	0.18	0.18	0.21
Average Hydrophobicity of all Protein-Protein Interfaces					
Buried Zone	0.81	0.75	0.67	0.69	0.76
Partially Buried Zone	0.22	0.19	0.16	0.19	0.20
Protomer Weight (Da)					
Min	6950	4500	12000	16800	4500
Max	96900	61600	68800	53800	96900
Mean	29000	26200	33000	35100	29700
SD	19150	12700	14530	13800	16900

Table 4.1: Summary of the structural characteristics of the datasets of obligate homo-complexes.

Characteristic: Obligate Hetero-Datasets				
	Dimers	Tetramers	Hexamers	All
Size of Dataset	10	7	3	20
Mean Δ ASA (\AA^2) ^{*a}				
Min	1250	1190	1650	1190
Max	7170	6830	11150	11150
Mean	3310	3070	5230	3730
SD	1690	1680	2790	2220
% ASA buried per protomer				
Min	14.2	7.9	15.0	7.9
Max	49.0	50.3	48.0	50.3
Mean	24.5	25.1	39.2	28.8
SD	9.1	12.3	9.2	12.2
Planarity of all interfaces within multimer (\AA) ^{*b}				
Min	2.4	0.4	0.6	0.4
Max	10.8	8.8	6.9	10.8
Mean	5.4	3.0	3.0	3.3
SD	2.6	1.9	1.5	2.0
Interface averaged amino acid composition (%) ^{*c}				
Charged	27.8	28.2	28.3	28.1
Polar	31.7	25.1	29.7	28.3
Hydrophobic	40.5	46.8	42.1	43.6
No of hydrogen bonds per 100 (\AA^2) Δ ASA ^{*d}				
Min	0.87	0.53	0.54	0.53
Max	1.58	1.33	1.48	1.58
Mean	1.14	0.97	1.05	1.05
SD	0.19	0.20	0.31	0.24
Average Hydrophobicity (Fauchere & Pliska scale) ^{*e}				
Interior	0.77	0.88	0.80	0.83
Interface	0.40	0.34	0.29	0.34
Exterior	0.26	0.25	0.20	0.24
Average Hydrophobicity of all Protein-Protein Interfaces				
Buried Zone	0.73	0.68	0.54	0.66
Partially Buried Zone	0.25	0.21	0.15	0.21
Protomer Weight (Da)				
Min	5240	14600	10600	5240
Max	80000	85500	60000	85500
Mean	34000	31000	31000	32000
SD	22600	20000	18000	20000

Table 4.2: Summary of the structural characteristics of the datasets of obligate hetero-complexes.

Characteristic: Non Obligate Hetero-Datasets					
Size of Dataset	Enzymes 16*	Inhibitors 20	Antibodies 15	Antigens 15	Signaling 10
Mean Δ ASA (\AA^2) * ^a					
Min	600	670	640	630	290
Max	1560	1760	1280	1330	2680
Mean	970	1040	870	880	1030
SD	300	320	200	200	600
% ASA buried per protomer					
Min	4.1	7.1	3.2	2.7	4.0
Max	19.1	50.8	6.7	16.8	36.9
Mean	9.3	22.3	4.5	9.6	10.9
SD	3.7	10.9	1.1	4.3	5.8
Planarity of all interfaces within multimer (\AA) * ^b					
Min	1.4	1.3	0.7	0.8	1.6
Max	5.1	4.8	2.9	3.0	5.4
Mean	3.3	2.7	1.7	1.7	3.0
SD	0.9	0.8	0.5	0.5	1.0
Interface averaged amino acid composition (%) * ^c					
Charged	16.0	22.6	18.1	34.4	29.6
Polar	49.8	37.5	57.9	35.4	32.6
Hydrophobic	34.3	39.9	23.9	30.2	37.8
No of hydrogen bonds per 100 (\AA^2) Δ ASA * ^d					
Min	0.55	0.50	0	0	0
Max	1.66	1.67	1.86	1.78	1.61
Mean	1.04	0.98	0.92	0.90	0.78
SD	0.29	0.30	0.55	0.53	0.40
Average Hydrophobicity (Fauchere & Pliska scale) * ^e					
Interior	0.88	1.05	1.01	0.98	1.06
Interface	0.34	0.39	0.34	0.06	0.29
Exterior	0.17	0.26	0.17	0.14	0.23
Average Hydrophobicity of all Protein-Protein Interfaces					
Buried Zone	0.60	0.71	0.68	0.40	0.78
Partially Buried Zone	0.21	0.32	0.24	-0.02	0.13
Protomer Weight (Da)					
Min	12200	3600	37100	9100	8000
Max	51100	49800	48300	56200	37300
Mean	29100	12000	46200	26100	18800
SD	12300	11000	2600	16800	9100

Table 4.3: Summary of the structural characteristics of the datasets of non-obligate hetero-complexes.

Characteristic: Summary (All Datasets)			
	Homo-Complexes (obligate)	Hetero-Complexes (obligate)	All Non-Obligate Datasets
Size of Dataset	142	20	45
Mean Δ ASA (\AA^2) ^{*a}			
Min	540	1190	290
Max	10040	11150	2680
Mean	2380	3730	980
SD	1470	2220	430
% ASA buried per protomer			
Min	4.2	7.9	2.7
Max	40.3	50.3	50.8
Mean	19.2	28.8	11.7
SD	8.2	12.2	8.3
Planarity of all interfaces within multimer (\AA) ^{*b}			
Min	0	0.4	0.7
Max	8.4	10.8	5.4
Mean	3.1	3.3	2.4
SD	1.7	2.0	1.0
Interface averaged amino acid composition (%) ^{*c}			
Charged	24.6	28.1	24.9
Polar	30.6	28.3	40.9
Hydrophobic	44.8	43.6	34.2
No of hydrogen bonds per 100 (\AA^2) Δ ASA ^{*d}			
Min	0	0.53	0
Max	2.15	1.58	1.86
Mean	0.94	1.05	0.89
SD	0.41	0.24	0.42
Average Hydrophobicity (Fauchere & Pliska scale) ^{*e}			
Interior	0.95	0.83	1.00
Interface	0.34	0.34	0.29
Exterior	0.21	0.24	0.20
Average Hydrophobicity of all Protein-Protein Interfaces			
Buried Zone	0.76	0.66	0.65
Partially Buried Zone	0.20	0.21	0.18
Protomer Weight (Da)			
Min	4500	5240	3600
Max	96900	85500	56200
Mean	29700	32000	23900
SD	16900	20000	15000

Table 4.4: Summary of the structural characteristics of the datasets obligate homo-complexes, obligate hetero-complexes, and non-obligate hetero-complexes.

4.2 Obligate Hetero-Multimers

4.2.1 Size (ASA) of Protein-Protein Interfaces

The accessible surface areas (ASA) of the protein-protein interfaces within each of the obligate hetero-complexes has been calculated using NACCESS as described in section 3.2 in chapter three.

The subunits from the obligate hetero-complexes generally bury a larger fraction of their ASA in protein-protein interfaces than the subunits of homo-complexes. The minimum, maximum and mean amounts of ASA buried in the datasets of hetero-complexes compared with the corresponding values for the homo-datasets shows this trend (see figure 4.1).

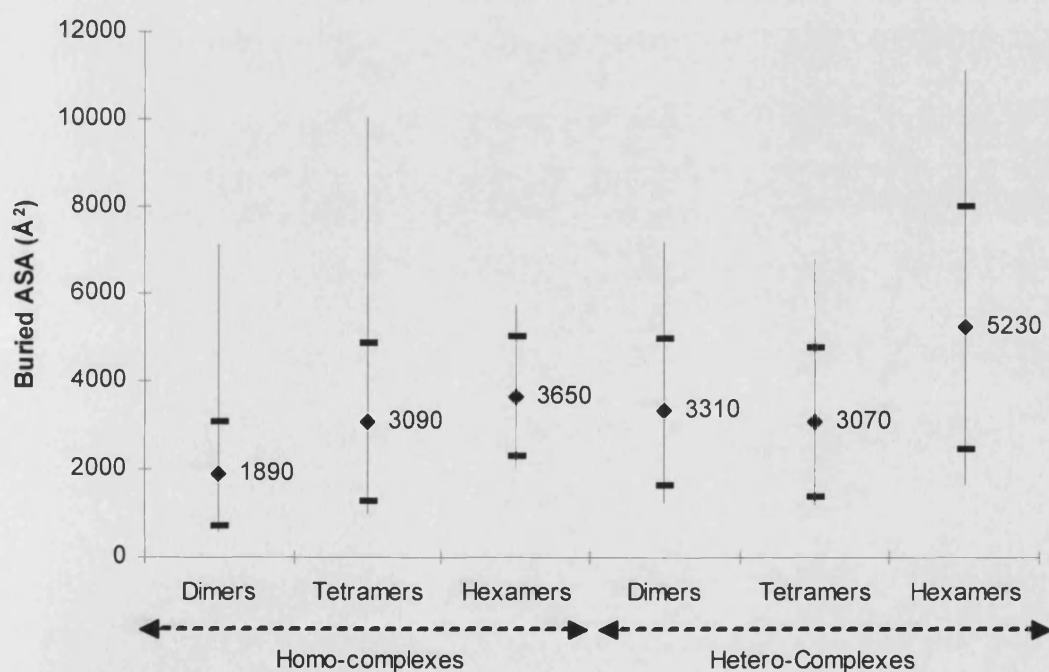


Figure 4.1: A chart showing the average minimum, average, and maximum amounts of ASA buried in protein-protein interfaces for the subunits of obligate hetero and homo dimers, tetramers, and hexamers. The average ASA buried for each class of multimer is marked by a diamond (♦). Horizontal bars mark one standard deviation away from the average ASA buried.

Obligate hetero-dimers on average bury 3310\AA^2 in protein-protein interfaces compared with 1890\AA^2 for homo-dimers. A t-test was carried out on these two mean

values to determine if they differ to a statistically significant degree. The results of the t-test show that a comparison of these mean values is significant to the 5% level. This result shows that these two mean values differ to a statistically significant degree. From looking at the diagrams of the obligate hetero-proteins in the appendix it is apparent that the reason why the total amount of ASA buried in these complexes is so large is due to the extended structures adopted by the constituent proteins within these complexes. Proteins with extended structures can be quite unstable due to their high surface area to volume ratios. Presumably this is also an explanation for why these complexes are all 'permanent' assemblages of proteins.

Penicillin acylase (1ajq) is a hetero-dimer of a light chain of 209 residues and a heavy chain of 557 residues. The interface between these two chains is the largest within the dataset of hetero-dimers at 7000\AA^2 .

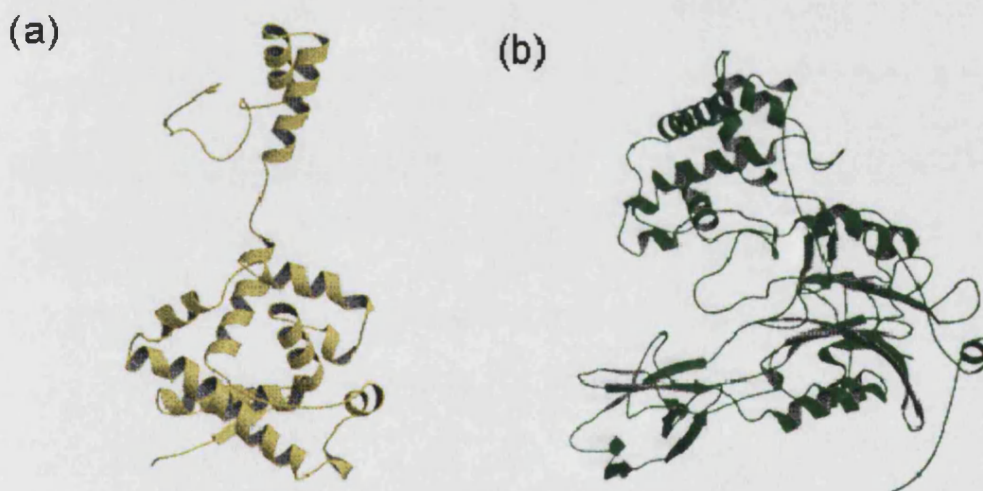


Figure 4.2: The small (a) and large (b) subunits of penicillin acylase (1ajq, Done et al., 1997). The dimer interface is largest within the dataset of hetero-dimers being 7000\AA^2 in size. The subunits of penicillin acylase are unlikely to be stable in isolation of each other due to their extended structures.

A diagram of the two subunits of penicillin acylase in isolation of each other in figure 4.2(a) and 4.2(b) reveals just how extended the structures of both subunits are, and the consequent unlikelihood of them existing independently of each other. The same is plainly true for other entries in the datasets, particularly the small subunit of hydrogenase (1hfe) a diagram of which is shown in chapter 5.

In contrast to homo-complexes the amount of ASA buried in the protein-protein interfaces of obligate hetero-complexes does not consistently increase from dimers to tetramers to hexamers (figure 4.1). This may be an artefact of the datasets. An alternative explanation is simply that hetero-complexes are by definition composed of non-identical subunits that can bind together in many more distinct configurations than can a number of identical subunits. As with homo-complexes the average fraction of ASA buried in protein-protein interfaces in the hetero-proteins increases the higher the multimer. The values go from 24.5% for dimers, to 25.1% for tetramers, and 39.2% for the dataset of hexamers. However in each dataset there are large variations about the average (see table 4.2) and the small size of the datasets calls for caution in attaching any particular significance to this trend. A special point of interest is the *maximum* fraction of ASA that is buried in protein-protein interfaces. No constituent part of any hetero-complex buries greater than 50% of its total ASA in protein interfaces. In fact no protein within any other dataset used in thesis buried more than half of its ASA in protein interfaces when part of a complex.

As with homo-complexes on the whole there is no strong correlation between the molecular weight of the protein subunits in hetero-complexes and the amount of ASA that it buries in interfaces with other proteins when part of a complex. But, curiously, the higher the multimer the better correlated the molecular weight of the protein and the buried ASA. The Pearson correlation coefficient for the hetero-dimers is 0.6, 0.77 for tetramers, and 0.81 for hexamers. For all the hetero-complexes the correlation coefficient is 0.5.

4.2.2 Planarity

The planarity of the protein-protein interfaces within the obligate hetero-complexes has been calculated using the method described in section 3.12 in chapter 3. As with the homo-complexes and as expected there is a good correlation between the size of a protein-protein interface and its planarity (see figure 4.3). In general the bigger the interface the less planar it is. This trend holds well for each of the datasets of obligate hetero-complexes and is amply illustrated in figure 4.3. The correlation coefficient for

the data in figure 4.3 is 0.87 revealing just how well correlated are interface size and planarity.

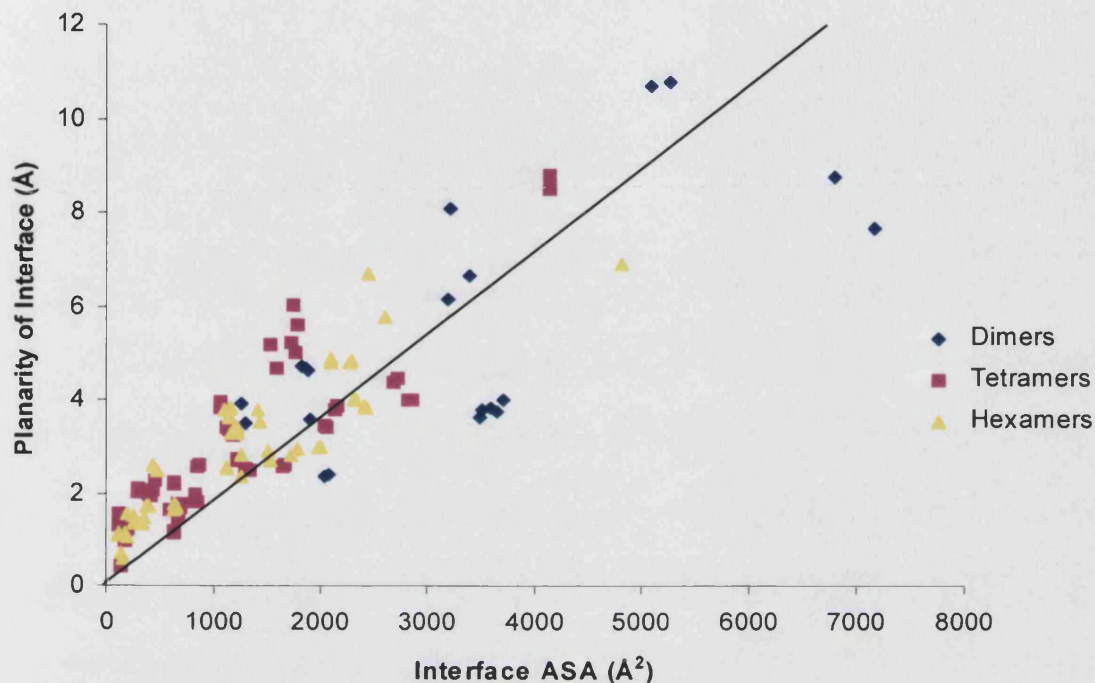


Figure 4.3: The size (ASA) of each protein-protein interface within the datasets of obligate hetero-dimers, tetramers, and hexamers plotted together with its planarity. The gradient of the line of best fit for the points in this plot is 0.002 with the correlation coefficient being 0.87. As with the obligate homo-complexes (see figure 3.22) there is a linear relationship between the size of a protein-protein interface and how planar it is.

Hetero Datasets	Mean Interface Size (Å ²)	Mean Planarity (Å)	Correlation coefficient
Dimers	3310	5.4	0.71
Tetramers	1410	3.0	0.90
Hexamers	1500	3.0	0.89
All Obligate Heteros	1720	3.3	0.87
All Obligate Homos	1450	3.1	0.84

Table 4.5: Mean planarities of the protein-protein interfaces of hetero-dimers, tetramers, and hexamers. Values are also given for all 142 obligate homo-complexes and all 20 obligate hetero-complexes.

The mean planarity of all the interfaces within each dataset separately, together with the Pearson correlation coefficient between the size of the interface and its planarity

are tabulated in table 4.5. The protein-protein interfaces within the obligate hetero-complexes are slightly less planar than those within homo-complexes. However this is apparently not a special property of interfaces formed from two different surfaces (a heterologous interface). As already shown in figure 4.3 the larger the protein-protein interface the less planar it is. The protein-protein interfaces within hetero-complexes are therefore *on average* less planar than the interfaces within homo-complexes because they are *on average* larger. Possible reasons why large interfaces are less planar than small interfaces are further set out in section 4.5.

The protein-protein interfaces within hetero-dimers are for the most part less planar than those found within any of the other datasets of obligate hetero-proteins. The component subunits of the hetero-dimers frequently adopt quite extended structures (see section 4.2.6). Since dimers only have one interface conceivably there is a need for subunits to interact extensively with each other via a large and interdigitated (or interlocked) interface to stabilise the complex. Interfaces in which subunits interlock with each other are almost invariably less planar than those that do not. One example is methylmalonyl-Coa mutase (1req) a hetero-dimer with the least planar interface in all three datasets with a planarity of 10.8Å. From the diagram of 1req in the appendix it can be seen the interface is large in size and the two subunits do interlock extensively with each other.

4.2.3 Hydrogen Bonding

The number of inter-subunit hydrogen bonds for the obligate hetero-complexes has been calculated using HBPLUS as described in section 3.8 in chapter 3. There is a good linear relationship between the size of a protein-protein interface and the numbers of direct hydrogen bonds across it (see figure 4.4). The Pearson correlation coefficient for all points in figure 4.4 is > 0.9 highlighting just how well correlated are the amount of buried ASA and the number of direct inter-subunit bonds in the obligate hetero-complexes. The term 'direct' is important as there are almost certainly hydrogen bonds between protein subunits that are mediated indirectly by solvent molecules. However it is difficult to accurately assess the number of indirect

hydrogen bonds due to the resolution of the some of the protein structures in these datasets.

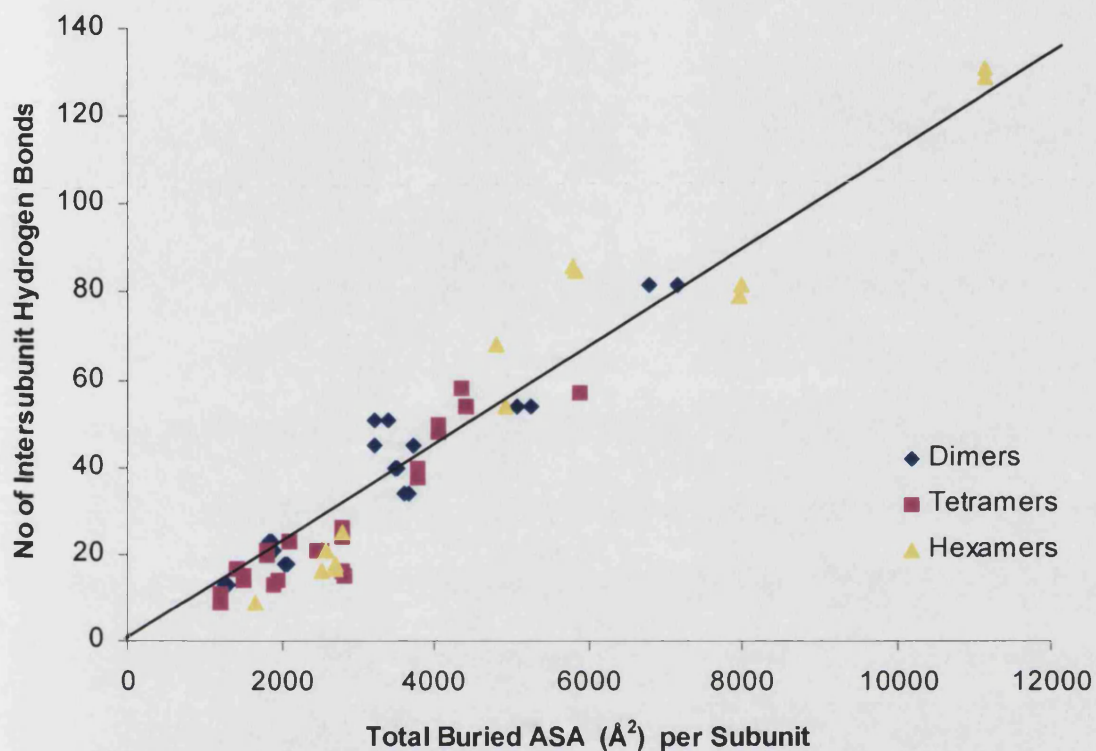


Figure 4.4: A plot of the number of inter subunit hydrogen bonds for each protein subunit against its total interface ASA. The gradient of the line of best fit of the plot is 0.01. The correlation coefficient for the plot is 0.96.

There is no real difference between the hydrogen bonding patterns found within any of the datasets. Hetero-dimers, tetramers, and hexamers all have very similar numbers of hydrogen bonds across their interfaces (although all datasets show some variation). Hetero-dimers on average have the greatest numbers of hydrogen bonds across their interfaces with 1.14 hydrogen bonds per 100\AA^2 of buried ASA. Tetramers have the least number with 0.97 for every 100\AA^2 of interface. Further details regarding these figures can be seen in table 4.2. Hetero-complexes appear to have slightly more hydrogen bonds at their interfaces than do homo-complexes. There is no obvious reason why this should be and it is almost certainly a statistical artefact.

	No of Hydrogen Bonds Per 100Å ² of Buried ASA	Min	Max	SD
All Homo-complexes	0.94	0	2.15	0.41
All Hetero-complexes	1.05	0.53	1.58	0.24

Table 4.6: The mean numbers of inter-subunit hydrogen bonds for every 100Å² of buried ASA in the datasets of all 142 obligate homo-complexes and all 20 obligate hetero-complexes.

The lines of best fit in figure 4.4 and figure 3.17(b) in chapter 3 are the same to two decimal places pointing to the linear relationship between interface ASA and the number of inter-subunit hydrogen bonds being very similar for both homo and hetero-complexes.

4.2.4 Amino Acid Composition

The amino acid composition of the protein-protein interfaces within the obligate hetero-complexes is not very different from those found within the homo-complexes studied in chapter 3. Furthermore there are no distinctive variations in the amino acid composition of any of the protein-protein interfaces in the hetero-dimers, tetramers, or hexamers. The mean frequency with which each of the twenty amino acids is found in a protein-protein interface is given for all three datasets of obligate hetero-complexes collectively in figure 4.6. The equivalent figures for the protein interfaces of all 142 homo-complexes are also given. As with the homo-complexes the residues that are most prevalent in the protein-protein interface of the hetero-complexes are either hydrophobic or charged. Arginine is the most commonly occurring amino acid at the interface making up 7.3% of all interfacial residues followed by leucine (6.9%), glycine (6.9%), and glutamic acid (6.7%).

The residues in the interior, exterior, and interface regions of the obligate hetero-complexes have also been analysed in terms of whether they are charged, polar, or hydrophobic. A chart showing the averaged make up of each of these three regions for all the obligate hetero-complexes together is shown in figure 4.5.

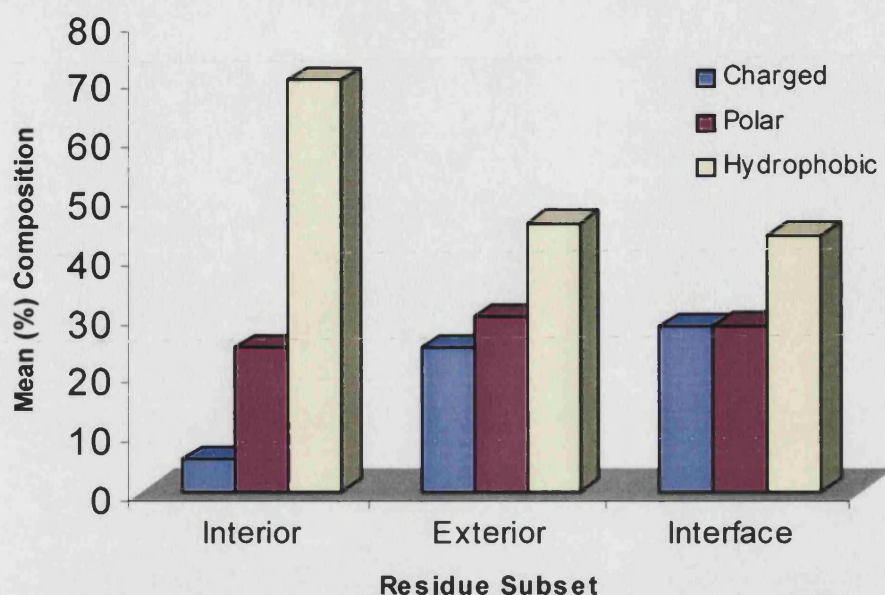


Figure 4.5: The mean percentage of residues that are charged, polar, and hydrophobic for the interior, exterior, and interface regions of all twenty obligate hetero-complexes. The protein-protein interfaces of the hetero-complexes on average contain roughly equal fractions of charged and polar residues.

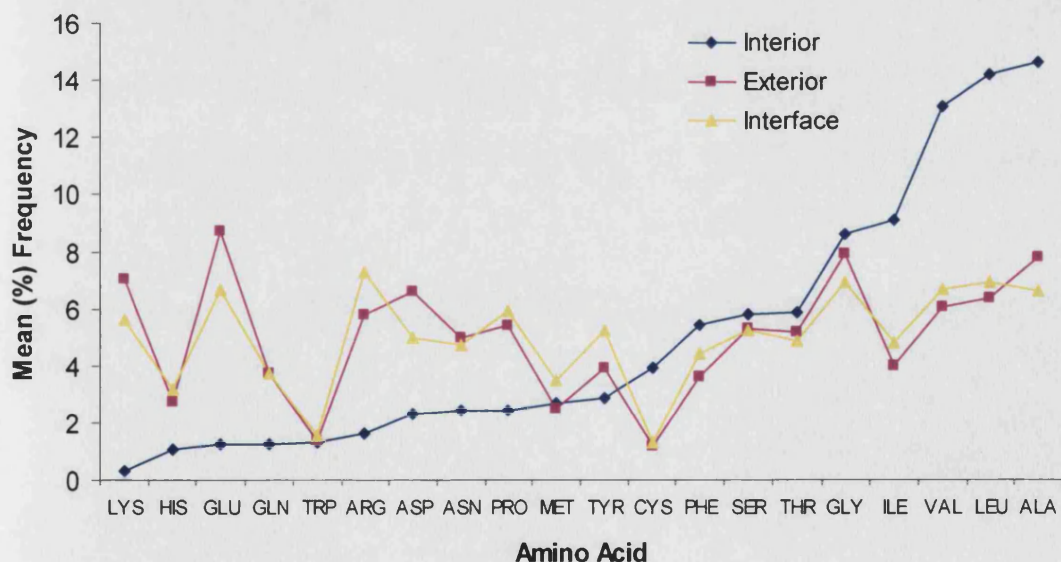


Figure 4.6: The mean percent frequencies of amino acids in the interior, exterior, and interface regions of all 20 obligate hetero-complexes. The amino acids are ordered according to the increasing % frequency in the interior region. As with the homo-complexes (see figure 3.10) the amino acid composition of the interface is well correlated with the protein exterior.

Generally the compositions of the interior, exterior, and interface regions follow the same trends as do the datasets of homo-complexes (see figure 3.10 in chapter 3) . The exterior and interfaces are well correlated as is illustrated in figure 4.6.

The interfaces within the hetero-complexes contain equal proportions of charged and polar residues each comprising 28% of interface residues. Hydrophobic amino acids make up the remaining 44% of all residues at the interface. The interiors of the hetero-complexes are dominated by hydrophobic residues just as is the case in every other dataset studied in this thesis.

The interface residue propensities for the hetero-complexes have been calculated as described in section 3.5 in chapter 3 and are shown later in figure 4.19. The five residues with the greatest interface propensities are tyrosine (1.45), methionine (1.37), arginine (1.34), phenylalanine (1.26), and proline (1.11). With the exception of proline these are the same residues that have large (>1) interface propensities in the homo-complexes (see table 3.5 for the residue interface propensities of the homo-complexes). It can be assumed that these residues are particularly favoured at the interfaces of the hetero-complexes for the same reasons that they are favoured in the interfaces of homo-complexes. As stated in the next section (4.2.5) the protein-protein interfaces of some of the hetero-complexes are quite extended being formed by loops and other irregular secondary structure elements. Proline has a uniquely restricted side-chain that is known to have a stabilizing effect on loops enabling them to maintain a distinct shape (Reiersen & Rees, 2001). This could explain why the interface propensity of proline in hetero-complexes is greater than that found for the homo-complexes.

4.2.5 Secondary Structure Content

In order to explain the high interface propensity of proline the secondary structure content of the protein-protein interfaces of the obligate hetero-complexes was studied. As in chapter 3 residues are classified into four categories of secondary structure:

helix, strand, turn, and coil. The mean content of the protein interfaces of each dataset in terms of these four divisions are summarised in table 4.7.

Dataset	Mean (%) Frequency Secondary Structure States			
	Helix	Strand	Turn	Coil
Dimers	33.5	18.9	22.9	24.7
Tetramers	39.2	13.7	24.4	22.7
Hexamers	27.7	16.4	17.5	38.4
All Hetero	34.0	16.5	21.9	27.7
All Homo	37.8	20.0	21.6	19.9

Table 4.7: The average percentages of interface residues that are helix, strand, turn, and coil for the datasets of hetero-dimers, tetramers, and hexamers. Values are also given for all 142 obligate homo-complexes and all 20 obligate hetero-complexes.

A comparison between the values in table 4.7 and the equivalent figures for the protein-protein interfaces within the homo-complexes (table 3.10 in chapter 3) do suggest that there are differences in the secondary structure content of the interfaces found within homo and hetero-complexes. The interfaces within hetero-complexes typically contain more residues that fall in the 'coil' and classification (regions of little or no secondary structure). For example, on average some 38% of interface residues in the dataset of hetero-hexamers are part of coil regions compared with 17.5% for the homo-hexamers. A t-test was carried out on these two mean values to determine if they differ to a statistically significant degree. The results of the t-test show that a comparison of these mean values is significant to the 5% level. This result shows that the two average values do differ to a statistically significant degree. The interfaces of the hetero-dimers and tetramers also contain slightly more coil residues than the homo-dimers and tetramers but the differences are much less. In terms of absolute numbers 34% of all the 5300 residues that make up the protein-protein interfaces of the twenty hetero-complexes are classed as being coil. Nevertheless these differences are probably sufficient to explain why proline has rather a high propensity to be found at the interface in the obligate hetero-complexes as compared with the homo-complexes. As already pointed out in section 4.2.4 loops are often rather rich in proline residues due to the stabilising effect of this residue. It follows therefore that the reason why the interfaces within the hetero-complexes are to some extent enriched with proline is simply because these proteins often interact with each other through loops and other regions of irregular secondary structure.

