PROTEIN-PROTEIN INTERACTIONS

Susan Jones

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

November 1995
Protein-protein interactions are fundamental to many biological processes. This work updates the current knowledge of the nature of these interactions using the 3-dimensional co-ordinates of protein complexes deposited in the Brookhaven Protein Data Bank (PDB). The characterisation of interfaces within a data set of 32 homo-dimers is presented, including the analysis of interface size, shape complementarity, amino acid composition, polarity and segmentation. The interfaces in these complexes were observed to be, in general, circular, hydrophobic patches on the surface of proteins, which were planar, and segmented with respect to the amino acid sequence.

Interface properties were also analysed in different types of protein-protein complex, including dimers, trimers, enzyme-inhibitors, and antibody-proteins. These were classified as permanent (where the components only occur and function in a complexed state), and 'optional' (where the components also exist independently). The 'optional' complexes were less hydrophobic, and contained more inter-molecular hydrogen bonds. These complexes also had smaller and less well packed interfaces than those that only exist in the complexed state.

Protein-protein interface properties were also used to differentiate the observed interface from other sites on the protein surface. Circular and contiguous patches of amino acid residues were defined on the surface of protein structures. For all surface patches defined, 6 properties (solvation potential, residue interface propensity, planarity, hydrophobicity, accessible surface area, and protrusion) were calculated, and compared to those of the observed interface. Using this method, homo-dimer interfaces were found to be those patches that were the most planar, hydrophobic, surface accessible and protruding. This method of patch analysis was used in a predictive algorithm, to locate putative interaction sites on the surface of isolated protomers. The algorithm proved successful in the prediction of such sites in homo-dimers and enzyme-inhibitor complexes.
ACKNOWLEDGEMENTS

I would like to thank my supervisor, Janet Thornton, for all her help and encouragement throughout the work that has gone into this thesis. Her suggestions, inspired thinking, and her ability to spot mistakes at a hundred paces have all been invaluable.

In addition, there are numerous BSM group members who have helped me along the way. Everyone deserves a mention for discussions and help at some time or other, but there is only room to single out a few. I would first like to thank Roman Laskowski for teaching me all I needed to know about things mathematical, and for allowing me to use his excellent graph plotting program and mathematical functions. Christine Mason also deserves a thank you for assisting me with my first programming efforts at UCL. Also I would like to thank Malcolm MacArthur who gave much time and effort to the proof reading of this thesis.

On a personal note I would like to thank Wendy for all her help and encouragement over the last 3 years, and in particular for her meticulous proof reading of this thesis. Glenn Myatt and Alex Sike also deserve a mention for their efforts in teaching me programming in the very beginning.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>TITLE</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABSTRACT</td>
<td>2</td>
</tr>
<tr>
<td>ACKNOWLEDGEMENTS</td>
<td>3</td>
</tr>
<tr>
<td>TABLE OF CONTENTS</td>
<td>4</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>8</td>
</tr>
<tr>
<td>LIST OF TABLES</td>
<td>13</td>
</tr>
</tbody>
</table>

## CHAPTER 1: Protein-Protein Interactions

1.1 Introduction ......................................................... 15
1.2 Protein Structure Hierarchy ........................................... 15
1.3 Multimeric Proteins .................................................... 17
  1.3.1 Definitions of Quaternary Structure ......................... 17
  1.3.2 Protein Stoichiometry ........................................... 18
  1.3.3 Protein Stereochemistry ......................................... 19
1.4 'Optional' Protein-Protein Interactions ............................ 20
1.5 Aspects of Protein-Protein Interactions ............................ 21
  1.5.1 Hydrophobicity .................................................. 21
  1.5.2 Electrostatic Interactions ....................................... 22
  1.5.3 Shape Complementarity ......................................... 22
1.6 Studies of Protein-Protein Interactions ............................ 23
  1.6.1 Structure Determination ......................................... 23
  1.6.2 Computational Studies .......................................... 24
  1.6.3 Significance and Application .................................... 26
1.7 Outline of Thesis .................................................... 27

## CHAPTER 2: Analysis of Protein-Protein Interactions in Protein Dimers

2.1 Introduction ......................................................... 28
2.2 Classification of Multimeric Proteins ............................... 29
2.3 The Non-Homologous Data Set of Protein Dimers .................. 30
  2.3.1 Generation of Co-ordinates ..................................... 30
Protein-Protein Interactions

Table of Contents

2.3.2 Validation of Co-ordinates ...........................................................................................30
2.4 Definitions of Protein-Protein Interfaces ...........................................................................32
  2.4.1 The Residue Level Definition ....................................................................................33
  2.4.2 The Atom Level Definition .........................................................................................34
2.5 Characterisation of the Interfaces .......................................................................................34
  2.5.1 Accessible Surface Area ..............................................................................................34
  2.5.2 Sphericity of Dimeric Proteins ....................................................................................37
  2.5.3 Shape of the Dimer Interfaces ......................................................................................38
  2.5.4 Amino Acid Composition ...........................................................................................41
  2.5.5 Distribution of Atom Types .........................................................................................45
  2.5.6 Secondary Structural States ........................................................................................48
  2.5.7 Hydrogen Bonding ......................................................................................................50
  2.5.8 Salt Bridges ................................................................................................................51
  2.5.9 Disulphide Bonding .....................................................................................................52
  2.5.10 Segmentation ............................................................................................................53
  2.5.11 Gap Volumes and Surface Complementarity ..........................................................57
  2.5.12 Protrusion of Interface Residues .................................................................................57
  2.5.13 Flexibility of Atoms in the Dimer Interface ...............................................................60
2.6 A Tool for the Analysis of Interfaces in Multimeric Proteins ............................................64
2.7 Conclusion ............................................................................................................................64

CHAPTER 3: Comparison of Protein-Protein Interactions in Different Categories of Protein
Complex .................................................................................................................................68

3.1 Introduction ........................................................................................................................68
3.2 Categories of Protein Complex ........................................................................................69
3.3 Definition of a Protein-Protein Interface ...........................................................................71
3.4 Characterisation and Comparison of Protein-Protein Interfaces ......................................71
  3.4.1 Interface Size ..............................................................................................................71
  3.4.2 Interface Shape ..........................................................................................................76
  3.4.3 Complementarity Between Surfaces ..........................................................................77
  3.4.4 Residue Interface Propensities ....................................................................................79
  3.4.5 Hydrophobicity Including Hydrogen Bonding ........................................................80
  3.4.6 Segmentation ............................................................................................................82
# Table of Contents

## CHAPTER 4: Patch Analysis of Protein-Protein Interaction Sites

4.1 Introduction ................................................................. 88
4.2 Definition of a Surface Patch ........................................ 89
4.3 Calculation of Patch Parameters .................................... 90
4.3.1 Solvation Potentials .................................................. 90
4.3.2 Planarity .................................................................. 92
4.3.3 Residue Interface Propensities .................................. 92
4.3.4 Hydrophobicity ....................................................... 93
4.3.5 Protrusion Index ...................................................... 94
4.3.6 Accessible Surface Area ........................................... 94
4.4 Evaluation of Patch Parameters ..................................... 95
4.4.1 Homo-Dimers .......................................................... 99
4.4.2 Enzymes ............................................................... 99
4.5 Sampling of the Observed Interface .............................. 100
4.6 Conclusions .............................................................. 102

## CHAPTER 5: Interface Prediction by Patch Analysis

5.1 Introduction .............................................................. 104
5.2 The Surface Residue Patch ........................................... 107
5.3 The Prediction Algorithm ............................................. 108
5.3.1 Individual Parameter Score ....................................... 108
5.3.2 Combined Parameter Score ...................................... 108
5.3.3 Residue Frequencies and Patch Profiles ..................... 109
5.3.4 Input Variables ...................................................... 110
5.4 Evaluation of the Prediction Algorithm .......................... 111
5.4.1 Sampling of the Observed Interface .......................... 111
5.4.2 Effectiveness of the Combined Score ......................... 116
5.5 Prediction of Interaction Sites in Enzymes ...................... 122
5.6 The PATCH Software ................................................ 124
5.7 Conclusions ........................................................... 124
CHAPTER 6: Applications of Patch Analysis in Proteins ...................................................... 127
  6.1 Introduction ............................................................................................................................ 127
  6.2 Multiple Multimeric States within Protein Families ........................................................... 127
    6.2.1 Classification of Oligomeric States in Protein Families ............................................. 127
    6.2.2 Haemoglobin (Hb): Dimeric and Tetrameric ............................................................... 129
  6.3 Crystal Contacts in Sickle Cell Haemoglobin ................................................................. 136
    6.3.1 Structure of Sickle Cell Haemoglobin: Crystal and Polymer ..................................... 136
    6.3.2 Characterisation of Crystal Contacts ........................................................................... 137
    6.3.3 Patch Analysis and Prediction of Crystal Contacts ..................................................... 138
  6.4 Prediction of Protein Antigen Epitopes ............................................................................. 141
    6.4.1 Characterisation of Structural Epitopes ....................................................................... 142
    6.4.2 Prediction of Structural Epitopes .................................................................................. 144
  6.5 Conclusions ............................................................................................................................ 152

CHAPTER 7: Conclusions ........................................................................................................... 154

REFERENCES.............................................................................................................................. 161

ENCLOSURE IN BACK COVER POCKET
LIST OF FIGURES

CHAPTER 1

Figure 1.1: The 20 amino acid residues that occur in proteins ......................................................16

Figure 1.2: MOLSCRIPT diagrams of monomeric and multimeric proteins: ..................................17

Figure 1.3: Modes of association in multimeric proteins: ..............................................................19

Figure 1.4: The calculation of accessible surface area and molecular surface area ....................25

CHAPTER 2

Figure 2.1: Classification of the proteins in the PDB (July 1993 version) in to their biological multimeric states .....................................................................................................................29

Figure 2.2: Percentage of overlapping atoms in 32 protein dimers. ................................................32

Figure 2.3: CPK models of 6 of the protein dimers in the non-homologous data set ..................35

Figure 2.4: Relationship between the molecular weight of the protomer and the interface accessible surface area ...........................................................................................................36

Figure 2.5: Accessible surface area / molecular weight ratio of 32 protein dimers and their protomers.. ..........................................................................................................................37

Figure 2.6: Rms of best fit planes fitted through the interface (one subunit) of 32 protein dimers. .......................................................................................................................................38

Figure 2.7: CPK models of non-planar interfaces: .........................................................................40

Figure 2.8: Shape of the dimer interfaces ......................................................................................41

Figure 2.9: Percentage frequencies of amino acid residues in the interior, interface and exterior of 32 protein dimers. .........................................................................................................42

Figure 2.10: Interface/protein (IP) propensities ..............................................................................43

Figure 2.11: Interface/surface (IS) propensities: ...........................................................................44

Figure 2.12: The percentage frequencies of amino acid residues in the interface and exterior of 32 protein dimers. .........................................................................................................45
Protein-Protein Interactions

List of Figures

Figure 2.13: Percentage frequency of 19 atom types in the interior, interface and exterior of protein dimers. .................................................................46

Figure 2.14: Method of defining 3 zones within a dimer interface. .................................................................47

Figure 2.15: Mean distribution of polar and non-polar atoms within 3 defined zones in dimer interfaces .............................................................................47

Figure 2.16: Percentage frequency of secondary structure motifs in interface residues in protein dimers. ......................................................................................48

Figure 2.17: Classification of dimer interfaces by secondary structure. ..........................................................49

Figure 2.18: MOLSCRIPT diagrams of extended and stacked β sheets ..........................................................50

Figure 2.19: The number of inter-subunit hydrogen bonds and the interface accessible surface area for 32 protein dimers. ........................................................52

Figure 2.20: The number of residue segments in 32 dimer interfaces. ..........................................................54

Figure 2.21: Interface ASA contributions of interface segments ........................................................................55

Figure 2.22: Bonding in dimer interface segments. ...........................................................................................56

Figure 2.23: Gap volumes and accessible surface areas of dimer interfaces. .................................................57

Figure 2.24: The distributions of relative accessibilities in the 20 amino acids ...........................................59

Figure 2.25: Liquorice bond diagrams of dominant residues in 2 dimer interfaces. .........................................60

Figure 2.26: Percentage frequency distributions of B-factors in 23 non-homologous protein dimers: ..............................................................................62

Figure 2.27: Percentage frequency ratios of atomic B-factors for a subset of 23 non-homologous protein dimers .............................................................................63

Figure 2.28: Postscript™ output produced by the PROFACE computational tool; designed to automatically analyse protein-protein interfaces. .............................................................................65
 CHAPTER 3

 Figure 3.1: Interface accessible surface area against molecular weight for homo-dimers and homo-trimers.................................................................75
 Figure 3.2: Interface accessible surface area against molecular weight for hetero-complexes...75
 Figure 3.3: Interface accessible surface area against molecular weight for all dimeric complexes.......................................................................................76
 Figure 3.4: Planarity of protein-protein interfaces: .................................................................................................................................77
 Figure 3.5: Comparison of gap volume indices between different categories of protein complexes..........................................................................................79
 Figure 3.6: Residue interface propensities for permanent and ‘optional’ dimeric complexes...80
 Figure 3.7: Relative hydrophobicity of interior, interface and exterior residues in different types of complex..................................................................................80
 Figure 3.8: Number of overlapping atoms per number of interface residues in protein-protein complexes..................................................................................84
 Figure 3.9: Protein-protein complexes in which one component is moved from its partner perpendicular to the line of the best fit plane through the interface atoms of both components..........................................................85

 CHAPTER 4

 Figure 4.1: Generation of solvent vectors for surface residues............................................90
 Figure 4.2: Solvation potentials for the 20 amino acids ..........................................................91
 Figure 4.3: Distribution of parameters for all patches in HIV protease................................96
 Figure 4.4: Patch analysis distributions for 29 homo-dimers................................................97
 Figure 4.5: Patch analysis distributions for 10 enzymes ........................................................98
 Figure 4.6: Distributions of percentage overlaps in 29 homo-dimers....................................101
CHAPTER 5

Figure 5.1: The relationship between the number of residues in the interface and the number of residues in the protomer. ................................................................. 107

Figure 5.2: Patch profile for subunit A of Mannose Binding Protein................................. 112

Figure 5.3: Flow diagram summarising the events in PATCH, the algorithm, designed to predict putative interaction sites on the surfaces of isolated proteins. .................. 113

Figure 5.4: Percentage differences in the predicted size of the surface patches and the size of the observed interface ................................................................................. 114

Figure 5.5: Distributions of percentage overlap (P1) in 29 homo-dimers. .......................... 117

Figure 5.6: Sampling the observed interface ......................................................................... 119

Figure 5.7: Percentage overlaps for the patch with the highest P1 overlap in the top 3 patches for protomers from 20 homo-dimers with correctly predicted interfaces. .... 120

Figure 5.8: The relationships between the percentage overlap with the observed interface (P1) and the combined rank for 29 protomers ......................................................... 121

Figure 5.9: The relationship between the percentage overlap with the observed interface (P1), and the combined rank for 3 enzymes ......................................................... 123

CHAPTER 6

Figure 6.1: MOLSCRIPT diagrams of the structures of dimeric deoxy-haemoglobin and human tetrameric deoxy-haemoglobin ................................................................. 130

Figure 6.2: Alignment of the amino acid sequence of one α subunit of haemoglobin from the arcid clam and the α subunit from human haemoglobin ............................. 131

Figure 6.3: CPK model of the Hb dimer and tetramer α subunit, indicating the position of the interfaces on each. ................................................................. 133

Figure 6.4: Schematic to show the equivalent interfaces between the α subunit of the Hb tetramer and the Hb dimer .................................................................................. 134

Figure 6.5: The arrangement of subunits in a polymer of sickle cell haemoglobin ............. 137

Figure 6.6: Interaction of tetramers of HbS in the asymmetric unit of the crystal ............... 137

Figure 6.7: Frequency profiles of the HbS and HbA tetramers ............................................. 140
Protein-Protein Interactions

List of Figures

Figure 6.8: Patch analysis distributions for 6 protein antigens ....................................................143
Figure 6.9: Prediction of structural epitope sites on lysozyme ....................................................145
Figure 6.10: MOLSCRIPT diagram of hetero-dimeric human chorionic gonadotropin ...........147
Figure 6.11: Patch profiles for human chorionic gonadotropin ...................................................149
Figure 6.12: Sequence alignment between hCG and LH ..............................................................150
Figure 6.13: Comparative patch profiles between hCG and LH β subunit structures ..........151
LIST OF TABLES