4.2.6 Subunit Shape

The shape of protomers that make up the hetero-complexes was investigated by fitting ellipsoids to the protein structures using the method of Taylor, Thornton, and Turnell, 1983. Using this method a protein structure is represented by an ellipsoid that encloses a given fraction of residues in the structure.

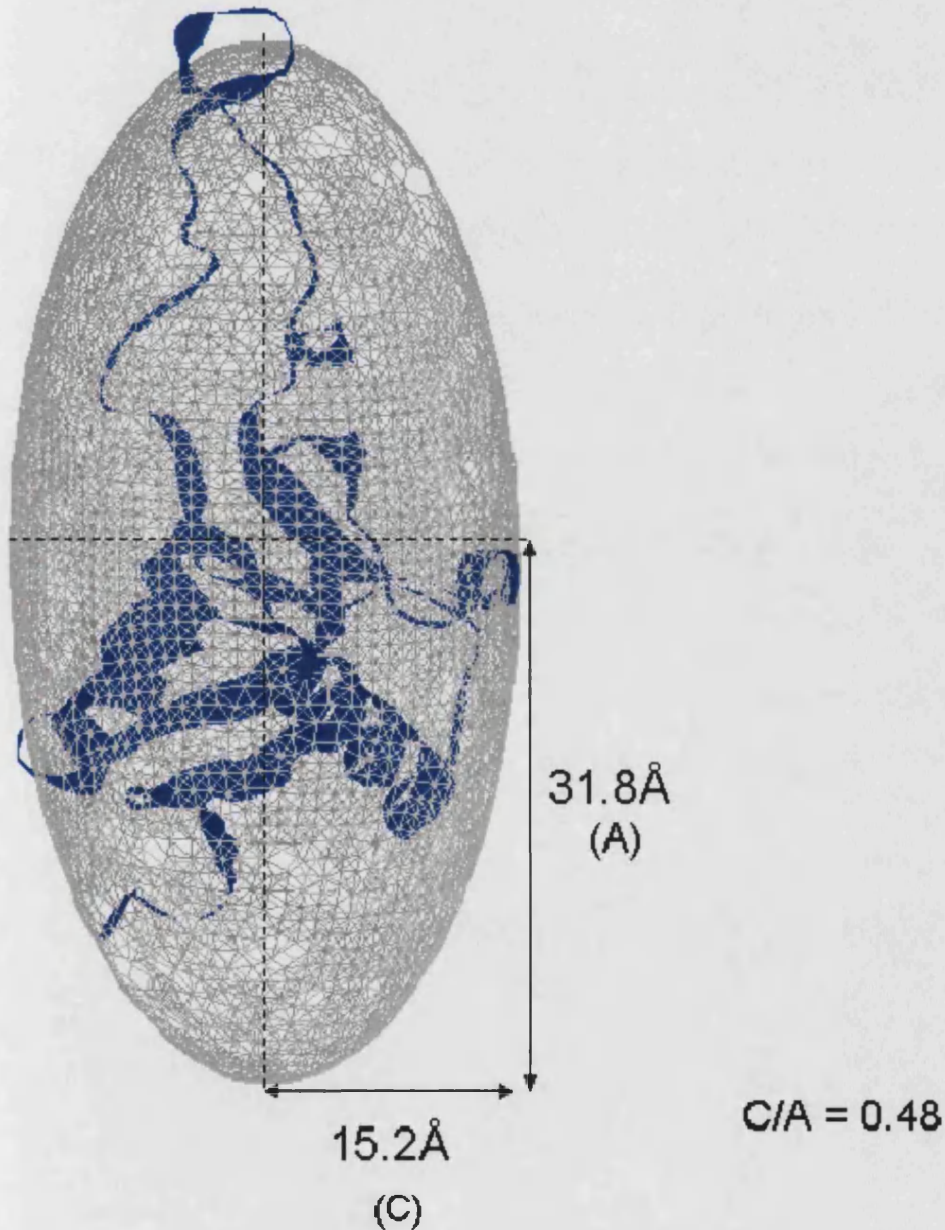


Figure 4.7: A diagram of the ellipsoid fitted to the large subunit of the Ix/X-bp coagulation factor protein (1ixx) using the method of Thornton, Taylor, and Turnell, 1983. The ratio of the smallest and largest semi-axial lengths of the ellipsoid (C/A) gives a quantitative measure of how ‘extended’ the structure is.

Currently ellipsoids have been fitted that contain 90% of the residues in the protomer structures. The ellipsoid fitted to the large subunit of 1ixx is shown in figure 4.7. In figure 4.7 the largest semi-axial length of the ellipsoid is marked 'A' and is 31.8Å. The smallest semi-axial length is marked as C and is 15.2Å long. The ratio C/A then gives a quantitative measure of how extended or elongated the structure is. The lower the value of C/A the more extended the structure. In the case of the large subunit of 1ixx the C/A ratio is 0.48. The statistics concerning the C/A quantity for the subunits in the hetero-dimer, tetramer, and hexamer datasets are shown in table 4.8

	Mean C/A	Min	Max	SD
Dimers	0.46	0.19	0.66	0.14
Tetramers	0.58	0.32	0.87	0.16
Hexamers	0.53	0.41	0.69	0.09
All Heteros	0.53	0.19	0.87	0.15
All Homos	0.55	0.13	0.89	0.15

Table 4.8: The ratios between the lowest semi-axial length (A) and the largest semi-axial length (C) for the ellipsoids fitted to the protomers of the obligate hetero-dimers, hetero-tetramers, and hetero-hexamers. Values are also given for all 142 obligate homo-complexes and all 20 obligate hetero-complexes.

From table 4.8 the mean C/A ratio for the hetero-dimers is lower than for any other dataset. In comparison the mean C/A ratio for the homo-dimers is 0.56. A t-test was carried out on the mean C/A ratios for the obligate homo-dimers and hetero-dimers to determine if they differ to a statistically significant degree. The results of the t-test show that a comparison of these mean values is significant to the 5% level. This result shows that the two mean values do differ to a statistically significant degree. That the subunits of the hetero-dimers adopt quite extended structures is quite plain from the diagrams of the obligate hetero-complexes in the appendix. In terms of numbers the protomers of obligate hetero-complexes more frequently have a lower C/A ratio than do those of homo-complexes. The inference is then that the protomers of hetero-complexes are more often elongated in shape than the protomers of homo-complexes. This is despite the mean C/A ratios for all the subunits in the homo and hetero-complex datasets being rather similar.

4.2.7 Hydrophobicity

The total hydrophobic content of all subunits within the hetero-complexes has been calculated (see figure 4.8). The averaged hydrophobic content of the subunits within each dataset is shown in tables 4.9 and 4.10.

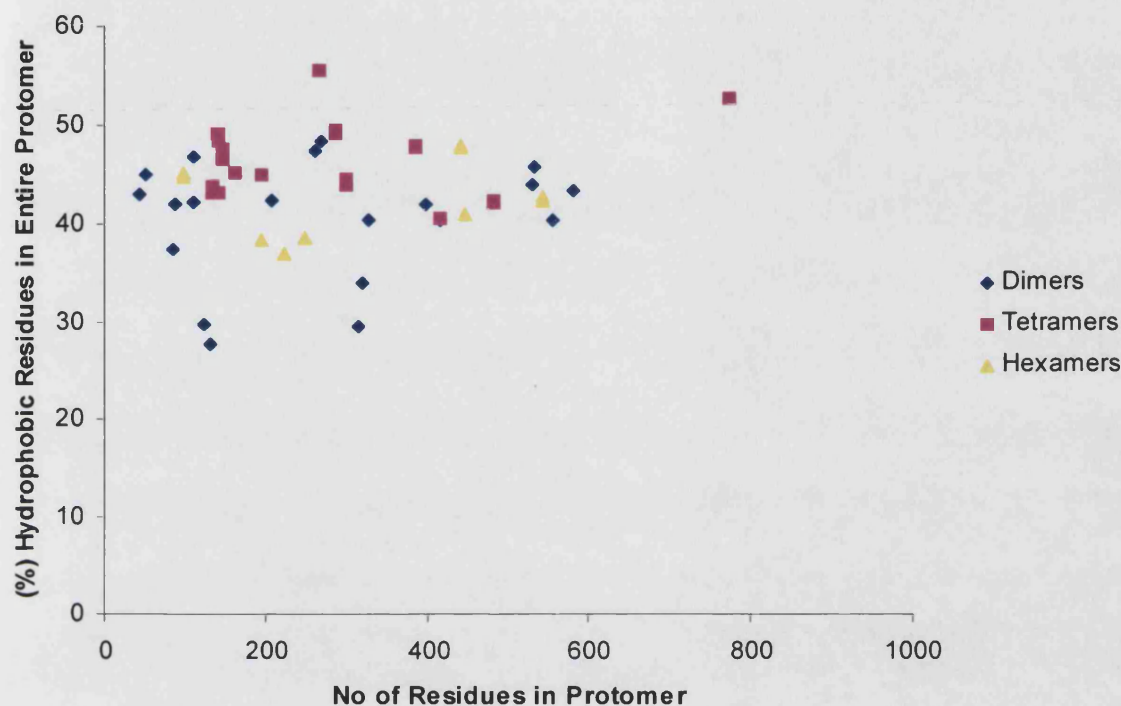


Figure 4.8: A plot of the number of residues in each subunit of each obligate hetero-complex against the percentage of residues that are hydrophobic.

Homo Complexes	Mean % hydrophobic residues in entire Protomer	Min	Max	SD
Monomers	47.4	32.3	59.4	5.1
Dimers	48.3	31.8	58.5	5.6
Trimers	49.4	37.1	57.3	5.0
Tetramers	49.9	40.3	59.6	4.4
Hexamers	52.0	43.9	57.3	4.2

Table 4.9: The mean hydrophobic content of the 92 monomers, and of the protomers from the datasets of homo-dimers, trimers, tetramers, and hexamers. This table is reproduced from chapter 3 for reference.

Hetero Complexes	Mean (%) Hydrophobic Residues in Protomer	Min	Max	SD
Dimers	40.5	27.8	48.3	6.0
Tetramers	46.7	40.6	55.7	4.1
Hexamers	42.1	36.9	48.0	3.5

Table 4.10: The mean hydrophobic content of the protomers from the datasets of hetero-dimers, tetramers, and hexamers. As with the homo-complexes there is no protomer from an obligate hetero-complex that has a hydrophobic content >60%.

As with the homo-complexes there is no protomer of an hetero-complex that has a hydrophobic content of >60%. From the data in figure 4.8 taken together with the data in table 4.9 reproduced from chapter 3 it appears that no protein that is part of an obligate homo or hetero protein-complex can have a hydrophobic content greater than 60%. It is possible that this figure represents a threshold beyond which solubility becomes an issue.

The figures in table 4.10 suggest that the obligate hetero-complexes typically have a lower hydrophobic content than the homo-complexes. With the exception of the hetero-dimers the standard deviations on the figures in table 4.10 are comparable with those in table 4.9. The hydrophobicity of the interior, exterior, and interface regions in the hetero-dimers, tetramers, and hexamers has been calculated using the Fauchere and Pliska scale in the same way as described in chapter 3. A summary chart showing the averaged hydrophobicities of the datasets together with those of some of the other datasets is given later in figure 4.19. The actual figures for the protein-protein interfaces of each dataset can be found in table 4.2. For each dataset the interface region is intermediate in hydrophobicity between the interior and exterior in the same way as is found for almost all other categories of proteins in this thesis. What is interesting is that the hydrophobicities of the exterior, and interface regions of the hetero-complexes do vary in a systematic way. The higher the multimer the more polar are its exterior and interfaces. The same trend is observed for the exteriors of these proteins but to a lesser degree. Hetero-dimers have an averaged hydrophobicity of 0.40, compared with tetramers (0.34), and hexamers (0.29). A t-test was carried out on the mean hydrophobicity of the hetero-dimers and the hetero-tetramers. The results of the t-test show that these two mean values differ to the 5% significance

level. A t-test was also carried out on the mean hydrophobicity of the hetero-tetramers and the hetero-hexamers and these two mean values also differ to the 5% significance level. The results of these two t-tests indicate that the trend that the higher the multimer the more polar are its protein-protein interfaces may be statistically significant for obligate hetero-complexes.

4.2.8 Subunit Organisation

All of the obligate hetero-complexes with the exception of 1mro are composed of two non-identical subunits denoted α and β . The subunit composition of the twenty hetero-complexes is summarised in tables 4.11 and 4.12. All of the hetero-dimers have a $\alpha\beta$ composition whereas the hetero-tetramers have a $\alpha_2\beta_2$ or $(\alpha\beta)_2$ subunit organization (the subunit organisation of haemoglobin is better described as $(\alpha\beta)_2$). From what structural information there is it is quite commonplace that hetero-complexes are composed of two different subunits. A number of different protein hetero-complexes with a $\alpha_m\beta_n$ composition can be found elsewhere to demonstrate the generality of this observation (see p103 of Price & Stevens, 1999). Also see, p279 of Pain, 2000. Further data concerning the subunits that are found within each hetero-complex is given in tables 4.11 and 4.12. These tables contain the data that is referred to in sections 4.2.9-4.2.11.

Key for Table 4.11


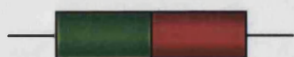
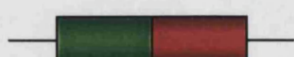


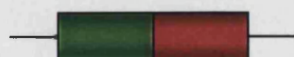
- (a) The protein is coded for by a single gene. The mature protein is a result of the proteolytic cleavage of a pre-cursor.
- (b) The genes coding for subunits of the protein are shown as red and green cylinders. The two genes are arranged consecutively with no gaps between them.
- (c) The genes coding for the α and β subunits of the protein are separated by a gene (denoted by a yellow cylinder) that codes for a protein of unknown function.
- (d) The genes coding for the α and β subunits of the protein are arranged with a small gap between them.
- (e) The three genes coding for the protein are shown as green, red and blue cylinders. These genes are separated by two genes that code for proteins of unknown function

*¹ SI: The sequence identity of two aligned sequences generated using the ALIGN program from the FASTA distribution.

*² The score produced by the ALIGN program for the sequence alignment. The higher the score the greater the statistical significance of the alignment.

*³ The number of equivalent residues (NEQ) when one protein is superimposed upon another using the ProSup program.

*⁴ The RMSD of the small subunit of the protein superimposed on its large subunit using the program ProSup (Lackner et al., 2000).

Protein	Subunit Structure	SI* ¹ (%)	Score* ²	NEQ* ³	RMSD * ⁴	Gene Structure
Dimers						
1ajq	$\alpha\beta$	10.2	-1160	34	2.80	(a) 
1ft1	$\alpha\beta$	12.7	-168	64	2.63	The genes coding for the α and β subunits are thought to be co-expressed
1h2a	$\alpha\beta$	14.8	-778	31	2.48	(b) 
1hcn	$\alpha\beta$	12.3	-189	61	2.09	Multiple genes on chromosomes 6 & 19
1hfe	$\alpha\beta$	8.6	-1062	23	2.91	(b) 
1ixx	$\alpha\beta$	45.8	291	115	1.48	Domain swapped protein
1luc	$\alpha\beta$	28.9	390	307	1.92	(b) 
1req	$\alpha\beta$	19.5	101	415	1.74	(b) 
2frv	$\alpha\beta$	12.7	-889	35	2.63	(b) 
4mon	$\alpha\beta$	9.8	-39	23	2.50	Thought to be coded for by a single gene.


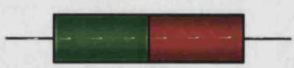







Tetramers						
lapy	$\alpha_2\beta_2$	11.1	-90	50	2.66	(a) 
lb7y	$\alpha_2\beta_2$	15.9	-1375	159	1.98	(b) 
lbou	$\alpha_2\beta_2$	7.9	-535	37	3.13	(b) 
lccw	$\alpha_2\beta_2$	7.7	-1249	76	2.62	(c) 
lqdl	$\alpha_2\beta_2$	13.3	-716	28	3.07	(d) 
lqsh	$(\alpha\beta)_2$	43.2	314	134	1.34	Multiple genes on chromosomes 11 & 16
2scu	$\alpha_2\beta_2$	15.5	-272	127	2.22	(b) 
Hexamers						
leg9	$\alpha_3\beta_3$	13.1	-851	45	2.98	(d) 
ltii	$\alpha\beta_5$	9.8	-499	26	2.13	(b) 
lmro	$\alpha_2\beta_2\gamma_2$					(e) 
	$\alpha w \beta$	14.2	-277	322	2.21	
	$\alpha w \gamma$	11.3	-992	42	2.51	
	$\beta w \gamma$	13.1	-593	46	2.56	

Table 4.11: A table showing the subunit composition and gene structure for each obligate hetero-complex together with a comparison between the subunits of each complex at both the sequence and structural levels.

Protein	SI (%)	Large Subunit			Small Subunit		
		No of Residues	No of Domains	CATH Architecture	No of Residues	No of Domains	CATH Architecture
Dimers							
1ajq	11.0	557	3	NYA	206	2	1.10.439.10 -> Mainly Alpha Orthogonal Bundle
1ft1	13.2	416	1	1.50.10.40 -> Mainly Alpha Alpha/alpha barrel	315	1	1.25.40.120 -> Mainly Alpha Horshoe
1h2a	14.8	534	1	1.10.645.10 -> Mainly Alpha Orthogonal Bundle	267	2	3.40.50.700 -> Alpha Beta 3-Layer(aba) Sandwich 4.10.480.10 -> Few Secondary Structures Irregular
1hcn	12.3	110	1	2.10.90.10 -> Mainly Beta Ribbon	85	1	2.10.90.10 -> Mainly Beta Ribbon
1hfe	8.6	397	3	3.40.50.1780 -> Alpha Beta 3-Layer(aba) Sandwich 3.30.70.20 -> Alpha Beta 2-Layer Sandwich 3.40.950.10 -> Alpha Beta 3-Layer(aba) Sandwich	88	1	4.10.260.20 -> Few Secondary Structures Irregular
lixx	45.8	129	1	3.10.100.10 -> Alpha Beta Roll	123	1	3.10.100.10 -> Alpha Beta Roll
1luc	28.9	326	1	3.20.20.30 -> Alpha Beta Barrel	320	1	3.20.20.30 -> Alpha Beta Barrel

1req	19.5	727	2	3.20.20.240 -> Alpha Beta Barrel 3.40.50.280 -> Alpha Beta 3-Layer(aba) Sandwich	619	2	3.20.20.240 -> Alpha Beta Barrel 3.40.50.280 -> Alpha Beta 3-Layer(aba) Sandwich
2frv	11.6	530	1	1.10.645.10 -> Mainly Alpha Orthogonal Bundle	261	2	3.40.50.700 -> Alpha Beta 3-Layer(aba) Sandwich 4.10.480.10 -> Few Secondary Structures Irregular
4mon	9.8	50	1	NYA	44	1	NYA
Tetramers							
1apy	11.1	161	1	3.30.426.10 -> Alpha Beta 2-Layer Sandwich	141	1	3.50.11.10 -> Alpha Beta 3-Layer(bba) Sandwich
1b7y	15.3	775	6	3.30.56.10 -> Alpha Beta 2-Layer Sandwich 2.40.50.30 -> Mainly Beta Barrel 3.50.40.10 -> Alpha Beta 3-Layer(bba) Sandwich 3.30.56.20 -> Alpha Beta 2-Layer Sandwich 3.40.690.10 -> Alpha Beta 3-Layer(aba) Sandwich	265	1	3.40.690.10 -> Alpha Beta 3-Layer(aba) Sandwich

1b77 (continued)				3.30.70.380 -> Alpha Beta 2-Layer Sandwich			
1bou	7.9	298	1	3.40.830.10 -> Alpha Beta 3-Layer(aba) Sandwich	132	1	1.10.700.10 -> Mainly Alpha Orthogonal Bundle
1ccw	7.0	483	2	3.20.20.290 -> Alpha Beta Barrel 6.1.81.1 -> Few Secondary Structures	137	1	3.40.50.280 -> Alpha Beta 3-Layer(aba) Sandwich
1qdl	13.3	416	2	NYA	195	1	3.40.50.880 -> Alpha Beta 3-Layer(aba) Sandwich
1qsh	43.2	146	1	1.10.490.10 -> Mainly Alpha Orthogonal Bundle	141	1	1.10.490.10 -> Mainly Alpha Orthogonal Bundle
2scu	15.9	385	3	3.30.470.20 -> Alpha Beta 2-Layer Sandwich 2.30.35.30 -> Mainly Beta Roll 3.40.50.261 -> Alpha Beta 3-Layer(aba) Sandwich	286	2	3.40.50.720 -> Alpha Beta 3-Layer(aba) Sandwich 3.40.50.261 -> Alpha Beta 3-Layer(aba) Sandwich
Hexamers							
1eg9	13.1	447	2	3.90.380.10 -> Alpha Beta Complex 2.102.10.10 -> Mainly Beta 3-layer Sandwich	193	1	3.10.30.90 -> Alpha Beta Roll

1tii	98	222	2	3.90.210.10 -> Alpha Beta Complex 1.20.5.200 -> Mainly Alpha Up-Down-Bundle	98	1	2.40.50.50 -> Mainly Beta Barrel
------	----	-----	---	--	----	---	----------------------------------

Protein	SI (%)	No of Residues	No of Domains	CATH Architecture	No of Residues	No of Domains	CATH Architecture
1mro		548 α subunit	3	3.90.390.10 -> Alpha Beta Complex 3.30.70.470-> Alpha Beta 2-Layer Sandwich 1.20.840.10-> Mainly Alpha Up-Down-Bundle	442 β subunit	2	3.30.70.470-> Alpha Beta 2-Layer Sandwich 1.20.840..10-> Mainly Alpha Up-Down-Bundle
1mro		247 γ subunit	1	3.90.320.2010 -> Alpha Beta Complex			

Table 4.12: The domain architecture of the 20 obligate hetero-complexes. Domain assignments are taken from the CATH database (Orengo et al., 1997). NYA denotes that a domain has not been assigned a CATH architecture.

4.2.9 Subunit Genetic Structure

The relative locations of the genes coding for the subunits of the most of the hetero-complexes has been established. Mostly this has been done by searching the literature cited in the SWISS-PROT entry of each protein. The physical locations of the genes coding for the subunits of haemoglobin and chorioinic gonadotropin were found using LocusLink (<http://www.ncbi.nlm.nih.gov/LocusLink>). As summarised in table 4.11 the genes coding for each subunit of a protein are usually adjacent to each other. Whether or not these two genes have evolved from a common ancestor is addressed in the next section. A large number of proteins in the three datasets of hetero-complexes are from prokaryotic organisms and their genes are organised and co-expressed as operons. There is also evidence that operons themselves may be found in close proximity to each other. Over 90% of the enzymes that form stable complexes in *Escherichia coli* metabolic pathways are adjacent on the *E. coli* chromosome (Ouzounis & Karp, 2000). The genes coding for the different subunits of haemoglobin and choroinic gonadotropin are found on different chromosomes revealing the more complex nature of mammalian genomes. Methy-coenzyme M reductase has the most complex subunit organisation ($\alpha_2\beta_2\gamma_2$) of any of the twenty hetero-complexes. The arrangement of the genes coding for the enzyme is correspondingly complex. There are five genes denoted mcrB, D, C, G, and A, arranged as an operon (Weil et al., 1989). The genes mcrD and mcrC code for two small proteins of 16 and 21 kDa of unknown functions. It could be that these two small proteins aid either the correct folding of the three enzyme subunits or in the assembly of the full hexamer (Allmansberger et al., 1989). This enzyme has the same genetic structure in five different methanogenic bacteria.

4.2.10 Subunit Homology

As detailed in tables 4.11 and 4.12 all the obligate hetero-complexes with the exception of methyl-coenzyme M reductase (1mro) are made up of two non-identical subunits. As established in the previous sections the genes coding for these two

subunits are usually in close proximity to one another. This observation leads to the question; are these two subunits homologous? To address this question two protein structures must be compared at the sequence and structural level. For each complex the structures of the two different subunits have been superimposed using the ProSup structural alignment tool (<http://www.ca-me.sbg.ac.at>, Lackner et al., 2000). ProSup is based upon the ridged body superimposition of the C_α coordinates of two proteins iteratively until a fitted structure is produced with the maximum possible number of structurally equivalent residues (NEQ). The RMSD and the number of structurally equivalent residues of the small subunit superimposed onto the large subunit of each complex (including methyl-coenzyme M reductase) are given in table 4.11.

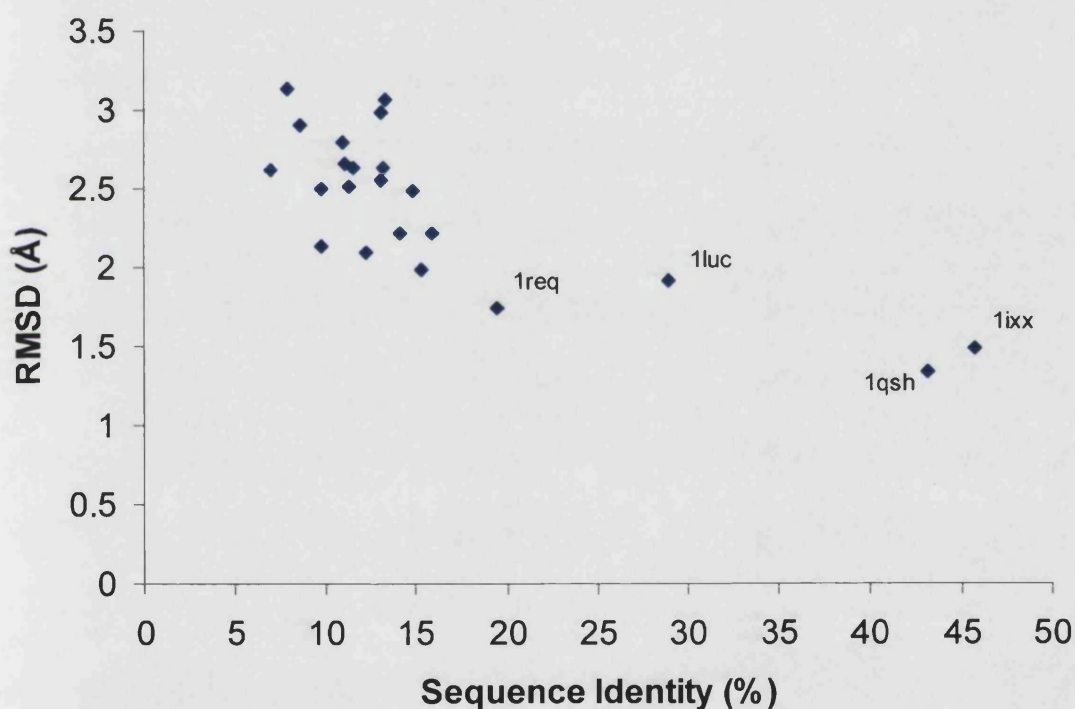


Figure 4.9: The sequence identity of the large and small subunit of each obligate hetero complex together with the RMSD of the small subunit of each complex structurally superimposed on its large subunit using ProSup (Lackner et al., 2000). Data is also plotted for the constituent subunits of methyl-coenzyme M reductase (1mro).

The domain structure of each subunit with its CATH classification code (Orengo et al., 1997) can be found in table 4.12. At the sequence level pairwise alignments of the large and small subunits of every hetero-complex have been produced using the

program ALIGN from the FASTA software distribution. The resulting sequence identity and global alignment scores are shown in table 4.11. A plot of the RMSD of the fitted structures produced by ProSup against the sequence identity of the two subunits is shown in figure 4.9. From figure 4.9 and tables 4.11 and 4.12 there are a few proteins whose subunits are unquestionably homologous to each other. These proteins include haemoglobin (1qsh) which is possibly the best known example of a protein complex that has arisen through gene duplication (Efstratiadis et al., 1980, Jeffreys et al., 1982 and references therein). A diagram of the small subunit of haemoglobin superimposed upon its large subunit is shown in figure 4.10(a).

The two subunits of bacterial luciferase (1luc) in figure 4.10(c) share a sequence identity of 29% have similar topologies and are certainly homologues (Fisher et al., 1996, Baldwin et al., 1979). The subunits of the blood coagulation factor IX/X-bp as set out in chapter 2 are homologues and have arisen through domain swapping (see figure 4.10(b), Mizuno, 1997). A probable example of a protein whose subunits are more remotely related to each other is human chorionic gonadotropin (1hcn, Wu et al., 1994). Although the subunits of hCG only have an SI of 12.3% the structural similarity is strong. The domain structure of the two subunits is identical and the RMSD of the small subunit fitted to the large subunit is 2.09Å. Another instance of a protein whose subunits are evolutionary related (albeit remotely) is CoA mutase from *Propionibacterium shermanii* (1req). Again the subunits of the enzyme share equivalent architectures and the RMSD of the fitted subunit structures is low at 1.74Å. The large subunit of CoA mutase in *P. Shermanii* (1req) has a SI of 60% with the human form of the enzyme which is a homo-dimer. The small β subunit of CoA mutase performs no major biological function, and both the *P. Shermanii* and the *Homo sapiens* enzyme most likely have evolved from a homo-dimeric ancestor (Mancia et al., 1996). There are also examples of proteins whose subunits may have evolved from a common ancestor. The subunits of 4,5-dioxygenase (1bou) have been observed to have a common structural core (Sugimoto et al., 1999). There are also examples of protein where there is local rather than global homology between them. That is to say proteins whose subunits both have domains with the same or similar architectures.

(a)



Haemoglobin

(b)



Blood coagulation factor IX/X -bp

(c)



Bacterial Luciferase

RMSD (Å)

(a) 1qsh 1.34

(b) 1ixx 1.48

(c) 1luc 1.92

Figure 4.10: Three protein complexes whose constituent subunits are homologous to each other. These are, (a) Haemoglobin (Miyazaki et al., 1999), (b) the blood coagulation factor IX/X-bp (Mizuno et al., 1998), (c) bacterial luciferase (Fisher et al., 1996). In each case the small subunit of each complex is superimposed on its large subunit using ProSup (Lackner et al., 2000). The RMSD for each of these three structures is also given.

In these proteins the gene for one subunit could have evolved from only part of the gene coding for the other subunit. For instance the small subunit of phenylalanyl-tRNA synthetase (1b7y) has the same fold as a domain of the large subunit that it makes contact with in the full tetrameric complex (see figure 4.11).



Figure 4.11: The small subunit of phenylalanyl-tRNA synthetase superimposed on the large subunit of the enzyme using ProSup (Lackner et al., 2000). The RMSD of the fitted structure is 1.98Å. The small subunit of the enzyme has the same architecture as a domain from the large subunit and could be the result of partial gene duplication.

The RMSD of the fitted structure in figure 4.11 is 1.98Å illustrating how closely related are these two structures. The role of the small subunit of 1b7y is to assist in the recognition and binding of tRNA^{Phe} molecules (see figure 5(c) from Mosyak et al., 1995). This example is particularly interesting since 1b7y belongs to one of the most ancient families of proteins. The phenylalanyl-tRNA synthetase complex could therefore be one of the earliest complexes to have evolved via some form of genetic duplication.

From the data shown in sections 4.2.8 and 4.2.9 the subunits within a hetero-complex are frequently homologous to each other. Eleven out of the twenty obligate hetero-complexes have subunits that are probably homologous to each other in whole or in part. There are five proteins whose subunits are unquestionably homologous to each other. These proteins are penicillin amidohydrolase (1ajq), the blood coagulation factor IX/X-bp (1ixx), bacterial luciferase (1luc), aspartylglucosaminidase (1apy), and haemoglobin (1qsh). The orthorhombic monellin dimer (4mon) is thought to be coded for by a single gene but there is no definitive evidence to prove this. Proteins whose subunits are more distantly related are CoA mutase (1req), chorionic gonadotropin (1hcn), and 4,5-dioxygenase (1bou). methyl-coenzyme M reductase (1mro) and phenylalanyl-tRNA synthetase (1b7y) are proteins whose subunits contain a common domain that may have been the result of partial gene duplication.

An example of a complex whose subunits are not homologous to each other is heat labile enterotoxin (1tii). The genes encoding the α and β subunits of heat labile enterotoxin are immediately next to each other but the subunits are probably not evolutionary related (Yamamoto et al., 1981). The subunits 1tii only have a SI of ~10% and the structural similarity between them is poor.

4.2.11 Subunit Assembly

Aside from questions of homology it is worth considering briefly why the genes coding for the subunits of the various hetero-complexes are frequently consecutively arranged. One very important advantage of this is to control the expression levels of the individual subunits. For example in prokaryotes it is much easier to control the

expression levels of individual genes if they lie on the same operon. The free large α subunit of p21^{ras} protein farnesyltransferase (1ft1) is toxic in rat cells in notable quantities (Chen et al., 1991). By synthesizing both the α and β subunits of the protein at the same levels to each other this can be avoided. Another reason why the genes coding for most of the hetero-complexes are consecutively arranged and expressed is connected with the probable instability of the subunits in isolation. By synthesising subunits consecutively the risk of any subunits with unstable structures becoming denatured is greatly diminished. This also has the affect of preventing the unwanted aggregation of free subunits. For instance the free subunits of bacterial luciferase (1luc) can form homo-dimers under certain conditions (Seckler in Pain, 2000).

4.3 Non Obligate Hetero-Multimers

In this section an analysis of non-obligate hetero-multimers is presented. An overview of the proteins in these three datasets can be found in chapter 2. Enzyme-inhibitor and antibody-antigen interactions have been intensively studied and the body of literature regarding these interactions is large. Accordingly, in this chapter the principle focus is on characterising the protein-protein interfaces within these protein complexes.

4.3.1 Enzyme-Inhibitors

Some of the highest resolution structures available are those of enzyme-inhibitor complexes and many of the principles governing protein-protein interaction have been elucidated by studying these complexes. In addition an appreciable number of enzymes and inhibitors have been solved in their free and bound states. In this chapter the enzyme and inhibitor parts of the complex are considered separately allowing their individual contributions to the protein-complex to be assessed. The principle object of this section is to look at the way in which enzymes bind to their inhibitors. Accordingly only enzyme-inhibitor interfaces are characterised and any protein interfaces internal to the enzyme or inhibitor are disregarded. As set out in Chapter 2 five out of the twenty entries in the dataset (1avw, 1ldt, 1slu, 1tab, and

3bth) are of trypsin in complex with various inhibitors. To make sure that the large numbers of trypsin complexes do not overly bias any of the characteristics that we use to describe the enzyme interface an average is taken across all the trypsin complexes before taking the average with the remaining enzymes in the dataset. For example the average interface size of the five trypsins is 793\AA^2 . A straight average is then taken over the remaining fifteen entries in the dataset together with the trypsin average of 793\AA^2 to give an average enzyme interface size of 970\AA^2 (a simple average over all twenty enzymes gives a value of 850\AA^2). Except where indicated a trypsin 'average' was used to represent the five trypsins in all the work presented in this section.

4.3.1.1 Size (ASA) of the Enzyme-Inhibitor Interface

In common with the other datasets of non-obligate protein-complexes the enzyme-inhibitor interface is typically rather small with the average interface being 1000\AA^2 . This figure is virtually identical to Conte, 1999 but is higher than the values of 720\AA^2 based upon 11 protease-inhibitor complexes used by Janin & Chothia, 1990, and 785\AA^2 determined by Jones & Thornton, 1995. The enzyme-inhibitor interfaces are significantly smaller than the interfaces to be found in any of the datasets of obligate protein-complexes. Enzymes bury between 4 and 20% of their total ASA in the enzyme-inhibitor interface. The five trypsins fall in the lower end of this range burying about 8% of their ASA at the interface. Three of the five enzymes that bury more than 10% of their ASA are angiogenin, barstar, and uracil DNA glycosylase. Barstar (1brs) as an example buries 13% of its surface area in complex with barnase.

The range for the inhibitors is much larger than for the enzymes ranging from 7% for 1a4y to 51% for 1dp5 reflecting the highly varied sizes of the inhibitors in the current dataset. Half of the inhibitors in the dataset bury >20% of their ASA. Small inhibitors like the IA₃ mutant inhibitor (1dp5), amylase inhibitor (1clv), hirudin (4htc), and hirustasin (1hia) all bury more than a quarter of their surfaces in the inhibitor interface. The small size of the enzyme-inhibitor interface underlies the fact that a large protein-protein interface is not a necessary pre-requisite for a high affinity interaction.

4.3.1.2 Planarity

The interfaces between enzymes and their inhibitors appear to be quite non-planar (see table 4.3). The reason for this is simple. This is because most enzyme active sites are located in clefts (Laskowski & Thornton, 1996). The catalytic active sites of all serine proteases (and most of the enzymes in the dataset are serine proteases) are all within similarly constituted concave pockets as illustrated in figure 2.5 in chapter 2. This makes the enzyme side of the enzyme-inhibitor interface non-planar with the average planarity of such interfaces being 3.3\AA with an average interface ASA of 970\AA^2 (see table 4.3). It should be pointed out that the enzyme side of the enzyme-inhibitor interface can only be said to be on average non-planar given the small average size of the interface (otherwise the average planarity of 3.3\AA for the enzyme dataset is comparable with other classes protein complexes studied in this thesis). For example, a protein-protein interface of 970\AA^2 should have a planarity of 1.9\AA according to the linear relationship between interface size and planarity for the obligate hetero-complexes shown in figure 4.3.

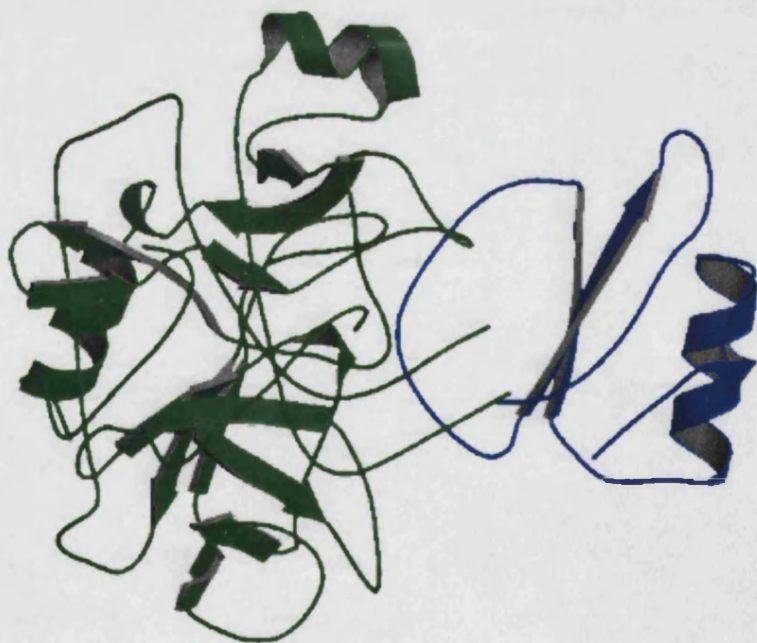


Figure 4.12: The chymotrypsin-Elgin C complex (1acb, Frigerio et al., 1992). The enzyme binding loop of Elgin C shown in blue does not extend fully into the active site of the enzyme and is slightly more planar than the enzyme side of the enzyme-inhibitor interface.

The inhibitor side of the interface is slightly more planar than the enzyme side with the average planarity being 2.7Å. An explanation of this should be made with reference to the fact that an inhibitor need only block access to an active site. A ‘perfect’ fit between enzyme and inhibitor is not necessary. For example the extended loop of elgin C adopts quite a rigid conformation in complex with chymotrypsin (1acb). It is apparent from fig 4.12 that the inhibitor binding loop of elgin C does not extend fully into the active site pocket of the enzyme. The binding loop of elgin C is in fact slightly more planar than the surrounding enzyme having a planarity of 2.05Å compared with 2.9Å for the chymotrypsin side of the interface (Frigerio et al., 1992).

4.3.1.3 Amino Acid Composition

Unsurprisingly the enzyme side of the interface is rich with residues associated with catalysis (see figure 4.13). The catalytic Ser-His-Asp triad of serine proteases is perhaps the best known paradigm of the residues that nature uses to catalyse chemical reactions.

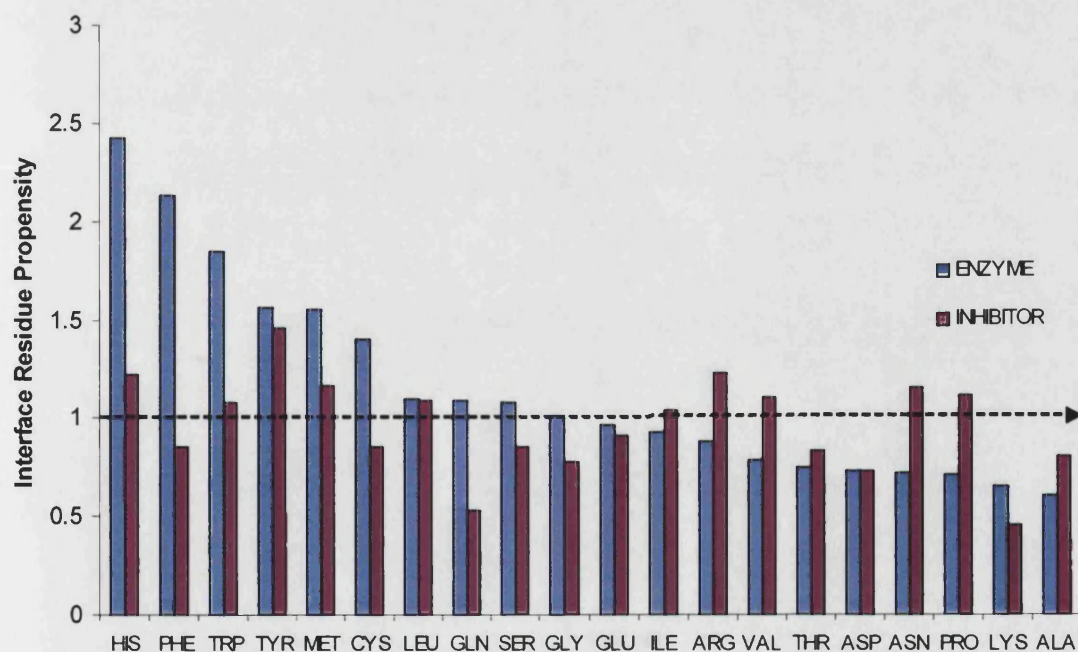


Figure 4.13: The interface residue propensities of the enzyme and inhibitor components of the enzyme-inhibitor dataset. Residues are placed in order of decreasing enzyme interface propensity.

The most abundant residue in the interface is serine which on average makes up 11.8% of all interface residues in the interface compared with 6.8% in proteins generally (Doolittle, 1989). Histidine is also present in large numbers at the interface on average comprising 6.5% of interface residues. The interface propensity of histidine is 2.4, greater than for any other residue. However there is still a prominent place for bulky residues like tyrosine, tryptophan, and phenylalanine in the enzyme interface (see figure 4.13). The interface propensities of these amino acids are all high being 1.6, 1.9, and 2.1 respectively. The occurrence of these bulky residues at the interface underlines the important role that these residues play in not only sticking large protein molecules together but also in binding to small substrates molecules during catalysis.

PDB Code	Enzyme interface				Inhibitor Interface				K _I (M)
	Charged (%)	Polar (%)	Hydrophobic (%)	Hydrophobicity	Charged (%)	Polar (%)	Hydrophobic (%)	Hydrophobicity	
1a4y	35.1	35.1	29.7	-0.05	34.9	51.2	14.0	0.17	7.1×10^{-16}
1acb	4.0	52.0	44.0	0.71	17.7	29.4	53.0	0.48	2.7×10^{-11}
1brs	31.8	45.5	22.7	0.14	27.8	33.3	38.9	0.43	1.0×10^{-13}
1clv	22.9	37.1	40.0	0.38	20.0	50.0	30.0	0.35	-
1dp5	13.5	34.6	51.9	0.58	27.6	31.0	41.4	0.19	$3 \pm 0.6 \times 10^{-9}$
1dtd	20.0	52.0	28.0	0.36	22.2	27.8	50.0	0.59	$0.17-0.78 \times 10^{-9}$
1fle	20.8	45.8	33.3	0.34	15.8	26.3	57.9	0.76	1.0×10^{-9}
1hia	3.6	53.6	42.9	0.66	31.3	31.3	37.5	0.51	13×10^{-9}
1smp	11.1	63.9	25.0	0.12	25.0	40.0	35.0	0.11	$\sim 10^{-6}$
1stf	6.7	50.0	43.3	0.54	9.5	33.3	57.1	0.43	120×10^{-12}
1ugh	16.7	50.0	33.3	0.19	13.8	41.4	44.8	0.45	1.3×10^{-6}
1viw	24.4	36.6	39.0	0.27	19.4	58.3	22.2	0.17	-
2sic	7.7	46.2	46.2	0.28	28.6	35.7	35.7	0.43	5×10^{-12}
4htc	37.8	33.3	28.9	0.29	27.3	36.4	36.4	0.28	2.0×10^{-14}
4sgb	12.5	41.7	45.8	0.21	6.7	53.3	40.0	0.64	-
Trypsins*	11.3	62.7	26.1	0.39					
1avw	10.3	58.6	31.0	0.40	33.3	33.3	33.3	0.32	350×10^{-9}
1ldt	4.4	65.2	30.4	0.49	35.7	14.3	50.0	0.45	0.9×10^{-9}
1slu	18.0	56.4	25.6	0.35	30.3	36.4	33.3	0.19	22×10^{-9}
1tab	14.8	66.7	18.5	0.24	21.4	57.1	21.4	0.19	-
3bth	4.2	70.8	25.0	0.53	15.4	30.8	53.9	0.56	6.2×10^{-6}
Average	16.0	49.8	34.3	0.34	22.6	37.5	39.9	0.39	

Table 4.13: The percentages of residues in the enzyme and inhibitor interface that are charged, polar, and hydrophobic. The hydrophobicity of the enzyme and inhibitor interfaces is given. The dissociation constant of each enzyme-inhibitor complex (where available) are also provided.

* The interface composition and hydrophobicity for the five trypsins (1avw, 1ldt, 1slu, 1tab, 3bth).

Table 4.13 shows the averaged composition of the enzyme interface in terms of charged, polar, and hydrophobic residues. Histidine is included in the set of charged residues in table 4.14 to allow comparison with the work of Bartlett et al., 2002. Bartlett analysed the catalytic residues in the active sites of some 178 enzymes. The results are included in table 4.13 to allow for comparison. Catalytic residues are so few in number (usually 3-4) compared with the total number of residues that make up the interface that there is no reason to expect a good correlation between the amino acid compositions of the active site and interface. In terms of numbers the enzyme interface is typically dominated by polar residues. Polar residues typically constitute some 43% of all enzyme interface residues with the figure for charged residues being 23% (table 4.3).

	Charged (%)	Polar (%)	Hydrophobic (%)
Enzyme Interface	23	43	34
Catalytic Residues	65	27	8
Inhibitor Interface	25	35	40

Table 4.14: The mean percentage of residues at the enzyme and inhibitor interfaces that are charged, polar, and hydrophobic. Histidine is included in set of charged residues. Values are also given for catalytic residues from 178 enzyme active sites (Bartlett et al., 2002).

It is interesting to note that the amino acid composition of the enzyme exterior (including interface residues) is also quite polar. This is in marked contrast to the datasets of obligate homo-proteins studied in chapter 3 in which the fraction of hydrophobic residues in the interface is usually higher. The utilization of such large numbers of polar and charged residues in and around the interface as well as being essential from the point of view of catalysing a reaction is also useful in marking out regions of the enzyme exterior involved in substrate binding from those that are not. A vivid example of this is D-dopachrome tautomerase (1dpt) shown in figure 4.14.



Figure 4.14: A diagram of the D-Dopachrome tautomerase trimer generated using GRASP (Nicholls et al., 1991). Regions of negative electrostatic potential are shown red with regions of positive potential being shown in blue. The three active sites of the trimer (marked with an *) are clearly distinguishable from the remainder of the enzymes surface.

Regions of positive potential are shown in blue and negative regions in red. The active sites of the enzyme are marked with yellow stars. The rate at which a substrate diffuses to the active site(s) of an enzyme can be considerably enhanced by the presence of a few charged residues around the periphery of the active site as are seen in 1dpt. These charged residues collectively create an electrostatic ‘funnel’ that draws a substrate towards an active site and holds it in the proper conformation during catalysis. Computer simulations of superoxide dismutase indicate that a region of positive potential surrounding its active site enhances the rate of association of the enzyme with its superoxide anion by a factor of 30 or more (Sharpe et al., 1987). Barnase is another well-studied example of this effect (Schreiber & Fersht, 1996).

The amino acid composition of the inhibitor side of the interface is fairly atypical when compared to the interfaces of the obligate homo and hetero-complexes (see figure 4.21). Loops are particularly favoured at the interfaces of the inhibitors within the current dataset, for instance the canonical binding loops of the eleven serine protease inhibitors (Apostoluk & Otewski, 1998). Accordingly the kinds of residues that are found at the interface in this dataset are those that are also found in loops

generally. Proline in one such residue. Proline makes the single biggest contribution to the inhibitor interface on average comprising 7.9% of all interface residues. Proline is the least flexible of all the twenty amino acids due to its side chain being bound to the main chain of the acid forming a ring structure (see figure 1.4 in chapter 1). Some of the inhibitors in the dataset are small in size and contain few secondary structures at the interface with their enzymes other than loops. It is probable that the inclusion of significant numbers of proline residues in parts of the inhibitor that make contact with the enzyme help to maintain the structural integrity of the interface. The inclusion of proline residues in the loop regions of oligo-1,6-glucosidase has been found to stabilise the enzyme in extreme environments and this could be true of proteins generally (Watanabe et al., 1991, 1997).

After proline, serine and arginine are the two most abundant amino acids comprising some 14% of all interface residues. Arginine residues at the interface either form salt bridges or interact with the aromatic groups of phenylalanine, tyrosine, or tryptophan. Serine residues help to maintain the extended network of hydrogen bonds that maintain the conformations of the loop regions at the interface. Cysteine is the fourth most commonly occurring residue at the interface. Most of these cysteine residues form disulphide bonds across the enzyme-inhibitor interface. These bonds enhance the stability of the enzyme-inhibitor complexes (the majority being extra-cellular proteins). A good example is the hirustasin-kallikrein complex in which 1 disulphide bond and two salt bridges between enzyme and inhibitor hold the complex together (Mittl et al., 1997).

The amino acids that have the highest interface propensities for the inhibitors are tyrosine (1.46) followed by arginine (1.23), histidine (1.22), and methionine (1.16) these can be seen in figure 4.13. Apart from methionine these are all polar or charged amino acids as are those on the enzyme side of the interface.

4.3.1.4 Hydrophobicity

The hydrophobicities of the interior, exterior, and interface regions of all the enzymes and inhibitors have been calculated separately using the scale of Fauchere and Pliska in the same way as described in the previous chapter. In line with nearly all other types of proteins both enzymes and inhibitors have a hydrophobic interior and an interface that is intermediate in hydrophobicity between the interior and exterior (see figure 4.19 shown in section 4.4). Hence the interfaces of both inhibitors and enzymes are formed from hydrophobic regions on the exterior of the protein.

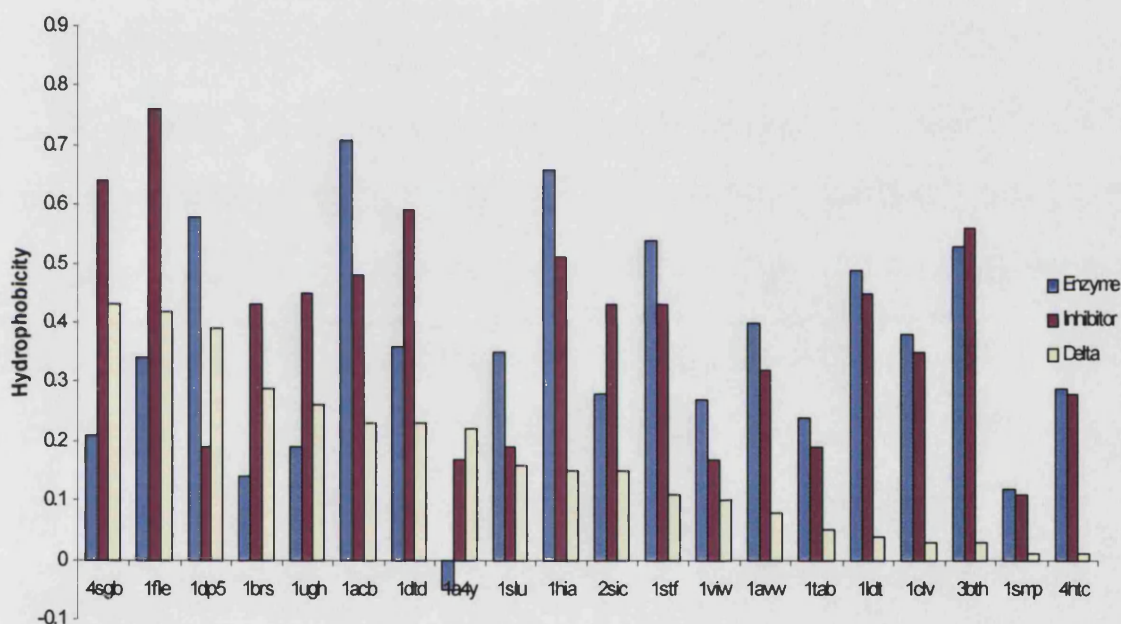


Figure 4.15: The hydrophobicities of the enzyme and inhibitor protein-protein interfaces of all 20 enzyme-inhibitor complexes. For each enzyme-inhibitor complex the difference in hydrophobicity between the enzyme and inhibitor side of the protein-protein interface is also shown as the yellow 'Delta' bar.

The hydrophobicities of each enzyme and inhibitor interface are shown in figure 4.15. What is striking from this plot is that there is quite frequently a significant difference in hydrophobicity between the enzyme and inhibitor sides of the interface. In twelve out of the nineteen complexes the enzyme interface is more hydrophobic than the inhibitor interface. In the remaining eight complexes the reverse is true. Interestingly some of the complexes in which there is a very large difference in hydrophobicity between enzyme and inhibitor interfaces are those that bind together with the

strongest affinities as measured by dissociation constants (K_d's). This is further discussed in section 4.3.1.7.

4.3.1.5 Hydrogen Bonding

The network of hydrogen bonds between the enzymes and their inhibitors has been analysed using HBPLUS as described in section 3.8 in chapter 3. The entries in the dataset are divided into three separate classes of enzyme-inhibitor complex (a) serine proteases, (b) other proteases, and (c) miscellaneous enzyme complexes as shown in table 4.15.

	PDB Codes
Serine Proteases	1acb, 1fle, 1hia, 2sic, 4htc, 4sgb, 1avw, 1ldt, 1slu, 1tab, and 3bth
Other Proteases	1dp5, 1dtd, 1smp, 1stf
Miscellaneous Enzyme Complexes	1a4y, 1brs, 1clv, 1ugh, 1viw

Table 4.15: The twenty enzyme inhibitor complexes classified as being (a) serine proteases (b), 'other' proteases, (c) and miscellaneous enzyme complexes.

The overall numbers and type of hydrogen bonds that exist between the enzymes and inhibitors in each of the three classes of enzyme-inhibitor complex are summarised in table 4.16

	Type of Hydrogen Bond					
	MM	(%)	SS	(%)	SM	(%)
Serine Protease	6.0	64.1	1.5	12.6	2.5	23.3
Other Proteases	1.3	23.6	2.0	24.3	4.0	52.1
Miscellaneous	0.2	2.5	6.6	59.5	4.4	38.0
All	3.0	34.7	3.2	30.2	3.5	35.1

Table 4.16: Inter-subunit hydrogen bonds analysed by type. The average numbers and percentage frequencies of inter-subunit hydrogen bonds between main-chain groups (MM), side-chain groups (SS), and between main-chain and side-chain groups (SM) are given for the enzyme-inhibitor dataset.

For the serine proteases there are typically 1.17 hydrogen bonds for every 100Å² of ASA buried at the interface. This figure is lower than some of the estimates obtained

with previously published data. Using a dataset of 15 protease-inhibitor complexes Janin & Chothia, 1990 found that the value to be 1.47. More recently Conte et al., 1999, found that there are 1.24 hydrogen bonds per 100Å² of buried ASA using a dataset of 19 protease-inhibitor complexes.

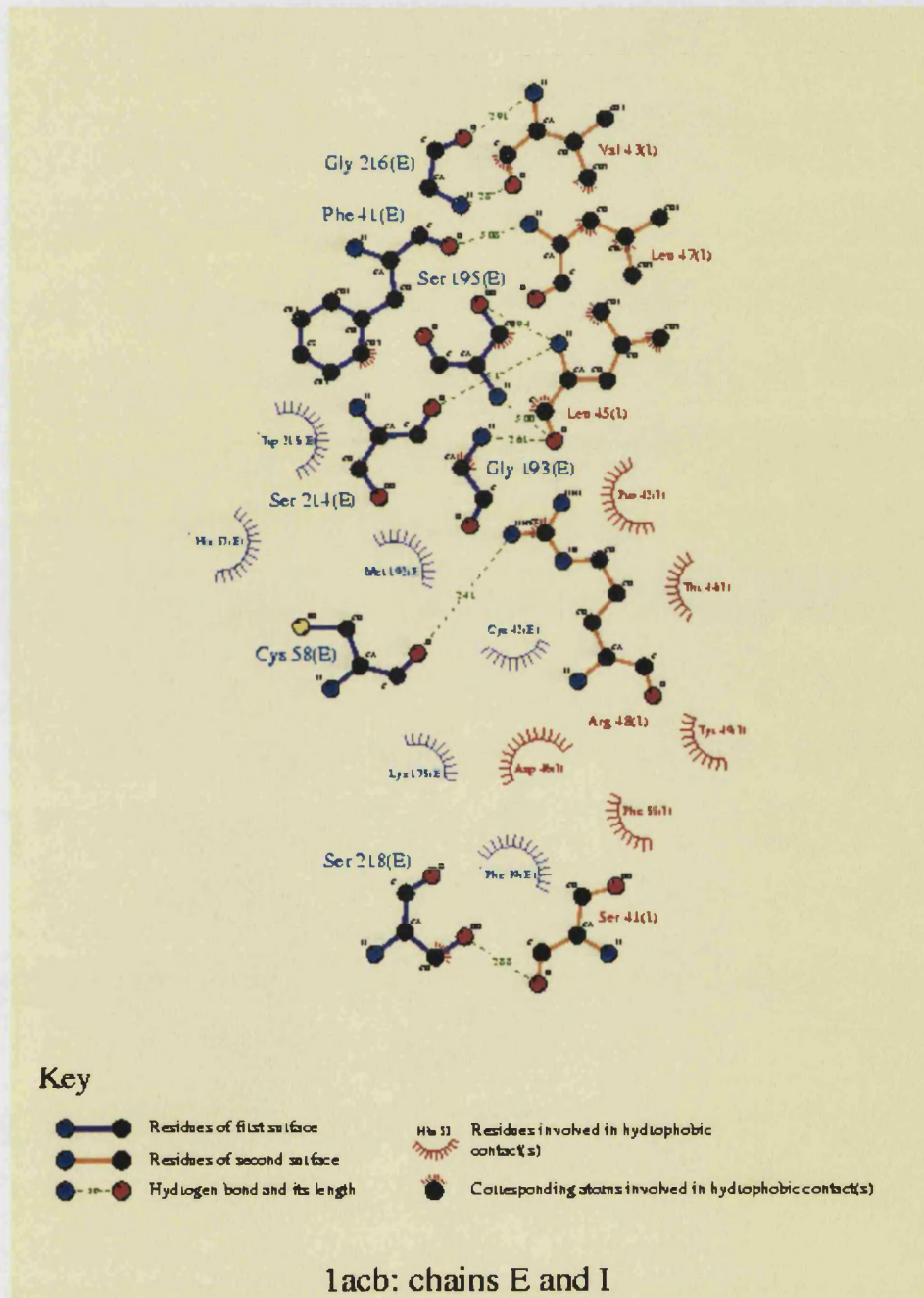


Figure 4.16: A diagram showing the pattern of hydrogen bonding at the chymotrypsin Elgin C interface generated using DIMPLLOT (Laskowski et al., 1996).

However it is clear that the interfaces between the serine protease and their various inhibitors are significantly enriched with hydrogen bonds. A striking property of the hydrogen bonds across the interface is that two thirds of them are between main chain atoms (table 4.16). This is because inhibitors bind to proteases in the same way that substrates do. A diagram (figure 4.16) of interactions between chymotrypsin and elgin C has been generated using DIMPLOT (Laskowski, 1996). The diagram illustrates that 6 out of a total of 9 hydrogen bonds between enzyme and inhibitor are between main chain atoms. The remaining 3 hydrogen bonds are between side chain and main chain atoms.

Overall the proteins falling into the 'other protease' classification seemingly contain fewer inter-subunit hydrogen bonds than the eleven serine proteases. On average these four proteins have 0.81 hydrogen bonds per 100\AA^2 of ASA at the interface. But this figure may be misleading and not representative of protease-inhibitor interactions. Two of the proteins (1dp5 and 1stf) have relatively few hydrogen bonds between enzyme and inhibitor. As set out in chapter 2, the IA₃ mutant inhibitor (1dp5) is little more than a truncated α -helix and the manner in which the inhibitor binds to aspartate proteinase A is highly unusual and possible unique (Li et al., 2000). In the papain and stefan B complex (1stf) water molecules mediate a large number of interactions between the enzyme and inhibitor. In fact there are a total of 17 hydrogen bonds between enzyme and inhibitor mediated by 13 solvent molecules (Stubbs et al., 1990).

The remaining miscellaneous enzyme-complexes on average contain 1.05 hydrogen bonds per 100\AA^2 of interface ASA. This is similar to that observed for the datasets of obligate proteins. Although the average is 1.05 the range is quite large varying from 0.73 for 1a4y to 1.66 for 1brs. Again the actual number of hydrogen bonds between enzyme and inhibitor is probably higher if indirect hydrogen bonds are considered. This is certainly the case for the angiogenin-RNase inhibitor complex (1a4y) in which 15 hydrogen bonds are mediated by water molecules at the enzyme-inhibitor interface (Papageorgiou et al., 1997).

4.3.1.6 Shape Complementarity

The two surfaces that make up the enzyme-inhibitor interface are generally highly complementary in shape. This is expected from the high affinities with which inhibitors and enzymes bind to one another (see section 4.3.1.7). The shape complementarity statistics (S_c) of the enzyme-inhibitor interfaces by and large reflect this. Lawrence & Colman, 1993, found that the average S_c of four enzyme-inhibitor complexes (all of them serine proteases) is 0.73. The averaged S_c of the eleven serine proteases in the present dataset is 0.72 with all but two of the serine proteases having a $S_c > 0.70$. This is unremarkable since as previously mentioned all serine proteases have similarly constituted active sites. The other four protease complexes have S_c values ranging from 0.66 to 0.69 with the average being 0.68. The remaining six enzyme complexes have enzyme-inhibitor interfaces that are generally less complementary in shape than the various proteases. The α -amylase *Phaseolus vulgaris* complex (1viw) has the lowest S_c value in the entire dataset of 0.55 but in contrast the barnase-barstar complex has one of 0.72.

4.3.1.7 Strength of Interaction

Enzymes and inhibitors bind together to form some of the most tightly bound protein complexes to be found in nature. Kinetic data is available for the majority of the enzyme-inhibitor complexes and the dissociation constants for each complex (where available) are listed in table 4.13. The dissociation constants of the enzymes and inhibitors vary over ten orders of magnitude from 10^{-6} to 10^{-16} M. However the binding constant of any complex can only be fully explained with reference to a highly detailed anatomy of the enzyme-inhibitor interface. The trypsin BPTI complex (3bth) is a good example of this. Helland et al., 1999 produced ten different variants of BPTI each with a single point mutation in the enzyme binding loops of the inhibitor. The binding constants of these ten mutant complexes varied tremendously from 1.5×10^4 to 1.7×10^{13} M⁻¹. This effectively shows that very minor changes in the enzyme or inhibitor interface can dramatically affect the binding constant of the complex. A generalised statistical survey of the enzyme-inhibitor interface such as is presented here is then unlikely to be sufficient to resolve the level of detail required to

explain either these huge variations in binding constant (or even the binding constant of the complex itself). Despite this the interfaces of some of the most tightly bound enzyme-inhibitor complexes in the dataset do display certain characteristics.

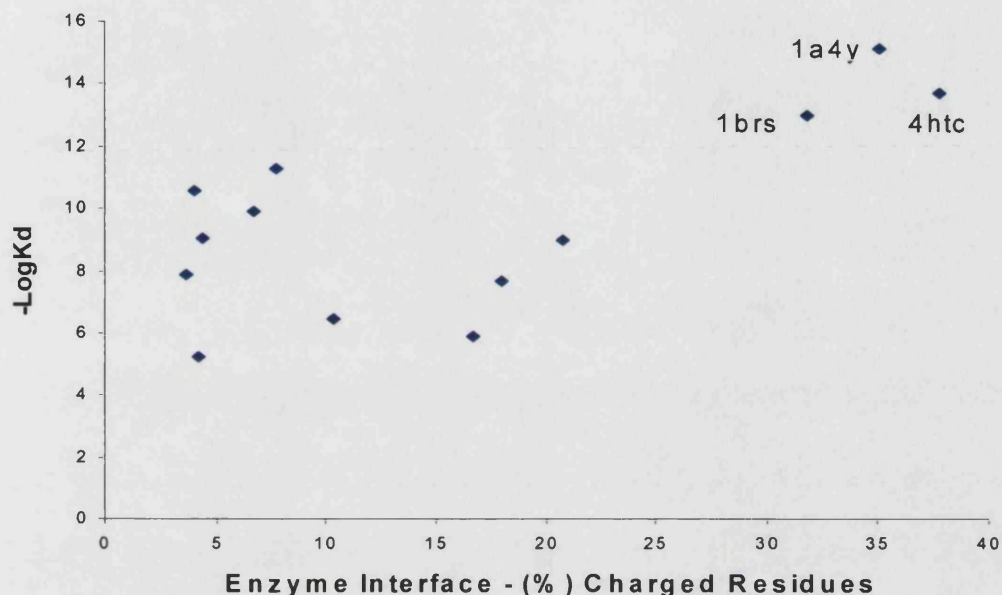


Figure 4.17: The logarithm of the dissociation constant (K_d) of 13 of the enzyme-inhibitor complexes plotted together with the percentage of residues that are charged in the enzyme side of the enzyme-inhibitor complexes. Complexes with very low dissociation constants (particularly 1brs, 1a4y, and 4htc) tend to have large fractions of charged residues in the enzyme side of the enzyme-inhibitor interface.