CHAPTER 2
Table 2.1: Non-homologous data set of 32 protein dimers. ........................................31
Table 2.2: Percentage frequencies of hydrophobic, polar and charged residues and ratios of
interface/interior and interface/ exterior for each set of residues. ...............................41
Table 2.3: Definitions used in the classification of 32 dimer interfaces by secondary structure.49
Table 2.4: Classification of 7 β interfaces into 3 types: extended, stacked and complex........50
Table 2.5: Mean number and percentage of inter- and intra- subunit hydrogen bonds in 32
protein dimers. ..............................................................................................................51
Table 2.6: Mean relative accessibilities for each of the 20 amino acids compared in the exterior
and the interface .............................................................................................................58

CHAPTER 3
Table 3.1: The data sets of 4 categories of protein-protein complex ................................70
Table 3.2: The division of the dimeric protein-protein complexes into permanent and ‘optional’
classes. ..............................................................................................................................71
Table 3.3: Results of structural analysis for 5 categories of protein-protein complexes ..........72
Table 3.4: Results of structural analysis for 2 classes of protein-protein complexes ..........73

CHAPTER 4
Table 4.1: (a) Natural Logs of the interface residue propensities (b) The Fauchère and Pliska
(1983) hydrophobicity scale. ............................................................................................93

CHAPTER 5
Table 5.1: Results of the (PATCH) prediction algorithm for protomers from a data set of 29
homo-dimers. ....................................................................................................................118
Table 5.2: Results of the (PATCH) prediction algorithm for enzymes from 3 enzyme-inhibitor
complexes. ....................................................................................................................123
CHAPTER 6

Table 6.1: Protein fold families with protein members in different oligomeric states ..........128

Table 6.2: Results of structural analysis on the protein-protein interfaces within dimeric (4sdh) and tetrameric (2hhb) deoxy-haemoglobins.................................................................132

Table 6.3: Comparison of relative ranking of the observed and ‘equivalent’ interfaces between the Hb dimer and tetramer........................................................................................134

Table 6.4: Results of structural analysis for the interface between the two tetramers in the asymmetric unit of sickle cell haemoglobin (PDB code 1hbs)...............................................138

Table 6.5: Relative rankings of the observed interface in each tetramer of the HbS crystal for 6 parameters......................................................................................................................139

Table 6.6: Percentage overlaps between the top 5 patches with the highest combined rank and 3 observed structural epitopes........................................................................................146

Table 6.7: Substitutions made in the β subunit of hCG and their position .........................148
Protein-Protein Interactions

1.1 Introduction
Interactions between macromolecules, which include the concordant association of proteins, are fundamental to many biological processes. Protein-protein interactions can be classified as ‘optional’ or permanent. In this context ‘optional’ describes interactions involving proteins that exist independently, as well as in complexes. For example, ‘optional’ interactions are central to the functioning of the vertebrate immune system in which antibodies recognise different protein antigens. Similarly ‘optional’ protein-protein interactions occur between different molecules for the processes of enzyme inhibition, neurotransmitter release in the nervous system, and cell adhesion. Permanent protein-protein interactions form the basis of the quaternary structure of multimeric proteins, and represent one of the most complex levels of structural organisation in biological molecules. Multimeric proteins are found in every cellular location, including the cell organelles, the cytosol and cell membranes. Such structures are of great importance biologically as they mediate biochemical phenomena such as enzyme co-operativity and signal transduction.

1.2 Protein Structure Hierarchy
Proteins are organised into a structural hierarchy. The polypeptide chain at the primary structural level, comprises a linear, non-covalently linked amino acid sequence. These amino acid residues (Figure 1.1) are the foundation of protein structure. Secondary structure is the level at which the linear sequences aggregate to form structural motifs such as helices and sheets. The tertiary structure is formed by the packing of the secondary structural elements into one or more compact globular domains. In many cases proteins are composed of only a single polypeptide chain that has tertiary structure as its highest level of organisation, e.g. lysozyme (Figure 1.2a). These are termed monomeric proteins. However, some proteins are composed of more than one polypeptide chain, associated into assemblies possessing a specific quaternary structure e.g. dimeric cytochrome C, and trimeric chorismate mutase (Figures 1.2b and 1.2c). This is the highest level of structural organisation. The quaternary structure of a protein describes the stoichiometry and stereochemistry of assemblies of covalently linked subunits, characterised by the lower levels of structural organisation (Jaenicke, 1987).
### Hydrophobic

<table>
<thead>
<tr>
<th>Residue</th>
<th>Structure</th>
<th>3 letter code</th>
<th>1 letter code</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alanine</td>
<td><img src="image" alt="Alanine" /></td>
<td>ALA</td>
<td>A</td>
</tr>
<tr>
<td>Isoleucine</td>
<td><img src="image" alt="Isoleucine" /></td>
<td>ILE</td>
<td>I</td>
</tr>
<tr>
<td>Leucine</td>
<td><img src="image" alt="Leucine" /></td>
<td>LEU</td>
<td>L</td>
</tr>
<tr>
<td>Methionine</td>
<td><img src="image" alt="Methionine" /></td>
<td>MET</td>
<td>M</td>
</tr>
<tr>
<td>Phenylalanine</td>
<td><img src="image" alt="Phenylalanine" /></td>
<td>PHE</td>
<td>F</td>
</tr>
<tr>
<td>Proline</td>
<td><img src="image" alt="Proline" /></td>
<td>PRO</td>
<td>P</td>
</tr>
<tr>
<td>Valine</td>
<td><img src="image" alt="Valine" /></td>
<td>VAL</td>
<td>V</td>
</tr>
</tbody>
</table>

### Charged

<table>
<thead>
<tr>
<th>Residue</th>
<th>Structure</th>
<th>3 letter code</th>
<th>1 letter code</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aspartic Acid</td>
<td><img src="image" alt="Aspartic Acid" /></td>
<td>ASP</td>
<td>D</td>
</tr>
<tr>
<td>Glutamic Acid</td>
<td><img src="image" alt="Glutamic Acid" /></td>
<td>GLU</td>
<td>E</td>
</tr>
<tr>
<td>Arginine</td>
<td><img src="image" alt="Arginine" /></td>
<td>ARG</td>
<td>R</td>
</tr>
<tr>
<td>Lysine</td>
<td><img src="image" alt="Lysine" /></td>
<td>LYS</td>
<td>K</td>
</tr>
</tbody>
</table>

### Polar

<table>
<thead>
<tr>
<th>Residue</th>
<th>Structure</th>
<th>3 letter code</th>
<th>1 letter code</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asparagine</td>
<td><img src="image" alt="Asparagine" /></td>
<td>ASN</td>
<td>N</td>
</tr>
<tr>
<td>Glutamine</td>
<td><img src="image" alt="Glutamine" /></td>
<td>GLN</td>
<td>Q</td>
</tr>
<tr>
<td>Threonine</td>
<td><img src="image" alt="Threonine" /></td>
<td>THR</td>
<td>T</td>
</tr>
<tr>
<td>Histidine</td>
<td><img src="image" alt="Histidine" /></td>
<td>HIS</td>
<td>H</td>
</tr>
<tr>
<td>Cysteine</td>
<td><img src="image" alt="Cysteine" /></td>
<td>CYS</td>
<td>C</td>
</tr>
<tr>
<td>Serine</td>
<td><img src="image" alt="Serine" /></td>
<td>SER</td>
<td>S</td>
</tr>
<tr>
<td>Tryptophan</td>
<td><img src="image" alt="Tryptophan" /></td>
<td>TRP</td>
<td>W</td>
</tr>
<tr>
<td>Tyrosine</td>
<td><img src="image" alt="Tyrosine" /></td>
<td>TYR</td>
<td>Y</td>
</tr>
</tbody>
</table>

**Figure 1.1:** The 20 amino acid residues that occur in proteins. They are divided into hydrophobic, polar and charged categories. (Glycine can be categorized as hydrophobic, or considered in a category of its own). The 3 letter and 1 letter codes of each residue are also indicated.
1.3 Multimeric Proteins

It was the pioneering dissociation experiments of Svedberg and Fahraeus (1926), and Svedberg (1929), which established that some proteins exist as complex associations. Combinations of association and hybridisation techniques have since led to the discovery of large numbers of proteins possessing quaternary structures.

1.3.1 Definitions of Quaternary Structure

In a discussion of protein quaternary structure it is important to adhere to a single set of definitions. Those widely used in the literature, and adopted here, were derived by Monod et al. (1965) in their theoretical model of allosteric effects in protein structures. An oligomer is defined as a protein assembly containing a finite, relatively small number of identical subunits. Protomers are defined as the identical subunits associated within an oligomeric protein. A

---

**Figure 1.2:** MOLSCRIPT (Kraulis, 1991) diagrams of monomeric and multimeric proteins: (a) monomeric lysozyme, (b) dimeric cytochrome C3, (c) trimeric chorismate mutase.
monomer is defined as the fully dissociated protomer, or any protein that is not made up of subunits. A subunit is purposely undefined, and may be used to refer to any chemically or physically identifiable submolecular entity within a protein, whether identical or different from, other components. Using these definitions the haemoglobin tetramer (comprised of 2 α and 2 β polypeptide chains) is defined as an oligomer consisting of 2 protomers, each consisting of 2 monomers i.e., one α and one β polypeptide chain. The definition of a subunit allows the term to be used for either the α or β monomer, or for the αβ-protomer. The term multimer is also widely used in the literature and is defined, in the current work, as a protein with a finite number of subunits that need not be identical: hence the term defines all proteins possessing a quaternary structure.

1.3.2 Protein Stoichiometry

The stoichiometry of proteins possessing a quaternary structure considers the number of subunits involved in the assemblies. Collated data on the quaternary structure of proteins, reviewed by Klotz et al (1975), revealed that dimeric and tetrameric proteins were the most prevalent. A similar distribution of multimeric states has more recently been observed in the proteins deposited in the Brookhaven Protein Data Bank (PDB) (Jones and Thornton, 1995a).

The way in which subunits associate to form aggregates has been defined as macro-association (Weber, 1992). Within the process of macro-association it is possible to distinguish 3 modes of association, namely, heterologous, isologous and pseudo-isologous (Monod et al., 1965; Matthews and Bernhard, 1973). Each can be defined using 2 terms; binding set (the residues of one protomer involved in binding to one other protomer) and domain of bonding (the 2 linked binding sets) (Monod et al., 1965). In heterologous associations the domain of bonding is made up of 2 different binding sets, and in isologous associations the 2 binding sets involved are identical (Figure 1.3). In pseudo-isologous associations, the domain of bonding comprises 2 almost identical binding sets. The definition of different modes of association raised the question of which one is most prevalent amongst proteins. Monod et al. (1965) proposed that the exclusive use of isologous associations would lead only to dimers and tetramers. In support of this is the prevalence of dimers and tetramer amongst the proteins solved, and by implication isologous associations (Matthews and Bernhard, 1973). However, thermodynamic calculations on the possibilities of all-isologous, all-heterologous and mixed structures gave no indication that isologous associations would be more energetically favourable than heterologous (Cornish-Bowden and Koshland, 1970).
1.3.3 Protein Stereochemistry

Stereochemistry (the spatial arrangements of subunits within a complex) at the quaternary structure level involves the concept of symmetry. Symmetry is the spatial property of a body, by which the body can be brought from an initial position, to another indistinguishable position, by means of a geometrical operation (Ladd, 1992). An initial understanding of the importance of symmetry in oligomeric proteins was derived principally from the comparative studies of myoglobin (Kendrew et al., 1960) and haemoglobin (Perutz and Rossman, 1960). The importance of symmetry in terms of protein structures was also introduced in the theoretical model of the allostERIC effects of enzymes (Monod et al., 1965). From these, and the increasing number of protein structures solved by X-ray crystallography, it was found that many subunits of oligomeric proteins were organised into stable arrays with high symmetry (Gust and Dirks, 1981).

The symmetry of an oligomeric protein affects its properties and functions, and can be crystallographic or non-crystallographic. A crystal can be considered as an infinite ordered array obtained by the repetition of the same object in a regular way throughout space (Jaffe and Orchin, 1965). The unit cell of a crystal is the basic building block, repeated infinitely by translation in 3 dimensions (Glusker and Domenicano, 1992). The asymmetric unit is the basic repeating unit which is related to all the other identical units in the unit cell by the operation of the symmetry elements, and to the contents of the other unit cells by the translations \( a, b \) and \( c \) (Blundell and Johnson, 1976). If the volume of the asymmetric unit of a crystal accommodates...
just one subunit of an oligomeric protein, then the other subunit(s) will be related to it by the 
same symmetry operation(s) that relate the asymmetric units to each other (Banaszak et al., 1981). In this case the symmetry of the oligomer is expressed in the crystallographic symmetry. Alternatively the asymmetric unit of the crystal may accommodate the whole oligomer or more 
than one oligomer. In this case the symmetry of the protein will not be determined by the 
crystallographic symmetry but by a non-crystallographic symmetry operation.

1.4 ‘Optional’ Protein-Protein Interactions

Protein-protein interactions have a fundamental role in many ‘optional’ recognition processes; these include protein-inhibitor and antibody-protein recognition. The structures of a number of 
enzyme-inhibitor complexes have been solved, including a large group of serine proteinases and 
their inhibitors (see review by Bode and Huber (1992)). The subtilisins, which are extracellular 
alkaline proteinases from *Baccillus*, are a further group of protein-inhibitor complexes for 
which structures have been solved (e.g. Bode et al., 1987; Dauter and Betzel, 1991; Takeuchi et al., 1991). The proteolytic functions of these enzymes are important as they have industrial 
applications, e.g. additives in laundry detergents (Dauter and Betzel, 1991). Protein-inhibitor 
complexes have been used to characterise protein-protein interactions in the past (Chothia and 
Janin, 1975; Janin and Chothia, 1990), and differences were observed between these ‘optional’ 
interactions and their permanent counterparts in multimeric structures.

‘Optional’ recognition also occurs between antibodies and antigenic proteins in the vertebrate 
immune system. The structures of a number of antibody-protein complexes have been solved 
(e.g. Sheriff et al., 1987; Padlan et al., 1989; Fischmann et al., 1991) and have been the target 
of many studies (see review by Wilson and Stanfield, 1994). One major area of controversy is 
how recognition is achieved in this specific system, and what part is played by the 
conformational changes that are observed on antigen binding (e.g. Herron et al., 1991; Rini et al., 1992; Friedman et al., 1994).
1.5 Aspects of Protein-Protein Interactions

Protein-protein association involves the specific complementary recognition of 2 macromolecules to form a stable assembly (Duquerroy et al., 1991). The recognition process involves factors favouring and opposing the stable association. Hydrophobic and electrostatic interactions favour the association. The loss of translational and rotational freedom of amino acids on binding opposes the association. The relative importance of each interaction is the matter of some controversy (e.g. Chothia and Janin, 1975). A brief summary of the most important concepts is presented below, but will be addressed in more detail in subsequent chapters.

1.5.1 Hydrophobicity

The hydrophobic interaction is considered to be the primary driving force in the stabilisation of protein associations (Chothia and Janin, 1975; Dill 1990a). The term hydrophobic interaction is used to describe the gain in free energy upon the association of non-polar residues of proteins in an aqueous environment (Kauzmann, 1959). The process of folding and protein-protein aggregation reduces the surface of a protein in contact with water. This is the structural basis of the hydrophobic effect in proteins. The folding of polypeptide chains and aggregation of subunits buries the hydrophobic residues of the proteins, and hence minimises the number of thermodynamically unfavourable solute-solvent interactions. The quantitative evaluation of exactly how much hydrophobic interactions contribute to the stabilisation of protein-protein associations is controversial (Privalov and Gill, 1988; Dill, 1990a; Lesser and Rose, 1990; Murphy et al., 1990). The controversy is based on different definitions and interpretations of the hydrophobic effect in proteins (Dill, 1990b; Muller, 1993). It has been estimated, from an empirical correlation, that for non-polar surfaces there is an energy gain of approximately 25 calories per Å² of accessible surface area (Chothia, 1974; 1975). However energy values as high as 72 calories per Å² of accessible surface area have been reported (see Sharp et al., 1991). Recently the significance of correlating the hydrophobic effect with the molecular surface area, rather than accessible surface area, has become apparent (e.g. Nicholls et al., 1991); and energy values of approximately 47 calories per Å² of molecular surface have been reported (Jackson and Sternberg, 1994). (The concept of accessible surface area and molecular surface area are described in section 1.6.2).