A plot of the percentage of charged residues in the enzyme interfaces against $-\log K_d$ of the complex is given in figure 4.17. From this chart the three enzyme-inhibitor complexes with the lowest dissociation constants are 1a4y, 4htc, and 1brs having K_d 's of 7.1×10^{-16} , 2×10^{-14} , and 10^{-13} M respectively. The enzyme interfaces within these three complexes all contain large numbers of charged residues ranging from 32% for 1brs to 38% for 4htc. As a point of comparison charged residues (excluding histidine) on average comprise 17.5% of enzyme interface residues. The inhibitor side of the interface of 1a4y, 4htc, and 1brs also contain significant numbers of charged residues. This leads to the perhaps unsurprising inference that interactions between charged residues from both enzyme and inhibitor do make an important contribution to the binding constant of the complex.

4.3.2 Antibody-Antigen Complexes

Antibodies are capable of recognising and binding to an almost infinite range of antigens in a highly specific manner. Accordingly the interactions between antibodies and antigens have been extensively reviewed (Wilson & Stanfield, 1994, Davies & Cohen, 1996, Thornton et al., 1996, and Decanniere et al., 2001). As with the enzyme-inhibitor complexes the antibody and antigen components of the antibody-antigen complex are considered separately. Only antibody-antigen interfaces are considered here. The interface between the heavy and light chains of each antibody is disregarded as are any interfaces internal to the antigen.

The contact area between antibody and antigen is on average around 870\AA^2 (table 4.3). This value is about 100\AA^2 higher than was reported by Jones & Thornton, 1995 but only 40\AA^2 less than Conte et al., 1999. The interfaces between antibodies and their protein-antigens are very planar with a mean planarity of 1.7\AA . This value is lower than any of the other category of protein-complex studied in this thesis. There are about 0.9 hydrogen bonds between antibodies and antigens for every 100\AA^2 of buried ASA. This is likely to be an underestimate since water molecules are known to pack at the antibody-antigen interface mediating hydrogen bonds between them.

The antibody interface includes large numbers of aromatic residues. Including histidine, aromatic residues on average comprise 29% of all residues that make up the antibody interface. This is slightly lower than the figure of 34% calculated using six different antibody-antigen complexes by Davies & Cohen (1996). Tyrosine and tryptophan are particularly prevalent in the antibody interface with interface propensities of 4.98 and 4.34 respectively. Other aromatic residues also prominent at the antibody-antigen interface are histidine having a propensity of 1.39 and phenylalanine with a propensity of 1.22. Numerically antibody interfaces are rather rich in polar residues which on average constitute a total of 57.9% of all residues in the interface. Serine and tyrosine residues together comprise almost 30% of the residues from the antibody side of the interface (see table 4.17). Despite this the antibody interface is still broadly hydrophobic as shown in figure 4.19.

PDB Code	Antibody interface				Antigen Interface			
	Charged (%)	Polar (%)	Hydrophobic (%)	Hydrophobicity	Charged (%)	Polar (%)	Hydrophobic (%)	Hydrophobicity
1ahw	20.6	55.9	23.5	0.21	41.4	34.5	24.1	-0.05
1dqj	12.0	84.0	4.0	0.23	36.4	45.5	18.2	-0.10
1e6j	5.6	77.8	16.7	0.42	26.7	20.0	53.3	0.22
1egj	19.1	42.9	38.1	0.33	31.3	46.7	26.7	0.21
1fdl	23.5	58.8	17.7	0.37	31.3	37.5	31.3	-0.11
1fns	21.1	57.9	21.1	0.40	41.7	33.3	25.0	-0.24
1g9m	26.7	26.7	46.7	0.49	18.8	31.3	50.0	0.62
1jrh	31.8	50.0	18.2	0.36	37.5	39.4	12.5	0.03
1kb5	14.8	63.0	22.2	0.28	27.3	39.4	33.3	0.30
1mlc	13.6	68.2	18.2	0.25	23.5	47.1	29.4	-0.01
1nca	20.0	63.3	16.7	0.32	37.5	25.0	37.5	0.19
1nfd	25.0	37.5	37.5	0.33	52.9	23.5	23.5	-0.10
1nsn	11.1	74.1	14.8	0.33	40.7	33.3	25.9	-0.11
1qfu	18.5	48.2	33.3	0.45	37.0	22.2	40.7	0.12
2jel	8.7	60.9	30.4	0.33	36.8	42.1	21.1	-0.12
Average	18.1	57.9	23.9	0.34	34.7	34.7	30.2	0.06

Table 4.17: The percentages of residues in the antibody and antigen interface that are charged, polar, and hydrophobic. The hydrophobicity of the antibody and antigen interfaces has also been calculated using the Fauchere and Pliska scale, 1983.

The single most striking characteristic of the antigen interfaces are how polar they are. In most of the antibody-antigen complexes the antigen interfaces are extremely polar in nature as can be seen in table 4.17. The interface has a hydrophobicity of 0.06 compared with 0.14 for the exterior, emphasizing this fact. In fact antigens are the only class of proteins in which the interface is more polar than any other part of the protein. Charged residues comprise a third of the residues at the interface. Arginine alone makes up 8.2% of all interface residues with the figure being 11.3% for lysine residues.

The rapid rate at which antibodies evolve in response to antigenic infection has been invoked as an explanation of the poor shape complementarity that exists between antibodies and antigens compared with other categories of protein complex (Lawrence & Colman, 1993). The shape complementarity statistic (S_c) has been calculated for every antibody-antigen interface. The averaged S_c statistic for all antibody-antigen complexes works out to be 0.66 compared with 0.69 for the enzyme-inhibitor complexes detailed in the previous section. Previous authors have pointed out that antibodies bind to antigens in an asymmetric way (Davies et al., 1990). The binding is asymmetric in the sense that antibodies frequently appear to bind to their antigens using more residues from the heavy than the light antibody chain. In the current dataset there are a number of instances of this. The complex between the monoclonal antibody D44.1 and hen egg-white lysozyme (1mlc) is a good example. The D44.1 antibody binds to lysozyme using almost exclusively residues from its heavy hyper-variable loop (Braden et al., 1994).

The presence of so many aromatic residues at the antibody interface and the polar character of the antigen interfaces points to interactions between aromatic residues and polar or charged groups from the antigen being of fundamental importance in antibody-antigen interactions. The importance of such interactions has previously been established in the HyHEL10 and hen egg-white lysozyme complex (Tsumoto et al., 1995) and other antibody-antigen complexes (Hofstadter et al., 1999 and references therein). The role of aromatic residues in molecular recognition and protein-protein interactions in general has been well studied (Gallivan & Dougherty, 1999 and references therein). Aromatic residues can be involved in both hydrophobic and polar

interactions. Tryptophan and phenylalanine are both bulky and hydrophobic amino acids. The burial of these residues at a protein-protein interface therefore produces a large amount of free energy. Both tyrosine (through its 4' hydroxyl group) and tryptophan can form hydrogen bonds. Both of these residues have a de-localised ring of π electrons that may take part in further interactions with the cation groups of neighbouring amino acids. Most of these interactions have quite substantial binding enthalpies and are most commonly observed to take place between the NH_4^+ of lysine and arginine and the aromatic rings of tryptophan, tyrosine, and phenylalanine. Since the antibody-antigen interface is enriched in all of these residues it is reasonable to assume that they interact in the ways just described. In short the abundance of aromatic residues (able to participate in hydrophobic and polar interactions) in the antigen binding loops of an antibody enable the antibody to bind effectively to a huge range of antigens with the high affinities that are observed in vivo.

4.3.3 Signalling Proteins

The dataset of signalling proteins contains an extremely diverse range of protein complexes. All of the proteins within this dataset are fragments. For this reason the conclusions that can be made regarding the nature of the binding surfaces within the signalling proteins are extremely limited. The amount of surface area buried amongst the signalling proteins varies widely from 286\AA^2 to 2678\AA^2 . This range is greater than that reported by Conte et al (1999) of 1000\AA^2 with the mean being 1250\AA^2 . The chemical compositions of the interface areas within the signalling proteins appear to be quite similar to those within the other datasets of non-obligate protein complexes (see figure 4.20 in section 4.5). The residues most likely to be found at the protein-protein interface all have quite large side-chains. Phenylalanine has the highest residue interface propensity (1.55) followed by arginine (1.32) and tyrosine (1.31). The protein interfaces are broadly hydrophobic and intermediate in hydrophobicity between the interior and the exterior of the protein. The protein-protein interfaces in the signalling dataset are among the least planar in any of the other non-obligate datasets. The planarity of the protein-protein interfaces in the complexes of signalling proteins ranges from 1.6 to 5.4\AA with the average being 3\AA (see table 4.3).

Blundell et al., 2000, show that conformational changes are commonplace in signalling complexes with protein-protein interfaces of 1500 to 2000Å² in size. For instance the conformational changes that take place during the formation of the phosphatidylinositol-3-OH kinase transducin-βγ complex are outlined in chapter 2.

The non-planar nature of some of these protein interfaces may be a result of the extended structures of some of the interacting proteins, and the conformational changes that take place between them on binding. In conclusion we still do not have a large enough number of sufficiently complete structures to determine any characteristic attributes of binding surfaces within the dataset of signalling proteins. Further structural information will be required to better characterise the proteins implicated in signalling processes.

4.4 Comparison of Obligate Vs Non-Obligate Hetero-Complexes

The bulk properties of all 40 non-obligate protein-complexes together are considered here. These forty complexes are the enzyme-inhibitors, antibody-antigens, and signalling proteins that were characterised in section 4.3. The protein-protein interfaces within these complexes are quite distinctive and are easily distinguishable from the protein-protein interfaces in the obligate datasets.

The protein-protein interfaces within the non-obligate protein complexes are much smaller than those within the obligate complexes. This is instantly apparent from tables 4.3 and 4.4. The actual average interface size for the non-obligate protein complexes is ~980Å² in size compared with 2380Å² for the obligate homo-complexes and 3730Å² for the obligate hetero-complexes. A small interface size appears to be the best indicator that the complex is non-obligate rather than obligate. As noted by Nooren & Thornton, 2003, it is probably much easier to control how two proteins associate or dissociate if there is only a small interface between them. With a small interface only a small number of contacts between proteins need be broken to break up a complex into its component proteins. Regulation of assembly is of obvious importance in interactions between enzymes and their inhibitors.

The interfaces in the non-obligate datasets are quite flat. Almost two thirds of all the interfaces within these complexes have a planarity $< 2.5\text{\AA}$. Most of the interfaces with a planarity that is higher than 2.5\AA are from the signalling proteins. The fact that the interfaces are relatively planar is an indication that the proteins within non-obligate complexes do not ‘interlock’ with each other as the obligate hetero-proteins do (see the diagrams of the obligate hetero-complexes in the appendix). This together with the small interface size allows complexes to be assembled or disassembled with relative ease, and with few structural changes in response to changes in the local environment (Noreen & Thornton, 2003).

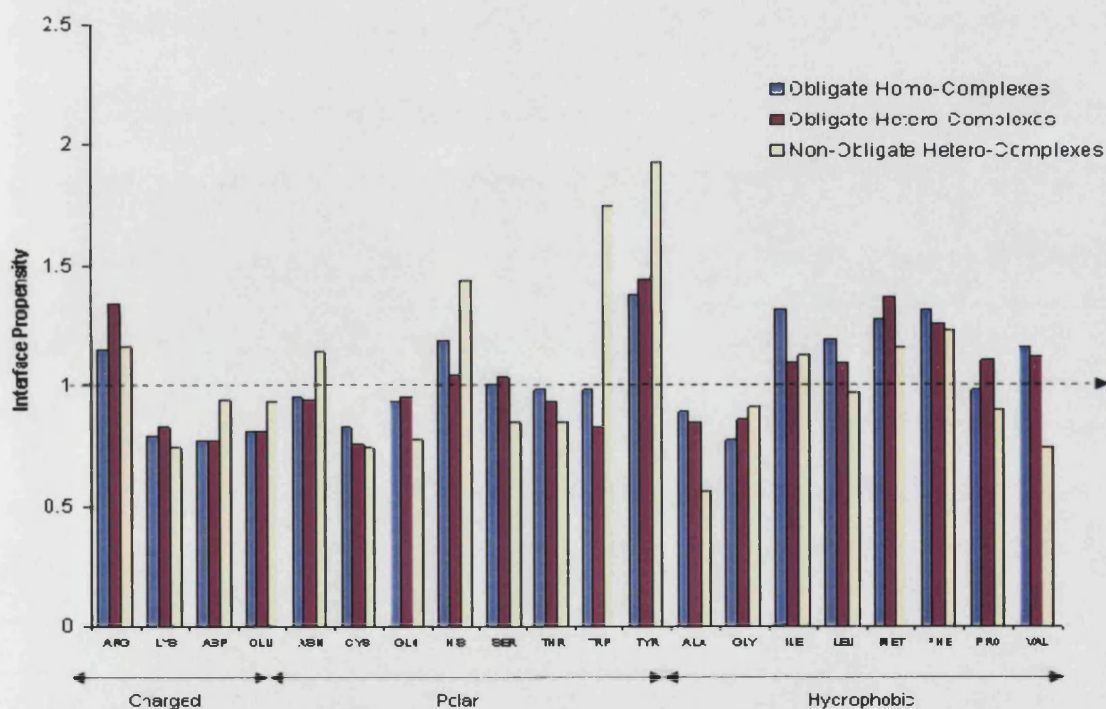


Figure 4.18: Mean residue interface propensities for the datasets of obligate homo-complexes, obligate hetero-complexes, and non-obligate hetero-complexes.

Polar and charged residues are particularly prevalent at the interface. Polar residues on average make up 41% of all interfacial residues compared with 30% for the obligate homo-complexes and 28% for the obligate hetero-complexes (see table 4.4). The interface propensities of the non-obligate datasets also reveal a marked preference for large, charged or polar residues at the interface (see figure 4.18).

Tyrosine has the highest interface propensity (1.93), followed by tryptophan (1.75), histidine (1.44), phenylalanine (1.24), and arginine (1.16). It is noteworthy that tyrosine, tryptophan, and arginine have been described as 'hot spots' of protein binding and are thought to contribute disproportionately to the free energy of binding (Bogan & Thorn, 1998). Furthermore the locations of these residues appear to be conserved in the interfaces of many protein-complexes particularly enzyme-inhibitors (Hu et al., 2000). That the interfaces of the non-obligate protein complexes are enriched in residues that apparently yield large amounts of free energy upon burial at the interface in part explains how the proteins within the non-obligate datasets can bind to each other so strongly through such small contact areas (Sheinerman et al., 2000).

There are numerous other explanations of why the interfaces of non-obligate complexes include quite large numbers of polar and charged residues. The most obvious reason is that proteins need to bind to each other with high affinity but in a reversible way. For the most part charged or polar residues interact across the interface forming salt bridges, hydrogen bonds (directly or via water molecules), or a variety of other electrostatic interactions. The strengths of any of these electrostatic interactions are highly dependent on the pH of the local environment. Thus simply moving a complex to regions within the cell with a different pH is an effective way of breaking up what would otherwise be a tightly bound protein complex (Price & Stevens, 1999).

The second reason is that the electric fields created by charged or polar residues on the protein exterior can considerably enhance the rate with which proteins associate with one another to form a complex. The most studied example of this is the barnase-barstar complex. A Brownian dynamics simulation of barnase and barstar compared with experimental data showed that the association rate of the enzyme and inhibitor is very much increased by varying the number of polar and charged residues at the interface (Gabdoulline & Wade, 1999, Schreiber & Fersht, 1995 and 1996). Subsequent studies of other proteins point to the 'electrostatic energy of interaction between proteins in a complex correlating strongly with the rate of association' (Sheinerman et al., 2000, Selzer & Schreiber, 1999).

The third reason is that the presence of charged and polar residues at the interface can and does contribute considerably to the specificity of an interaction. Most interactions between enzymes and inhibitors and antibody interactions are highly specific due in no small way to clusters of polar and charged residues at the interface. The final reason is stability. As described in detail in chapter 3 there is good evidence that salt bridges and other electrostatic interactions between residues across the interface stabilise protein-complexes in harsh environments.

The hydrophobicities of the interior, exterior, and interface regions of the non-obligate complexes follow the same trends as the datasets of obligate protein-complexes. With the notable exception of antigens the interface has a hydrophobicity that is intermediate between that of the interior and exterior as shown in figure 4.19.

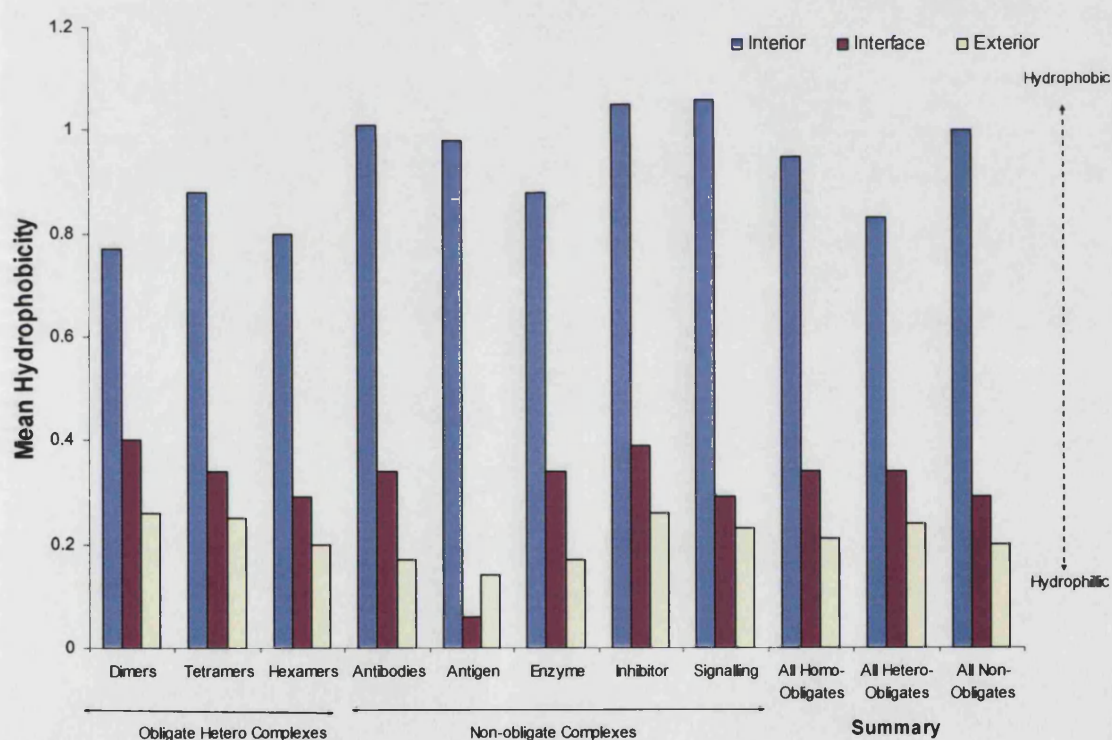


Figure 4.19: Mean hydrophobicities of the interior, interface, and exterior regions of the datasets used in this thesis. With the exception of the dataset of antigens the interface is intermediate in hydrophobicity between the protein interior and exterior.

The interfaces of all the non-obligate proteins have an averaged hydrophobicity of 0.29 compared with 0.34 for the obligate hetero and homo-complexes. From figure

4.19 the exteriors of the non-obligate proteins both individually and collectively are also less hydrophobic than those of the obligate complexes. This all serves to underline the hydrophilic nature of the exteriors and interfaces of the non-obligate proteins compared with obligate protein complexes.

Averaged across all non-obligate proteins there are 0.89 hydrogen bonds per 100\AA^2 of buried ASA. This figure is actually slightly lower than the obligate homo-complexes (0.94) and hetero-complexes (1.05). But the standard deviations on these values are large (see table 4.4). What is significant is that for all types of protein complex the number of direct hydrogen bonds across an interface is directly proportional to its size. As previously mentioned, water molecules are known to mediate indirect hydrogen bonds across the interface. From studies of high resolution structures it is known that solvent plays a vital role in protein-protein recognition. For instance in the D1.3-HEL structure there are 25 water molecules that mediate hydrogen bonds between the antibody and antigen (Branden & Pojak, in Kleanthous, 2000). Using 36 protein structures solved to a resolution $< 2.4\text{\AA}$ Conte et al., 1999 estimate that water mediated polar interactions are more numerous than direct hydrogen bonds and that there is ~ 1 water molecule per 100\AA^2 of interface. Further examination of high resolution protein structures will be needed to determine the details of water mediated hydrogen bonds at protein interfaces.

4.5 Comparison of Homo Vs Hetero Obligate-Complexes

There are some distinct differences between the protein-protein interfaces found in the obligate homo and hetero-complexes. However, again the overriding difficulty in making this comparison is the small number of hetero-complexes (20) compared with the homo-complexes (142). Although the evidence (see table 2.12 in chapter 2, Godsell & Olson, 2000) is that obligate hetero-complexes are much less prevalent in nature than homo-complexes it does seem probable that hetero-complexes are under-represented in the current work due to the lack of protein structures in the PDB. More structures of obligate hetero-complexes will be needed to make sure of the conclusions presented in this section.

Hetero-complexes on average bury a much larger amount of ASA in protein-protein interfaces than homo-complexes (3730 versus 2380Å²). The fraction of ASA buried in interface regions is also typically larger for the hetero-complexes (29%) than homo-complexes (19%). But the standard deviations on both of these figures in table 4.4 are large meaning that no definite significance can be attached to this. The hetero-complexes appear to have extensive interfaces because of the quite extended and inherently unstable structures adopted by the constituent protomers of many of the hetero-complexes. The protomers of the homo-complexes of course do adopt a wide variety of conformations, but they are for the most part more globular in shape than those that make up the hetero-complexes. Consequently there is less of a need for a large interface to make a stable complex within homo-complexes than there is for hetero-complexes.

In all classes of protein complex the planarity of the interface varies linearly with the size of the interface. Following this trend the obligate hetero-complexes have an average planarity of 3.3Å with the value for the homo-complexes being 3.1Å. The standard deviations on both of these populations are comparable making this comparison valid. That the interfaces within hetero-complexes are generally non-planar is readily observable from the diagrams of the obligate hetero-complexes in the appendix. There are good reasons why large interfaces must be less planar than small ones. In vivo a protein with extensive and reasonably flat hydrophobic interaction sites will aggregate indiscriminately producing unwanted aggregates. Assuming a flat broadly hydrophobic surface the potential for unwanted aggregation will increase with the size of the interface. In contrast a protein with a large interaction site that has a reasonably intricate shape can only bind to a protein with a complementary shape and chemical character thus preventing unwanted aggregates. As with planarity the number of direct hydrogen bonds across a protein-protein interface varies linearly with interface size, with on average one hydrogen bond for every angstrom squared of buried ASA. This relationship is equally true for obligate homo or hetero-complexes.

Averaged amino acid composition of protein-protein interfaces

Summary

	Enzyme (%)	Inhibitor (%)	Antibody (%)	Antigen (%)	Signalling (%)		All obligate Homo's (%)	All obligate Hetero's (%)	All Nob-Obligate Hetero's (%)
ARG	4.4	6.8	4.2	8.2	7.8	} Charged	6.4	7.3	6.6
LYS	4.5	4.6	3.1	11.3	6.5		5.8	5.6	5.9
ASP	4.1	4.7	6.4	7.1	6.8		5.5	5	5.9
GLU	3.1	6.5	4.5	7.9	8.5		6.5	6.7	6.5
ASN	5.8	5.4	9.1	6.4	4.2	} Polar	4.4	4.8	6
CYS	4.1	6.5	0	0.8	2		1.2	1.3	2.8
GLN	5.8	2.8	1.9	5.2	3.9		4.2	3.7	3.8
HIS	6.5	2.6	2	2.3	4.4		3	3.2	3.5
SER	11.8	7.5	10.4	6.1	5.4		5.6	5.2	7.7
THR	4.4	5.4	9.3	8.5	5.7		5.7	4.9	6.4
TRP	3.9	2.3	5.9	2.7	1.5		1.3	1.5	3
TYR	7.6	5	19.4	3.3	5.5		4.7	5.2	7.6
ALA	2.9	5.3	1.5	3.5	3.3	} Hydrophobic	6.5	6.6	3.4
GLY	11.5	5.2	8.8	6.7	4.9		6.1	6.9	7
ILE	2.6	4.8	2.9	4.8	5.1		4.9	4.8	4.2
LEU	4.7	5.4	3.9	3.7	7.2		8.6	6.9	5.3
MET	0.9	2.7	1.1	1.7	2.5		2.8	3.5	2
PHE	5.1	2.3	2.1	1.1	4.9		4.1	4.4	3.2
PRO	3.2	7.9	1.9	5.6	5.4		5.2	5.9	5.1
VAL	3.4	6.3	1.9	3.1	4.6	6.2	6.7	4.1	
Charged	16	22.6	18.1	34.4	29.6		24.6	28.1	24.9
Polar	49.8	37.5	57.9	35.4	32.6		30.6	28.3	40.9
Hydrophobic	34.3	39.9	23.9	30.2	37.8		44.8	43.6	34.2

Figure 4.20: a summary chart showing the mean amino acid composition of the protein-protein interfaces of the different categories of protein-complex studied in this thesis.

The amino acid compositions of the interfaces within the obligate homo and hetero-complexes are quite similar. The percentages of hydrophobic and polar residues at the interfaces of the hetero and homo datasets are the same to within 2% (see figure 4.20). The only significant difference appears to be that the protein-protein interfaces of the hetero-complexes on average contain slightly higher numbers of charged residues than in the homo-complexes (24% compared with 28% in the hetero-complexes). The interface propensities of both the homo and hetero-datasets are also alike with only rather minor differences between them (see figure 4.18). It appears that obligate hetero and homo-complexes are essentially indistinguishable from each other in terms of their interface amino acid composition. But obligate homo and hetero-complexes might be distinguishable by looking at the amino acid make up of the subunits within the complex.

From table 4.9 and table 4.10 hetero-complexes do contain smaller percentages of hydrophobic residues than homo-complexes. The origin of this fact may be related to the non-globular nature of a significant number of the subunits within the hetero-complexes. An 'open' structure exposed to solvent has to have a lower hydrophobic content in order to be soluble than the more compact globular structure adopted by the majority of the subunits within the homo-complexes. Whether this is actually true in the most general sense (and not a statistical artefact) will require the analysis of larger numbers of hetero-complexes than the twenty structures used in the present study.

The interiors of the hetero-complexes are more hydrophilic than those of the homo-complexes (see figure 4.19). The averaged hydrophobicity of the interiors of the hetero-complexes is 0.83 while that of the homo-complexes is 0.95. The reason for this is again probably the fact that several of the subunits within the hetero-complexes only have small or poorly defined hydrophobic cores. The internal structure of these proteins is maintained through a combination of hydrophobic contacts and disulphide bridges and interactions between charged and polar groups. In contrast averaged over all proteins in the datasets the exteriors of the hetero-complexes are more hydrophobic than the homo-complexes (0.24 compared with 0.21). The hydrophobicities of the protein-protein interfaces within homo and hetero-complexes are virtually identical. Again this indicates that there are no great differences in the chemical character of the contact areas between subunits within obligate homo or hetero-complexes.

Chapter 5

Prediction of Protein-Protein Interaction Sites Using a Neural Network

5.1 Introduction

Given a protein of known structure what does it bind to? Answering this question has been one of the most important goals of structural biology for the last twenty years. From a predictive point of view there are two aspects of any protein-protein interaction that are of interest. The first is predicting which residues are involved in binding other proteins or ligands. The second is having predicted the residues that make up a binding site predicting what kind of protein or ligand binds there. Aside from experimental methods there are broadly speaking two major approaches that have been used to predict protein-protein interactions. The first approach is to take two protein structures that are thought to interact and to physically fit or 'dock' them together to give a model of the two proteins bound together in a complex. Usually, the individual proteins are treated as being rigid bodies although some recent docking methods do treat proteins as being flexible to some degree (Taylor & Burnett, 2000). The surfaces of the two proteins are then compared with each other in order to detect regions that are complementary to each other. The two proteins are then docked together at the points where they are most complementary in shape and/or electrostatics. Some of the earliest docking methods were based upon docking proteins together at points where they are most complementary in shape (Lee & Rose, 1985, Norel et al., 1994). Other methods have included docking proteins together based upon detecting clusters of hydrophobic residues on protein surfaces (Korn &

Burnett, 1991, Young et al., 1994). Comprehensive reviews of docking methods have been written by Smith & Sternberg, 2002, and Halperin et al., 2002.

The CAPRI experiment (Critical Assessment of Predicted Interactions) is an attempt to benchmark the accuracy of the various protein-docking methods that are used to predict protein-protein interactions (<http://capri.ebi.ac.uk>). At present four out of the seven targets provided by organisers of CAPRI are correctly predicted by the nineteen participating groups (Janin et al., 2003). One of the drawbacks of most docking algorithms is that they tend to be quite computationally intensive limiting the practical usefulness of such methods.

The second major approach that is used to predict protein-protein interactions is by looking at sequence data. Intuitively residues at ligand or protein-protein binding sites should be conserved (Valdar & Thornton, 2001). In theory by identifying surface residues that are conserved protein or ligand binding sites can be identified (Lichtarge & Sowa, 2002). Phylogenetic based methods show promise in identifying both binding-site residues and possible interaction partners (Valencia & Pazos in Bourne & Weissig, 2003).

The focus of this chapter is on improving on an existing method known as Patch Analysis to locate protein-protein interfaces. Patch Analysis is a method that is based upon defining patches of residues on the surface of a protein. The physical and chemical characteristics of each patch are then encoded in the form of six parameters such as hydrophobicity and planarity. By comparing the distribution of these six parameters with those of known protein-protein interaction sites the likelihood of any patch corresponding to a protein-protein binding site can be assessed. Patch Analysis was devised in 1997 by Jones & Thornton and is described in section 5.2. The method was benchmarked as being ~66% accurate having correctly located the protein-protein interfaces of 39 out of 59 proteins that form homo or hetero-complexes (Jones & Thornton, 1997). One of the weaknesses of the patch analysis method is that the six parameters used to characterise the surface patches are treated as being of equal importance in the predictive algorithm. To address this problem a neural network is used in conjunction with the original patch analysis method. The work presented in this chapter should serve to highlight some of the problems

associated with predicting the location of protein-protein interfaces using surface patches and provide a basis for future work in this area.

5.2 Patch Analysis

A concise outline of the patch analysis method is given in the following sections. A full treatment can be found in two papers, Jones & Thornton, 1997a and 1997b. The notation used in this section is that used in these two papers. The patch analysis methodology can be summarised as follows:

- (a) Defining a number of patches over the surface of a protein.
- (b) Encoding physical and chemical information about the residues in each patch in the form of six parameters such as size (accessible surface area, ASA) and hydrophobicity.
- (c) Assessing the likelihood of a patch corresponding to a protein-protein interface on the basis of averaged values of these 6 parameters.
- (d) Checking to what extent patches that are thought to correspond to likely protein interaction sites actually overlap with a protein-protein interface.

(d) is only carried out when attempting to benchmark the accuracy of the patch analysis method.

5.2.1 Definition of a Surface Patch

The way in which surface patches are defined is relatively simple. A patch is defined about a single exterior residue with n nearest surface accessible residues (the value of n is calculated using equation 1). This definition ensures that patches are contiguous.

The size of the interface region is roughly correlated to the size of the protein (Jones & Thornton, 1997b). Correlation of the number of residues in the observed interface region and the number of residues in the protomer for 28 homo-dimers gave rise to the following relation:

$$NR_i = 1.9NR_p^{0.6} \quad (1)$$

where NR_i is the number of residues in the observed interface and NR_p is the number of residues in the protomer. Patch sizes for each protein are calculated using the above relation. This is to allow for comparison between results obtained using the original patch analysis method with the enhanced neural network based method. Figure 5.5 in section 5.4.1 shows one of the 183 patches defined over the surface of the bacteriophage rb69 sliding clamp monomer.