A number of scales have been proposed which quantify the hydrophobic effect in proteins. To date, more than 40 'hydrophobicity scales', involving the assignment of relative hydrophobicity
values to each of the 20 amino acid residues, have been devised. Such scales can be divided into 3 types: solution scales (e.g. Nozaki and Tanford, 1971; Wolfenden et al., 1981; Fauchère and Pliska, 1983), empirical scales (e.g. Rose et al., 1985; Janin et al., 1988) and combination scales (which combine solution and empirical scales) (e.g. Kyte and Doolittle, 1982). The solution scales are based on the measurement of the chemical behaviour of each amino acid, such as their solubility in water or other organic solvents (Cornette et al., 1987). The empirical scales are based on the positions of the amino acids observed in the tertiary structure of proteins solved by X-ray crystallography. Correlations between the numerous scales have been made (Rose et al., 1985; Cornette et al., 1987; Ponnuswamy, 1993), and poor correlation is usually observed between pairs of solution scales, or a pairing between solution and experimental scales (Cornette et al., 1987). The number of scales available makes it difficult to decide which one is appropriate, and it is probably the case that different scales are appropriate in different situations.

1.5.2 Electrostatic Interactions

Electrostatic interactions, including hydrogen bonds and van der Waals interactions, are considered of secondary importance in protein associations (Chothia and Janin, 1975). However, the hydrogen bond (a polar interaction between donor and acceptor electronegative atoms) is an intrinsic component of protein-protein interactions. Hydrogen bonds between protein molecules are more favourable than those made with water (Fersht et al., 1985), and hence inter-molecular hydrogen bonds contribute to the binding energy of association. It has been proposed that whilst hydrophobic forces drive protein-protein interactions, hydrogen bonds and salt bridges confer specificity (Fersht, 1984; 1987).

Van der Waals interactions occur between all neighbouring atoms, but these interactions at the interface are not more energetically favourable than those made with the solvent. However they are more numerous, as the tightly packed interfaces are more dense than the solvent (Duquerroy et al., 1991). Hence these interactions also contribute to the binding energy of association.

1.5.3 Shape Complementarity

The complementarity of protein interfaces is derived from both electrostatic interactions and shape. Shape complementarity has been characterised by the size of buried surface, and the packing density of interface atoms. The close packing of protein interiors has been analysed (Richards, 1974), and this method has been used to measure the packing density of protein-
protein interfaces, revealing that such sites are tightly packed (Chothia and Janin, 1975). The close packed nature of interfaces in antibody-antigen complexes has also been observed using volume calculations (Walls and Sternberg, 1992). More recently, a shape correlation statistic (Sc) has been defined to measure packing and shape complementarity in protein-protein interfaces (Lawrence and Colman, 1993). This method revealed differences in interface packing between different types of complex; for example antibody-protein interfaces were found to be less well packed than those of protein-inhibitor complexes.

1.6 Studies of Protein-Protein Interactions

Studies of protein-protein interactions have only been possible due to the determination of protein structures using such methods as X-ray crystallography. The 3-dimensional atomic coordinates of proteins derived from such techniques have laid the foundation for computational studies.

1.6.1 Structure Determination

If a protein can be crystallised then its structure can potentially be solved using X-ray crystallography. The determination of atomic co-ordinates by X-ray crystallography requires a number of stages; crystallisation of the protein, data collection and processing, location of heavy atoms, calculation of phases and an electron density map, chain tracing and interpretation, model building, and refinement.

X-rays have a wavelength about the order of a bond length (≈1.5 Å), and are diffracted by individual atoms in a molecule. Diffraction by a crystal can be considered as the reflection of the X-rays by sets of parallel planes through the unit cell of the crystal. The way in which these diffracted beams relate to the position of the atoms in the crystal is defined in Bragg’s equation;

\[ 2d \sin \theta = n \lambda \]  

where \( n \) is an integer, \( \lambda \) is the wavelength, \( d \) is the spacing in the crystal, and \( \theta \) the angle made by the X-ray and the reflecting plane. In order to construct the molecular image, the amplitude of the reflections and the phases (the relative relationship of the wavefronts) are required. The amplitude is measured directly from the observed intensity, but the phase information is lost.
This is the phase problem in crystallography. For new protein structures, multiple isomorphous replacement (MIR) is commonly used to overcome the phase problem. Heavy atoms, usually heavy metals, are introduced into the unit cell of the crystal. These atoms make a significant contribution to the diffraction pattern, but they do not change the structure of the molecule or the crystal. The intensity differences of the X-ray reflections contain information about the phases. From these intensity differences it is possible to find the positions of the heavy atoms in the unit cell of the crystal, and hence determine the phases. The native data from a crystal without heavy atoms, the derivative data from the crystal with heavy atoms, and the knowledge of the position of the heavy atoms, are combined to calculate the phases for the native protein.

In practice it is common that 3, 4 or 5 heavy atom derivatives are incorporated into the phase calculation. The amplitude and phases of the diffraction data from the unit cell of the crystal are then used to calculate an electron density map. This is a contoured representation of the electron density at various points in the crystal structure. Building the initial model of the protein structure is somewhat subjective, but computer graphics, together with automated procedures, are used for initial chain tracing and the building of side chains in particular positions. The initial model is then refined in an iterative process until the discrepancies between the observed structure amplitudes and those back-calculated from the model are minimised. Thus the 3-dimensional co-ordinates of a protein structure can be determined.

Nuclear magnetic resonance (NMR) has emerged in the past 15 years as a second approach to the determination of protein structures. The method is based on the magnetic properties of atomic nuclei. However it is currently limited to the determination of small structures (<20kDaltons) and hence, to date, X-ray crystallography is the only method for the elucidation of structures of large proteins, such as oligomeric proteins and protein-inhibitor complexes.

1.6.2 Computational Studies

The deposition of the 3-dimensional co-ordinates of protein structures (solved by X-ray crystallography and NMR) in the Brookhaven Protein Data Bank (Bernstein et al., 1977), has permitted the analysis of relatively large numbers of multimeric proteins, and other protein complexes, for the purposes of the structural characterisation of protein interfaces. Studies have principally detailed the structural properties of the interfaces, including hydrophobicity, amino acid composition, hydrogen bonding, secondary structure and shape (Argos, 1988; Janin et al., 1988; Korn and Burnett, 1991; Jones and Thornton, 1995a). Such studies have analysed the properties of the interfaces of multimeric proteins in comparison with the protein exterior.
and protein interior. In addition the principles of protein-protein interactions have been analysed in transient complexes (Chothia and Janin, 1975; Janin and Chothia, 1990; Duquerroy et al., 1991; Jones and Thornton, 1995b). Such analyses rely upon a method of defining those parts of the protein that are involved in the protein-protein interactions. This has been achieved by the use of accessible surface area or molecular surface area definitions.

The native structure of proteins exists only in the presence of water, and the accessible surface area (ASA) describes the extent to which protein atoms can form contacts with water. Lee and Richards (1971) were the first to propose the concept of accessible surface area, defining it as the area of a sphere of radius $R$, on each point of which the centre of a solvent molecule can be placed in contact with an atom without penetrating any other atoms of the molecule. The radius $R$, is given by the sum of the van der Waals radius of the atom and the chosen radius of the solvent molecule (Figure 1.4). Extensions and refinements to this initial algorithm have been made by Shrake and Rupley (1973). Also comparable (except for the atomic radii it assigns) is the geodesic sphere integration algorithm designed by Kabsch and Sander (1983).

![Figure 1.4: The calculation of accessible surface area and molecular surface area of a protein. (Taken from Jackson and Sternberg, 1994)](image-url)

Subsequent to the initial definition of surface accessibility, Richards (1977) proposed an alternative: that of the molecular surface. The molecular surface consists of the parts of the van der Waals surface of the atoms that are accessible to a probe sphere (termed contact surfaces), connected by a network of surfaces (termed re-entrant surfaces) that smooth over the crevices
and pits between the atoms. This surface is the boundary of the volume from which a probe sphere is excluded if it is not to experience van der Waals overlap with the atoms (Figure 1.4). Connolly (1983b) developed an algorithm (MS), combining the calculation of the contact and re-entrant surfaces with the Shrake and Rupley (1973) method for calculating accessible surface areas. This algorithm calculates a smooth 3-dimensional contour about a molecule by placing dots over the solvent accessible surface of the molecule. The computer graphics representation of this system then shows only the atoms that are accessible to the solvent (Connolly, 1983a). The methods of accessible surface area and molecular surface area calculation imply that the system is static; they do not account for any movement or flexibility an atom or group may possess within the molecule.

1.6.3 Significance and Application

The many examples of protein-protein interactions in different systems give a clear indication of the importance of these interactions in biological processes. As well as the characterisation of protein-protein interactions, the long term aim is to predict the location of such interaction sites on the surfaces of proteins. These predictions can be divided into two main problems. The first problem is the docking of two proteins of known structure. The second problem is the identification of putative interaction sites upon the surface of an isolated protein, known to be involved in protein-protein interactions, but for which the structure of the complex is not available. Many algorithms have been designed to solve both prediction problems (e.g. Kuntz et al., 1982; Walls and Sternberg, 1992; Zielenkiewicz and Rabczenko, 1984); and these are discussed in more detail in Chapter 5 (section 5.1).

Interface prediction methods can act as a prerequisite for drug design or for the optimisation of drugs therapies already in use. In this respect it has been suggested that a possible blueprint for a drug designed to combat the HIV virus would be a molecule that could disrupt the dimerization of the HIV protease (Blundell and Pearl, 1989; Wlodawer et al., 1989). The characterisation of protein interfaces would be a preliminary step in such a drug design regime. Such a step would also prove important to current work concerned with the development of dimerization inhibitors as anti-parasitic drugs (Nordhoff et al., 1993). Possibilities for the application of interface characterisation are also at the forefront of work into insulin and diabetes. The current therapy for diabetes involves, in simple terms, the injection of insulin in its hexameric state. To be biologically active the protein is required as a dimer. The dissociation of the hexameric insulin is a relatively slow process, and it is thought that this slow release of
active protein is one cause of the detrimental health effects which result from long term diabetes. Work has already started on the alteration of the interface of the insulin multimer, by protein engineering methods, to allow injection of the protein as the monomer or dimer (Dodson et al., 1993). This follows in the path of similar work in which protein engineering methods have been used to control the thermostability of oligomeric enzymes (Ahern et al., 1987).

1.7 Outline of Thesis

The above discussion highlights the complexity of protein-protein interactions, and briefly refers to some of the many and varied elements that contribute to successful macromolecular associations. The work described in this thesis has been conducted with the aim of characterising protein-protein interfaces, and developing a method capable of predicting putative interaction sites. Chapter 2 describes the analysis of protein-protein interactions in one structural system; homo-dimer proteins. The analysis includes characterisation of the hydrophobic nature of the interactions as well as electrostatic and shape complementarity. The work culminates in the production of a computational approach to automatically characterise a protein-protein interaction from the 3-dimensional co-ordinates of the complex. Chapter 3 describes the comparison of protein-protein interactions in a number of different structural systems; including homo-dimers, homo-trimers, hetero-dimers, enzyme-inhibitor complexes and antibody-protein complexes. The work highlights significant differences between different types of complex and relates this to their biological function. In Chapter 4, a method is devised for the comparison of patches of residues on the surface of proteins. The patches defined are compared for a number of different structural parameters to find those characteristics that distinguish the observed interface from the remainder of the protein surface. In Chapter 5 the knowledge gained from the characterisation of protein-protein interactions in different systems is used to develop a prediction algorithm. The algorithm is used to predict the interfaces in homo-dimers and enzyme inhibitor complexes. Chapter 6 concludes the analysis of protein-protein interactions by applying the computational tools, developed in Chapters 2, 4 and 5, to a number of different protein systems. This includes the analysis of the interfaces in dimeric and tetrameric forms of haemoglobin, the analysis of crystal contacts in sickle cell haemoglobin, and the prediction of structural epitopes on the surface of two antigenic proteins. The conclusions from the work are drawn in Chapter 7.
CHAPTER 2

Analysis of Protein-Protein Interactions in Protein Dimers

2.1 Introduction

Protein-protein interactions are complex, with many fundamental factors contributing to a stable association. The aim is to characterise protein-protein interfaces, and this chapter concentrates on interactions in just one structural system: dimeric proteins. A number of studies have previously been conducted upon protein-protein interactions in multimeric proteins. Janin et al. (1988) analysed subunit interactions in 23 oligomeric proteins (including dimers, tetramers, hexamers and octamers), and Argos (1988) conducted a similar study on a data set of 24 oligomeric proteins (including dimers, trimers and tetramers). These studies included analysis of accessible surface area buried on oligomerization, distributions of amino acid residues, hydrophobicity and interface secondary structure. The analysis of secondary structural motifs at protein-protein interfaces has also been the subject of specific analysis (Miller, 1989; Slingsby et al., 1992a). In addition, Korn and Burnett (1991) have analysed the hydrophobicity of protein interfaces in 40 oligomeric proteins and 2 protein-protein complexes. The studies of Argos (1988), Janin et al. (1988), and Korn and Burnett (1991), all divide proteins into interface, interior and surface subsets of residues and atoms; and compare the interface subset characteristics with those of the protein interior and surface. This is the method implemented in the current analysis.

This chapter begins with analysis of the numbers of biological multimeric structures recorded in the Brookhaven Protein Data Bank (PDB). It is followed by a discussion of the methods used to generate a non-homologous data set and to define the protein-protein interfaces. The main section describes the characterisation of the interfaces, including analysis of the residue and atom composition, the secondary structure, segmentation, planarity and ionic interactions. The underlying theme has been the differentiation of the interface from the remainder of the protein surface. To this end the characteristics of the interface have been compared with those of the protein surface and the protein interior. The chapter concludes with a discussion of the implications of the work and its application to the prediction of protein-protein interfaces.
2.2 Classification of Multimeric Proteins

This study is concerned with the classification of proteins into their *biological* multimeric states. The process was conducted by consulting the individual entry for each protein in the PDB. If no information could be found there, the original journal article in which the structure was published was consulted. This classification was based on the July 1993 version of the PDB that contained 974 protein structures, determined experimentally. These proteins have been classified with reference to the total number of subunits in their biological multimeric state (Figure 2.1). Four proteins were unclassified as they had no information pertaining to their biological multimeric state recorded in their PDB entries, and the references cited for each were recorded in the PDB as unpublished.

![Figure 2.1: Classification of the proteins in the PDB (July 1993) into their biological multimeric states.](image)

The distribution of biological multimeric states in the PDB showed the prevalence of dimers and tetramers, and the relatively small number of odd numbered multimers. Such a distribution is consistent with previous observations on cumulative crystallographic data of oligomeric proteins (Matthews and Bernhard, 1973; Klotz *et al*., 1975). This distribution is biased as it reflects only those proteins whose structures have been solved and, therefore, probably over-represents the small monomers. However, it is clear that trimers are relatively rare compared with tetramers, and that the number of structures in the higher multimeric states falls markedly, with the obvious exception of the viral coat proteins that comprise high numbers (e.g. 60, 180, 240) of subunits.
2.3 The Non-Homologous Data Set of Protein Dimers

The protein dimers constitute the largest set of proteins in one biological multimeric state, with 148 entries in the July 1993 release of the PDB. These entries represented 81 different protein structures, as 30 dimers had multiple file entries, which recorded proteins crystallised in different solutions, in different crystal classes, or with different inhibitors bound.

In this analysis only non-homologous proteins were selected for the data set. To define the non-homologous proteins both the sequence relationship and the structural similarity were considered, as often homology can only be detected from the 3-dimensional structure. Proteins were selected for inclusion on the basis that they had a sequence identity of < 35% and were structurally different. The structural similarity of the proteins was measured using a method of direct structural alignment (SSAP (Taylor and Orengo, 1989)). Proteins were selected for the data set if they had a SSAP score of ≤80. In the process of selection, only dimers with homologous subunits were considered. This selection resulted in a non-homologous data set of 32 protein dimers, each belonging to a different homologous protein family (e.g. only one immunoglobulin was included in the data set) (Table 2.1).

2.3.1 Generation of Co-ordinates

Of the 32 non-homologous dimers in the data set, 31 had been solved by crystallography. The asymmetric units of 14 protein crystals contained only one subunit. For these structures the rotation and translation matrices provided in the PDB files were applied to produce the complete molecule, using an algorithm implemented by T. Flores.

2.3.2 Validation of Co-ordinates

The structure of the 14 dimers generated in the method described in section 2.3.1 were inspected visually using the graphics software Quanta (Quanta 4.0: Molecular Simulations Inc.). This ensured that the structures produced were valid structures: i.e. that the dimers generated had an area of interface between their subunits.
In addition to visual inspection, the generated structures were validated systematically by checking for the presence of atoms with overlapping van der Waals radii ($vdr$) (the radii used were those assigned by Chothia (1976)). The distance ($d$) between each atom in one subunit ($p$) and each atom in the second (generated) subunit ($q$) was calculated. If the equation:

$$d < vdr_p + vdr_q + 1\AA$$

holds, then the two subunits are considered to be overlapping.

### Table 2.1: Non-homologous data set of 32 protein dimers.

<table>
<thead>
<tr>
<th>PDB Code</th>
<th>Protein</th>
<th>Source (as recorded in the PDB)</th>
<th>Resolution (Å)</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>1cds</td>
<td>Cardiotoxin</td>
<td>Naja mossambica mossambica</td>
<td>2.5</td>
<td>Rees et al (1990)</td>
</tr>
<tr>
<td>1hc1</td>
<td>Fc Fragment (Immunoglobulin)</td>
<td>Human pooled serum</td>
<td>2.9</td>
<td>Deisenhofer (1981)</td>
</tr>
<tr>
<td>1iL8</td>
<td>Interleukin 8</td>
<td>Human neutrophil recombinant in E.Coli</td>
<td>-----</td>
<td>Clore et al (1990)</td>
</tr>
<tr>
<td>1mub</td>
<td>Mannose Binding Protein</td>
<td>Rat expressed in E.Coli</td>
<td>2.3</td>
<td>Weis et al (1991)</td>
</tr>
<tr>
<td>1ppb</td>
<td>P-Hydroxybenzoate Hydroxylase</td>
<td>Pseudomonas fluorescens</td>
<td>2.3</td>
<td>Schreuder et al (1988)</td>
</tr>
<tr>
<td>1ppp</td>
<td>Inorganic Pyrophosphatase</td>
<td>Western Diamondback Rattlesnake</td>
<td>2.5</td>
<td>Brunie et al (1985)</td>
</tr>
<tr>
<td>1sdo</td>
<td>Haemoglobin (Clam)</td>
<td>Arcid clam</td>
<td>3.0</td>
<td>Cohen et al (1978)</td>
</tr>
<tr>
<td>1utg</td>
<td>Uteroglobin</td>
<td>Rat female genital tract</td>
<td>2.4</td>
<td>Royer et al (1989)</td>
</tr>
<tr>
<td>1vg</td>
<td>Variant Surface Glycoprotein</td>
<td>Trypanosoma brucei</td>
<td>1.35</td>
<td>Morize et al (1987)</td>
</tr>
<tr>
<td>1yp1</td>
<td>Triose Phosphate Isomerase</td>
<td>Yeast</td>
<td>2.9</td>
<td>Freymann et al (1990)</td>
</tr>
<tr>
<td>2ccy</td>
<td>Cytochrome C'</td>
<td>Rhodospirillum molischianum</td>
<td>1.9</td>
<td>Lolis et al (1990)</td>
</tr>
<tr>
<td>2cts</td>
<td>Citrate Synthase C</td>
<td>Porcine heart</td>
<td>1.9</td>
<td>Finzel et al (1985)</td>
</tr>
<tr>
<td>2gs5</td>
<td>Gene 5 DNA Binding Protein</td>
<td>Filamentous bacteriophage FD(M13)</td>
<td>2.3</td>
<td>Brayer and McPherson (1983)</td>
</tr>
<tr>
<td>2orl</td>
<td>434 Repressor</td>
<td>Phase 434</td>
<td>2.5</td>
<td>Anderson et al (1987)</td>
</tr>
<tr>
<td>2rhe</td>
<td>Bence - Jones Protein</td>
<td>Human myeloma patient RHE urine</td>
<td>1.6</td>
<td>Furey et al (1983)</td>
</tr>
<tr>
<td>2rus</td>
<td>Rubisco</td>
<td>Rhodospirillum rubrum in E.coli</td>
<td>2.3</td>
<td>Lundqvist and Schneider (1991)</td>
</tr>
<tr>
<td>2rve</td>
<td>ECO RV Endonuclease</td>
<td>E.Coli recombinant form in E.coli</td>
<td>3.0</td>
<td>Winkler et al (1991)</td>
</tr>
<tr>
<td>2sod</td>
<td>Copper, Zinc Superoxide Dismutase</td>
<td>Bovine Erythrocyte</td>
<td>2.0</td>
<td>Tainer et al (1982)</td>
</tr>
<tr>
<td>2ssi</td>
<td>Subtilisin Inhibitor</td>
<td>Streptomyces albogriseolus S-3253</td>
<td>2.6</td>
<td>Mitsui et al (1979)</td>
</tr>
<tr>
<td>2ts1</td>
<td>Tyrosyl-Transfer RNA Synthetase</td>
<td>Bacillus steinoreuthophilus NCA</td>
<td>2.3</td>
<td>Brick et al (1988)</td>
</tr>
<tr>
<td>2tsr</td>
<td>Thymidylate Synthase</td>
<td>E.coli</td>
<td>1.97</td>
<td>Montfort et al (1990)</td>
</tr>
<tr>
<td>2wpp</td>
<td>Trp Repressor</td>
<td>E.coli</td>
<td>1.65</td>
<td>Schevitz et al (1985)</td>
</tr>
<tr>
<td>3aat</td>
<td>Aspartate Aminotransferase</td>
<td>E.coli</td>
<td>2.8</td>
<td>Smith et al (1983)</td>
</tr>
<tr>
<td>3enl</td>
<td>Enolase</td>
<td>Baker’s Yeast</td>
<td>2.25</td>
<td>Stec and Leboda (1990)</td>
</tr>
<tr>
<td>3gap</td>
<td>Catabolite Gene Activator Protein</td>
<td>E.coli</td>
<td>2.5</td>
<td>Weber and Steitz (1987)</td>
</tr>
<tr>
<td>3gs2</td>
<td>Glutathione Reductase</td>
<td>Human Erythrocyte</td>
<td>1.54</td>
<td>Schulz et al (1982)</td>
</tr>
<tr>
<td>3icd</td>
<td>Isocitrate Dehydrogenase</td>
<td>E.coli</td>
<td>2.5</td>
<td>Harley et al (1989)</td>
</tr>
<tr>
<td>3sdp</td>
<td>Iron Superoxide Dismutase</td>
<td>Pseudomonas ovalis wild type</td>
<td>2.1</td>
<td>Stoddard et al (1990)</td>
</tr>
<tr>
<td>4msdh</td>
<td>Cytoplasmic Malate Dehydrogenase</td>
<td>Porcine heart</td>
<td>2.5</td>
<td>Birktoft et al (1989)</td>
</tr>
<tr>
<td>5adh</td>
<td>Alcohol Dehydrogenase</td>
<td>Horse liver</td>
<td>2.9</td>
<td>Ekland et al (1976)</td>
</tr>
<tr>
<td>5hvp</td>
<td>HIV Protease</td>
<td>N75 strain of human immunodeficiency virus type 1 in E.Coli</td>
<td>2.0</td>
<td>Navia et al (1989)</td>
</tr>
</tbody>
</table>

...
was true then the 2 atoms were considered as overlapping. The 1Å error margin was incorporated to account for small errors in the crystallographic co-ordinates. The number of overlapping pairs was calculated in each generated dimer, and in each dimer in which both subunits were contained in the asymmetric unit. The number of overlapping atoms was calculated as a percentage of the number of atoms in each dimer (Figure 2.2). The mean percentage of overlapping atoms in the 14 generated dimers was 0.04% and in the 18 dimers, that were complete in the asymmetric unit, 0.01%. The higher percentage of overlapping atoms in the generated dimers may indicate inaccuracies derived from the refinement stage of the crystallographic process. In this study, it was considered that dimers with \( \leq 0.25\% \) overlapping atoms were 'valid' structures. On this basis all 32 dimers were classed as valid and included in the data set.

![Figure 2.2: Percentage of overlapping atoms in 32 protein dimers. (Two atoms were defined as "overlapping" if the distance between them was less than the sum of their van der Waals radii plus 1Å). The vertical line at 0.25\% indicates the cut-off used to select structures for the data set of protein dimers.](image)

### 2.4 Definitions of Protein-Protein Interfaces

Protein interfaces have been defined in different ways in previous studies of multimeric proteins. Argos (1988) and Janin et al. (1988) base their definitions on accessible surface area
calculations, using the Shlake and Rupley (1973) algorithm and the Kabsch and Sander (1983)
algorithms respectively. Korn and Burnett (1991) base their definitions on molecular surface area
calculations using an adaptation of Connolly's MS algorithm (Connolly, 1983a).

In this current work protein-protein interfaces were defined at two levels, at the residue level and
at the atom level. At both levels the definitions were based on the concept of solvent accessible
surface area (ASA). The native structure of a protein exists in the presence of water, and the
ASA describes the extent to which a protein can form contacts with water (see chapter 1, section
1.6.2). The ASAs of the dimers were calculated using an implementation of the Lee and
Richards (1971) algorithm, developed by Hubbard (1990). With a probe sphere, of radius 1.4Å,
the accessible surface was defined as the surface mapped out by the centre of the probe if it
were rolled around the van der Waals surface of the protein. Both atomic and residue
accessibilities for a given protein were evaluated. In addition, the relative ASA for each residue
was calculated by summing the atomic ASAs in a residue and comparing these totals to the
ASA that each of the 20 residues possess in a standard extended state. The states used were
those with the residue X in a tripeptide ALA-X-ALA, where $\phi = 140^\circ$; $\psi = 135^\circ$; $\omega = 180^\circ$
(except for proline) (Hubbard, 1992). The program was used to calculate the ASAs of atoms
and residues in each dimer, and each subunit as a discrete structure.

2.4.1 The Residue Level Definition

The interface definitions were based on the change in ASA ($\Delta$ASA) on dimerization. On the
basis of ASA calculations the interface residues were defined as those residues with side chains
possessing an ASA that decreased by $>1\AA^2$ on dimerization. The $1\AA^2$ error factor was used to
account for small errors in the crystallographic co-ordinates and computational inaccuracies in
the calculation of the ASAs. The exterior and interior residue definitions were based on the
relative ASA of each residue, which ranged from 0% for residues with no atom contact with the
solvent, to 100% for fully accessible residues. On this basis, the exterior residues were defined
as having relative accessibilities of $>5\%$ and the interior residues as those with relative
accessibilities of $\leq 5\%$. This 5% cut-off was devised and optimised by Miller et al. (1987), who
used it to define residues buried in monomeric proteins.
The definition of the exterior/interior residues was based on relative accessibilities, and the definition of the interface residues was based on absolute ASA. Thus all the interface residues were originally defined as interior or exterior residues. At this stage the subset of interface residues was excluded from the subsets of interior and exterior residues. This resulted in 3 discrete sets of residues for each of the dimers in the non-homologous data set.

2.4.2 The Atom Level Definition

The atom level definition was also based on ΔASA. The interface was defined as those atoms that showed a decrease > 0.01Å² in ASA on dimerization. The interior was defined as those atoms with an atomic ASA of zero, and the exterior as those atoms with an atomic ASA of greater than zero. As with the residue level definition, the interface atoms were removed from the exterior and interior subsets, to produce 3 discrete sets of atoms. The interface atoms of 6 of the 32 dimers in the data set are shown as Corey-Pauling-Koltun (CPK) models in Figure 2.3.

2.5 Characterisation of the Interfaces

The interfaces of the 32 dimers have been analysed with respect to a number of characteristics identified as being potentially important for molecular recognition in multimeric proteins. The analysis uses the interface definitions at both the atom and the residue level (section 2.4).

2.5.1 Accessible Surface Area

The ASA for the interface \( (A_i) \) of one subunit in each dimer was calculated as:

\[
\text{Interface } \text{ASA}(A_i) = \frac{(A_a + A_b) - A_{ab}}{2} \tag{2.2}
\]

where

- \( A_a \) = Accessible surface area of the first subunit dissociated
- \( A_b \) = Accessible surface area of the second subunit dissociated
- \( A_{ab} \) = Accessible surface area of the subunits associated in the dimer
Figure 2.3: CPK models of 6 of the protein dimers in the non-homologous data set. One subunit of each dimer is shown and the interface atoms are coloured red and the remainder of the subunit atoms are coloured in blue. Each structure is labelled using its PDB code; for the full name of the structure refer to Table 2.1. Each subunit is viewed looking face onto the interface surface, but the scales between the different structures are not comparable.
The percentage ASA buried, per subunit, on dimerization was calculated as:

\[
\text{Percentage ASA Buried (}\%A_i\text{)} = \frac{A_i}{A_{as}}
\]  

[2.3]

where \(A_i\) = Interface ASA

\(A_{as}\) = ASA of the first subunit

The ASA buried, per subunit, on dimerization ranged from 368.1\(\text{Å}^2\) in the 434 repressor (2orl) to 4746.1\(\text{Å}^2\) in citrate synthase C (2cts). The percentage ASA buried, per subunit, upon dimerization ranged from 6.5\% in inorganic pyrophosphatase (1ppy), to 29.4\% in the trp repressor (2wrp). There is a positive linear relationship between the interface ASA (\(ASA_i\)) and the molecular weight (Mr) of the protomer (Figure 2.4). A regression line, of the form \(y = ax\), fitted to the data gives the equation:

\[
ASA_i = 0.06\text{Mr}
\]  

[2.4]

The approximate nature of the relationship was indicated by the very high estimated variance (6.97 \(\times\) 10\(^5\)), and a correlation coefficient (r) of data to model of 0.69.

![Figure 2.4: Relationship between the molecular weight (in Daltons) of the protomer and the interface ASA (in \(\text{Å}^2\)) (per subunit). The solid line is the fitted line of the equation \(y = 0.06x\) and the dashed lines are the 95\% confidence limits of the individual y values.](image)
2.5.2 Sphericity of Dimeric Proteins

The folding of polypeptide chains, and aggregation of subunits, buries the hydrophobic residues of a protein and hence minimises the number of thermodynamically unfavourable solute-solvent interactions. In terms of physical shape, a reduction of a protein’s surface exposed to solvent is achieved by the protein adopting a structure that is spherical. A sphere has the lowest surface area to volume ratio, hence the globular nature of proteins.

In this work the globular (spherical) nature of the dimer was compared with that of the protomer. An approximation of the surface area to volume ratio was calculated for each protein (as a dimer and as a protomer), using the ratio of ASA to molecular weight (ASA/Mr).

In Figure 2.5 all proteins lie below the x=y line as expected (as ASA (but not Mr) is lost in the dimerization process).

![Figure 2.5: Accessible surface area (ASA) / molecular weight (Mr) ratio of 32 protein dimers and their protomers. The solid line traces the ASA/Mr ratio of 1:1. All the data points fall below this line indicating that on changing from monomer to dimer the structures effectively reduce the amount of surface area exposed to solvent.](image)

A difference is observed in the relationship between monomer ASA/Mr and protomer ASA/Mr ratios in the larger sized monomers. Monomers with a larger ASA/Mr ratio have a relatively small dimer ASA/Mr ratio. This indicates that the larger monomers have to bury proportionally more ASA to form a stable dimer than the smaller monomers.
2.5.3 Shape of the Dimer Interfaces

The shape (the relative length and breadth) and the planarity of the interfaces were analysed by calculating a plane of best fit through the 3-dimensional co-ordinates of the atoms in the interface.

A program (implemented by R. Laskowski) was used to calculate the best fit plane by principal component analysis. The program calculated the equation for the best fit plane and the root mean squared (rms) of the plane (Figure 2.6).

**Figure 2.6:** Rms (Å) of best fit planes fitted through the interface (one subunit) of 32 protein dimers. The dashed line indicates a cut-off of 6Å, below which the interfaces were classed as flat. The structures have been placed in order of increasing interface size.

In the data set, 84% of the interfaces were relatively flat, with a rms of ≤6Å. This was consistent with the observations of Argos (1988) who found that 83% of the protein-protein interfaces he studied were flat. Argos (1988) argued that an overall flatness would be expected when considering the symmetry involved in the associations. Whilst the presence of symmetry does not dictate that the interface will be flat, it does require that any 'protrusion' is complemented by a 'hollow'. This requirement suggests that the gross surface of the interface can be approximated by a plane. However, in the current study, 5 proteins had comparatively high rms values (> 6Å) when planes were fitted. These proved to be proteins in which the 2
subunits were twisted together across the interface (e.g. isocitrate dehydrogenase (3icd)), or proteins that had subunits with 'arms' apparently clasping the 2 halves of the structure together (e.g. aspartate aminotransferase (3aat)). These 2 examples can be seen in Figure 2.7.