The overlap of each patch with a protein-protein interface (if its location is known) is evaluated using two measures:

(a) Absolute Overlap

This is a straight comparison of the number of residues to be found in the protein-protein interface with the number of residues in a given patch that are also to be found in the interface. The absolute overlap (%) is given by:

$$Absolute\ Overlap\ (P1) = \frac{NrO \cap NrC}{NrO} \times 100 \quad (2)$$

where NrO is the number of residues in the observed interface patch and NrC is the number of residues in the calculated surface patch. For any given protein looking at the patch with the highest absolute (P1) overlap gives an idea as to how well the surface patches actually resemble the protein-protein interaction site. A histogram of the maximum absolute overlap value (P1) for each protein on which the neural-network based patch analysis method is tested is shown in figure 5.1. The average maximum P1 value for the dataset of 76 homo-dimers is 68%. The figures for the homo-trimers and tetramers are 74 and 77% respectively. For hetero-dimers the

average maximum absolute overlap value is 51% whilst that of the hetero-tetramers is 72%.

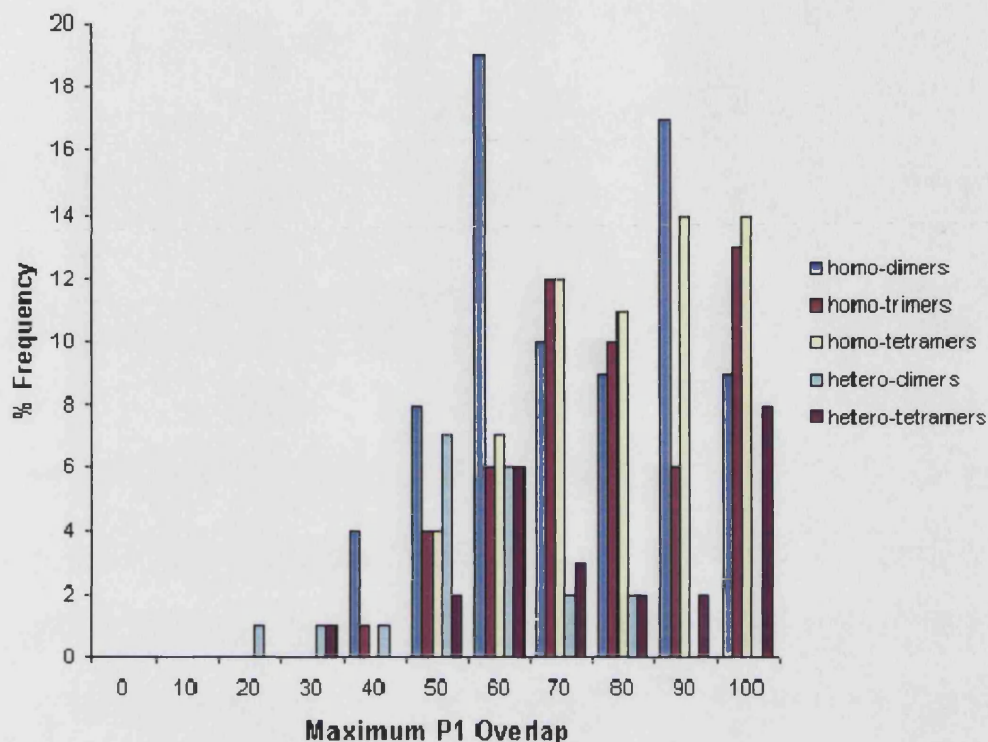


Figure 5.1: A histogram of the maximum absolute overlap values (P1) for each protein on which the neural network based patch analysis method is tested.

Generally speaking therefore, the surface patches as defined using patch analysis do generally match the size of the protein-protein interface quite well. Notable exceptions to this are non-globular proteins with extended structures. Many such proteins are hetero-dimers explaining the low average maximum P1 value for that dataset.

(b) Relative Overlap

This is a comparison of the number of interface residues that are to found in the patch *that contains the maximum number of interface residues* with the number of residues in a given patch that are also to be found in the interface. The relative overlap (%) is given by equation (3).

$$\text{Relative Overlap (P2)} = (P1 / \text{Maximum P1}) \times 100 \quad (3)$$

The relative overlap measure therefore takes into account the fact that surface patches are often not the same size and shape as the protein-protein interface.

5.2.2 Definition of Patch Parameters

Here the six patch parameters used to encode physical and chemical characteristics of each surface patch are described. All of these parameters with the exceptions of solvation potential and protrusion index have been defined and used to describe protein-protein interfaces elsewhere in this thesis. The explanations and equations given in this section are adapted from those given in the two papers (Jones & Thornton, 1997a and 1997b).

Hydrophobicity. The Fauchere & Pliska scale (1983) is used to calculate a value of the average hydrophobicity of each patch just as described in chapter 3.

$$\text{Patch hydrophobicity} = \frac{\sum_{i=1}^{N_p} (HV_{AA}(i))}{N_p} \quad (4)$$

where N_p is the total number of residues in each patch and HV_{AA} is the hydrophobicity value assigned to the amino acid residue.

The **protrusion index (PI)** gives a quantitative measure of how protruding a residue is from the surface of a protein.

$$\text{Patch PI} = \frac{\sum_{i=1}^{N_p} PI_{AA}(i)}{N_p} \quad (5)$$

where N_p is the total number of residues in each patch and $PI_{AA}(i)$ is the protrusion index of an amino acid residue evaluated using C_α co-ordinates. The PI of each residue in the patch is calculated and then averaged to produce a PI for the patch as a whole.

Residue interface propensity. The interface residue propensity of an amino acid is a measure of how frequently it is observed in the interface relative to the protein exterior and has been defined in section 3.5. The amino acid propensities of each amino acid were calculated using a data set of 63 assorted protein-protein complexes (Jones & Thornton, 1997a). The interface propensity of each residue in the patch is calculated and averaged to produce a value for the entire patch.

$$\text{Patch Interface propensity} = \frac{\sum_{i=1}^{N_p} (\ln IP_{AA}(i))}{N_p} \quad (6)$$

where N_p is the number of residues in the patch and $IP_{AA}(i)$ is the amino acid interface propensity.

Planarity. The mean planarity of each patch is calculated using the RMS deviation of each atom in the patch from a least-squares plane fitted through all the atoms in the patch.

Solvation potential. Solvation potentials “measure the propensity of each amino acid type for a certain degree of solvation, approximated by the residue solvent-accessible surface area, ASA” (Jones & Taylor, 1992).

$$\text{Patch } \Delta SP = \frac{\sum_{i=1}^{N_p} SP(AA(i))_{ASA_m} - SP(AA(i))_{ASA_o}}{N_p} \quad (7)$$

where ΔSP is the difference in solvation potential between a patch that is exposed to solvent and a patch which is not. $SP_{AA} (AA)_{ASAm}$ is the solvation potential of the amino acid AA (i) with an ASA of $ASAm$ in the protein's monomeric form. $SP_{AA} (AA(i))_{ASAO}$ is the solvation potential of the amino acid residue with an ASA of zero. ΔSP can take negative or positive values. "The larger and more positive the ΔSP value the greater the preference for burial" (Jones & Thornton, 1997a).

Accessible surface area (ASA). The relative ASA (rASA) of each residue in each patch is calculated using NACCESS (Hubbard, 1989). The mean rASA of all residues in the patch is then defined as:

$$\text{Patch rASA}(\text{\AA}^2) = \frac{\sum_{i=1}^{N_p} \text{rASA}_{AA}(i)}{N_p} \quad (8)$$

where rASA_{AA} = the relative ASA of a residue in the patch.

5.2.3 The Scoring Algorithm

For a protein the six parameters detailed in section 5.2.2 are calculated for each surface patch. For any one of these six parameters there will be a range of values across all the patches over the protein surface. For each protein the minimum and maximum values of each of the six parameters is noted. The lowest parameter value is denoted as having a score of 1 and the highest a score of 100. In the case of the planarity parameter the lowest parameter value is denoted as having a score of 100 whilst the highest has a score of 1. By doing this a score between 1 and 100 can be assigned for each of the six parameters for each patch. Based on previous observations patches are expected to have a high residue interface propensity, be hydrophobic, planar and protruding (Jones & Thornton, 1997a). The combined score of a patch, P_j is then defined as:

$$\text{Combined Score } P_j = \frac{S_{sp} + S_{rp} + S_{hy} + S_{pi} + S_{asa} + S_{pl}}{N_{Parameters}} \quad (9)$$

where $N_{\text{parameters}}$ is the number of parameters calculated. S_{sp} is the combined score of patch P_j for the salvation potential distribution. S_{rp} is the combined score of patch P_j for the interface residue propensity distribution. S_{hy} is the combined score of patch P_j for the hydrophobicity distribution. S_{pi} is the combined score of patch P_j for the protrusion index distribution. S_{asa} is the combined score of patch P_j for the accessible surface area distribution. S_{pl} is the combined score of patch P_j for the planarity distribution.

The combined score of a patch is a measure of the probability of a patch corresponding to a protein-protein interaction site. Each patch is then listed in order of its combined score (P_j) from highest to lowest. The three patches with the highest combined score are then selected as the three 'best patches'. If any of these three 'best patches' have a relative overlap value $\geq 70\%$ then the prediction is defined as being correct.

5.3 Neural Networks

A neural network is essentially a collection of artificial neurons connected together in a particular way. The 'strength' of the connections between neurons are known as weights. Neural networks have the property that without any prior programming by altering the strength (or weights) of the connections between each neuron a neural network can 'learn' certain features of the data that is presented to it. Thus a neural network can be trained to classify and extract important information from data without necessarily having any previous knowledge of it.

The patch analysis technique treats each of the six parameters as being equally important in the scoring of each surface patch. However, some parameters are clearly more important than others in the predictive process. A neural network should be able to adjust its weights to effectively use the predictive power of the data that is presented to it. An analysis of the weights of a neural network after training will allow the relative importance of each individual parameter in the predictive process to be assessed. This is most easily done with neural networks known as single layer

perceptrons since all the weights are directly connected to the network's inputs, and thus their significance in determining the overall output is relatively clear.

5.3.1 Feed Forward Neural Networks

The networks that are used to produce the interface prediction data that is presented here are single and multi-layer perceptron feed-forward neural networks. Feed forward neural networks have a layered structure as shown in figure 5.2.

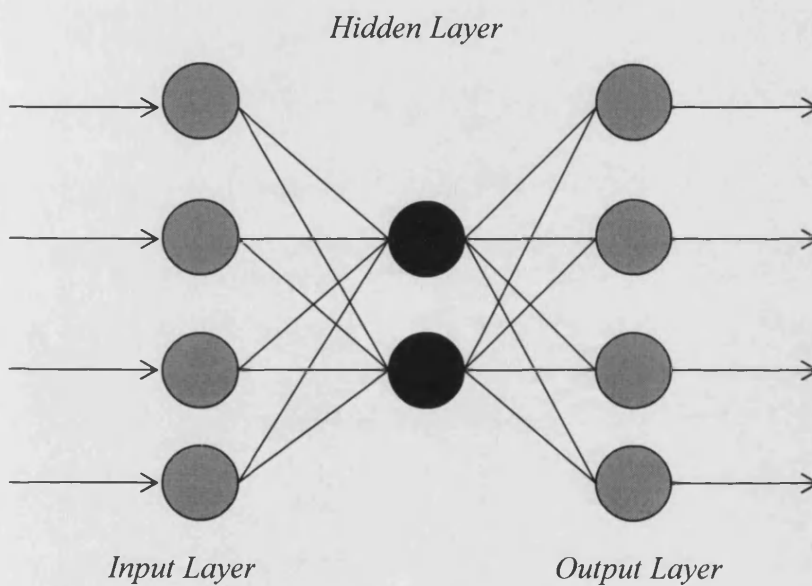


Figure 5.2: A representation of the structure of a feed-forward neural network. This kind of neural network is also known as a multilayer perceptron.

A hidden layer is a general term for layers that do not contain neurons belonging to either the output or input layer. For most purposes it is not necessary to use neural networks with more than one hidden layer.

A feed-forward neural network has a layered structure as shown figure 5.2. Information passes through the network in one direction only from one layer to the layer immediately above it and so on, hence the terminology 'feed forward'.

The diagram below shows a single layer perceptron neural network. As explained above, these neural networks have the advantage that the significance of their weights can be more directly assessed. The computational power of such neural networks however, may be insufficient for the problem at hand, as single layer networks can only address ‘linearly separable’ problems and most interesting problems are unfortunately not of this class.

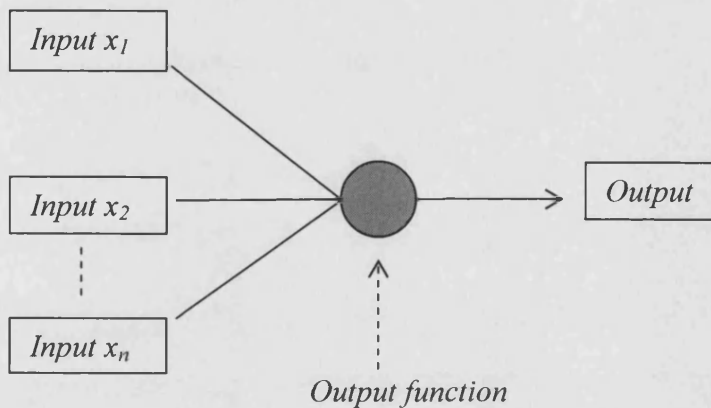


Figure 5.3: The basic model of an artificial neuron. The neuron takes the weighted sum of its inputs and compares it with some threshold value and then produces a suitable output using an output function.

For the single neuron of the network shown in figure 5.3 the total output is given by:

$$Total\ Input = weight\ on\ line\ 1 \times Input\ x_1 + weight\ on\ line\ 2 \times Input\ x_2 + \dots \\ weight\ on\ line\ n \times Input\ x_n$$

$$= \sum_{i=1}^n w_i x_i$$

Let the output be denoted by y . The total input is then compared to a threshold value S . If the total input is greater than the threshold value then the output y will conventionally be close to one. Otherwise the output will be closer to zero. In biological neurons the threshold is non-adaptive, but in artificial neural networks it is allowed to ‘learn’ appropriate values and is treated in effect as another ‘bias’ weight. The output can then be written:

$$y = f\left[\sum w_i x_i - S\right] \text{ or } y = f(a) \text{ where } a = \sum w_i x_i - S \quad (10)$$

where $f(a)$ is the threshold function. The neural network that was used has a sigmoid output function. A sigmoid function has the form as shown in figure 5.4.

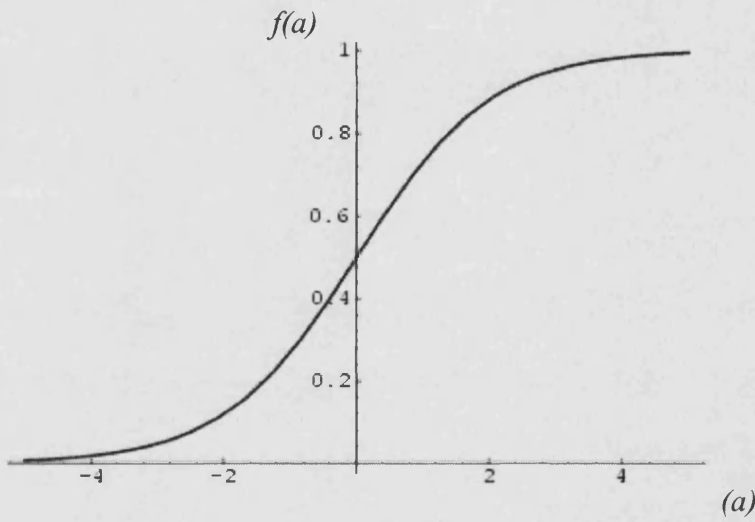


Figure 5.4: A plot showing a sigmoidal output function used by the neural network.

where the function $f(a)$ has the form:

$$f(a) = \frac{1}{1 + e^{-\beta a}} \quad (11)$$

β is a constant that relates to the gradient of the function and is usually set to one. In the perceptron model the threshold is usually represented as minus the value of a weight w_{i0} that has $x_0=1$, the 'bias weight' referred to above.

The output is then:

$$y_i(t) = f\left[\sum_{i=0}^n w_i(t)x_i(t)\right] \quad (12)$$

where $S = -w_o$ and the weights are time dependant over multiple epochs. An epoch is simply a single presentation of all patterns in the training dataset to the neural network, followed by the application of an appropriate learning procedure.

5.3.2 Supervised Learning

Supervised learning is a method whereby the neural network is presented with the desired output for a given input during training. This is usually done a large number of times. The neural network's weights adjust themselves during this process to best reproduce the desired output for a given input.

It was therefore hoped that the use of a neural network would significantly improve the original patch analysis method for the prediction of protein-protein interaction sites. A multi-layer perceptron neural network was used in the work that is presented in this chapter. The neural network was one that was used in the prediction of beta turns in proteins (Shepherd et al., 1999). At the beginning of the training process the weights of the neural network are set to random values. The outline below describes the learning of a single output response, within one epoch of training.

The neural network is presented with the input $x_0, x_1 x_2 \dots x_n$ and the desired output (the overlap of each patch with the interface patch).

Each neuron i within the network, in both the hidden and output layers computes its output according to equation 13.

$$y_i(t) = f\left(\sum_{j=0}^n w_{ij}(t)x_j(t)\right) \quad (13)$$

The output of the neuron in the final layer is then compared with the target. The weights of the network are then adapted according to the equation 14:

$$w_{ij}(t+1) = w_{ij}(t) + \eta\delta_i x_j(t) \quad (14)$$

where η is the training rate, $\eta > 0$ is the only restriction, and the training rate is normally much less than 1.0. The term δ_i represents the error for pattern p on node i and $w_{ij}(t)$ is the weight from node j to node i . δ_i is calculated for the output layer by a straightforward comparison of desired and target outputs; for a hidden layer neuron the ‘error back-propagation’ process can be used.

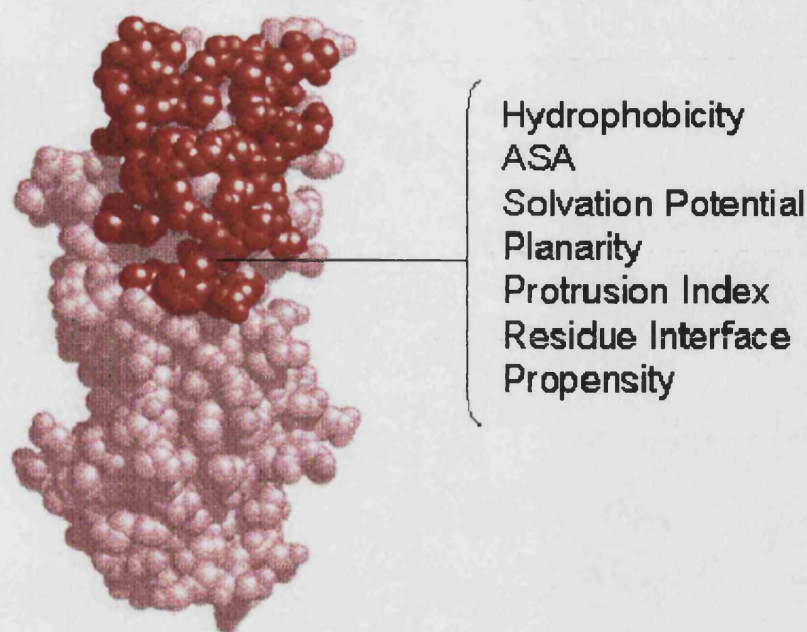
This process is repeated for all other patterns in the training set until the current epoch’s weight updates are completed, if necessary then re-presenting all the patterns for a further epoch of training.

5.4 Neural Network based Patch Analysis

5.4.1 Training and Testing the Neural Network

This section deals with the way that the neural network is trained and tested using data calculated using the original patch analysis method. As explained in section 5.2.1 a number of surface patches are defined over the surface of each protein. There are 183 such patches defined over the surface of the bacteriophage rb69 sliding clamp monomer.

The neural networks that were used are trained according to the following process. As inputs the neural network is presented with average values of the six patch analysis parameters previously defined in section 5.2.2 for each of the 183 surface patches. The neural network is given as a target output the absolute overlap of each patch with the protein-protein interface that the patch covers most. In the cases of trimeric and tetrameric proteins, only the two largest protein-protein interfaces in the complex are considered. This is because the third largest protein-protein interface in tetramers is usually small in size (some 11% of the total buried surface area on average) reflecting subunit packing effects. In trimeric proteins each subunit only has two protein-protein interfaces.



Patch	Centre	Solv	Plane	Prop	Hydro	PI	ASA	Overlap
0	1	-0.51	4.103	0.006	0.264	5.341	38.22	50.00
1	2	-0.55	4.432	0.037	0.339	4.902	36.57	58.33
2	3	-0.37	4.657	0.084	0.462	4.073	32.03	50.00
3	4	-0.65	5.264	-0.083	0.081	4.488	42.38	33.33
4	5	-0.64	4.770	-0.074	0.098	4.488	41.94	58.33
5	6	-0.74	4.719	-0.114	0.009	4.122	42.02	25.00

etc

Figure 5.5: The bacteriophage rb69 sliding clamp monomer with residues from one surface patch coloured in red (1b77, Shamoo et al., 1999). There are a total of 183 surface patches defined over the surface of the protein. As inputs the neural network is presented with average values of the six patch analysis parameters previously defined in section 5.2.2 for each of the 183 surface patches.

After a presentation of the averaged patch analysis values for a surface patch the neural network predicts the overlap of the patch with the desired protein-protein interface. The neural network output is then compared with the actual overlap and the weights of the network are updated accordingly. This iterative training process is summarised in figure 5.6. In the case of training a network using the bacteriophage rb69 protein a total of 183 patterns are presented corresponding to the 183 surface patches.

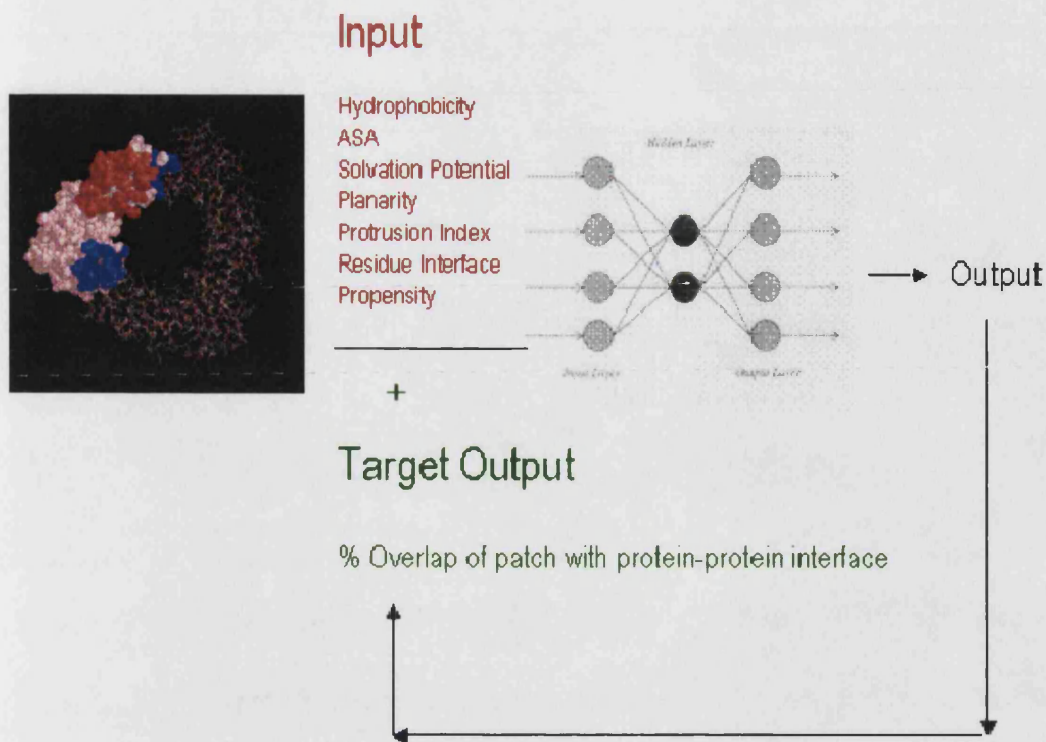


Figure 5.6: A summary of the procedure used to train the neural network. For each protein in the training dataset the neural network is presented with the six patch analysis parameters for each surface patch. The neural network then produces an output for each patch and compares it with the target output and updates its weights accordingly.

There are a number of different algorithms that are commonly used to update the weights of a neural network during training. Two of the most widely used are back-propagation (mentioned in the earlier outline of multilayer perceptron learning) and conjugate gradients. The conjugate gradients method was used in preference to back-propagation as weights are generally optimised relatively quickly using this algorithm (Shepherd et al., 1997).

When it comes to assessing the performance of the neural network the output is just a list of patches together with a predicted interface overlap value.

One of the major disadvantages of neural networks is that they require large amounts of data to train them properly. With the exception of the dataset of homo-dimers all the other datasets used in this thesis are comparatively small. Indeed, the datasets of obligate hetero-proteins are too small to train the neural network at all. In view of this

a homo-dimer trained neural network is used when testing the neural network on hetero-dimeric and tetrameric proteins. The neural network was not tested on the datasets of homo or hetero hexamers. For hexameric proteins such a large fraction of surface area (~40% on average for hetero-hexamers) is buried in subunit interfaces that it is not possible to 'predict' the location of subunit interfaces with any degree of statistical significance.

For the datasets of homo-dimers, trimers, and tetramers the neural-network was trained and tested using a procedure known as jack-knife testing. This procedure is straightforward. For a dataset of n proteins the neural network is trained on all but one of the proteins in the dataset and tested on the single protein that is left out of the original dataset. This procedure is then repeated a total of n times each time testing the network on a different protein and training it on the remaining proteins in the dataset. Thus for the dataset of 76 homo-dimers the network is trained and tested 76 times each time using a different training and test dataset.

When working with neural networks it is important to avoid over-training. It is almost inevitable that the training dataset contains a certain amount of 'noise' or information that is unrepresentative of the data as a whole. There is consequently a danger that the network will incorrectly adjust its weights in response to this 'noise' over every epoch that the network is trained. The cumulative effect of this can result in a neural network's error rate with respect to a test data set actually increasing over time. The neural network was trained using the conjugate gradients training method for 200 epochs using 57 proteins chosen at random from the 76 homo-dimers (or three quarters of the total dataset). The network is tested at each epoch on the remaining 19 dimers. A plot of the error rate on the training and test datasets over a training period of 200 epochs is shown in figure 5.7. The error rate on the training dataset (and the test dataset) in figure 5.7 levels out after ~60 epochs of training and does not increase thereafter. Further testing of the neural network with different training and test datasets confirmed that over 200 epochs no over-training occurs. All the neural networks that were used to locate protein-protein interfaces were trained for a total of 200 epochs.

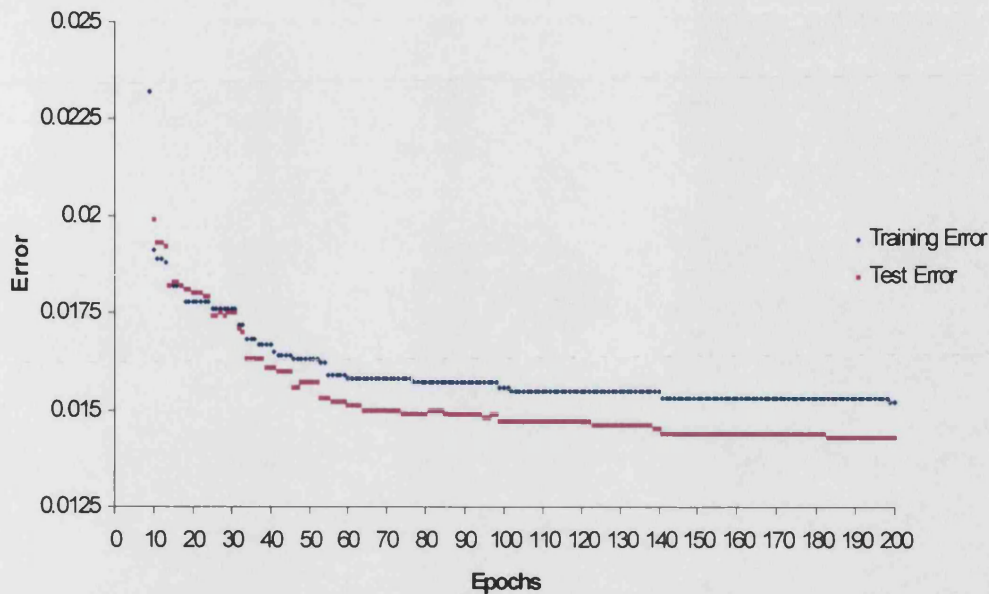


Figure 5.7: A plot showing the training and test errors for a two hundred epoch run of the neural network. The training dataset consisted of 57 homo-dimers chosen at random from the full dataset of 76 homo-dimers. The test dataset consisted of the remaining 19 homo-dimers. Both the training and test errors level out over a period of two hundred epochs indicating that no over training occurs.

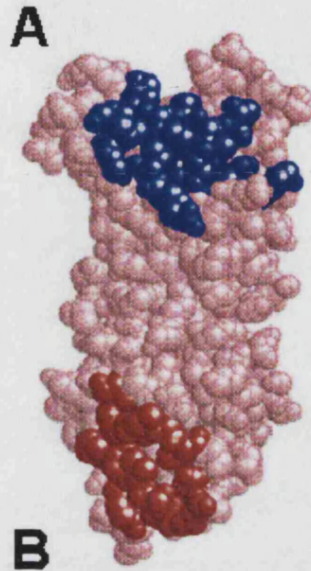
5.4.2 Evaluating the Results

The definition of what is a correct or incorrect prediction is inevitably somewhat arbitrary. Originally the patch analysis method was only tested on proteins that form binary complexes such as homo-dimers and hence only have a single protein-protein interface. In the case of such proteins a prediction was defined as being correct by Jones and Thornton if any of the three highest ranked patches have a relative overlap $\geq 70\%$. A certain justification for this criterion was based on the observation that the three highest ranking patches of 28 homo-dimers all tend to cluster around the same physical location on the proteins exterior. The ‘highest ranked patch’ is the patch with the highest predicted overlap value with a protein-protein interface. The second highest ranked patch is the patch with the second highest predicted overlap value and so on. The three highest ranked patches for most of the 76 homo-dimers also cluster together around the same region on the surface of the protein exterior sharing $>50\%$ or more residues in common with each other. In light of this for a protein with only one protein-protein interface such as homo or hetero-dimers if any of the three highest

ranking patches have a relative overlap $\geq 70\%$ then a prediction is defined as being correct.

(a)

Patch No	Predicted Overlap (%)
0	23.5
1	56.1
2	10.2
etc	etc



(b)

Patch No	Predicted Overlap (%)	Actual Relative Overlap (%)	
		Interface A	Interface B
101	80.0	75.0	10.5
0	23.5	80.1	24.9
9	70.0	10.5	25.9

Figure 5.8: A summary of the procedure used to assess the neural-network results for a protein with two protein-protein interfaces. The bacteriophage rb69 sliding clamp monomer is shown with its two protein-protein interfaces labelled A and B. As usual for each patch defined over the protein surface the neural network output is a list of predicted overlap values. The three highest ranked patches are then determined as described in section 5.4.2. The actual relative overlap of the top three highest ranking patches with interface A and B is then determined. If any two of the three patches have a combined relative overlap $\geq 140\%$ with interface A and B then a prediction is defined as being correct. In this case patch numbers 0 and 9 have a combined relative overlap with interface A and B of $80.1 + 25.9 = 111\%$. The prediction is incorrect.

A different criterion for defining what is a correct prediction has to be employed for a protein with more than one protein interface. The monomer of the trimeric bacteriophage rb69 sliding clamp protein is shown in figure 5.8(a) with its two

protein-protein interfaces labelled A and B. The list of predicted overlap values for each patch is sorted as usual from the highest to lowest as in figure 5.8(b). The patch with the highest predicted overlap value is retained as the highest ranking patch. The second highest ranking patch is found by finding the patch with the next highest predicted overlap value that contains <50% of the residues contained in the first highest ranking patch. The third highest ranking patch is the patch with the highest predicted overlap value that contains <50% of the residues in the second or first highest ranking patches. In this way the three patches corresponding to three semi-distinct locations on a protein exterior are selected. This is reasonable since the subunits of the trimeric and tetrameric proteins on which the neural network is tested have 2-3 protein-protein interfaces

The relative overlap of each of these three patches with the two protein-protein interfaces labelled A and B in figure 5.8(b) is then calculated. If any two of the three patches have a combined relative overlap of $\geq 140\%$ with the two protein interfaces then the prediction is defined as being correct.