To give a measure of the shape of the interfaces, the standard deviations of the interface atom distributions from the centre of gravity of the best fit planes were calculated. The standard deviations in the x and y dimensions were compared (Figure 2.8). In the data set 31 of the 32 dimers had interfaces that were in general circular: having standard deviations of the atom distributions in the x dimension (sd_x) very close to those in the y dimension (sd_y). The ratios of sd_y / sd_x for the 31 proteins lie between 0.48 and 1.00. The only exception was the variant surface glycoprotein (1vsg) which had an elongated interface, with the sd_y being 0.25 times the sd_x, which reflects the elongated nature of the protein as a whole (the protomer has a sd_y / sd_x ratio of 0.33). Variant surface glycoprotein of Trypanosoma brucei forms a coat on the surface of the parasite (Freymann et al., 1990) and thus its shape may reflect its function. It was also found that, for the 32 dimers, the ratio of sd_y / sd_x was independent of the size of the protein (correlation coefficient (r) = 0.13) and was very poorly correlated to the shape of the protomer as a whole (r = 0.45). The shapes of 6 of the 32 interfaces can be seen as Corey-Pauling-Koltun (CPK) models in Figure 2.3.
Figure 2.7: CPK models of non-planar interfaces: In 1 (a)(b) and 2 (a)(b) 2 subunits are shown, in each one subunit is red with yellow interface atoms, and one blue with green interface atoms. (1). Aspartate Aminotransferase Mutant (3aat). (a) and (b) Views along the subunit interface with the arms of each subunit shown stretched around its partner. (c) Dimer interface only, viewed along the interface. (2). Isocitrate Dehydrogenase (3icd). (a) View along the subunit interface (b) View looking down the subunit interface showing the 2 subunits twisted together. (c) Dimer interface only, viewed along the interface.
Figure 2.8: Shape of the dimer interfaces; measured as the standard deviations of the interface atom distributions (in the x and y dimensions), from the centre of gravity of the best fit plane. The solid line depicts a ratio of 1:1.

2.5.4 Amino Acid Composition

The total number of each of the 20 amino acids present in the interface was calculated and expressed as a percentage of the total number of interface residues. For the purposes of comparison, the percentage frequencies were also calculated for the interior and the exterior of the proteins (Figure 2.9). These results, and the percentage frequency ratios calculated in Table 2.2, clearly indicate that the interfaces show a greater similarity to the exterior of the protein than the interior, in terms of their amino acid composition. This characteristic is exemplified in the charged residues, of which the interface has 3.54 times the percentage frequency compared with that of the interior, but only 0.74 times that of the exterior.

<table>
<thead>
<tr>
<th>Amino Acids</th>
<th>Percentage Frequencies</th>
<th>Percentage Frequency Ratios</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Exterior</td>
<td>Interface</td>
</tr>
<tr>
<td>Hydrophobic</td>
<td>40.90</td>
<td>46.68</td>
</tr>
<tr>
<td>Polar</td>
<td>29.48</td>
<td>31.38</td>
</tr>
<tr>
<td>Charged</td>
<td>29.53</td>
<td>21.94</td>
</tr>
</tbody>
</table>

Table 2.2: Percentage frequencies of hydrophobic, polar and charged residues and ratios of interface/interior and interface/exterior for each set of residues. In this study Ala, Gly, Ile, Leu, Met, Phe, Pro, Val were classed as hydrophobic; Asp, Arg, Lys, Glu as charged; and Asn, Cys, Gln, His, Ser, Thr, Trp, Tyr as polar.
Figure 2.9: Percentage frequencies of amino acid residues (indicated by their one letter codes) in the interior, interface and exterior of 32 protein dimers.

These results are in contrast to the work of Janin et al. (1988) who concluded that the interface amino acid composition was more closely related to the interior of the protein. However, Korn and Burnett (1991) found that the hydropathy of the average interface of a multimeric protein lies between that of the exterior and the interior. Argos (1988) also made similar conclusions, stating that the subunit interfaces display characteristics that lie between those of the hydrophobic nature of the protein interior and the hydrophilic properties of the protein exterior.

Propensities have also been calculated from the percentage frequencies of the 20 amino acid residues in the interface. Two propensities for each amino acid (AAj) were calculated:

\[
\text{Interface / Protein (IP) Propensity } AA_j = \frac{\%\text{Freq } AA_j^{p}}{\%\text{Freq } AA_j^{(p)}} \quad [2.5]
\]

\[
\text{Interface / Surface (IS) Propensity } AA_j = \frac{\%\text{Freq } AA_j^{(i)}}{\%\text{Freq } AA_j^{(s)}} \quad [2.6]
\]

Where

- \(\%\text{Freq } AA_j^{(i)}\) = Percentage frequency of amino acid j in the protein interface
- \(\%\text{Freq } AA_j^{(p)}\) = Percentage frequency of amino acid j in the whole protein
- \(\%\text{Freq } AA_j^{(s)}\) = Percentage frequency of amino acid j in the protein surface
The 2 types of propensities calculated give the ‘relative probability’ of an amino acid being present in the interface, with respect to the probability of it being present in the protein as a whole (equation 2.5); and with respect to the probability of it being present in the protein surface (interface and exterior residues) (equation 2.6). A propensity of >1 denotes that a residue occurs more frequently in the interface than in the protein as a whole (or than in the protein surface). A propensity of <1 denotes that a residue occurs less frequently.

The interface/protein (IP) propensities (Figure 2.10) reveal that the charged and polar residues, especially arginine and asparagine, show an increased affinity for the interface. In addition, the hydrophobic residues methionine and proline, show a small increased affinity for the interface. This supports the results of Janin et al. (1988) and Argos (1988), who found that arginine and methionine have a high affinity for the interface. The 2 polar residues showing less affinity for the interface are cysteine and tryptophan. It is possible that the double aromatic ring in the side chain of tryptophan is too bulky to be accommodated in the interface of a protein-protein interaction. However in general the aromatics have an affinity for the interface; 2 out of the 3 single aromatic residues showed an increased affinity. This affinity was also found by Argos (1988) who suggested that aromatics make particularly good ‘glue’ for sticking together protein subunits.

![Figure 2.10: Interface/protein (IP) propensities. The propensities give the ‘relative probability’ of an amino acid being present in the interface with respect to the probability of it being present in the protein as a whole.](image)

---

**Figure 2.10**: Interface/protein (IP) propensities. The propensities give the ‘relative probability’ of an amino acid being present in the interface with respect to the probability of it being present in the protein as a whole.
The interface/surface (IS) propensities (Figure 2.11) confirm the results of the IP propensities. The 3 single aromatic residues and arginine all show an increased affinity for the interface relative to the protein surface. In addition the propensities clearly show that the interface contains a greater percentage of hydrophobic residues than the protein surface as a whole. This characteristic is further illustrated in Figure 2.12 where the percentage frequency of residues in the interface and the exterior are shown with the residues in order of increasing hydrophobicity; based on the Wolfenden hydrophobicity scale (Wolfenden et al., 1981). This hydrophobicity scale is one closely linked to experimental methods, with the equilibria of distribution of amino acid side chains, between their dilute aqueous solutions and the vapour phase, being determined by dynamic vapour pressure measurements. Thus, the scale gives a direct measure of the free energy for transfer from water to the hydrophobic phase of amino acid side chains.

![Diagram of amino acid propensities](image)

**Figure 2.11:** Interface/surface (IS) propensities: The propensities give the ‘relative probability’ of an amino acid being present in the interface with respect to the probability of it being present in the surface of the protein.

The analysis of the percentage frequencies of amino acid residues relative to their hydrophobicities revealed that the interfaces represent hydrophobic patches on the surfaces of proteins. This equates with the theory that the hydrophobic interaction is fundamental to the stabilisation of protein associations (Chothia and Janin, 1975).
2.5.5 Distribution of Atom Types

The percentage frequencies of 19 atom types in the protein interface, exterior, and interior, have been calculated. The atom types are those defined by Warme and Morgan (1978), in their studies of long range atomic interaction in proteins, and includes 15 different side chain atoms and 4 backbone atoms. The distribution is shown in Figure 2.13, in which the atom types have been ordered according to their percentage frequency in the interior subset. This clearly shows that the distribution of the 19 atom types in the interface closely resembles those in the exterior. There is a relatively large percentage frequency of polar CH₂ atoms in the exterior and interface of proteins (CH₂ atom types are those occurring in charged amino acid side chains). This shows the increased affinity for polar atoms on the protein surface.

Atom Distributions and Hydrophobicity

The percentage of polar and non-polar atoms in the interface and the exterior have been calculated for the 32 protein dimers. It was found that the interfaces were slightly more hydrophobic than the exterior: the interface comprising 31.9% polar and 68.1% non-polar atoms, and the exterior 36.7% polar and 63.3% non-polar. In all, 88% of the dimers displayed this distribution. This gives additional support to the results of the residue distributions (section 2.5.4), indicating that the interfaces represent hydrophobic patches on the surface of proteins.
Protein-Protein Interactions: Chapter 2

Figure 2.13: Percentage frequency of 19 atom types in the interior, interface and exterior of 32 protein dimers. The atom types are those defined by Warme and Morgan (1978).

To further investigate the nature of the hydrophobicity of the interfaces each interface was divided into 3 zones. To achieve this each set of interface atoms were projected onto the best fit plane centred at the origin (see section 2.5.3), and the maximum and minimum $x$ and $y$ co-ordinates used as boundaries to define a 2-dimensional box containing all the interface atoms. A second box was then defined, using 25% of the maximum and minimum $x$ and $y$ co-ordinates as boundaries. A third box was also defined, using 50% of the maximum and minimum $x$ and $y$ co-ordinates as boundaries. In this way 3 zones (inner, middle and outer) were defined within each interface (Figure 2.14).

The percentage of polar and non-polar atoms in each zone was calculated for each dimer, and a mean percentage calculated for the whole data set (Figure 2.15). It can be seen that the interfaces are areas in which the non-polar atoms are concentrated in the centre, with the distribution of these atoms decreasing as the outer parts of the interface are reached. This could be described as a round ‘sticky plaster’ arrangement, with the interface representing the plaster with the non-polar atoms in the centre protected from the outside solvent by the ‘sticky’ polar interactions. However it should be noted that this distribution is not marked. The outer zone has 61.81% non-polar atoms and the inner zone 72.31%. In addition when considering the distributions in individual dimers only 41% show this ‘sticky plaster’ arrangement of polar and non-polar atoms.
**Figure 2.14:** Method of defining 3 zones within a dimer interface. The dots represent the interface atom positions as transformed into 2 dimensions centred at the origin. The 4 black dots, shown with their co-ordinates, are the maximum and minimum $x$ and $y$ co-ordinates used to define the outer boundary of zone 1 (outer). The outer boundary of zone 2 (middle) is defined by taking 25% from each of the maximum and minimum $x$ and $y$ co-ordinates. The outer boundary of zone 3 (inner) is defined by taking 50% from each of the maximum and minimum $x$ and $y$ co-ordinates. The atoms are coloured in three shades to show that they fall into different zones.

**Figure 2.15:** Mean distribution of polar and non-polar atoms within 3 defined zones from 32 dimer interfaces. The zones are defined as described in the legend to Figure 2.14.
2.5.6 Secondary Structural States

The secondary structural states of those residues in the interface of the dimers have been analysed. The secondary structure definitions (coil, helix, sheet and turn) were based on those of Kabsch and Sander (1983). The number of interface residues in each type of secondary structural conformation was calculated as a percentage of the total number of interface residues. For comparison the percentage distributions were also calculated for the interior and the exterior residues (Figure 2.16).

![Figure 2.16: Percentage frequency of secondary structure motifs in interface residues in 32 protein dimers. The secondary structure motifs are those defined by Kabsch and Sander (1983).](image)

All 4 secondary structural states were represented in the interfaces; with the helical secondary structural configuration being the most common. In the coil, sheet and turn regions, the secondary structural percentages were more closely correlated with those of the exterior residues. These findings are comparable to the analysis of Argos (1988), who calculated percentage distributions of secondary structural states (helix, strand, turn and coil) of interfaces according to their surface contribution.

**Interface Classification by Secondary Structure**

The interfaces in the non-homologous dimers were classified into 4 types, $\alpha$, $\beta$, $\alpha/\beta$ and coil (Figure 2.17). The classification was based on the percentage frequency of $\alpha$ and $\beta$ secondary structure in the interface residues (Table 2.3). In the data set 53.2% of the protein interfaces were classified as $\alpha$, 21.8% as $\beta$, 12.5% as $\alpha/\beta$ and 12.5% as coil.
Table 2.3: Definitions used in the classification of 32 dimer interfaces by secondary structure. Each dimer interface, defined at the residue level was classified according to its percentage content of $\alpha$ and $\beta$ secondary structure.

<table>
<thead>
<tr>
<th>Classification</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha$</td>
<td>$\alpha &gt; 20%$ &amp; $\beta &lt; 20%$</td>
</tr>
<tr>
<td>$\beta$</td>
<td>$\alpha &lt; 20%$ &amp; $\beta &gt; 20%$</td>
</tr>
<tr>
<td>$\alpha/\beta$</td>
<td>$\alpha &gt; 20%$ &amp; $\beta &gt; 20%$</td>
</tr>
<tr>
<td>Coil</td>
<td>$\alpha \leq 20%$ &amp; $\beta \leq 20%$</td>
</tr>
</tbody>
</table>

Figure 2.17: Classification of dimer interfaces by secondary structure. The solid lines indicate the 20% cut-offs used in the classification (see Table 2.3).

The 7 $\beta$ interface proteins were further classified into 3 groups (extended, stacked and complex) according to the arrangement of the interface sheets between the 2 subunits in the protein. The classification was conducted by visually inspecting the structures using the graphics software Quanta (Quanta 4.0: Molecular Simulations Inc.). The interfaces were classified as extended if the sheets in the 2 interfaces formed a continuous sheet between the 2 subunits. Interfaces where the sheets in one subunit were stacked on top of the sheets in the second subunit were classified as stacked. Interfaces where the sheets did not clearly stack or form extended conformations were classified as complex. The results of the classifications are shown in Table 2.4, and 2 examples of the classifications are shown as MOLSCRIPT (Kraulis, 1991) diagrams in Figure 2.18. Of the 7 $\beta$ interfaces, 2 were defined as extended, 3 as stacked and 2 as complex. Superoxide dismutase (2sod) was one of the structures classified as complex, having the first and the last strands in a flattened $\beta$ barrel in each subunit interacting across the dimer interface (Tainer et al., 1982).
Table 2.4: Classification of 7 β interfaces into 3 types: extended, stacked and complex.

<table>
<thead>
<tr>
<th>Extended</th>
<th>Stacked</th>
<th>Complex</th>
</tr>
</thead>
<tbody>
<tr>
<td>1cdd</td>
<td>2ssi</td>
<td>1fc1</td>
</tr>
<tr>
<td>5hvp</td>
<td>2rhe</td>
<td>2sod</td>
</tr>
<tr>
<td></td>
<td>2tsc</td>
<td></td>
</tr>
</tbody>
</table>

2.5.7 Hydrogen Bonding

A hydrogen bond is a polar interaction between 2 electronegative atoms, a donor and an acceptor. To identify these polar interactions between the subunits in the dimers, hydrogen bonds were calculated using the program HBPLUS (McDonald and Thornton, 1994). The program involves 2 stages; the generation of a set of possible positions for a hydrogen (H) attached to a donor, and secondly a search between all the atoms to find donor (D) and acceptor
(A) pairs that fit specified geometric criteria. The criteria used for the generation of inter-subunit hydrogen bonds were as follows:

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>D — A</td>
<td>distance &lt; 3.9Å</td>
<td>D — H — A angle &gt; 90°</td>
</tr>
<tr>
<td>H — A</td>
<td>distance &lt; 2.5Å</td>
<td>D — A — AA angle &gt; 90°</td>
</tr>
<tr>
<td>H — A</td>
<td>AA angle &gt; 90°</td>
<td></td>
</tr>
</tbody>
</table>

where AA is the atom attached to the acceptor.

The mean number and percentage of inter-subunit hydrogen bonds in each protein are shown in Table 2.5 with those for intra-subunit hydrogen bonds shown for comparison.

<table>
<thead>
<tr>
<th>Type of Hydrogen Bond</th>
<th>Inter-Subunit H-bonds</th>
<th>Intra-Subunit* H-bonds</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>%</td>
</tr>
<tr>
<td>Main-chain — Main-chain</td>
<td>3.1</td>
<td>23.6</td>
</tr>
<tr>
<td>Side-chain — Side-chain</td>
<td>5.0</td>
<td>38.2</td>
</tr>
<tr>
<td>Main-chain — Side-chain</td>
<td>5.0</td>
<td>38.2</td>
</tr>
<tr>
<td>All</td>
<td>13.1</td>
<td>—</td>
</tr>
</tbody>
</table>

Table 2.5: Mean number and percentage of inter- and intra-subunit hydrogen bonds in 32 protein dimers. *The numbers for the intra-subunit hydrogen bonds are for a single subunit.

The number of inter-subunit hydrogen bonds ranged from 0–46. Phospholipase (1pp2), gene 5/DNA binding protein (2gn5), uteroglobin (lutg) and inorganic pyrophosphatase (1ppy) were found to have no inter-subunit hydrogen bonds. The percentage of each type of hydrogen bond was very different between inter- and intra-subunit bonds. Side-chains were involved in 76.4% of the inter-subunit interactions, but only 32.5% of the intra-subunit interactions. There is a clear positive correlation between the size of the interface and the number of hydrogen bonds present (correlation coefficient (r) = 0.77) (Figure 2.19). Interfaces covering ≥ 1500Å² (per subunit) (13 of the 32 dimers) have on average 0.88 (± 0.40) hydrogen bonds per 100Å² buried (per subunit).

2.5.8 Salt Bridges

In protein structures there are positive (N-terminal amide, histidine, arginine and lysine side chains) and negatively charged (C-terminal carboxyl, aspartic acid and glutamate side chains) groups. Within a protein oppositely charged atoms in close proximity are defined to form a salt bridge if they are ≤4.0Å apart (Barlow and Thornton, 1983). The numbers of inter-subunit salt
bridges were calculated for the dimers, and they were found to range from 0–5. However there was no clear correlation with the size of the interface, with 56% of the proteins having no salt bridges between their subunits.

![Graph showing inter-subunit hydrogen bonds and interface accessible surface area (Å²) for 32 protein dimers. The solid line is the straight line regression of the equation \( y = 0.01x - 4.33 \). The correlation coefficient \( r \) is 0.77.]

**Figure 2.19:** The number of inter-subunit hydrogen bonds and the interface accessible surface area (Å²) (per subunit) for 32 protein dimers. The solid line is the straight line regression of the equation \( y = 0.01x - 4.33 \). The correlation coefficient \( r \) is 0.77.

### 2.5.9 Disulphide Bonding

The number of disulphide bonds between each of the subunits was calculated using the program, HBPLUS (McDonald and Thornton, 1994). The program treats cysteines specifically; if any two cysteines were found with their sulphur atoms within 3 Å of each other they were defined as cystines. Disulphide bonds are formed in oxidizing environments, and therefore occur in extracellular proteins where they contribute to the structural properties of the protein (Schulz and Schirmer, 1979). In the data set 71.8% of the proteins were intracellular, and hence a very small number of inter-subunit disulphide bonds were expected. In fact only 2 proteins with inter-subunit disulphide bonding were found; uteroglobin (lutg), which had 2 disulphide bonds (cys3:cys69' and cys69:cys3') and glutathione reductase (3grs), which had a single bond (cys58:cys63'). The small number of inter-subunit disulphides was expected as such bonds are rare (Thornton, 1981). However when inter-subunit disulphides do occur in proteins they often play an important role in structural stabilisation. Recent protein engineering experiments on 2 structures, platelet-derived growth factor-B (Prestrelski et al., 1994) and thymidylate synthase
Protein-Protein Interactions: Chapter 2

Protein Dimers

(Gokhale et al., 1994) have shown that the introduction of inter-subunit disulphides into multimeric proteins increases the stability of the proteins.

Uteroglobin (lutg) is an extracellular protein, in which the 2 disulphide bonds hold the 2 protomers of the structure together (Morize et al., 1987). Glutathione reductase (3grs) is one of the few intracellular proteins that contain a disulphide bond. In general disulphides in intracellular proteins have functions other than structure stabilisation, and in glutathione reductase the disulphide bond has an active catalytic function (Schulz et al., 1982).

2.5.10 Segmentation

To analyse the discontinuous nature of the interfaces, in terms of the amino acid sequence, the numbers of segments in the interfaces were calculated. It was defined that interface residues separated by more than 5 residues were allocated to different segments. For example in the following set of interface residues (indicated by their 1 letter code) 4 segments were defined:

<table>
<thead>
<tr>
<th>Residue Number</th>
<th>EG GY PI WP RT EE L</th>
</tr>
</thead>
<tbody>
<tr>
<td>Segment</td>
<td>1 1 1 2 2 2 2 3 4 4 4 4</td>
</tr>
</tbody>
</table>

In the homo-dimer data set the number of segments ranged from 2-15 (with a mean of 5.5) and there was a very weak correlation with the size of the interface (correlation coefficient \( r \) is 0.59) (Figure 2.20). Alcohol dehydrogenase (5adh) had a comparatively large number of segments for the size of its interface, that is probably reflective of the coiled nature of the interface in this structure (section 2.5.6).
Figure 2.20: The number of residue segments in 32 dimer interfaces. It was defined that interface residues separated by more than 5 residues were assigned to different segments. The data point symbols divide the interfaces into their secondary structural definitions: circle = α, triangle = β, square = αβ, star = coil. The continuous line is the fitted line of the equation $y=0.003x$ and the dashed lines are the 90% confidence limits of the individual y values. The correlation coefficient (r) is 0.59.

The contribution of each segment to the ASA of the interface (Figure 2.21) and the number of hydrogen bonds and salt bridges in each segment of a protein were also calculated (Figure 2.22). All the dimers had particular segments that dominated in terms of their ASA contribution. For example, those proteins with more than 10 segments generally had 1–4 dominating segments, e.g. alcohol dehydrogenase (Sadhi), which had 15 segments, had a single predominant segment that contributed 54% of the ASA of the protein's interface. As expected, the distribution of hydrogen bonds and salt bridges in the segments correlated closely with the ASA contributions of the segments. The larger the ASA contribution of the segment, the greater the number of hydrogen bonds and salt bridges within that segment. These predominant segments may represent exploitable 'hot spots' in the interface, which could prove important targets for mutagenesis.
Figure 2.21: Interface ASA contributions of interface segments. The graphs indicate the percentage ASA contributed by each segment defined in each dimer interface.
Figure 2.22: Bonding in dimer interface segments. The graphs show the number of hydrogen bonds (dark blocks) and salt bridges (light blocks) occurring in each segment defined in each dimer interface. Those dimers with no inter-subunit bonds are not shown.
2.5.11 Gap Volumes and Surface Complementarity

The complementarity of the interacting surfaces was evaluated by quantifying the volume of the gaps existing between the subunits in each of the dimers, using the program SURFNET (Laskowski, 1991). This program considered each pair of subunit atoms in turn (a pair consists of one atom from the first subunit and one from the second), placing a sphere (maximum radius 5.0 Å) half-way between the surfaces of the 2 atoms, such that its surface just touched the surfaces of the atoms in the pair. Checks were then made to test if any other atoms intercepted this sphere and each time an intercept was found, the size of the sphere was reduced accordingly. If at any time the size of the sphere fell below a minimum radius of 1Å the sphere was discarded. If the sphere remained after all the checks its size was recorded. The sizes of all the allowable gap-spheres were then used to calculate the gap volume between the 2 subunits. It was found that there was a positive correlation between the size of the interface (per dimer) and the gap volume (Figure 2.23).

![Graph showing the correlation between gap volume and interface ASA](image)

**Figure 2.23**: Gap volumes and accessible surface areas of dimer interfaces. The gaps between the interacting subunits of each dimer were calculated using SURFNET (Laskowski, 1991). The continuous line is the fitted line of the equation $y=1.75x$. The correlation coefficient (r) is 0.82.

2.5.12 Protrusion of Interface Residues

Residues involved in protein-protein interactions could be 'preformed', in that key surface residues (the interface residues) could be specifically exposed on the protomer surface (possibly anchored by intra-subunit bonding) offering favourable interaction sites for a second protomer. To investigate this hypothesis, residue accessibilities (see section 2.4) were used as a measure of the 'protrusion' of residues from the surface of the dimer. The distributions of relative
accessibilities of the 20 amino acids were compared between the interface and the exterior in isolated subunits (Figure 2.24). The distributions clearly showed that the residues in the interface were more accessible than those in the exterior (in an isolated subunit), and the mean of each distribution (Table 2.6) supported this result. Their increased accessibility implies that the interface residues have fewer intramolecular contacts and hence are more flexible. This could indicate that the interfaces are not preformed; that the residues are not fixed into a specific conformation until the actual dimerization takes place.

Table 2.6: Mean relative accessibilities for each of the 20 amino acids compared in the exterior and the interface.

<table>
<thead>
<tr>
<th>Residue</th>
<th>Mean Relative Accessibilities</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Exterior</td>
</tr>
<tr>
<td>Ala</td>
<td>39.11</td>
</tr>
<tr>
<td>Asn</td>
<td>45.89</td>
</tr>
<tr>
<td>Arg</td>
<td>39.79</td>
</tr>
<tr>
<td>Asp</td>
<td>48.35</td>
</tr>
<tr>
<td>Cys</td>
<td>19.86</td>
</tr>
<tr>
<td>Gln</td>
<td>43.46</td>
</tr>
<tr>
<td>Glu</td>
<td>48.27</td>
</tr>
<tr>
<td>Gly</td>
<td>44.27</td>
</tr>
<tr>
<td>His</td>
<td>35.73</td>
</tr>
<tr>
<td>Ile</td>
<td>26.37</td>
</tr>
<tr>
<td>Leu</td>
<td>25.35</td>
</tr>
<tr>
<td>Lys</td>
<td>51.09</td>
</tr>
<tr>
<td>Met</td>
<td>31.99</td>
</tr>
<tr>
<td>Pro</td>
<td>46.20</td>
</tr>
<tr>
<td>Phe</td>
<td>20.06</td>
</tr>
<tr>
<td>Ser</td>
<td>44.63</td>
</tr>
<tr>
<td>Trp</td>
<td>19.35</td>
</tr>
<tr>
<td>Tyr</td>
<td>25.44</td>
</tr>
<tr>
<td>Thr</td>
<td>38.90</td>
</tr>
<tr>
<td>Val</td>
<td>27.66</td>
</tr>
<tr>
<td>Mean</td>
<td>36.09</td>
</tr>
</tbody>
</table>

To further investigate the presence of 'protruding' residues in the interface, dominant interface residues were defined. This definition was based on the criterion that dominant residues had an absolute ASA of more than 2 standard deviations above the mean ASA of all the residues in the dimer. Using this definition the number of dominant residues in the 32 dimers ranged from 0–7. As expected, from their definition, the larger residues, such as the aromatics, were more highly represented than the smaller linear residues.
Figure 2.24 The distributions of relative accessibilities in the 20 amino acids; compared between the exterior residues (dark bars) and the interface residues (light bars) of 32 non-homologous dimers.
Visual inspection of the dominant residues revealed that they protruded from one subunit into the other and in some interfaces were interdigitating (Figure 2.25). The overall impression was that each dimer had a small number of residues that protruded from one subunit to the other, often at dispersed sites over the interface, thus physically anchoring the 2 subunits together.

Figure 2.25: Liquorice bond diagrams of dominant residues in 2 dimer interfaces. In each diagram one subunit is coloured blue and one red, the dominant residues are displayed in contrasting colours, the green residues are in the 'blue' subunit and the yellow residues are in the 'red' subunit. (a) Inorganic Pyrophosphatase (1pyp). (b) P-Hydroxybenzoate Hydroxylase (1phh).

2.5.13 Flexibility of Atoms in the Dimer Interface

Crystallographic temperature factors (B-factors) have been used to analyse the flexibility of atoms and residues in dimer interfaces. The determination of atomic motion (and hence temperature factors) by crystallography involves the application of a physical model. The Debye-Waller model assumes that the probability of finding an atom a given distance $x$ from its
equilibrium position $x_0$ is Gaussian (Blundell and Johnson, 1976). When the motion is assumed to be isotropic (where atoms are free to move uniformly in all directions) the motion in any direction can be analysed in terms of the mean square displacement. The X-ray scattering from each atom is modified by a Gaussian function that is related to the mean square displacement of that atom (Petsko and Ringe, 1984). The Gaussian function is

$$\exp[-B(\sin^2 \theta / \lambda^2)]$$  \hspace{1cm} [2.7]

where $\lambda$ is the wavelength of the incident radiation and $\theta$ is the Bragg angle (half the angle between the directions of the incident and diffracted beams of the radiation). $B$, termed the isotropic temperature factor, or the Debye-Waller factor, is related to the mean square displacement of the atom from its mean position ($u$) by the expression

$$B = 8\pi^2 <u^2>$$  \hspace{1cm} [2.8]

The B-factor calculated for each atom, as the result of a restrained least squares refinement, can be viewed as an attempt to fit a Gaussian to the spread of the electron density about the average position of that atom. Any factor that contributes to the spreading of this electron density will contribute to the B-factor. A major contributor is atomic motion which can be individual motion (e.g. vibrations) or collective motion (e.g. movements of groups of atoms that are covalently linked and hence move as one entity).

The distributions of B-factors in the 23 protein dimers (single subunits only) (Figure 2.26) reveal their diverse nature, both within single protein structures and between different proteins. Nine proteins (1cdt, 1il8, 1pyp, 2or1, 2ssi, 3aat, 3gap, 3sdp and 5adh) did not have valid B-factors assigned, and in the subsequent analysis a subset of 23 protein dimers was therefore used.

**Comparisons of Atomic Temperature Factors Within Dimeric Proteins**

The percentage frequencies of B-factors in the interior, interface and exterior atoms of the 23 dimers and the ratios of these frequencies for interface/interior, interface/exterior and exterior/interior have been calculated (Figure 2.27).
Figure 2.26: Percentage frequency distributions of B-factors in 23 non-homologous protein dimers. The mean (mn) and standard deviation (sd) for each distribution are shown on each graph.
Protein-Protein Interactions: Chapter 2  Protein Dimers

**Figure 2.27:** Percentage frequency ratios of atomic B-factors for a subset of 23 non-homologous protein dimers.

In the data set all of the proteins had an exterior/interior ratio of > 1.0 (Figure 2.27c). This indicated that the interior residues had atoms that were less flexible than those on the exterior of the protein. This was expected as protein interiors are less solvent accessible than those in the exterior, and have more steric hindrance, due to the closely packed nature of atoms.

58% of the proteins had an interface/interior ratio of > 1.0 (Figure 2.27a), indicating that the interface was composed of residues with atoms slightly more flexible than those in the interior. In the comparison of interface to exterior, all proteins had an interface/exterior ratio of < 1.0 (2.27b). A ratio of < 1.0 indicated that the interface residues have atoms that are less flexible than those in the exterior of the structure. This was to be expected as the interface residues were involved in protein-protein interactions.
2.6 A Tool for the Analysis of Interfaces in Multimeric Proteins

The analysis discussed in this chapter has lead to the development of a computational tool (named PROFACE) to analyse the interfaces of protein complexes. The program was designed to automatically generate interface data, including size, planarity, polarity, inter-subunit bonding, segmentation and amino acid composition. It operates on the 3-dimensional crystallographic co-ordinates of multimeric structures (and other protein-protein complexes, including enzyme-inhibitor and antibody-protein complexes) recorded in the current (1995) PDB file format. A series of Postscript™ tables and graphs are produced, as well as a number of flat data files which can be used as input to other programs. An example of the output for the interface in interleukin 8 (II8) is shown in Figure 2.28. This computational tool makes it possible to compare interfaces in dimers and in higher multimeric structures relatively quickly.