5.5 Results

Training neural networks so as to produce optimal results is often difficult. Aside from the data that is used to train the neural network (and way that the data is presented to it) both the method used to train the network and its architecture affects the quality of the results produced by a neural network. As yet, there is no standard way of determining either the training method or neural network architecture that will produce the best results for any given set of data. Finding the neural network architecture that will produce the best results for any given set of data is to all intents and purposes a matter of trial and error. Consequently the best neural network architectures for the homo-dimer, homo-trimer, and homo-tetramer data must be determined by experiment and it is entirely possible that for each class of homo (or hetero) multimer different neural network architectures will produce the best results. For different datasets of proteins neural networks with different architectures will produce the best results. Indeed, as will be seen later, for homo-dimers the optimal architecture was found to be a neural network with no hidden units was optimal

whereas a homo-dimer trained network with four hidden units was found to be best for the obligate hetero-dimers. Feed forward neural networks with 0, 2, 3, and 4 hidden units were used to train to predict the locations of the protein-protein interfaces in the homo-dimers, trimers, and tetramers. Four sets of results are thus obtained for each class of multimer. The results are given in tables 5.1 to 5.3. In each case the results given are for the neural network architectures that produces the best interface predictions. In all cases the neural networks were trained for 200 epochs using the conjugate gradients training method and the jack-knife procedure described in section 5.4.1 was used.

For the obligate hetero-dimers and tetramers a different procedure had to be used to train the neural network. Quite simply, there are not enough obligate hetero-dimers or tetramers to train a neural network properly. To deal with this the neural networks that were used were firstly trained using the 76 homo-dimers and then tested on the obligate hetero-dimers (and separately) the obligate hetero-tetramers. As discussed earlier the architecture of the neural network that is used does affect the quality of the networks predictions. There is as yet no way to predict in advance which neural network architecture will produce the best results for any given dataset. Neural networks with 0, 2, 3, and 4 hidden units were trained using the 76 homo-dimers and then tested on the hetero-dimers (the same procedure was used to produce predictions for the obligate hetero-tetramers). Consequently, four sets of results for the obligate hetero-dimers and tetramers were generated. The results are given in tables 5.4 to 5.7. In each case the results given are for the homo-dimer trained neural network architecture that produces the best interface predictions.

Key for Tables 5.1 to 5.7

The PDB codes of proteins whose interface predictions are incorrect are shaded in green. In each case the patches labelled 1st, 2nd, and 3rd denote the three patches most highly ranked as covering protein-protein interaction sites

No of patches – the total number of patches that are defined over the surface of the protein subunit

Patch size – the number of residues in each of the surface patches that are defined over the surface of the protein.

Absolute overlap (P1) – the absolute overlaps of each the three highest ranked patches with the dimer interface.

Max P1 – the absolute overlap of the patch that covers the given protein-protein interface best.

Relative overlap (P2) – the relative overlaps of each of the three highest ranked patches with the given protein-protein interface.

5.5.1 Homo-Complexes

PDB Code	No of Patches	Patch Size	Overlap (P1)			MaxP1 (%)	Relative Overlap (P2)		
			1 st	2 nd	3 rd		1 st	2 nd	3 rd
1a3c	124	35	65	56	30	78	83	72	39
1ad3	314	61	45	33	43	45	100	74	96
1af5	100	30	4	0	4	83	5	0	5
1afw	239	56	41	46	63	75	55	61	84
1ajs	284	58	38	38	39	47	81	80	83
1alk	288	61	49	49	37	52	94	94	72
1alo	542	91	100	23	91	100	100	23	91
1amk	175	44	75	77	77	83	90	93	93
1aom	287	60	6	6	6	92	6	6	6
1aor	346	72	30	0	39	93	29	0	42
1aq6	179	43	44	65	49	65	67	100	76
1auo	155	40	80	65	90	90	89	72	100
1bam	144	38	88	0	0	100	88	0	0
1bif	325	59	46	35	27	85	54	41	32
1bsr	110	29	44	35	42	51	86	68	82
1buo	101	29	48	48	48	55	88	88	88
1cg2	282	56	82	84	89	92	89	91	97
1chm	265	57	39	49	45	51	77	95	88
1cmb	94	27	61	59	45	64	96	92	71
1cp2	181	46	33	54	67	87	38	62	77
1csh	297	60	47	47	44	47	100	100	94
1ctt	194	48	34	59	57	73	46	81	78
1czj	107	27	83	88	44	94	88	94	47
1daa	210	46	62	50	54	64	97	78	84
1fip	67	22	50	47	39	53	95	89	74
1fro	156	36	34	32	35	35	97	91	100
1gvp	78	24	83	54	71	83	100	65	85

PDB Code	No of Patches	Patch Size	Relative Overlap: Interface One			Max P1 (%) Interface One	Relative Overlap: Interface Two			Max P1 (%) Large Interface
			1 st	2 nd	3 rd		1 st	2 nd	3 rd	
1aa0	111	28	80	95	55	42	58	68	58	40
1b77	183	41	0	25	67	100	23	0	0	100
1bro	185	46	8	25	8	100	79	79	26	90
1bvp	271	53	9	12	76	47	68	35	45	51
1ca4	135	35	18	82	27	92	100	93	53	71
1cbo	188	45	83	67	67	88	7	3	7	94
1cbu	143	36	0	100	81	80	36	64	59	88
1ce0	33	14	58	58	33	67	64	71	43	78
1cjd	268	54	0	59	68	63	30	87	97	79
1dpt	99	28	73	30	33	79	10	5	5	68
1dun	101	29	41	45	68	58	77	73	55	69
1e2a	83	26	88	81	100	55	29	86	71	61
1fqj	421	65	35	10	30	53	14	78	97	63
1nif	242	51	92	95	59	74	5	5	3	67
1nks	156	38	32	0	64	92	76	47	24	81
1ppr	296	49	73	67	33	94	0	0	91	100
1qex	241	47	41	28	21	43	42	39	10	49
1qlm	203	50	29	96	58	83	74	26	4	90
1rla	200	50	94	31	94	94	4	50	8	77
2chs	92	28	33	28	0	69	87	87	75	70
2pii	102	28	61	52	48	64	0	4	4	77
2std	130	34	65	65	53	77	35	65	82	74
3cla	168	40	43	81	43	91	86	11	4	85
3csu	191	47	50	41	45	100	88	82	65	94
3tdt	213	46	10	3	3	54	73	79	79	54
4bcl	325	53	17	76	21	69	0	47	9	70

Table 5.2: Neural network results for the dataset of homo-trimers. Fifty four percent of the predictions are correct. A neural network with no hidden units was used.

PDB Code	No of Patches	Patch Size	Relative Overlap: Small Interface			Max P1 (%) Small Interface	Relative Overlap: Large Interface			Max P1 (%) Large Interface
			1 st	2 nd	3 rd		1 st	2 nd	3 rd	
1a0l	159	41	98	98	89	74	98	100	88	88
1a2z	152	41	64	96	59	79	68	58	95	68
1a4e	371	64	83	100	73	43	78	82	76	46
1ado	254	54	0	0	60	89	4	4	100	81
1az9	300	60	7	33	27	75	95	43	25	68
1b25	344	72	0	0	0	80	79	100	84	91
1bfd	346	66	6	6	15	76	96	86	82	54
1bsm	154	39	89	78	100	90	24	42	24	60
1buc	259	56	52	41	72	66	44	66	34	73
1bvq	114	31	19	75	31	73	61	22	61	69
1cs1	263	56	14	20	26	88	85	66	46	62
1cuk	155	37	89	73	85	70	50	23	73	73
1dco	78	26	100	31	0	100	44	94	0	89
1e5a	94	28	100	77	54	87	84	100	37	86
1euh	309	63	50	25	54	83	100	66	40	59
1ftr	220	48	0	0	0	67	66	74	74	46
1gp1	132	37	56	0	39	95	15	70	5	95
1gsh	211	48	39	0	0	100	69	28	62	62
1ith	111	32	0	0	9	85	0	36	14	100
1mpy	205	49	39	30	52	96	94	78	78	68
1mxb	257	55	100	94	100	94	39	52	34	60
1nhk	114	32	9	55	0	100	0	15	10	65
1nhp	336	61	31	0	0	93	79	87	15	59
1sml	178	45	42	16	26	91	48	89	96	84
1toh	234	51	95	0	16	100	92	0	66	69
1uox	242	48	69	28	78	49	32	76	11	53
1xva	228	48	78	100	0	69	22	89	5	69
2fua	156	40	86	18	64	79	8	84	72	89
2izg	108	29	92	83	58	92	46	50	33	52
4pga	219	51	44	33	78	93	36	10	12	74
5pgm	164	42	32	64	100	88	20	60	35	87

Table 5.3: Neural network results for the dataset of homo-tetramers. Fifty eight percent of the predictions are correct. A single layer neural network was used.

5.5.2 Hetero-Complexes

PDB Code	No of Patches	Patch Size	Overlap (P1)			Max P1 (%)	Relative Overlap (P2)		
			1 st	2 nd	3 rd		1 st	2 nd	3 rd
1ajq	187	39	40	41	41	44	91	94	94
1ft1	240	50	38	31	37	49	78	63	75
1h2a	211	45	51	54	46	55	92	98	84
1hcn	82	24	36	48	50	52	69	92	96
1ixx	101	29	53	56	53	60	89	93	89
1luc	224	50	58	52	58	72	80	72	80
1req	441	73	44	42	30	50	87	84	60
2frv	196	45	56	58	42	64	87	91	66
4mon	39	16	52	48	55	55	94	88	100

Table 5.4: Neural network results for the small subunits of the dataset of hetero-dimers. The small subunit of hydrogenase (1hfe) has been excluded from the dataset due to its non-globular shape. All of the predictions are correct. A homo-dimer trained neural network with 4 hidden units was used.

PDB Code	No of Patches	Patch Size	Overlap (P1)			Max P1 (%)	Relative Overlap (P2)		
			1 st	2 nd	3 rd		1 st	2 nd	3 rd
1ajq	424	69	11	17	17	19	59	91	91
1ft1	277	58	20	11	30	42	47	26	71
1h2a	345	67	20	29	10	38	53	75	25
1hcn	106	27	30	42	30	46	65	91	65
1hfe	267	57	15	15	17	22	68	68	77
1ixx	106	30	54	54	49	56	96	96	87
1luc	241	51	42	36	63	67	63	54	94
1req	494	80	53	50	53	53	100	94	100
2frv	337	67	21	30	17	48	43	62	36
4mon	49	18	47	38	25	47	100	80	53

Table 5.5: Neural network results for the large subunits of the dataset of hetero-dimers. Nine out of the ten predictions are correct. A homo-dimer trained neural network with 4 hidden units was used.

PDB Code	No of Patches	Patch Size	Relative Overlap: Small Interface			Max P1 (%) Small Interface	Relative Overlap: Large Interface			Max P1 (%) Large Interface
			1 st	2 nd	3 rd		1 st	2 nd	3 rd	
1apy	111	32	27	13	40	88	97	71	97	48
1b7y	203	45	66	75	72	89	88	91	93	52
1bou	111	30	91	55	9	100	55	82	73	73
1qdl	129	58	35	24	59	100	79	88	48	94
1qsh	188	32	100	0	64	93	0	57	5	95
2scu	191	47	63	63	100	100	92	66	34	66

Table 5.6: Neural network results for the small subunits of the dataset of hetero-tetramers. The small subunit of glutamate mutase (1ccw) has been excluded from the dataset due to it only having one protein-protein interface. Five out of the six predictions are correct. A homodimer trained neural network with 2 hidden units was used.

PDB Code	No of Patches	Patch Size	Relative Overlap: Small Interface			Max P1 (%) Small Interface	Relative Overlap: Large Interface			Max P1 (%) Large Interface
			1 st	2 nd	3 rd		1 st	2 nd	3 rd	
1apy	140	34	0	0	0	62	94	94	94	43
1b7y	596	83	25	54	0	51	74	71	29	30
1bou	192	48	100	43	62	70	43	11	71	56
1ccw	289	63	53	6	0	51	29	75	75	55
1qsh	120	32	41	53	0	100	35	45	10	91
2scu	286	56	38	33	19	72	71	57	100	60

Table 5.7: Neural network results for the large subunits of the dataset of hetero-tetramers. The large subunit of anthranilate synthase (1qdl) has been excluded from the dataset due to it only having one protein-protein interface. One out of the six predictions are correct. A homodimer trained neural network with 2 hidden units was used.

5.6 Rationalising the Results

In this section the neural network predictions for each dataset of proteins are analysed. It was unexpected and interesting that in many cases the best performing network was one with no hidden units. However this does not necessarily mean the interface prediction problem is truly a linearly separable one in these cases; it might well be that an enlarged training dataset would allow a hidden layer to pick up and use subtle non-linear correlations between the input parameters. As pointed out in section 5.3.1 when using a single layer neural network with no hidden units a comparison of the magnitude of each weight then allows the relative importance of each parameter in the predictive process to be assessed. The six weights for a neural network with no hidden units trained on the seventy six homo-dimers are shown in table 5.8.

Parameter	Weight
Solvation Potential	1.89
Planarity	0.07
Residue Interface Propensity	2.44
Hydrophobicity	-0.56
Protrusion Index	0.09
ASA	0.03

Table 5.8: The weights of a neural network with no hidden units trained using the seventy-six homo-dimers.

As can be seen from table 5.8 the neural network weights the solvation potential, residue interface propensity, and hydrophobicity parameters as being of particular importance when predicting the location of the dimer interface. In contrast the planarity, protrusion index, and ASA parameters are not treated by the neural network as being particularly good indicators of a protein-protein interface. This result shows that for homo-dimers the neural network could probably work with fewer input parameters than the six used here with little loss in predictive power.

5.6.1 Understanding Incorrect Predictions

The neural network results for the dataset of homo-dimers are encouraging. Using a single layer neural network 58 out of the 76 predictions are correct giving a success rate of 76%. The predictions using the original patch analysis method are 63% correct (results not shown). These results show that a neural network can considerably enhance the performance of the original patch analysis method. Most of the 18 homo-dimers for which the neural network fails to locate the dimer interface fall into two classes

- Homo-dimers that interact with other proteins and therefore have additional interaction sites.

and/or

- Proteins whose dimer interfaces are to some extent atypical of protein-protein interfaces in general.

One of the proteins for which the neural network completely fails to locate the dimer interface is 1-crel endonuclease. The endonuclease recognizes a section of DNA approximately twenty base pairs in length and was the first example of a protein coded for by a gene within an intron (Heath, 1997). A diagram of the endonuclease when bound to the DNA helix is shown in figure 5.9(a). Figure 5.9(b) shows all residues contained within the patch with the greatest relative overlap with the protein-protein interface as predicted by the neural network coloured in red and the rest of the protein coloured in blue. The patch that is selected clearly contains residues that make contact with the backbone of the DNA double helix as well as the base pairs themselves. The loops that make contact with the DNA helix are actually more hydrophobic than the dimer interface and this may be the reason why the neural network locates the DNA binding site in preference to the dimer interface. Other endonucleases for which the neural network locates the DNA binding site but not the dimer interface include the *Serratia* endonuclease (1smn) and RuvC resolvase (1hjr). Another protein with a major interaction site aside from the dimer interface is the α -amylase inhibitor, 1hss. The neural network appears to locate the inhibitor interface rather than the dimer interface.

(a)



(b)



Figure 5.9: (a) The 1-creI endonuclease dimer bound to the DNA double-helix. This diagram is taken from Heath, 1997. (b) The endonuclease with the residues from the patch most highly ranked by the neural network as corresponding to the dimer interface coloured in red. Whilst the neural network fails to locate the dimer interface, the DNA binding sites of the protein are correctly located.

These examples demonstrate that even in cases where the neural network fails to locate the dimer interface a biologically relevant interaction site is often located instead. Half of the incorrect predictions relate to homo-dimers whose interfaces are quite polar in character as is shown in table 5.9.

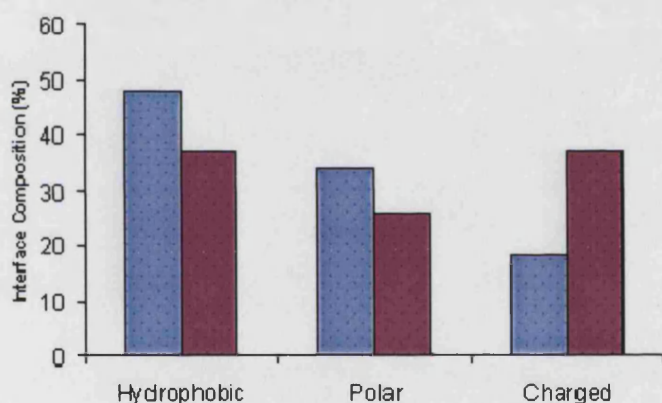
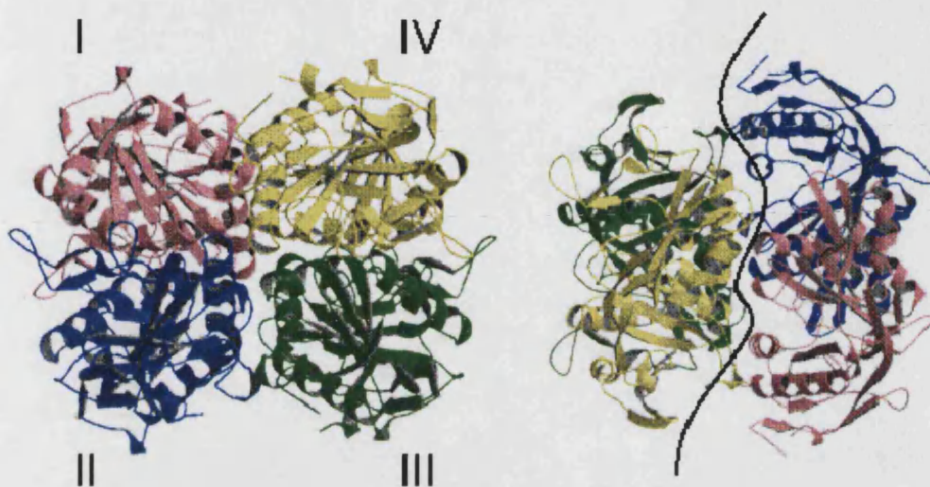
PDB Code	Protein	Hydrophobicity
1aom	Nitrite Reductase	-0.08
1aor	Aldehyde Ferredoxin Oxidoreductase	-0.01
1bif	6-Phosphofructo-2-Kinase/Fructose-2,6-Bisphosphatase	-0.11
1hss	Alpha-Amylase Inhibitor	0.20
1icw	Interleukin-8 Mutant	0.12
1moq	Glucosamine 6-Phosphate Synthase	0.07
1pre	Proaerolysin	0.15
1smn	Serratia Endonuclease	0.17
1tox	Diphtheria Toxin	0.16

Table 5.9: The interface hydrophobicities of nine homo-dimers for which the neural network fails to locate the dimer interface and for which the dimer interface is quite hydrophilic in character. For comparison the average hydrophobicity of the seventy six homo-dimers is 0.37.

Because the majority of interfaces are hydrophobic, the neural network is trained to recognise hydrophobic regions as being representative of protein-protein interfaces. It is therefore unsurprising that the neural network has difficulty in predicting the location of polar interfaces.

The neural network does seem to have difficulty in locating the dimer interface of some electron transfer proteins, and proteins with metal ion clusters at or near the dimer interface. It has been observed that the protein-protein interfaces of electron transfer proteins are often significantly more polar than those found within other categories of oligomeric proteins (Mathews et al in Kleantous, 2000). It is this characteristic that endows electron transfer proteins with the ability to associate with each other rapidly in response to environmental conditions. This fact helps to explain why the neural network has difficulty in locating the dimer interface of proteins such as nitrate reductase (1aom). At first sight the success rate for the dataset of homo-tetramers is disappointing with only 58% of all predictions being correct using a neural network with four hidden units, though in ten out of the thirteen proteins for

which the predictions are incorrect one protein-protein interface is correctly located. In three cases the neural-network fails to locate either of the two protein-protein interfaces.



	Size (Å)	Hydrophobicity
Interface between subunits I and II	2970	0.47
Interface between subunits I and IV	800	-0.11

Figure 5.10: The Formylmethanofuran tetramer with its four subunits labelled I, II, III, and IV. The tetramer can be considered to be a ‘dimer of dimers’ with one dimer being subunits I and II with the second dimer being subunits III and IV. The neural network correctly locates the large, hydrophobic interface between subunits I and II. The neural network fails to locate the small, polar interface between subunits I and IV.

The key to understanding some of the incorrect predictions lies in the observation made in chapter 3 that many of the tetramers are ‘dimers of dimers’. In such composite structures there are usually two quite differently constituted protein-protein interfaces. Formylmethanofuran (1ftr) shown in figure 5.10 is a good example. The interface between subunits I and II of 1ftr is quite large and closely packed being 2965\AA^2 in size. As can be seen from figure 5.10 the interface is quite hydrophobic almost half of all residues at the interface being hydrophobic. In contrast the interface between subunits I and IV is quite atypical of the protein-protein interfaces found with obligate protein-complexes. The interface is small in size at 800\AA^2 . The interface is also quite polar with 37% of all residues at the interface being charged compared with 18% for the interface between subunits I and II (Ermler et al., 1997). As can be seen from table 5.3 the neural network locates the large interface between subunits I and II but completely fails to find the interface between subunits I and IV. Another protein for which the neural network selects patches covering the primary interface over the secondary interface is glutathione synthetase (1gsh, Matsuda et al., 1996). The implications of this are that the neural network based patch analysis method can locate large broadly hydrophobic protein-protein interfaces of the kind found in homo-dimers to a much higher degree of accuracy than the smaller, more polar, and less closely packed interfaces found in homo-tetramers (and higher multimers).

To investigate this more closely the neural network was given as a target for each homo-tetramer:

(a) The overlap of each patch with the largest or ‘primary’ protein-protein interface within the tetramer.

And separately:

(b) The overlap of each patch with the second largest or ‘secondary’ interface within the tetramer.

A 31 fold cross validation was then performed on the 31 homo-tetramers with the target overlap values set as described in (a) and (b). A single layer neural network was used. As with the homo and hetero-dimers a prediction is defined as being correct if any of the top three patches has a relative overlap of $\geq 70\%$ with the interface.

With the target overlap values set as the overlap with the primary interface the predictions are 71% accurate. With the target overlap values set as the overlap with the secondary protein-protein interface the neural network's predictions are 48% correct. These results confirm that the neural network can locate large well defined protein-protein interfaces reasonably well but not the smaller less well defined interfaces such as are found within tetramers and some transitory protein complexes.

The results for the dataset of obligate hetero-dimers although encouraging demonstrates the limitations of the patch analysis method. Firstly, the subunits of hetero-dimers bury a larger fraction of their total surface area in protein-protein interfaces than do those of homo-complexes. On average some 30% of the surface patches of a hetero-dimer subunit have a relative overlap with the dimer interface $\geq 70\%$. This means that there is a significant chance that completely at random a patch will be selected that covers the dimer interface (see section 5.6.2). The second problem is the non-globular nature of some of the hetero-dimers. The small subunit of hydrogenase from *Desulfovibrio desulfuricans* in figure 5.11(a) is so non-globular in shape and so much of its surface is involved in the dimer interface that it is completely unsuitable for interface prediction using patch analysis.

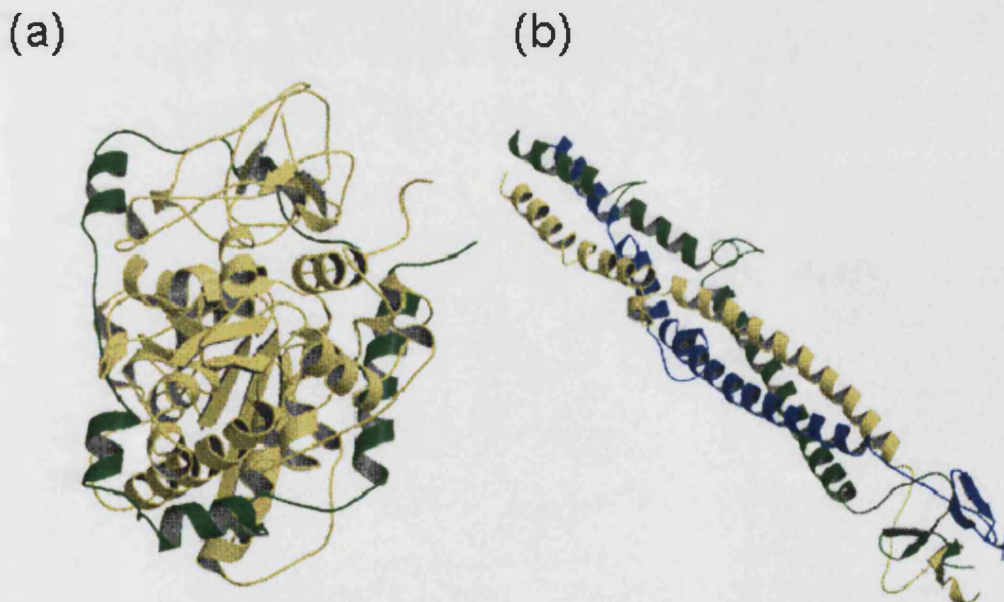


Figure 5.11: Two proteins that are unsuitable for patch analysis. (a) The small subunit of hydrogenase (1hfe). (b) The fibritin trimer from *Bacteriophage T4* (1aa0).

Four out of the twenty six homo-trimers are coiled-coil proteins of the kind shown in figure 5.11(b) and can be considered another class of protein inappropriate for patch analysis. Nevertheless, the fact that a homo-dimer trained neural network can succeed at all in predicting the location of dimer interfaces within hetero-dimers and tetramers is still encouraging.

The results for the dataset of hetero-tetramers are quite mixed. The interface predictions for the small subunits of hetero-tetramers are generally better than those of the large subunits. There are some indications that large homo-dimer like interfaces are correctly located with a higher degree of accuracy than smaller less well defined interfaces. This is to be expected since a homo-dimer trained neural network was used to locate the protein-protein interfaces of the hetero-tetramers. One protein where this is the case is succinyl-coa synthetase from *Escherichia coli* (2scu). A diagram of the full tetramer is shown in the appendix of this thesis. The interface between the two large subunits of 2scu is small only being 864Å in size. Additionally, the interface is non-planar and quite hydrophilic in character. In contrast the interface between the large and small subunit of 2scu is large, planar and hydrophobic. It is consequently no surprise that the interface between the large and small subunits of 2scu is correctly located but the interface between the two large subunits is not.

5.6.2 Assessing the Statistical Significance of the Results

In this section the statistical significance of some of the neural network results is assessed. This is done by calculating for each homo and hetero-dimer the probability of the dimer interface being correctly located completely by chance (the p-value of the protein). The first stage in assigning a p-value to say a homo-dimer is to randomly select three patches from the surface of the protein. The next step is to look at the relative overlap of each of these three patches with the dimer interface. If any one of these three patches has a relative overlap with the dimer interface of $\geq 70\%$ then the dimer interface is regarded to have been correctly located. The p-value of a protein is then calculated from counting the number of times the dimer interface is correctly located after selecting three patches at random from the protein a total of

10000 times. The p-value of the protein is then the total number of times the protein-protein interface is correctly located divided by 10000.

The average p-value for all 76 homo-dimers is 0.39. This shows that on average there is a 39% chance that a homo-dimer interface will be correctly located completely at random. The average p-value for the dataset of hetero-dimers is 0.59. This value is high and is a consequence of the fact that hetero-dimers on average bury ~25% of their surface area in the dimer interface (see table 4.2). Tables 5.1, 5.4, and 5.5 show the relative overlaps of the three patches for each homo and hetero-dimer that have been ranked by the neural network as most likely corresponding to protein-protein interaction sites. Of these three patches, the relative overlap of the patch with the highest relative overlap (the 'best' patch) is taken and plotted together with the p-value of the protein in figure 5.12.

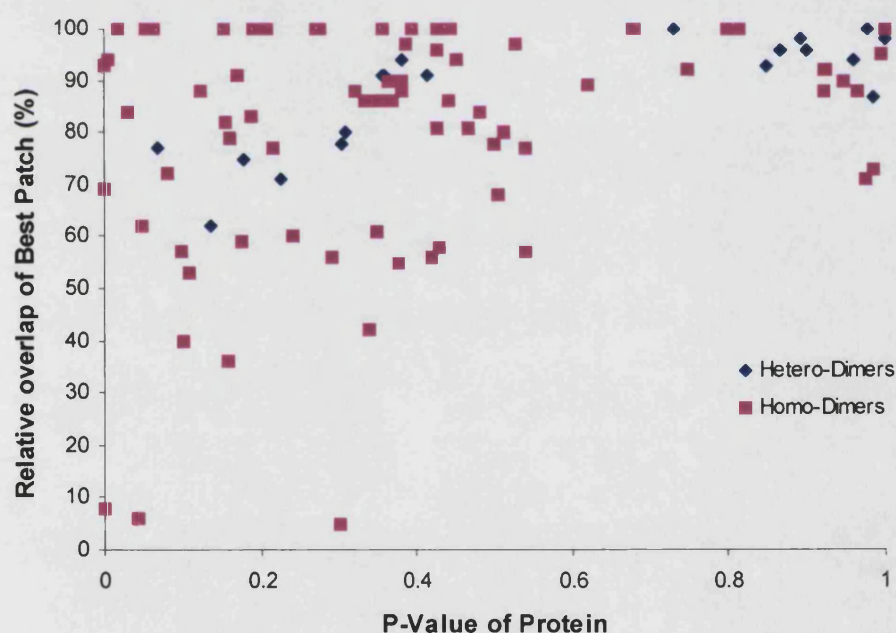


Figure 5.12: A chart showing the p-value of each homo and hetero-dimer together with the relative overlap of the 'best patch' with the dimer interface. Of the three patches most highly ranked by the neural-network as corresponding to the dimer interface the 'best patch' is taken to be the patch with the highest relative overlap.

The relative overlap of the 'best' patch is taken to be a measure of the quality of the neural network prediction. If the neural network is simply selecting patches completely at random then the quality of the neural networks predictions should scale

in a linear way with the p-value of the protein. It is clear that as the p-value of the protein increases on average the relative overlap of the ‘best patch’ increases. However, as can be seen from figure 5.12 the relative overlap of the best patch for each protein is often quite high even if the protein has quite a small p-value. This shows that the neural network is not simply selecting patches at random and the neural network results for the homo-dimers are meaningful. The neural network predictions for the homo-tetramers (analyzed in section 5.6.1) also show that the neural network consistently selects patches with distinct characteristics as corresponding to protein-protein interaction sites. For instance the neural network usually locates hydrophobic protein-protein interfaces correctly but not smaller more polar interfaces. This indicates that the results for the homo-tetramers are also meaningful.

5.7 *Future Work*

One of the most serious constraints on the current work has been the relatively small sizes of the datasets of homo and hetero-complexes. It may be that the datasets used in this work are simply too small to train a neural network properly and fully utilise the non-linear modelling abilities of a multi-layer network. The rapid rate at which protein structures are being determined should allow this problem to be addressed in the near future. However, at present it is still not possible to reliably extract proteins of a particular oligomer type from the PDB in an automated way. Obtaining high quality datasets of proteins that can be used to train the neural network will therefore still require a great deal of effort.

As noted in the previous section many multimeric proteins (for example homo-tetramers) contain a number of different protein-protein interfaces of varying sizes and chemical character. It would therefore be prudent to analyse a protein using several neural networks each trained to locate a different kind of protein-protein interface. For instance it could be possible to analyse a single protein using one neural network trained to locate large hydrophobic protein-protein interfaces of the kind found within homo-dimers and a different network trained to locate the smaller

more polar protein-protein interfaces such as are found in antibody-antigen complexes. Potentially with enough data this approach could be taken even further. For example it may be possible to use a network to distinguish between a protein-protein interface and a protein-DNA binding site.

At present only the structural characteristics (i.e size and planarity) of the residues in each surface patch are considered when predicting the location of protein-protein interaction sites. It could be useful to consider sequence data together with structural data when locating protein-protein interfaces. To investigate this conservation scores have been used in addition to the six original patch analysis parameters to locate the dimer interfaces of the 76 homo-dimers. Conservation scores give a quantitative measure as to how conserved a given residue is at the sequence level (Valdar et al., 2001). Intuitively if the multimer is biologically relevant the residues at the protein-protein interfaces of the complex should be conserved and have high conservation scores. For each residue in a surface patch a conservation score is calculated. A mean conservation score is then calculated for all the residues in each surface patch. The mean conservation score of each patch is used as a seventh parameter when training the network (the others being hydrophobicity, protrusion index, residue interface propensity, planarity, salvation potential, and accessible surface area all as defined in section 5.2.2). There are fifty three homo-dimers for which there is sufficient sequence data available to calculate conservation scores. A 53 fold cross validation was then performed on these 53 homo-dimers with the mean patch conservation scores as a seventh input in addition to the usual six patch analysis parameters. A single layer network was used. Overall, 89% of the predictions are correct compared to a success rate of 84% when not using conservation scores (data not shown). This result does show that there may be some additional value in using conservation scores in predicting the location of protein-protein interfaces. One disadvantage of using conservation scores is that the speed of the predictive process would be somewhat reduced.