2.7 Conclusion

The structural properties of dimer protein-protein interfaces are intermediate of the exterior and the interior; but are more closely correlated to the exterior than the interior. Specifically the following characteristics describe the observations made in this study:

- The ASA is approximately linearly related to the molecular weight of the protomer
- The ASA buried (per subunit) in the interface ranges from 368.1Å² - 4746.1Å²
- The ASA/Mr ratio decreases from protomer to dimer
- The interface is more hydrophobic than the exterior but less hydrophobic than the interior
  - atoms: 68.1% non polar, 31.9% polar
  - residues: 46.7% hydrophobic, 31.4% polar, 21.9% charged
- The interface has an affinity for hydrophobic residues, as well as arginine and single aromatic residues (histidine, tyrosine, phenylalanine)
- On average there are 0.88 hydrogen bonds per 100Å² of ASA buried (for interfaces covering >1500Å² per subunit)
- Interfaces occur between helix, sheet and coil motifs, with both like and non-like interactions across the interface
- Interfaces are discontinuous, segmented surfaces, with 2 - 15 segments and a mean of 5.5
- The gap volume between dimer subunits is approximately proportional to the ASA of the interface (per dimer)
### Interleukin 8 (1il8 A) A B INTERFACE

<table>
<thead>
<tr>
<th>PROPERTY</th>
<th>DATA</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Interface ASA (Å²)</td>
<td>A+B 1579.70 (13.32%)</td>
</tr>
<tr>
<td></td>
<td>Subunit A 789.53 (13.32%)</td>
</tr>
<tr>
<td>2 Size/Shape of Interface (Å)</td>
<td>x 33.5 Å</td>
</tr>
<tr>
<td></td>
<td>y 20.8 Å</td>
</tr>
<tr>
<td></td>
<td>sdx/sdy 0.48</td>
</tr>
<tr>
<td>3 RMS of atoms from best fit plane (Å) (A)</td>
<td>2.89</td>
</tr>
<tr>
<td>4 Molecular Weight</td>
<td>Dimer 16749.86</td>
</tr>
<tr>
<td></td>
<td>Subunit A 8374.93</td>
</tr>
<tr>
<td></td>
<td>Subunit B 8374.93</td>
</tr>
<tr>
<td>5 &quot;Sphericity&quot; (ASA/Mr ratio)</td>
<td>Subunit A 0.71</td>
</tr>
<tr>
<td></td>
<td>Dimer 0.61</td>
</tr>
<tr>
<td>6 Number of segments in interface (Å)</td>
<td>4</td>
</tr>
<tr>
<td>Segment Limits (residue numbers)</td>
<td>(22-39) (53-53) (59-59) (65-72)</td>
</tr>
<tr>
<td>7 Percentage contribution to the ASA (Å):</td>
<td>Polar 47.61</td>
</tr>
<tr>
<td>Interface Atoms</td>
<td>Non-Polar 52.39</td>
</tr>
<tr>
<td></td>
<td>Exterior Atoms 47.48</td>
</tr>
<tr>
<td></td>
<td>52.52</td>
</tr>
<tr>
<td>8 Interface Secondary Structure (A)</td>
<td>Alpha/Beta</td>
</tr>
<tr>
<td>9 Numbers of Hydrogen Bonds</td>
<td>Mc-Mc 6</td>
</tr>
<tr>
<td>Inter-Subunit (Subunits A and B)</td>
<td>Sc-Sc 0</td>
</tr>
<tr>
<td>Intra-Subunit (Subunit A)</td>
<td>Sc-Mc 4</td>
</tr>
<tr>
<td></td>
<td>Total 40</td>
</tr>
<tr>
<td>10 Number of Inter-Subunit Salt Bridges</td>
<td>0</td>
</tr>
<tr>
<td>11 Number of Inter-Subunit Disulphide Bonds</td>
<td>0</td>
</tr>
<tr>
<td>12 Gap Volume between A and B (Å³)</td>
<td>4764.75</td>
</tr>
<tr>
<td>13 Gap Volume Index</td>
<td>3.02</td>
</tr>
</tbody>
</table>

**Interface Segment Properties (Subunit A)**

**ASA Contributions**

**Hydrogen Bonding**

**Figure 2.28:** Postscript™ output produced by the PROFACE computational tool; designed to automatically analyse protein-protein interfaces. The example shown is for the interface of Interleukin 8 (subunit A).
This study of protein-protein interactions in dimers has consolidated the results previously presented by Argos (1988) and Janin et al. (1988). It has provided additional data on the surfaces involved in protein-protein interactions in multimeric proteins. The interfaces have been shown to be in general circular, planar, hydrophobic patches on the surface of proteins. This work has also identified the dominant segments of the polypeptide chain, which could prove to be exploitable 'hot spots' for mutagenesis.

Multimeric proteins represent just one system in which protein-protein interactions are important, and interactions have been studied in many protein-inhibitor, protein-ligand and antibody/antigen complexes (e.g. Janin and Chothia, 1990). These different systems represent different levels of interaction. The interactions between subunits in dimeric proteins are amongst the strongest and most extensive (ASA up to 4746.1\,\AA^2 per subunit). Dimer interactions are long lived, with isolated oligomer subunits rarely achieving their biological function in the monomeric state. These features distinguish dimer interactions from other protein-protein interactions, which in comparison are often weaker temporary contacts. Comparisons between different types of protein-protein complex are discussed in Chapter 3.

The evolutionary pathway followed to achieve such effective and optimal interactions has been sought, and a number of possible modes of evolution of oligomerization have been considered (Riddihough, 1994). The most recent hypothesis, termed 'domain swapping' (Bennett et al., 1994), has been observed between the monomeric \( \gamma B \) and the homodimeric \( \beta B_2 \) crystallins. Structural analysis has shown that whilst the N- and C-terminal domains interact within the same molecule in the \( \gamma B \)-crystallin structure, the corresponding interaction in \( \beta B_2 \) is between domains from different subunits (Slingsby et al., 1991). Hence, in the \( \beta \)- and \( \gamma \)-crystallin family pairs of domains associate either through intramolecular contacts to form monomeric proteins, or through intermolecular contacts to form oligomers (Slingsby et al., 1992b). In effect interfaces that have evolved to be optimal in the monomer are 'hijacked' to stabilise the oligomeric structure. However, it should be considered that many dimers are not multidomain proteins and solid evidence for domain swapping is rare, so although it may occur occasionally, it is unlikely to be the dominant evolutionary pathway for forming stable dimers.

Whatever the mode of evolution, protein oligomers have evolved to fulfil a wide variety of biological functions, controlling many cellular processes including signal transduction and
enzyme co-operativity. Protein oligomers are observed as allosteric enzymes and multienzyme complexes, and in many such structures the catalytic or functional units are found repeated in each subunit. In some structures the binding of substrates occurs at the protein-protein interface; for example, the trimeric aspartate transcarbamylase from *E.coli* has an active site that lies at the interface between adjacent subunits. The functional role of oligomer interfaces is also seen in the dimeric DNA binding proteins, such as gene 5 DNA binding protein (Brayer and McPherson, 1983). This protein has symmetry related DNA binding sites that include interactions with both subunits of the structure.

This current study of dimer-dimer interfaces has effectively produced a rule base of interface characteristics including size, residue composition, hydrophobicity, planarity, segmentation and complementarity. With this information it should be possible to formulate a new knowledge-based predictive algorithm for the identification of interface forming regions on the surfaces of protomers. The development of such a predictive algorithm is discussed in Chapter 5.
CHAPTER 3

Comparison of Protein-Protein Interactions in Different Categories of Protein Complex

3.1 Introduction

Many biological functions involve the formation of protein-protein complexes. In this chapter the factors that influence the formation of these complexes are explored and compared in 5 different categories of complex. The complexes are, homo-dimeric proteins, hetero-dimeric proteins, homo-trimeric proteins, enzyme-inhibitor complexes, and antibody-protein complexes. Within these 5 categories, 2 classes of complex are identified; permanent and 'optional'. The permanent complexes are those in which the components only exist and function in the complexed state. The 'optional' complexes are those in which the components exist independently, in addition to their complexed state. These include structures which once complexed, remain complexed; and structures which form and break interactions depending upon external factors. The homo-dimers, homo-trimers, and some hetero-dimers fall into the permanent class. Most homo-dimers are only observed in the multimeric state, and it is often impossible to separate them without denaturing the individual monomer structures. Many homo-dimers also have 2-fold symmetry, which places additional constraints on their inter-subunit relationship. The enzyme-inhibitors, antibody-protein complexes and some hetero-complexes fall into the 'optional' class. Many enzyme-inhibitor complexes are strongly associated, with binding constants ranging from $10^7$ mol$^{-1}$ to $10^{13}$ mol$^{-3}$ (Laskowski and Kato, 1980), yet these molecules also exist independently as stable entities in solution. The comparisons between the complexes, highlight differences which reflect their biological roles.

Whilst the homo-dimers and the homo-trimers are both classed as permanent complexes they represent different modes of association. Monod et al. (1965) classified 2 modes of association in oligomeric proteins, isologous and heterologous, and Matthews and Bernhard (1973) defined a third mode of association, pseudo-isologous (see Chapter 1, section 1.3.2). The homo-dimers analysed in this chapter included structures with isologous associations (those with 2-fold rotational symmetry) and pseudo-isologous associations (those lacking 2-fold rotational symmetry). The homo-trimers represent structures with heterologous associations. Thus comparisons were also made between structures with different modes of association.
Comparison of Protein-Protein Interactions

The factors influencing protein-protein interactions in specific oligomeric complexes have previously been examined (Argos 1988; Janin et al 1988; Jones and Thornton, 1995a). In addition the comparison of different types of complex (enzyme inhibitor and antibody-antigen) in terms of interface size and hydrophobicity has been addressed (Janin and Chothia, 1990; Duquerroy et al., 1991). The work described in this chapter utilizes methods implemented in Chapter 2, extending the work to make a detailed comparison between different categories of protein-protein complex.

3.2 Categories of Protein Complex

In this chapter 63 complexes found in the PDB have been divided into the following 5 categories:

- 32 non-homologous homo-dimers (proteins with 2 homologous subunits) (Chapter 2, Table 2.1).
- 4 non-homologous homo-trimers (proteins with 3 homologous subunits) (Table 3.1a)
- 10 enzyme-inhibitor complexes (Table 3.1b)
- 6 antibody-protein (antigen) complexes (Table 3.1c)
- 11 ‘other’ hetero-complexes (Table 3.1d). These include complexes comprised of non-identical subunits which do not fit into either the enzyme-inhibitor or the antibody-protein categories. This category contains 10 structures, but 11 protein-protein interfaces, as 2 hormone-receptor interfaces are included for the human growth hormone-receptor complex (3hhr). This category includes 4 structures that occur only as hetero-complexes (permanent), and 6 structures that occur as both monomers and hetero-complexes (‘optional’).

In addition, the 59 dimeric complexes (i.e. the complete data set of 63 complexes excluding the 4 trimers) have been assigned to 2 classes; permanent and ‘optional’ (Table 3.2).
Protein-Protein Interactions: Chapter 3

Comparison of Protein-Protein Interactions

(a) Trimers

<table>
<thead>
<tr>
<th>PDB Code</th>
<th>Protein</th>
<th>Resolution (Å)</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>1afn</td>
<td>Nitrite Reductase</td>
<td>2.6</td>
<td>Godden (1991)</td>
</tr>
<tr>
<td>1tnf</td>
<td>Tumour Necrosis Factor-α</td>
<td>2.6</td>
<td>Eck and Sprang (1989)</td>
</tr>
<tr>
<td>2chs</td>
<td>Chorismate Mutase</td>
<td>1.9</td>
<td>Chook et al (1993)</td>
</tr>
<tr>
<td>3ela</td>
<td>Type III Chloramphenicol Acetyl Transferase</td>
<td>1.75</td>
<td>Leslie (1990)</td>
</tr>
</tbody>
</table>

(b) Enzyme-Inhibitors

<table>
<thead>
<tr>
<th>PDB Code</th>
<th>Complex Components</th>
<th>Resolution (Å)</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>1acb</td>
<td>Alpha-Chymotrypsin / Eglin C</td>
<td>2.0</td>
<td>Frigerio et al (1992)</td>
</tr>
<tr>
<td>1cho</td>
<td>Alpha-Chymotrypsin / Ovomucoid Third Domain</td>
<td>1.8</td>
<td>Fujinaga et al (1987)</td>
</tr>
<tr>
<td>1ces</td>
<td>Subtilisin Carlsberg / Eglin C</td>
<td>1.2</td>
<td>Rode et al (1987)</td>
</tr>
<tr>
<td>1mcn</td>
<td>Trypsin / Inhibitor from Bitter Gourd</td>
<td>1.6</td>
<td>Huang et al (1993)</td>
</tr>
<tr>
<td>1meo</td>
<td>Peptidyl Peptide Hydrolase / Eglin C</td>
<td>2.0</td>
<td>Dauter and Betzel (1991)</td>
</tr>
<tr>
<td>1sfh</td>
<td>Papain / Inhibitor Stefan B Mutant</td>
<td>2.37</td>
<td>Stubbs et al (1990)</td>
</tr>
<tr>
<td>1tab</td>
<td>Trypsin / Bowman-Birk Inhibitor</td>
<td>2.3</td>
<td>Tsunogae et al (1986)</td>
</tr>
<tr>
<td>1tgs</td>
<td>Trypsinogen / Pancreatic Secretory Trypsin Inhibitor</td>
<td>1.8</td>
<td>Bolognesi et al (1982)</td>
</tr>
<tr>
<td>2ptc</td>
<td>Beta Trypsin / Pancreatic Trypsin Inhibitor</td>
<td>1.9</td>
<td>Marquart et al (1983)</td>
</tr>
<tr>
<td>2sii</td>
<td>Subtilisin / Streptomyces Subtilisin Inhibitor</td>
<td>1.8</td>
<td>Takeuchi et al (1991)</td>
</tr>
</tbody>
</table>

(c) Antibody-Protein Complexes

<table>
<thead>
<tr>
<th>PDB Code</th>
<th>Complex Components</th>
<th>Resolution (Å)</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>1fdl</td>
<td>D1.3 Fab / Hen Egg White Lysozyme</td>
<td>2.5</td>
<td>Fischmann et al (1991)</td>
</tr>
<tr>
<td>1je1</td>
<td>JE142Fab / Histidine Containing Protein</td>
<td>2.8</td>
<td>Prasad et al (1993)</td>
</tr>
<tr>
<td>1jhl</td>
<td>D11.15 Fv / Pheasant Egg Lysozyme</td>
<td>2.4</td>
<td>Chitarra et al (1993)</td>
</tr>
<tr>
<td>1mca</td>
<td>NC41Fab / Influenza Virus N9 Neuraminidase</td>
<td>2.5</td>
<td>Tulip et al (1992)</td>
</tr>
<tr>
<td>2hfl</td>
<td>HYHEL-5 Fab/Chicken Lysozyme</td>
<td>2.54</td>
<td>Sheriff et al (1987)</td>
</tr>
<tr>
<td>3hfl</td>
<td>HYHEL-10 Fab/Chicken Lysozyme</td>
<td>3.0</td>
<td>Padlan et al (1989)</td>
</tr>
</tbody>
</table>

(d) 'Other' Hetero-Complexes

<table>
<thead>
<tr>
<th>PDB Code</th>
<th>Complex Components</th>
<th>Resolution (Å)</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>1atn</td>
<td>Decoribonuclease I/Actin</td>
<td>2.8</td>
<td>Kabsch et al (1990)</td>
</tr>
<tr>
<td>1gla</td>
<td>Glycerol Kinase/Glucose Specific Factor III</td>
<td>2.6</td>
<td>Hurley et al (1993)</td>
</tr>
<tr>
<td>1hpa</td>
<td>Human Chorionic Gonadotropin</td>
<td>3.0</td>
<td>Lathorn et al (1994)</td>
</tr>
<tr>
<td>1pya</td>
<td>Lipase/Colipase</td>
<td>3.04</td>
<td>van Tilbeorgh et al (1993)</td>
</tr>
<tr>
<td>1lya</td>
<td>Cathepsin D</td>
<td>2.5</td>
<td>Baldwin et al (1993)</td>
</tr>
<tr>
<td>2btf</td>
<td>Beta-Actin / Profilin</td>
<td>2.55</td>
<td>Schutt et al (1993)</td>
</tr>
<tr>
<td>2pcb</td>
<td>Yeast Cytochrome C Peroxidase / Horse Cytochrome C</td>
<td>2.8</td>
<td>Pelletier and Kraut (1992)</td>
</tr>
<tr>
<td>3hhr</td>
<td>Human Growth Hormone/Human Growth Hormone Receptor</td>
<td>2.8</td>
<td>de Vos et al (1992)</td>
</tr>
<tr>
<td>3hvr</td>
<td>Reverse Transcriptase</td>
<td>2.9</td>
<td>Wang et al (1994)</td>
</tr>
<tr>
<td>6rix</td>
<td>Relaxin</td>
<td>1.5</td>
<td>Eigenbrot et al (1991)</td>
</tr>
</tbody>
</table>

(# contributes 2 hormone-receptor interfaces, † derived from a single chain precursor).

Table 3.1: The data sets of 4 categories of protein-protein complex. (The data set for the fifth category, the homo-dimers, is recorded in Table 2.1, Chapter 2).
Table 3.2: The division of the dimeric protein-protein complexes into permanent and ‘optional’ classes.