Another way the neural-network based patch analysis method might be improved is to see how the six patch analysis parameters vary within each surface patch. One way of doing this would be to define a patch within a patch or to split a patch into two zones.

For proteins whose interface(s) have a hydrophobic core the centre of each patch covering the interface will be quite hydrophobic with the outer rim of the patch being relatively polar in character.

5.8 Conclusions

In this chapter a feed-forward neural network has been used to improve on the performance of the original patch analysis method. The results for the dataset of seventy six homo-dimers are encouraging. Using a neural network the protein-protein interfaces of around three quarters of the homo-dimers are correctly located, an improvement of around thirteen per cent on the original patch analysis method.

The neural network does less well when locating the protein-protein interfaces of trimers, and tetramers. In such proteins (especially homo or hetero-tetramers) there are often two very different types of interface. The first type are the large, well defined, and broadly hydrophobic interfaces of the kind found in homo-dimers. These interfaces are usually correctly located by the neural network. The second type are the small and often polar interfaces similar to those found within transitory protein-complexes. These interfaces are as often as not incorrectly located by the network. As discussed in the previous section it will probably prove necessary to use several different neural networks each trained to locate different types of binding site.

One of the advantages of using a neural network together with patch analysis is the speed with which predictions can be obtained. Using a previously trained neural network interface predictions can take under a minute. This compares with docking methods that can take many hours to run using a standard desktop computer.

A serious constraint on the current work is the relatively small amount of data available. The dataset of hetero-tetramers for instance consists of just seven proteins. Neural networks require quite large amounts of data to train them properly and it is probable that the accuracy of the interface predictions are seriously affected by this lack of data. The rapid rate at which protein structures (especially complexes) are being determined should go some way towards alleviating this problem. The

increasing levels of annotation seen in some protein databases will also enable datasets of proteins to be compiled with more ease than has previously been possible.

In conclusion while there is much work still to be done the accuracy of the patch analysis method does seem to have been improved by using a neural network. In cases where the neural network fails to locate the protein-protein interface of interest another binding site is often located instead (for example a ligand binding site). The integration of sequence conservation data may improve the performance of the neural network further. Once this is done it is hoped that the neural network based patch analysis method (or some variant of it) will provide a useful foundation for further work towards answering the question first posed at the beginning of this chapter: given a protein of known structure what does it bind to?

Chapter 6

Conclusions

What makes a binding site a binding site? (Ringe,1995). For nearly thirty years authors have examined the protein-protein interfaces of the protein-complexes that were available to them in order to answer this question (Chothia & Janin, 1975, Argos, 1988, Miller, 1989, Jones & Thornton, 1996, Conte et al., 1999). The work presented in this thesis updates the work of these authors and makes use of the ever increasing numbers of protein structures being deposited in the PDB.

In chapter 2 the procedure that was used to compile the datasets of obligate and non-obligate protein complexes was outlined. These datasets include 142 obligate homo-complexes, 20 obligate hetero-complexes, 20 enzyme-inhibitor complexes, 15 antibody-antigen complexes, and 10 complexes involved in signaling processes. A brief description of the content of some of the datasets was also given.

In chapter 3 the protein-protein interfaces of obligate homo-dimers, trimers, tetramers, and hexamers were analyzed. The average fraction of surface area buried in protein-protein interfaces ranges from 16% for homo-dimers to 26% for homo-hexamers. Aside from size there appears to be few differences between the protein-protein interfaces found within the four different types of homo-complex. In each case protein-protein interfaces are relatively hydrophobic when compared with the entire protein exterior. The three residues that the protein-protein interfaces of the homo-complexes are most enriched in compared with the entire protein exterior are tyrosine, phenylalanine, and isoleucine. The residues that make up the protein-protein interfaces of homo-complexes were also found to be almost as closely packed as the protein interior in agreement with previous studies (Conte et al., 1999).

The number of inter-subunit hydrogen bonds scales in a linear way with the size of the protein-protein interface. For all classes of multimer there is approximately one inter-subunit hydrogen bond for every 100\AA^2 of buried ASA. One notable aspect of the homo-complexes is symmetry. All of the one hundred and forty two homo-complexes are symmetrical, possessing various kinds of symmetries (single rotational axes such as 2-folds in most homo-dimers, or combinations of intersecting axes, such as 222 in most tetramers, or 32 in many hexamers). Asymmetric homo-complexes are comparatively rare in nature. As predicted by Cornish-Bowden & Koshland in 1972, the vast majority of the thirty one homo-tetramers are complexes with only isologous protein-protein interfaces. Heterologous interfaces are only to be found in two of the homo-complexes. Why isologous interfaces appear to be preferred over heterologous interfaces in homo-tetramers is a matter for further investigation.

In chapter 4 datasets of obligate and non-obligate hetero-complexes were studied. As was found for the obligate homo-complexes the protein-protein interfaces of the obligate hetero-dimers, tetramers, and hexamers have a similar chemical composition. The protein-protein interfaces of both the obligate homo and hetero-complexes are intermediate in hydrophobicity between the protein interior and the protein exterior. This fact underlines the central role of the hydrophobic effect in protein-protein interactions. The major difference between the protein-protein interfaces of obligate homo-complexes and obligate hetero-complexes is the size of the interface. The protein-protein interfaces within obligate hetero-complexes are on average larger than those found within obligate homo-complexes. For instance the protein-protein interface within homo-dimers is 1890\AA^2 in size compared with 3310\AA^2 for the obligate hetero-dimers and a t-test shows that these two means do differ to a statistically significant degree (5%). However, the standard deviations on both these two mean values are quite large meaning that it cannot be said that the protein-protein interfaces of obligate hetero-complexes are larger than those found within obligate homo-complexes with absolute certainty.

With the sole exception of methyl-coenzyme M reductase all the obligate hetero-complexes are composed of a small and a large protein subunit. In most cases the genes coding for the small and large subunits of each complex are adjacent to each other on the genome. It was also shown that for at least eleven out of the twenty

obligate hetero-complexes the small and large protein subunits are homologous to each other in whole or part.

In comparing the protein-protein interfaces of obligate and non-obligate protein complexes the following conclusions can be made. As can be seen from tables 4.1-4.4 the protein-protein interfaces within the non-obligate protein-complexes are generally smaller than those within obligate complexes. A small interface size is one of the best indicators that the protein complex in question is non-obligate rather than obligate. The chemical composition of the protein-protein interfaces within obligate protein complexes is quite different from those found within non-obligate protein-complexes. In general, the protein-protein interfaces of non-obligate protein-complexes contain larger numbers of polar and charged residues than do those from obligate protein-complexes. As an example on average 34% of the residues in the protein-protein interfaces of non-obligate protein-complexes are hydrophobic compared with 45% for obligate homo-complexes (see table 4.4).

One class of protein-complex that has not been studied in this thesis is large protein complexes such as the ribosome, ATP synthase, and GROEL. The GROEL chaperone deserves special attention due to its role in facilitating protein folding. An examination of the protein-protein interfaces of ATP synthase may help to explain how the protein subunits of the molecular motor move together in a highly coordinated way. The spatial distribution of polar and non-polar residues across protein-protein interfaces has also not been analyzed. It would be useful to see how many of the protein-protein interfaces of the complexes studied in this thesis have a recognizable hydrophobic core and how many do not. The work of Larsen et al., 1998, suggests that only a minority of homo-dimers have protein-protein interfaces with a single hydrophobic core surrounded by polar and charged residues.

The role of water molecules at protein-protein interfaces has been well studied for a number of high resolution structures (Levitt, 1993, Janin, 1999). It would be useful to look at the role of water molecules at the protein-protein interfaces of the protein-complexes examined in this thesis. Although the properties of protein-protein interaction sites have been studied in this thesis the nature of protein-ligand binding sites has not been considered. It would be worthwhile in future to determine the

properties of ligand binding sites and contrast them with the survey of protein-protein binding sites presented in chapters 3 and 4. An analysis of the protein-protein interfaces of homo-dimers show that residues at the dimer interface are conserved at the sequence level (Valdar & Thornton, 2001). It would be a logical extension of this work to see if residues in the protein-protein interfaces of other categories of protein-complex are also conserved.

In chapter 5 a feed forward neural network was used together with the patch analysis method of Jones & Thornton (1997) to predict the location of protein-protein interfaces in obligate homo and hetero-complexes. In the original patch analysis method a number of patches are defined over the surface of a protein. The physical and chemical characteristics of each patch are encoded in the form of six parameters (hydrophobicity, protrusion index, residue interface propensity, planarity, protrusion index, and accessible surface area). By comparing average values of these six parameters with those of known protein-protein interfaces the likelihood of a patch corresponding to a protein-protein interface can be assessed. The neural network correctly locates the dimer interface of seventy six percent of the seventy six homo-dimers compared with a success rate of sixty three percent using the original patch analysis method alone. This shows that for homo-dimers a neural network improves the performance of the original patch analysis method by around thirteen percent. In addition to homo-dimers the neural network was also tested on homo-trimers, homo-tetramers, hetero-dimers, and hetero-tetramers. In cases where the neural network fails to locate a protein-protein interface a ligand binding site is often located instead. For example, in homo-dimeric endonucleases the DNA binding site is often located rather than the dimer interface.

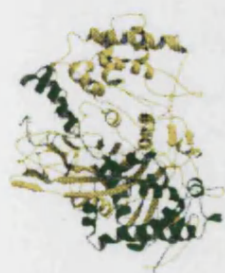
In proteins with more than one protein-protein interaction site (for example homo-tetramers) it is often the case that the neural network can locate large hydrophobic protein-protein interfaces of the kind found in homo-dimers to a much higher degree of accuracy than smaller, more polar, and less closely packed interfaces. A significant constraint on the current work has been the rather small sizes of the datasets of proteins that are used to train and test the neural network. Indeed there are so few hetero-dimers and tetramers that a homo-dimer trained neural network had to be used when testing the neural network on these proteins. It is quite possible that by training

the neural network with larger datasets of proteins the accuracy of the neural network predictions would be considerably enhanced.

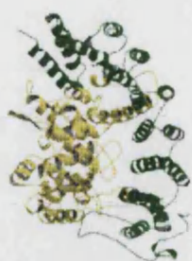
In future it may prove necessary to use a number of different neural networks each trained to locate a different class of interaction site. For example a protein could be examined using a neural network trained to locate DNA binding sites and another network trained to locate homo-dimer like interfaces. It may also be of value to consider the sequence conservation of the residues in each surface patch in addition to the six patch analysis parameters. For fifty three out of the seventy six homo-dimers the use of sequence conservation data improves the accuracy of the neural network predictions by five per cent. Another way the patch analysis method may be improved is to look at the way the six patch analysis parameters vary across each surface patch. For protein-protein interfaces with a hydrophobic core the center of the patch covering the interface will be relatively hydrophobic with the outer rim of the patch being comparatively polar.

As biology moves towards high-throughput methods and proteomics develops, the importance of studying protein-protein complexes is increasingly being recognized. New methods (such as the two-hybrid method used with yeast) are revealing the presence of many previously unrecognized complexes. Consequently, improving on our ability to recognize protein-protein interfaces and predict the geometries of protein-protein complexes remains an important goal. It is clear that to understand the processes of life at the molecular level we will need to better understand the energetics and specificity of protein-protein interactions. This thesis has made some progress towards this goal but much work still needs to be done.

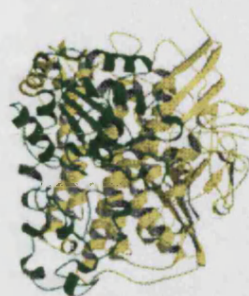
Appendix



1ajq



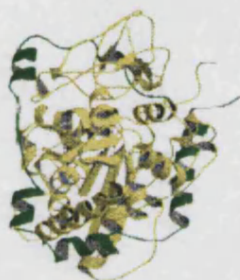
1ft1



1h2a



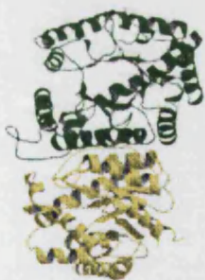
1hcn



1hfe



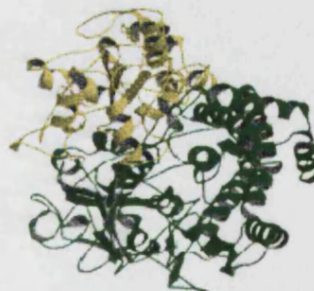
1ixx



1luc



1req

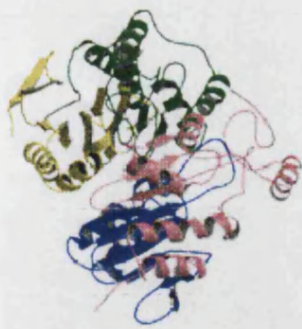


2frv

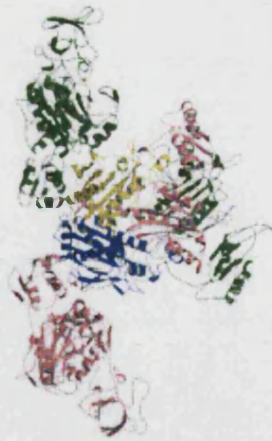


4mon

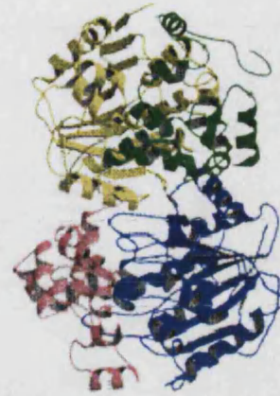
Figure A1: Diagrams of the ten obligate hetero-dimers. In each case the PDB code of the complex is indicated.



1apy



1b7y



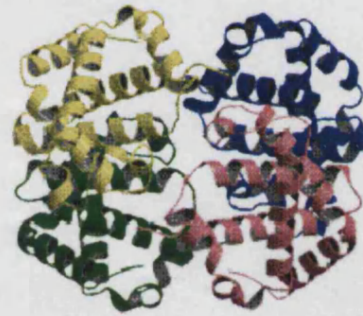
1bou



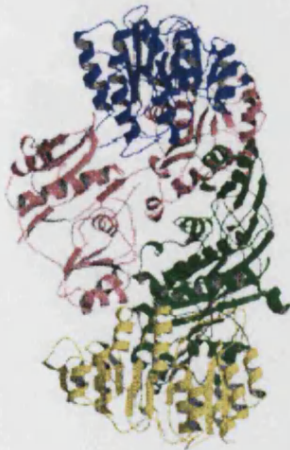
1ccw



1qdl

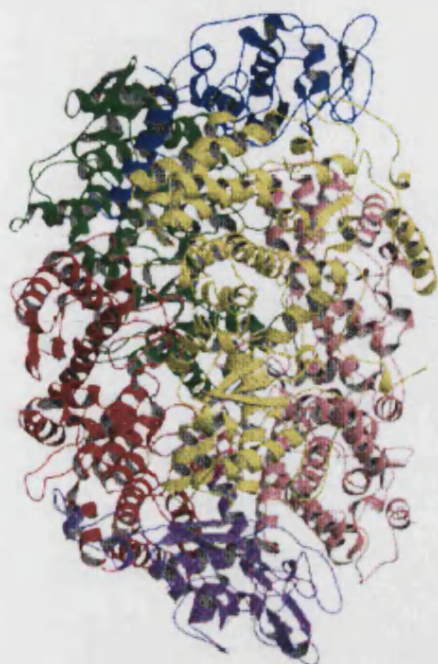


1qsh

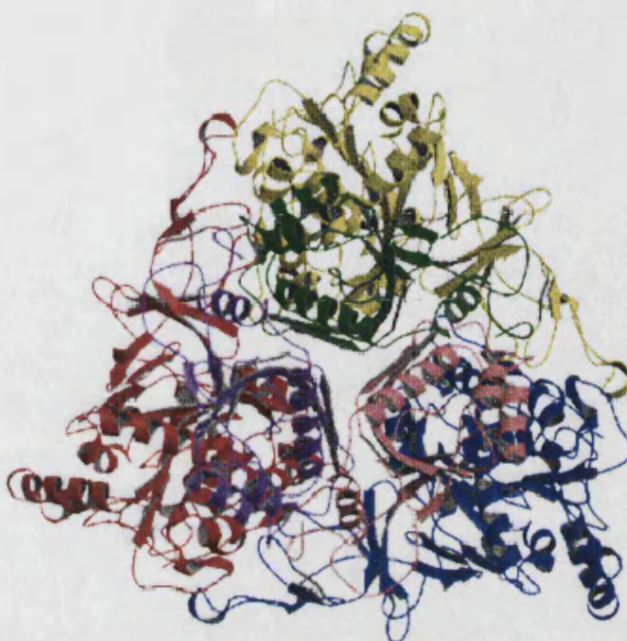


2scu

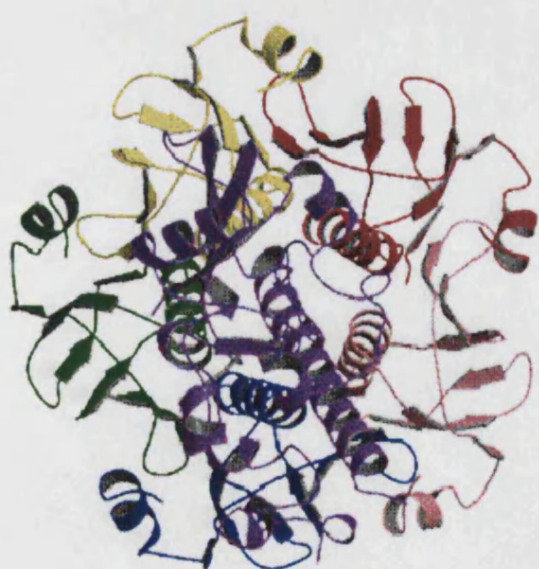
Figure A2: Diagrams of the seven obligate hetero-tetramers. In each case the PDB code of the complex is indicated.



1mro



1eg9



1tii

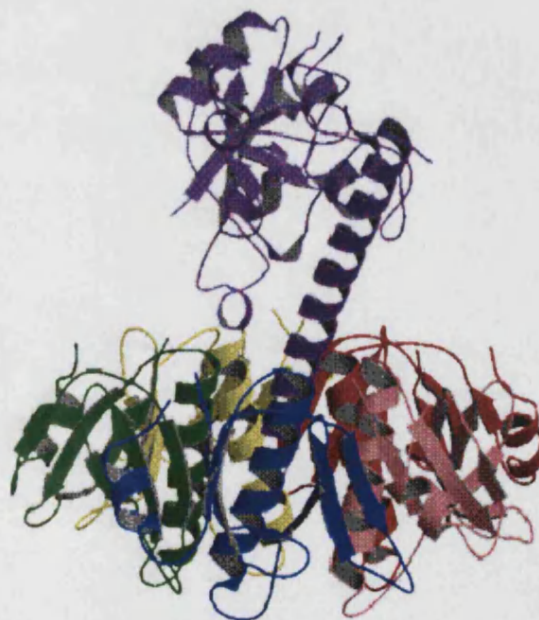
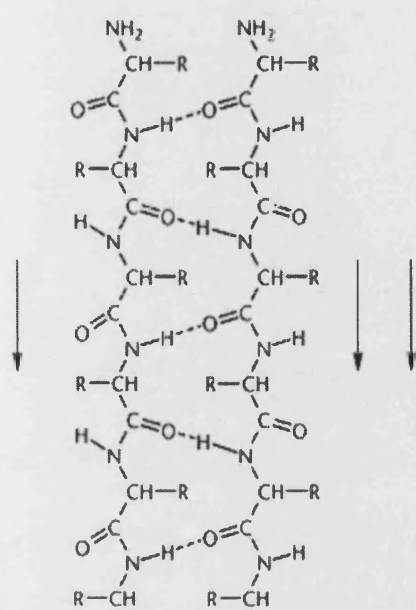
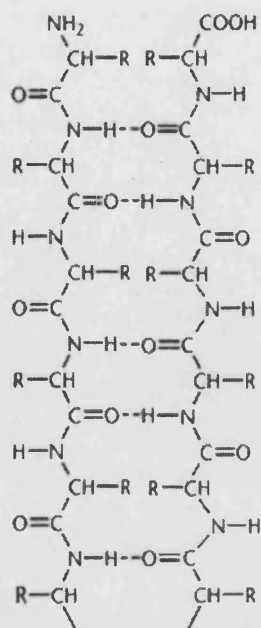


Figure A3: Diagrams of the three obligate hetero-hexamers. In each case the PDB code of the complex is indicated.

(a)

Parallel β pleated sheetAntiparallel β pleated sheet

(b)

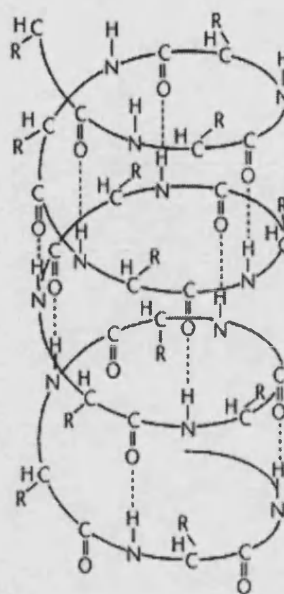
Right-handed α helix

Figure A4: (a) Parallel and anti-parallel β -sheets. (a) A right handed α -helix. Both types of secondary of secondary structure are maintained through hydrogen bonds between C=O and N-H groups. This diagram is taken from 'Proteins: Fundamental Chemical Properties' by Alain Cozzone (2002) and is available at www.els.net.

References

1. Allmansberger, R., Bokranz, M., Krockel, L., Schallenberg, J., & Klein, A. (1989). Conserved gene structures and expression signals in methanogenic archaeobacteria. *Can. J. Microbiol.* **35**, 52-57.
2. Apostoluk, W. & Otlewski, J. (1998). Variability of the canonical loop conformations in serine proteinases inhibitors and other proteins. *Proteins* **32**, 459-474.
3. Argos, P. (1988). An investigation of protein subunit and domain interfaces. *Protein Eng* **2**, 101-113.
4. Auld, D. S., Kornecook, T. J., Bastianetto, S., & Quirion, R. (2002). Alzheimer's disease and the basal forebrain cholinergic system: relations to beta-amyloid peptides, cognition, and treatment strategies. *Prog. Neurobiol.* **68**, 209-245.
5. Bairoch, A. & Apweiler, R. (2000). The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.* **28**, 45-48.
6. Baldwin, T. O., Ziegler, M. M., & Powers, D. A. (1979). Covalent structure of subunits of bacterial luciferase: NH₂-terminal sequence demonstrates subunit homology. *Proc. Natl. Acad. Sci. U. S. A* **76**, 4887-4889.
7. Barlow, D. J. & Thornton, J. M. (1983). Ion-pairs in proteins. *J. Mol. Biol.* **168**, 867-885.
8. Bartlett, G. J., Porter, C. T., Borkakoti, N., & Thornton, J. M. (2002). Analysis of catalytic residues in enzyme active sites. *J. Mol. Biol.* **324**, 105-121.
9. Berman, H. M. (2002). Protein Structures from Famine to Feast. *Scientific American* **90**, 350-359.

10. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., & Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Res.* **28**, 235-242.
11. Bhat, T. N., Bourne, P., Feng, Z., Gilliland, G., Jain, S., Ravichandran, V., Schneider, B., Schneider, K., Thanki, N., Weissig, H., Westbrook, J., & Berman, H. M. (2001). The PDB data uniformity project. *Nucleic Acids Res.* **29**, 214-218.
12. Blundell, T. L. & Srinivasan, N. (1996). Symmetry, stability, and dynamics of multidomain and multicomponent protein systems. *Proc. Natl. Acad. Sci. U. S. A* **93**, 14243-14248.
13. Blundell, T. L., Burke, D. F., Chirgadze, D., Dhanaraj, V., Hyvonen, M., Innis, C. A., Parisini, E., Pellegrini, L., Sayed, M., & Sibanda, B. L. (2000). Protein-protein interactions in receptor activation and intracellular signalling. *Biol. Chem.* **381**, 955-959.
14. Bogan, A. A. & Thorn, K. S. (1998). Anatomy of hot spots in protein interfaces. *J. Mol. Biol.* **280**, 1-9.
15. Bourne, P. & Wessig, H. (2003). *Structural Bioinformatics*, Wiley.
16. Braden, B. C. & Tooze, J. (1998). *Introduction to Protein Structure*, Garland Science.
17. Braden, B. C., Souchon, H., Eisele, J. L., Bentley, G. A., Bhat, T. N., Navaza, J., & Poljak, R. J. (1994). Three-dimensional structures of the free and the antigen-complexed Fab from monoclonal anti-lysozyme antibody D44.1. *J. Mol. Biol.* **243**, 767-781.
18. Brenner, S. E. (2001). A tour of structural genomics. *Nat. Rev. Genet.* **2**, 801-809.
19. Chakrabarti, P. & Janin, J. (2002). Dissecting protein-protein recognition sites. *Proteins* **47**, 334-343.
20. Chan, H. & Dill, K. (1997). Solvation: how to obtain microscopic energies from partitioning and solvation experiments. *Annu. Rev. Biophys. Biomol. Struct.* **26**, 425-459.

21. Chan, H. (2002). Amino acid side-chain hydrophobicity. *Nature Electronic Encyclopedia of Life Sciences (www. els. net)*.
22. Chan, M. K., Mukund, S., Kletzin, A., Adams, M. W., & Rees, D. C. (1995). Structure of a hyperthermophilic tungstopterin enzyme, aldehyde ferredoxin oxidoreductase. *Science* **267**, 1463-1469.
23. Chen, W. J., Andres, D. A., Goldstein, J. L., Russell, D. W., & Brown, M. S. (1991). cDNA cloning and expression of the peptide-binding beta subunit of rat p21ras farnesyltransferase, the counterpart of yeast DPR1/RAM1. *Cell* **66**, 327-334.
24. Chook, Y. M., Ke, H., & Lipscomb, W. N. (1993). Crystal structures of the monofunctional chorismate mutase from *Bacillus subtilis* and its complex with a transition state analog. *Proc. Natl. Acad. Sci. U. S. A* **90**, 8600-8603.
25. Chothia, C. (1975). Structural invariants in protein folding. *Nature* **254**, 304-308.
26. Chothia, C. & Janin, J. (1975). Principles of protein-protein recognition. *Nature* **256**, 705-708.
27. Chothia, C. (1992). Proteins. One thousand families for the molecular biologist. *Nature* **357**, 543-544.
28. Conte, L. L., Chothia, C., & Janin, J. (1999). The atomic structure of protein-protein recognition sites. *J. Mol. Biol.* **285**, 2177-2198.
29. Cornish-Bowden, A. J. & Koshland, D. E., Jr. (1971). The quaternary structure of proteins composed of identical subunits. *J. Biol. Chem.* **246**, 3092-3102.
30. Crick, F. H. C. & Watson, J. D. (1957). Virus structure: general principles: CIBA Foundation Symposium: "The Nature of Virus's", 5-13.
31. D'Alessio, G. (1999). The evolutionary transition from monomeric to oligomeric proteins: tools, the environment, hypotheses. *Prog. Biophys. Mol. Biol.* **72**, 271-298.

32. Davies, D. R., Padlan, E. A., & Sheriff, S. (1990). Antibody-antigen complexes. *Annu. Rev. Biochem.* **59**, 439-473.
33. Davies, D. R. & Cohen, G. H. (1996). Interactions of protein antigens with antibodies. *Proc. Natl. Acad. Sci. U. S. A* **93**, 7-12.
34. Decanniere, K., Transue, T. R., Desmyter, A., Maes, D., Muyldermans, S., & Wyns, L. (2001). Degenerate interfaces in antigen-antibody complexes. *J. Mol. Biol.* **313**, 473-478.
35. Dietmann, S., Park, J., Notredame, C., Heger, A., Lappe, M., & Holm, L. (2001). A fully automatic evolutionary classification of protein folds: Dali Domain Dictionary version 3. *Nucleic Acids Res.* **29**, 55-57.
36. Doolittle, R. F. (1989). Protein structure and the principles of protein conformation, Plenum Press.
37. Dougherty, D. A. (1996). Cation-pi interactions in chemistry and biology: a new view of benzene, Phe, Tyr, and Trp. *Science* **271**, 163-168.
38. Dreyer, M. K. & Schulz, G. E. (1996). Catalytic mechanism of the metal-dependent fuculose aldolase from *Escherichia coli* as derived from the structure. *J. Mol. Biol.* **259**, 458-466.
39. Efstratiadis, A., Posakony, J. W., Maniatis, T., Lawn, R. M., O'Connell, C., Spritz, R. A., DeRiel, J. K., Forget, B. G., Weissman, S. M., Slightom, J. L., Blechl, A. E., Smithies, O., Baralle, F. E., Shoulders, C. C., & Proudfoot, N. J. (1980). The structure and evolution of the human beta-globin gene family. *Cell* **21**, 653-668.
40. Ellis, R. J. (2001). Macromolecular crowding: an important but neglected aspect of the intracellular environment. *Curr. Opin. Struct. Biol.* **11**, 114-119.
41. Ermler, U., Merckel, M., Thauer, R., & Shima, S. (1997). Formylmethanofuran: tetrahydromethanopterin formyltransferase from *Methanopyrus kandleri* - new insights into salt-dependence and thermostability. *Structure.* **5**, 635-646.

42. Fauchere, J. L. & Pliska, V. (1983). Hydrophobic parameters π of amino-acid side chains from the partitioning of N-acetyl-amino-acid amides. *Eur. J. Biochem.* **18**, 369-375.
43. Fersht, A. R. (1987). The hydrogen bond in molecular recognition. *Trends Biochem. Sci.* **12**, 301-304.
44. Fiaux, J., Bertelsen, E. B., Horwich, A. L., & Wuthrich, K. (2002). NMR analysis of a 900K GroEL GroES complex. *Nature* **418**, 207-211.
45. Fischmann, T. O., Bentley, G. A., Bhat, T. N., Boulot, G., Mariuzza, R. A., Phillips, S. E., Tello, D., & Poljak, R. J. (1991). Crystallographic refinement of the three-dimensional structure of the FabD1.3-lysozyme complex at 2.5-A resolution. *J. Biol. Chem.* **266**, 12915-12920.
46. Fisher, A. J., Thompson, T. B., Thoden, J. B., Baldwin, T. O., & Rayment, I. (1996). The 1.5-A resolution crystal structure of bacterial luciferase in low salt conditions. *J. Biol. Chem.* **271**, 21956-21968.
47. Frigerio, F., Coda, A., Pugliese, L., Lionetti, C., Menegatti, E., Amiconi, G., Schnebli, H. P., Ascenzi, P., & Bolognesi, M. (1992). Crystal and molecular structure of the bovine alpha-chymotrypsin-eglin c complex at 2.0 A resolution. *J. Mol. Biol.* **225**, 107-123.
48. Furey, W., Jr., Wang, B. C., Yoo, C. S., & Sax, M. (1983). Structure of a novel Bence-Jones protein (Rhe) fragment at 1.6 A resolution. *J. Mol. Biol.* **167**, 661-692.
49. Gabdouliline, R. R. & Wade, R. C. (1999). On the protein-protein diffusional encounter complex. *J. Mol. Recogniton.* **12**, 226-234.
50. Gallivan, J. P. & Dougherty, D. A. (1999). Cation- π interactions in structural biology. *Proc. Natl. Acad. Sci. U. S. A* **96**, 9459-9464.
51. Gaudet, R., Bohm, A., & Sigler, P. B. (1996). Crystal structure at 2.4 angstroms resolution of the complex of transducin betagamma and its regulator, phosducin. *Cell* **87**, 577-588.

52. Gerstein, M., Tsai, J., & Levitt, M. (1995). The volume of atoms on the protein surface: calculated from simulation, using Voronoi polyhedra. *J. Mol. Biol.* **249**, 955-966.
53. Gerstein, M. & Chothia, C. (1996). Packing at the protein-water interface. *Proc. Natl. Acad. Sci. U. S. A* **93**, 10167-10172.
54. Glaser, F., Steinberg, D. M., Vakser, I. A., & Ben Tal, N. (2001). Residue frequencies and pairing preferences at protein-protein interfaces. *Proteins* **43**, 89-102.
55. Goodsell, D. S. & Olson, A. J. (1993). Soluble proteins: size, shape and function. *Trends Biochem. Sci.* **18**, 65-68.
56. Goodsell, D. S. & Olson, A. J. (2000). Structural symmetry and protein function. *Annu. Rev. Biophys. Biomol. Struct.* **29**, 105-153.
57. Goodwill, K. E., Sabatier, C., Marks, C., Raag, R., Fitzpatrick, P. F., & Stevens, R. C. (1997). Crystal structure of tyrosine hydroxylase at 2.3 Å and its implications for inherited neurodegenerative diseases. *Nat. Struct. Biol.* **4**, 578-585.
58. Halperin, I., Ma, B., Wolfson, H., & Nussinov, R. (2002). Principles of docking: An overview of search algorithms and a guide to scoring functions. *Proteins* **47**, 409-443.
59. Heath, P. J., Stephens, K. M., Monnat, R. J., Jr., & Stoddard, B. L. (1997). The structure of I-Crel, a group I intron-encoded homing endonuclease. *Nat. Struct. Biol.* **4**, 468-476.
60. Hecht, H. J., Erdmann, H., Park, H. J., Sprinzl, M., & Schmid, R. D. (1995). Crystal structure of NADH oxidase from *Thermus thermophilus*. *Nat. Struct. Biol.* **2**, 1109-1114.
61. Helland, R., Otlewski, J., Sundheim, O., Dadlez, M., & Smalas, A. O. (1999). The crystal structures of the complexes between bovine beta-trypsin and ten P1 variants of BPTI. *J. Mol. Biol.* **287**, 923-942.

62. Henrick, K. & Thornton, J. M. (1998). PQS: a protein quaternary structure file server. *Trends Biochem. Sci.* **23**, 358-361.
63. Hofstadter, K., Stuart, F., Jiang, L., Vrijbloed, J. W., & Robinson, J. A. (1999). On the importance of being aromatic at an antibody-protein antigen interface: mutagenesis of the extracellular interferon gamma receptor and recognition by the neutralizing antibody A6. *J. Mol. Biol.* **285**, 805-815.
64. Holm, L. & Sander, C. (1996). Mapping the protein universe. *Science* **273**, 595-603.
65. Hu, Z., Ma, B., Wolfson, H., & Nussinov, R. (2000). Conservation of polar residues as hot spots at protein interfaces. *Proteins* **39**, 331-342.
66. Hubbard, R. (1990). NACCESS, *University College London*.
67. Hubbard, R. E. (2001). Hydrogen Bonds in Proteins: Role and Strength, *Nature Electronic Encyclopedia of Life Sciences* (www.els.net).
68. Ibba, M. & Soll, D. (1999). Quality control mechanisms during translation. *Science* **286**, 1893-1897.
69. Igarashi, N., Moriyama, H., Fujiwara, T., Fukumori, Y., & Tanaka, N. (1997). The 2.8 Å structure of hydroxylamine oxidoreductase from a nitrifying chemoautotrophic bacterium, *Nitrosomonas europaea*. *Nat. Struct. Biol.* **4**, 276-284.
70. Janin, J. & Chothia, C. (1990). The structure of protein-protein recognition sites. *J. Biol. Chem.* **265**, 16027-16030.
71. Janin, J. (1995). Principles of protein-protein recognition from structure to thermodynamics. *Biochimie* **77**, 497-505.
72. Janin, J. (1999). Wet and dry interfaces: the role of solvent in protein-protein and protein-DNA recognition. *Structure. Fold. Des* **7**, R277-R279.

73. Janin, J., Henrick, K., Moult, J., Eyck, L. T., Sternberg, M. J., Vajda, S., Vakser, I., & Wodak, S. J. (2003). CAPRI: a Critical Assessment of PRedicted Interactions. *Proteins* **52**, 2-9.
74. Janvier, B., Mallet, F., Cheynet, V., Dalbon, P., Vernet, G., Besnier, J. M., Choutet, P., Goudeau, A., Mandrand, B., & Barin, F. (1993). Prevalence and persistence of antibody titers to recombinant HIV-1 core and matrix proteins in HIV-1 infection. *J. Acquir. Immune. Defic. Syndr.* **6**, 898-903.
75. Jeffreys, A. J. & Harris, S. (1982). Processes of gene duplication. *Nature* **296**, 9-10.
76. Jeong, H., Mason, S. P., Barabasi, A. L., & Oltvai, Z. N. (2001). Lethality and centrality in protein networks. *Nature* **411**, 41-42.
77. Jones, D. T., Taylor, W. R., & Thornton, J. M. (1992). A new approach to protein fold recognition. *Nature* **358**, 86-89.
78. Jones, S. (1996), PhD Thesis, University of London.
79. Jones, S. & Thornton, J. M. (1995). Protein-protein interactions: a review of protein dimer structures. *Prog. Biophys. Mol. Biol.* **63**, 31-65.
80. Jones, S. & Thornton, J. M. (1996). Principles of protein-protein interactions. *Proc. Natl. Acad. Sci. U. S. A* **93**, 13-20.
81. Jones, S. & Thornton, J. M. (1997a). Analysis of protein-protein interaction sites using surface patches. *J. Mol. Biol.* **272**, 121-132.
82. Jones, S. & Thornton, J. M. (1997b). Prediction of protein-protein interaction sites using patch analysis. *J. Mol. Biol.* **272**, 133-143.
83. Jones, S., Marin, A., & Thornton, J. M. (2000). Protein domain interfaces: characterization and comparison with oligomeric protein interfaces. *Protein Eng* **13**, 77-82.

84. Kabsch, W. & Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**, 2577-2637.
85. Kaldor, S. W., Kalish, V. J., Davies, J. F., Shetty, B. V., Fritz, J. E., Appelt, K., Burgess, J. A., Campanale, K. M., Chirgadze, N. Y., Clawson, D. K., Dressman, B. A., Hatch, S. D., Khalil, D. A., Kosa, M. B., Lubbehusen, P. P., Muesing, M. A., Patick, A. K., Reich, S. H., Su, K. S., & Tatlock, J. H. (1997). Viracept (nelfinavir mesylate, AG1343): a potent, orally bioavailable inhibitor of HIV-1 protease. *J. Med. Chem.* **40**, 3979-3985.
86. Kankare, J., Neal, G. S., Salminen, T., Glumoff, T., Glumhoff, T., Cooperman, B. S., Lahti, R., & Goldman, A. (1994). The structure of E.coli soluble inorganic pyrophosphatase at 2.7 Å resolution. *Protein Eng* **7**, 823-830.
87. Kim, K. K., Song, H. K., Shin, D. H., Hwang, K. Y., Choe, S., Yoo, O. J., & Suh, S. W. (1997). Crystal structure of carboxylesterase from *Pseudomonas fluorescens*, an alpha/beta hydrolase with broad substrate specificity. *Structure*. **5**, 1571-1584.
88. Kleanthous, C. E. (2000). Principles of Protein-Protein Recognition, Oxford University Press.
89. Kolatkar, P. R., Ernst, S. R., Hackert, M. L., Ogata, C. M., Hendrickson, W. A., Merritt, E. A., & Phizackerley, R. P. (1992). Structure determination and refinement of homotetrameric hemoglobin from *Urechis caupo* at 2.5 Å resolution. *Acta Crystallogr. B* **48** (Pt 2), 191-199.
90. Korn, A. P. & Burnett, R. M. (1991). Distribution and complementarity of hydrophathy in multisubunit proteins. *Proteins* **9**, 37-55.
91. Korn, A. P. & Burnett, R. M. (1991). Distribution and complementarity of hydrophathy in multisubunit proteins. *Proteins* **9**, 37-55.
92. Koronakis, V., Sharff, A., Koronakis, E., Luisi, B., & Hughes, C. (2000). Crystal structure of the bacterial membrane protein TolC central to multidrug efflux and protein export. *Nature* **405**, 914-919.

93. Kresge, N., Vacquier, V. D., & Stout, C. D. (2000). 1.35 and 2.07 Å resolution structures of the red abalone sperm lysin monomer and dimer reveal features involved in receptor binding. *Acta Crystallogr. D. Biol. Crystallogr.* **56** (Pt 1), 34-41.
94. Kumar, S., Ma, B., Tsai, C. J., & Nussinov, R. (2000). Electrostatic strengths of salt bridges in thermophilic and mesophilic glutamate dehydrogenase monomers. *Proteins* **38**, 368-383.
95. Lackner, P., Koppensteiner, W. A., Sippl, M. J., & Domingues, F. S. (2000). ProSup: a refined tool for protein structure alignment. *Protein Eng* **13**, 745-752.
96. Lakey, J. H. & Gokce, I. (2001). Protein-Protein Interactions. *Nature Electronic Encyclopedia of Life Sciences* (www.els.net).
97. Lander, E. S et al., (2001). Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921.
98. Larsen, T. A., Olson, A. J., & Goodsell, D. S. (1998). Morphology of protein-protein interfaces. *Structure.* **6**, 421-427.
99. Laskowski, M., Jr. Kato, I. (1980). Protein inhibitors of proteinases. *Annu. Rev. Biochem.* **49**, 593-626.
100. Laskowski, R. A. (1995). SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions. *J. Mol. Graph.* **13**, 323-328.
101. Laskowski, R. A., Luscombe, N. M., Swindells, M. B., & Thornton, J. M. (1996). Protein clefts in molecular recognition and function. *Protein Sci.* **5**, 2438-2452.
102. Lavie, A., Ostermann, N., Brundiers, R., Goody, R. S., Reinstein, J., Konrad, M., & Schlichting, I. (1998). Structural basis for efficient phosphorylation of 3'-azidothymidine monophosphate by *Escherichia coli* thymidylate kinase. *Proc. Natl. Acad. Sci. U. S. A* **95**, 14045-14050.

103. Lawrence, M. C. & Colman, P. M. (1993). Shape complementarity at protein/protein interfaces. *J. Mol. Biol.* **234**, 946-950.
104. Lebioda, L., Stec, B., & Brewer, J. M. (1989). The structure of yeast enolase at 2.25-Å resolution. An 8-fold beta + alpha-barrel with a novel beta beta alpha alpha (beta alpha)₆ topology. *J. Biol. Chem.* **264**, 3685-3693.
105. Lee, B. & Richards, F. M. (1971). The interpretation of protein structures: estimation of static accessibility. *J. Mol. Biol.* **55**, 379-400.
106. Lee, R. H. & Rose, G. D. (1985). Molecular recognition. I. Automatic identification of topographic surface features. *Biopolymers* **24**, 1613-1627.
107. Leslie, A. G. (1990). Refined crystal structure of type III chloramphenicol acetyltransferase at 1.75 Å resolution. *J. Mol. Biol.* **213**, 167-186.
108. Levitt, M. & Park, B. H. (1993). Water: now you see it, now you don't. *Structure*. **1**, 223-226.
109. Li, M., Philip, L. H., Lees, W. E., Winther, J. R., Dunn, B. M., Wlodawer, A., Kay, J., & Gustchina, A. (2000). The aspartic proteinase from *Saccharomyces cerevisiae* folds its own inhibitor into a helix. *Nat. Struct. Biol.* **7**, 113-117.
110. Li, Y., Li, H., Smith-Gill, S. J., & Mariuzza, R. A. (2000). Three-dimensional structures of the free and antigen-bound Fab from monoclonal antilysozyme antibody HyHEL-63. *Biochemistry* **39**, 6296-6309.
111. Lichtarge, O. & Sowa, M. E. (2002). Evolutionary predictions of binding surfaces and interactions. *Curr. Opin. Struct. Biol.* **12**, 21-27.
112. Lijnzaad, P. & Argos, P. (1997). Hydrophobic patches on protein subunit interfaces: characteristics and prediction. *Proteins* **28**, 333-343.

113. Liu, Y. & Eisenberg, D. (2002). 3D domain swapping: as domains continue to swap. *Protein Sci.* **11**, 1285-1299.
114. Lohse, D. L. & Fitzpatrick, P. F. (1993). Identification of the intersubunit binding region in rat tyrosine hydroxylase. *Biochem. Biophys. Res. Commun.* **197**, 1543-1548.
115. MacCallum, R. M., Martin, A. C., & Thornton, J. M. (1996). Antibody-antigen interactions: contact analysis and binding site topography. *J. Mol. Biol.* **262**, 732-745.
116. Mancina, F., Keep, N. H., Nakagawa, A., Leadlay, P. F., McSweeney, S., Rasmussen, B., Bosecke, P., Diat, O., & Evans, P. R. (1996). How coenzyme B12 radicals are generated: the crystal structure of methylmalonyl-coenzyme A mutase at 2 Å resolution. *Structure.* **4**, 339-350.
117. Mate, M. J., Zamocky, M., Nykyri, L. M., Herzog, C., Alzari, P. M., Betzel, C., Koller, F., & Fita, I. (1999). Structure of catalase-A from *Saccharomyces cerevisiae*. *J. Mol. Biol.* **286**, 135-149.
118. Matsuda, K., Mizuguchi, K., Nishioka, T., Kato, H., Go, N., & Oda, J. (1996). Crystal structure of glutathione synthetase at optimal pH: domain architecture and structural similarity with other proteins. *Protein Eng* **9**, 1083-1092.
119. Matthews, B. W. (2001). Hydrophobic Interactions In Proteins. *Nature Electronic Encyclopedia of Life Sciences (www. els. net)*.
120. McCoy, A. J., Chandana, E., V, & Colman, P. M. (1997). Electrostatic complementarity at protein/protein interfaces. *J. Mol. Biol.* **268**, 570-584.
121. McDonald, I. K. & Thornton, J. M. (1994). Satisfying hydrogen bonding potential in proteins. *J. Mol. Biol.* **238**, 777-793.
122. McGaughey, G. B., Gagne, M., & Rappe, A. K. (1998). pi-Stacking interactions. Alive and well in proteins. *J. Biol. Chem.* **273**, 15458-15463.

123. Miles, E. W. (2001). Tryptophan synthase: a multienzyme complex with an intramolecular tunnel. *Chem. Rec.* **1**, 140-151.
124. Miller, S., Lesk, A. M., Janin, J., & Chothia, C. (1987). The accessible surface area and stability of oligomeric proteins. *Nature* **328**, 834-836.
125. Miller, S., Janin, J., Lesk, A. M., & Chothia, C. (1987). Interior and surface of monomeric proteins. *J. Mol. Biol.* **196**, 641-656.
126. Miller, S., Lesk, A. M., Janin, J., & Chothia, C. (1987). The accessible surface area and stability of oligomeric proteins. *Nature* **328**, 834-836.
127. Miller, S. (1989). The structure of interfaces between subunits of dimeric and tetrameric proteins. *Protein Eng* **3**, 77-83.
128. Mitsui, Y., Satow, Y., Watanabe, Y., Hirono, S., & Iitaka, Y. (1979). Crystal structures of *Streptomyces subtilisin* inhibitor and its complex with subtilisin BPN'. *Nature* **277**, 447-452.
129. Mittl, P. R., Di Marco, S., Fendrich, G., Pohlig, G., Heim, J., Sommerhoff, C., Fritz, H., Priestle, J. P., & Grutter, M. G. (1997). A new structural class of serine protease inhibitors revealed by the structure of the hirustasin-kallikrein complex. *Structure*. **5**, 253-264.
130. Miyazaki, G., Morimoto, H., Yun, K. M., Park, S. Y., Nakagawa, A., Minagawa, H., & Shibayama, N. (1999). Magnesium(II) and zinc(II)-protoporphyrin IX's stabilize the lowest oxygen affinity state of human hemoglobin even more strongly than deoxyheme. *J. Mol. Biol.* **292**, 1121-1136.
131. Mizuno, H., Fujimoto, Z., Koizumi, M., Kano, H., Atoda, H., & Morita, T. (1997). Structure of coagulation factors IX/X-binding protein, a heterodimer of C-type lectin domains. *Nat. Struct. Biol.* **4**, 438-441.

132. Monaco-Malbet, S., Berthet-Colominas, C., Novelli, A., Battai, N., Piga, N., Cheynet, V., Mallet, F., & Cusack, S. (2000). Mutual conformational adaptations in antigen and antibody upon complex formation between an Fab and HIV-1 capsid protein p24. *Structure. Fold. Des* **8**, 1069-1077.
133. Monod, J., Wyman, J., & Changeux, J. P. (1965). On the nature of allosteric transitions: a plausible model. *J. Mol. Biol.* **12**, 88-188.
134. Morize, I., Surcouf, E., Vaney, M. C., Epelboin, Y., Buehner, M., Fridlansky, F., Milgrom, E., & Mornon, J. P. (1987). Refinement of the C222(1) crystal form of oxidized uteroglobin at 1.34 Å resolution. *J. Mol. Biol.* **194**, 725-739.
135. Mosyak, L., Reshetnikova, L., Goldgur, Y., Delarue, M., & Safro, M. G. (1995). Structure of phenylalanyl-tRNA synthetase from *Thermus thermophilus*. *Nat. Struct. Biol.* **2**, 537-547.
136. Neefjes, J. J., Dierx, J., & Ploegh, H. L. (1993). The effect of anchor residue modifications on the stability of major histocompatibility complex class I-peptide interactions. *Eur. J. Immunol.* **23**, 840-845.
137. Nicholls, A., Sharp, K. A., & Honig, B. (1991). Protein folding and association: insights from the interfacial and thermodynamic properties of hydrocarbons. *Proteins* **11**, 281-296.
138. Nooren, I. M. & Thornton, J. M. (2003). Structural characterisation and functional significance of transient protein-protein interactions. *J. Mol. Biol.* **325**, 991-1018.
139. Norel, R., Lin, S. L., Wolfson, H. J., & Nussinov, R. (1994). Shape complementarity at protein-protein interfaces. *Biopolymers* **34**, 933-940.
140. Ofran, Y. & Rost, B. (2003). Analysing six types of protein-protein interfaces. *J. Mol. Biol.* **325**, 377-387.

141. Okada, K., Hirotsu, K., Sato, M., Hayashi, H., & Kagamiyama, H. (1997). Three-dimensional structure of Escherichia coli branched-chain amino acid aminotransferase at 2.5 Å resolution. *J. Biochem. (Tokyo)* **121**, 637-641.
142. Orengo, C. A., Michie, A. D., Jones, S., Jones, D. T., Swindells, M. B., & Thornton, J. M. (1997). CATH--a hierarchic classification of protein domain structures. *Structure*. **5**, 1093-1108.
143. Ouzounis, C. A. & Karp, P. D. (2000). Global properties of the metabolic map of Escherichia coli. *Genome Res.* **10**, 568-576.
144. Padlan, E. A., Silverton, E. W., Sheriff, S., Cohen, G. H., Smith-Gill, S. J., & Davies, D. R. (1989). Structure of an antibody-antigen complex: crystal structure of the HyHEL-10 Fab-lysozyme complex. *Proc. Natl. Acad. Sci. U. S. A* **86**, 5938-5942.
145. Pain, R. H. (2000). *Mechanisms of Protein Folding*, Oxford University Press.
146. Papageorgiou, A. C., Shapiro, R., & Acharya, K. R. (1997). Molecular recognition of human angiogenin by placental ribonuclease inhibitor--an X-ray crystallographic study at 2.0 Å resolution. *EMBO J.* **16**, 5162-5177.
147. Parge, H. E., Arvai, A. S., Murtari, D. J., Reed, S. I., & Tainer, J. A. (1993). Human CksHs2 atomic structure: a role for its hexameric assembly in cell cycle control. *Science* **262**, 387-395.
148. Pereira, P. J., Bergner, A., Macedo-Ribeiro, S., Huber, R., Matschiner, G., Fritz, H., Sommerhoff, C. P., & Bode, W. (1998). Human beta-tryptase is a ring-like tetramer with active sites facing a central pore. *Nature* **392**, 306-311.
149. Phylip, L. H., Lees, W. E., Brownsey, B. G., Bur, D., Dunn, B. M., Winther, J. R., Gustchina, A., Li, M., Copeland, T., Wlodawer, A., & Kay, J. (2001). The potency and specificity of the interaction between the IA3 inhibitor and its target aspartic proteinase from *Saccharomyces cerevisiae*. *J. Biol. Chem.* **276**, 2023-2030.

150. Ponstingl, H., Kabir, T., & Thornton, J. M. (2003). Automatic inference of protein quaternary structure from crystals. *Journal of Applied Crystallography* **36**, 1116-1122.
151. Ponstingl, H., Henrick, K., & Thornton, J. M. (2000). Discriminating between homodimeric and monomeric proteins in the crystalline state. *Proteins* **41**, 47-57.
152. Price, N. & Stevens, L. (1999). *Fundamentals of Enzymology - Cell and Molecular Biology of Catalytic Proteins*, Oxford University Press.
153. Rafferty, J. B., Sedelnikova, S. E., Hargreaves, D., Artymiuk, P. J., Baker, P. J., Sharples, G. J., Mahdi, A. A., Lloyd, R. G., & Rice, D. W. (1996). Crystal structure of DNA recombination protein RuvA and a model for its binding to the Holliday junction. *Science* **274**, 415-421.
154. Ramakrishnan, V. (2002). Ribosome structure and the mechanism of translation. *Cell* **108**, 557-572.
155. Raschke, T. M., Tsai, J., & Levitt, M. (2001). Quantification of the hydrophobic interaction by simulations of the aggregation of small hydrophobic solutes in water. *Proc. Natl. Acad. Sci. U. S. A* **98**, 5965-5969.
156. Reiersen, H. & Rees, A. R. (2001). The hunchback and its neighbours: proline as an environmental modulator. *Trends Biochem. Sci.* **26**, 679-684.
157. Reshetnikova, L., Moor, N., Lavrik, O., & Vassilyev, D. G. (1999). Crystal structures of phenylalanyl-tRNA synthetase complexed with phenylalanine and a phenylalanyl-adenylate analogue. *J. Mol. Biol.* **287**, 555-568.
158. Richards, F. M. (1974). The interpretation of protein structures: total volume, group volume distributions and packing density. *J. Mol. Biol.* **82**, 1-14.
159. Ringe, D. (1995). What makes a binding site a binding site? *Curr. Opin. Struct. Biol.* **5**, 825-829.

160. Rittinger, K., Walker, P. A., Eccleston, J. F., Smerdon, S. J., & Gamblin, S. J. (1997). Structure at 1.65 Å of RhoA and its GTPase-activating protein in complex with a transition-state analogue. *Nature* **389**, 758-762.
161. Russo, A. A., Jeffrey, P. D., Patten, A. K., Massague, J., & Pavletich, N. P. (1996). Crystal structure of the p27Kip1 cyclin-dependent-kinase inhibitor bound to the cyclin A-Cdk2 complex. *Nature* **382**, 325-331.
162. Scandurra, R., Consalvi, V., Chiaraluce, R., Politi, L., & Engel, P. C. (1998). Protein thermostability in extremophiles. *Biochimie* **80**, 933-941.
163. Schreiber, G. & Fersht, A. R. (1996). Rapid, electrostatically assisted association of proteins. *Nat. Struct. Biol.* **3**, 427-431.
164. Schwikowski, B., Uetz, P., & Fields, S. (2000). A network of protein-protein interactions in yeast. *Nat. Biotechnol.* **18**, 1257-1261.
165. Selzer, T. & Schreiber, G. (1999). Predicting the rate enhancement of protein complex formation from the electrostatic energy of interaction. *J. Mol. Biol.* **287**, 409-419.
166. Shamoo, Y. & Steitz, T. A. (1999). Building a replisome from interacting pieces: sliding clamp complexed to a peptide from DNA polymerase and a polymerase editing complex. *Cell* **99**, 155-166.
167. Sharp, K., Fine, R., & Honig, B. (1987). Computer simulations of the diffusion of a substrate to an active site of an enzyme. *Science* **236**, 1460-1463.
168. Shaw, A., Fortes, P. A., Stout, C. D., & Vacquier, V. D. (1995). Crystal structure and subunit dynamics of the abalone sperm lysin dimer: egg envelopes dissociate dimers, the monomer is the active species. *J. Cell Biol.* **130**, 1117-1125.
169. Sheinerman, F. B., Norel, R., & Honig, B. (2000). Electrostatic aspects of protein-protein interactions. *Curr. Opin. Struct. Biol.* **10**, 153-159.

170. Shepherd, A. J., Gorse, D., & Thornton, J. M. (1999). Prediction of the location and type of beta-turns in proteins using neural networks. *Protein Sci.* **8**, 1045-1055.
171. Smith, G. R. & Sternberg, M. J. (2002). Prediction of protein-protein interactions by docking methods. *Curr. Opin. Struct. Biol.* **12**, 28-35.
172. Smyth, C., Kalsi, G., Brynjolfsson, J., O'Neill, J., Curtis, D., Rifkin, L., Moloney, E., Murphy, P., Sherrington, R., Petursson, H., & Gurling, H. (1996). Further tests for linkage of bipolar affective disorder to the tyrosine hydroxylase gene locus on chromosome 11p15 in a new series of multiplex British affective disorder pedigrees. *Am. J. Psychiatry* **153**, 271-274.
173. Somers, W., Ultsch, M., De Vos, A. M., & Kossiakoff, A. A. (1994). The X-ray structure of a growth hormone-prolactin receptor complex. *Nature* **372**, 478-481.
174. Somers, W., Ultsch, M., De Vos, A. M., & Kossiakoff, A. A. (1994). The X-ray structure of a growth hormone-prolactin receptor complex. *Nature* **372**, 478-481.
175. Song, H. K. & Suh, S. W. (1998). Kunitz-type soybean trypsin inhibitor revisited: refined structure of its complex with porcine trypsin reveals an insight into the interaction between a homologous inhibitor from *Erythrina caffra* and tissue-type plasminogen activator. *J. Mol. Biol.* **275**, 347-363.
176. Sprang, S., Yang, D., & Fletterick, R. J. (1979). Solvent accessibility properties of complex proteins. *Nature* **280**, 333-335.
177. Stock, D., Leslie, A. G., & Walker, J. E. (1999). Molecular architecture of the rotary motor in ATP synthase. *Science* **286**, 1700-1705.
178. Strater, N., Schnappauf, G., Braus, G., & Lipscomb, W. N. (1997). Mechanisms of catalysis and allosteric regulation of yeast chorismate mutase from crystal structures. *Structure.* **5**, 1437-1452.

179. Stryer, L., Berg, J. M., & Tymoczko, J. L. (2002). *Biochemistry* W H Freeman.
180. Stubbs, M. T., Laber, B., Bode, W., Huber, R., Jerala, R., Lenarcic, B., & Turk, V. (1990). The refined 2.4 Å X-ray crystal structure of recombinant human stefin B in complex with the cysteine proteinase papain: a novel type of proteinase inhibitor interaction. *EMBO J.* **9**, 1939-1947.
181. Sugimoto, K., Senda, T., Aoshima, H., Masai, E., Fukuda, M., & Mitsui, Y. (1999). Crystal structure of an aromatic ring opening dioxygenase LigAB, a protocatechuate 4,5-dioxygenase, under aerobic conditions. *Structure. Fold. Des* **7**, 953-965.
182. Sweet, R. M., Wright, H. T., Janin, J., Chothia, C. H., & Blow, D. M. (1974). Crystal structure of the complex of porcine trypsin with soybean trypsin inhibitor (Kunitz) at 2.6-Å resolution. *Biochemistry* **13**, 4212-4228.
183. Taylor, J. S. & Burnett, R. M. (2000). DARWIN: a program for docking flexible molecules. *Proteins* **41**, 173-191.
184. Taylor, W., Thornton, J. M., & Turnell, W. G. (1983). An ellipsoidal approximation of protein shape. *Journal of Molecular Graphics* **1**, 30-38.
185. Teller, D. C. (1976). Accessible area, packing volumes and interaction surfaces of globular proteins. *Nature* **260**, 729-731.
186. Tsai, C. J., Lin, S. L., Wolfson, H. J., & Nussinov, R. (1997). Studies of protein-protein interfaces: a statistical analysis of the hydrophobic effect. *Protein Sci.* **6**, 53-64.
187. Tsai, J., Taylor, R., Chothia, C., & Gerstein, M. (1999). The packing density in proteins: standard radii and volumes. *J. Mol. Biol.* **290**, 253-266.
188. Tsumoto, K., Ogasahara, K., Ueda, Y., Watanabe, K., Yutani, K., & Kumagai, I. (1995). Role of Tyr residues in the contact region of anti-lysozyme monoclonal antibody HyHEL10 for antigen binding. *J. Biol. Chem.* **270**, 18551-18557.

189. Tucker, C. L., Gera, J. F., & Uetz, P. (2001). Towards an understanding of complex protein networks. *Trends Cell Biol.* **11**, 102-106.
190. Valdar, W. S. & Thornton, J. M. (2001). Protein-protein interfaces: analysis of amino acid conservation in homodimers. *Proteins* **42**, 108-124.
191. Valdar, W. S. & Thornton, J. M. (2001). Conservation helps to identify biologically relevant crystal contacts. *J. Mol. Biol.* **313**, 399-416.
192. van den, A. F., Sarfaty, S., Twiddy, E. M., Connell, T. D., Holmes, R. K., & Hol, W. G. (1996). Crystal structure of a new heat-labile enterotoxin, LT-IIb. *Structure.* **4**, 665-678.
193. Venter, J. C et al., (2001). The Sequence of the Human Genome. *Science* **291**, 1304-1351.
194. Vihinen, M. (1987). Relationship of protein flexibility to thermostability. *Protein Eng* **1**, 477-480.
195. Vonrhein, C., Bonisch, H., Schafer, G., & Schulz, G. E. (1998). The structure of a trimeric archaeal adenylate kinase. *J. Mol. Biol.* **282**, 167-179.
196. Vrana, K. E., Walker, S. J., Rucker, P., & Liu, X. (1994). A carboxyl terminal leucine zipper is required for tyrosine hydroxylase tetramer formation. *J. Neurochem.* **63**, 2014-2020.
197. Wade, R. C., Gabdoulline, R. R., Ludemann, S. K., & Lounnas, V. (1998). Electrostatic steering and ionic tethering in enzyme-ligand binding: insights from simulations. *Proc. Natl. Acad. Sci. U. S. A* **95**, 5942-5949.
198. Wallace, A. C., Laskowski, R. A., & Thornton, J. M. (1995). LIGPLOT: a program to generate schematic diagrams of protein-ligand interactions. *Protein Eng* **8**, 127-134.

199. Watanabe, K., Chishiro, K., Kitamura, K., & Suzuki, Y. (1991). Proline residues responsible for thermostability occur with high frequency in the loop regions of an extremely thermostable oligo-1,6-glucosidase from *Bacillus thermoglucosidasius* KP1006. *J. Biol. Chem.* **266**, 24287-24294.
200. Watanabe, K., Hata, Y., Kizaki, H., Katsube, Y., & Suzuki, Y. (1997). The refined crystal structure of *Bacillus cereus* oligo-1,6-glucosidase at 2.0 Å resolution: structural characterization of proline-substitution sites for protein thermostabilization. *J. Mol. Biol.* **269**, 142-153.
201. Watson, H. C., Walker, N. P., Shaw, P. J., Bryant, T. N., Wendell, P. L., Fothergill, L. A., Perkins, R. E., Conroy, S. C., Dobson, M. J., Tuite, M. F., & . (1982). Sequence and structure of yeast phosphoglycerate kinase. *EMBO J.* **1**, 1635-1640.
202. Weil, C. F., Sherf, B. A., & Reeve, J. N. (1989). A comparison of the methyl reductase genes and gene products. *Can. J. Microbiol.* **35**, 101-108.
203. Williams, J. C., Zeelen, J. P., Neubauer, G., Vriend, G., Backmann, J., Michels, P. A., Lambeir, A. M., & Wierenga, R. K. (1999). Structural and mutagenesis studies of leishmania triosephosphate isomerase: a point mutation can convert a mesophilic enzyme into a superstable enzyme without losing catalytic power. *Protein Eng* **12**, 243-250.
204. Wilson, I. A. & Stanfield, R. L. (1994). Antibody-antigen interactions: new structures and new conformational changes. *Curr. Opin. Struct. Biol.* **4**, 857-867.
205. Wu, H., Lustbader, J. W., Liu, Y., Canfield, R. E., & Hendrickson, W. A. (1994). Structure of human chorionic gonadotropin at 2.6 Å resolution from MAD analysis of the selenomethionyl protein. *Structure.* **2**, 545-558.
206. Xu, D., Tsai, C. J., & Nussinov, R. (1997). Hydrogen bonds and salt bridges across protein-protein interfaces. *Protein Eng* **10**, 999-1012.

207. Yamamoto, T. & Yokota, T. (1981). Escherichia coli heat-labile enterotoxin genes are flanked by repeated deoxyribonucleic acid sequences. *J. Bacteriol.* **145**, 850-860.
208. Yang, D. S., Hon, W. C., Bubanko, S., Xue, Y., Seetharaman, J., Hew, C. L., & Sicheri, F. (1998). Identification of the ice-binding surface on a type III antifreeze protein with a "flatness function" algorithm. *Biophys. J.* **74**, 2142-2151.
209. Young, L., Jernigan, R. L., & Covell, D. G. (1994). A role for surface hydrophobicity in protein-protein recognition. *Protein Sci.* **3**, 717-729.
210. Zutshi, R., Brickner, M., & Chmielewski, J. (1998). Inhibiting the assembly of protein-protein interfaces. *Curr. Opin. Chem. Biol.* **2**, 62-66.