### 3.3 Definition of a Protein-Protein Interface

Protein-protein interfaces have been defined based on the change in their solvent accessible surface area (ΔASA) when going from a monomeric to a dimeric state. The ASAs of the complexes were calculated using the program ACCESS (Hubbard, 1990), an implementation of the Lee and Richards (1971) algorithm. The interface residues were defined as those having ASAs that decreased by >1Å² on complexation (see Chapter 2, section 2.4.1). The interface atoms were defined as those having ASAs that decreased by >0.01Å² on complexation (see Chapter 2, section 2.4.2).

### 3.4 Characterisation and Comparison of Protein-Protein Interfaces

In Chapter 2 it was shown that there were several fundamental properties (calculated from the 3-dimensional co-ordinates of a complex) that characterise a protein-protein interface. In the current chapter a selection of these characteristics have been calculated, and the results compared between the 5 categories, and 2 classes, of protein-protein complexes defined in section 3.2.

#### 3.4.1 Interface Size

The size of protein interfaces can be measured simply in absolute dimensions (Å) or, more accurately, in terms of the loss of solvent accessible surface area (ΔASA) on complexation. The ΔASA method was used as there is a known correlation between the hydrophobic free energy of transfer from a polar to a hydrophobic environment and the solvent ASA (Chothia, 1974). Thus, calculating ΔASA may provide a measure of the binding strength.
<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Homo-Dimers</th>
<th>Enzyme-Inhibitor (EI)</th>
<th>Antibody-Protein (AP)</th>
<th>Other Hetero Complexes (AB)</th>
<th>Trimers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Examples</td>
<td>32</td>
<td>10</td>
<td>6</td>
<td>7</td>
<td>optional permanent 4</td>
</tr>
<tr>
<td>(i) Δ ASA (Å²)</td>
<td>Mean</td>
<td>1685</td>
<td>785</td>
<td>777</td>
<td>848</td>
</tr>
<tr>
<td></td>
<td>σ</td>
<td>1101</td>
<td>74</td>
<td>135</td>
<td>243</td>
</tr>
<tr>
<td>(ii) Planarity (Å)</td>
<td>Mean</td>
<td>3.46</td>
<td>2.70</td>
<td>2.21</td>
<td>2.54</td>
</tr>
<tr>
<td></td>
<td>σ</td>
<td>1.72</td>
<td>0.45</td>
<td>0.39</td>
<td>0.61</td>
</tr>
<tr>
<td>(iii) Circularity</td>
<td>Mean</td>
<td>0.71</td>
<td>0.73</td>
<td>0.55</td>
<td>0.75</td>
</tr>
<tr>
<td></td>
<td>σ</td>
<td>0.17</td>
<td>0.05</td>
<td>0.10</td>
<td>0.12</td>
</tr>
<tr>
<td>Minimum</td>
<td>0.25</td>
<td>0.62</td>
<td>0.43</td>
<td>0.62</td>
<td>0.62</td>
</tr>
<tr>
<td>Maximum</td>
<td>1.00</td>
<td>0.78</td>
<td>0.71</td>
<td>0.92</td>
<td>0.91</td>
</tr>
<tr>
<td>(iv) Segmentation</td>
<td>Mean</td>
<td>5.22</td>
<td>7.80</td>
<td>2.7</td>
<td>3.81</td>
</tr>
<tr>
<td></td>
<td>σ</td>
<td>2.55</td>
<td>1.03</td>
<td>0.95</td>
<td>1.83</td>
</tr>
<tr>
<td>Minimum</td>
<td>2</td>
<td>6</td>
<td>2</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Maximum</td>
<td>11</td>
<td>9</td>
<td>5</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>(v) Hydrogen Bonds (per 100Å² ΔASA)</td>
<td>Mean</td>
<td>0.70</td>
<td>1.37</td>
<td>1.06</td>
<td>0.85</td>
</tr>
<tr>
<td></td>
<td>σ</td>
<td>0.46</td>
<td>0.37</td>
<td>0.51</td>
<td>0.37</td>
</tr>
<tr>
<td>Minimum</td>
<td>0</td>
<td>0.60</td>
<td>0.47</td>
<td>0.39</td>
<td>0.37</td>
</tr>
<tr>
<td>Maximum</td>
<td>1.7</td>
<td>1.87</td>
<td>1.88</td>
<td>1.47</td>
<td>1.47</td>
</tr>
<tr>
<td>(vi) Gap Index (Å)</td>
<td>Mean</td>
<td>2.20</td>
<td>2.20</td>
<td>3.02</td>
<td>3.02</td>
</tr>
<tr>
<td></td>
<td>σ</td>
<td>0.87</td>
<td>0.47</td>
<td>0.80</td>
<td>1.13</td>
</tr>
<tr>
<td>Minimum</td>
<td>0.57</td>
<td>1.38</td>
<td>2.04</td>
<td>2.03</td>
<td>2.03</td>
</tr>
<tr>
<td>Maximum</td>
<td>4.43</td>
<td>2.86</td>
<td>3.96</td>
<td>5.17</td>
<td>3.42</td>
</tr>
</tbody>
</table>

Table 3.3: Results of structural analysis for 5 categories of protein-protein complex.

i. ΔASA: The change in accessible surface area for one subunit on complexation. In the heterodimers, enzyme-inhibitor and ‘other’ hetero complexes, the mean ASA is that buried by each subunit on complexation. For the antibody-protein complexes the mean ASA is that on the protein (antigen) surface buried on complexation. *In the trimers the mean ASA shown is the mean ASA buried by each pair of interacting subunits.

ii. Planarity: Root mean squared (rms) deviation of all the interface atoms from the least squares plane through all these atoms.

iii. Circularity: Ratio of the lengths of the principal axes of the least squares plane through the atoms in the interface.

iv. Segmentation: It was defined that interface residues separated by more than 5 residues were allocated to different segments.

v. Inter-molecular Hydrogen Bonds: The number of inter-molecular hydrogen bonds per 100Å² ΔASA were calculated using HBPLUS (McDonald and Thornton, 1994), in which hydrogen bonds were defined according to standard geometric criteria.

vi. Gap Volume Index: The gap volume between the two components of the complexes was calculated using SURFNET (Laskowski, 1991).
Table 3.4: Results of structural analysis for 2 classes of protein-protein complexes. For details of each characteristic analysed see legend of Table 3.3.

The mean ΔASA on complexation (going from a monomeric state to a dimeric state) was calculated as half the sum of the total ΔASA for both molecules for each type of complex (Table 3.3). In the trimers the ΔASA was calculated for each of the 3 protein-protein interfaces individually (i.e. if the subunits are designated A, B and C, one interface occurs between A and B, a second between A and C, and a third between B and C). The ΔASA was calculated for each pair of subunits forming an interface (AB, AC, BC), as half the sum of the total ΔASA for both subunits. To give a guide as to how much of a protein subunit’s surface was buried on complexation, the ΔASA for individual complexes was compared with the molecular weight of the constituent subunits (Figures 3.1, 3.2, 3.3). For the hetero-complexes the molecular weights are different for each component; therefore the relationship between the molecular weight and the ΔASA was made using the molecular weight of the smallest component, as this would limit
the maximum size of the interface. In the trimers the ΔASA for each pair of interacting subunits was calculated.

In the homo-dimers the ΔASA varied widely from 368 Å² to 4746 Å², and there was a clear, though scattered, relationship with the molecular weight of the subunit (correlation coefficient (r) = 0.69), with the larger molecules, in general, having larger interfaces (Figure 3.1). The ΔASA for each pair of interacting subunits in the trimers ranged from 1002 Å² to 2083 Å². These values were in a similar range to those of the homo-dimers with a similar subunit molecular weight (Figure 3.1). In addition, the total ΔASA for each subunit and the total ΔASA for the whole complex (i.e. going from 3 monomers to a trimer complex) was calculated for each trimer. The mean total ΔASA for each subunit was 2616 Å² (σ = 866), and the mean of the total ASA buried on passing from the monomeric state into the trimeric state was 7849 Å² (σ = 2870). The size of the ASA buried by each subunit (mean = 2616 Å²) was higher than that buried by the subunits in the homo-dimers (mean = 1685 Å²).

The range of ΔASA in the hetero-complexes was smaller than for the homo-dimers, with all the structures having interfaces between 600–1100 Å², with 3 exceptions (Figure 3.2). This relatively small range probably reflects 2 factors; the limited nature of the PDB and the average size of protein domains. In addition, it should be noted that all the enzyme-inhibitor complexes involve proteinases, and with the exception of papain and subtilisin, all are related to trypsin, although the corresponding inhibitors are non-homologous.

Three ‘other’ hetero-complexes (1lya, 3hvt, 1hrp) had relatively large interfaces for their molecular weights (Figure 3.2). These 3 structures were classified as permanent complexes (see Table 3.2) and the size of their interfaces were more comparable with the distribution observed in the homo-dimers (Figure 3.1) than the hetero-complexes. In Figure 3.3 the relationship between molecular weight and ΔASA has been plotted for the permanent and ‘optional’ classes of dimeric complex. It was observed that the ‘optional’ complexes tended to have smaller interfaces than the permanent complexes. This was particularly true of those with subunits of higher molecular weight (Mr > 2000 Daltons).
Figure 3.1: Interface accessible surface area (ASA) (Å²) against molecular weight (Mr) (Daltons) for homo-dimers and homo-trimers. The accessible surface area is that buried by one subunit on dimerization, and the molecular weight is that of the monomer. For the trimers the mean ASA has been calculated pairwise, for each of the 3 interfaces in the oligomer. The dashed line is the straight line regression, of the form $y=ax$, for the homo-dimer data set ($r = 0.69$).

Figure 3.2: Interface accessible surface area (ASA) (Å²) against molecular weight (Mr) (Daltons) for hetero-complexes. The ΔASA and the molecular weight are both from the smallest subunit. The outlying points are labelled with the PDB code of the protein.
3.4.2 Interface Shape

The planarity of the surfaces between 2 components of a complex was analysed by calculating the root mean squared deviation (rms) of all the interface atoms from the least squares plane through the atoms. Means were calculated for each of the 5 categories of complex (Table 3.3), and the 2 classes (Table 3.4). Figure 3.4 shows that the hetero-complexes had interfaces which were more planar than the homo-dimers. However, the higher rms of the homo-dimers results from the 5 proteins which had comparatively high rms values (> 6 Å). These were dimers in which the 2 subunits were twisted together across the interface (e.g. isocitrate dehydrogenase), or proteins that had subunits with 'arms' apparently clasping the 2 halves of the structure together (e.g. aspartate aminotransferase), (see Chapter 2, section 2.5.3). It can be seen from Table 3.4 that the ‘optional’ complexes had interfaces that were more planar than their permanent counterparts.

To provide a rough guide to the shape of the interface, the ‘circularity’ of the interfaces was calculated as the ratio of the lengths of the principal axes of the least squares plane through the atoms in the interface. A ratio of 1.0 indicated that an interface was approximately circular. The shape of the interface region (Table 3.3) varied little between the homo-dimers, the antigens and the enzyme component of the enzyme-inhibitor complex, each type being relatively circular with a mean ratio of between 0.71 to 0.75. In comparison the inhibitors of the enzyme-inhibitor complexes had less circular interfaces, with a mean ratio of 0.55. The trimer interfaces (when
analysed pairwise) also had less circular interfaces, with a mean ratio of 0.58. Each subunit in the trimers have 2 interfaces, and the requirement to accommodate 2 sets of interactions could influence the shape of each interface. The homo-dimers show the largest variation, with the elongated interface of variant surface glycoprotein of *Trypanosoma brucei* at one extreme (ratio 0.25). The interface of this structure, which forms a coat on the surface of the parasite (Freymann *et al.*, 1990), reflects the elongated nature of the protein as a whole.

![Figure 3.4: Planarity of protein-protein interfaces: the root mean squared deviation (rms) (Å) of atoms from the least squares plane through these atoms is shown for one subunit of the homo-dimers, both subunits of the enzyme-inhibitor complexes and ‘other’ hetero-complexes, and for the antigenic protein component of the antibody-protein complexes. Within each group of complexes the proteins have been placed in ascending order of rms. The mean of each data set is indicated by a solid horizontal line.](image)

### 3.4.3 Complementarity Between Surfaces

Many authors have commented on the electrostatic and the shape complementarity observed between associating molecules (e.g. Morgan *et al.*, 1979; Janin and Chothia, 1990; Lawrence and Colman, 1993). In the current work the shape complementarity of the interacting surfaces in the protein-protein complexes was evaluated by defining a gap index as:

\[
\text{Gap Index (Å)} = \frac{\text{Gap Volume Between Molecules (Å}^3\text{)}}{\text{Interface ASA (Å}^2\text{) per complex}}
\]  

[3.1]

The gap volume was calculated using a novel procedure developed by Laskowski (1991), which estimates the volume enclosed between any 2 molecules, delimiting the boundary by defining a
maximum allowed distance from both interfaces (see Chapter 2, section 2.5.11). A mean gap index was calculated for each of the 5 categories of complex (Table 3.3).

The ‘interface ASA’ parameter in equation 3.1 is the ASA of interface atoms defined on the basis of ΔASA upon complexation (see section 3.4.1). However, the interface atom pairs used in the gap volume calculations comprised all pairs (one atom from each subunit) whose surfaces could be touched by a sphere of maximum radius 5.0 Å (see Chapter 2, section 2.5.11). The interface defined on this basis had a much larger ASA than that defined on the basis of ΔASA on complexation. Hence, it must be noted that the data shown should not be used to directly derive a value of gap volume per ASA of interface. The gap volume indices were calculated as a means of comparing the gap volumes between complexes of different categories.

The results indicate that the interacting surfaces in the homo-dimers, the enzyme-inhibitor complexes and the permanent hetero-complexes (a subset of the ‘other’ hetero-complex data set), were the most complementary, whilst the antibody-protein complexes and the ‘other’ hetero-complexes, which were classified as ‘optional’, were the least complementary. However, all 4 distributions overlap considerably (Figure 3.5). These data agree with the conclusions drawn by Lawrence and Colman (1993) using their shape complementarity statistic (Sc). When a mean gap index was calculated for the 2 classes of complex (permanent and ‘optional’) (Table 3.4), the permanent complexes were found to have a greater surface complementarity than the ‘optional’ complexes.

The relative size of the gap volume indices, can be explained if the complexes are considered with respect to their binding constants and evolutionary history. The interactions between antibodies and antigens are the molecular equivalent of a ‘first encounter’. Most antibody-antigen interactions initially have a binding constant of around $10^9$ mol$^{-1}$, but in subsequent immune responses this may increase substantially as somatic mutations improve recognition and the strength of binding. In contrast protein-inhibitor, dimeric, trimeric and some ‘other’ hetero-interactions have been subject to selective evolutionary pressures. Many dimeric interactions, which can have binding constants as high as $10^{16}$ mol$^{-1}$, are so strong that the monomers have to be denatured to separate the subunits. In the enzyme-inhibitor interactions, both partners must also exist independently and be soluble in water. Many of these interactions have binding constants ranging from $10^7$ mol$^{-1}$ to $10^{-13}$ mol$^{-1}$; e.g. the protein inhibitors of proteinases (Laskowski and Kato, 1980).
3.4.4 Residue Interface Propensities

The relative importance of different amino acid residues in the interfaces of complexes can give a general indication of the hydrophobicity. Such information can only be interpreted if the distribution of residues occurring in the interface are compared with the distribution of residues occurring on the protein surface as a whole. Residue interface propensities were calculated for each amino acid (AAj) as the fraction of ASA that AAj contributed to the interface compared with the fraction of ASA that AAj contributed to the whole surface (exterior residues + interface residues) i.e.

\[
\text{Residue Interface Propensity } \text{AAj} = \frac{\sum_{i=1}^{N_i} \text{ASA}_{\text{AAj}(i)}}{\sum_{i=1}^{N_i} \text{ASA}(i)}\]

\[
\text{where } \sum_{\text{ASA}_{\text{AAj}(i)}} = \text{sum of the ASA (in the monomer) of amino acid residues of type } j \text{ in the interface}
\]

\[
\sum_{\text{ASA}(i)} = \text{sum of the ASA (in the monomer) of all amino acid residues of all types in the interface}
\]

\[
\sum_{\text{ASA}_{\text{AAj}(i)}} = \text{sum of the ASA (in the monomer) of amino acid residues of type } j \text{ on the surface (exterior + interface residues)}
\]
**Comparison of Protein-Protein Interactions**

\[ \Sigma \text{ASA}(s) = \text{sum of the ASA (in the monomer) of all amino acid residues of all types on the surface} \]

\[ N_i = \text{Number of residues in the protein interface} \]

\[ N_s = \text{Number of residues on the protein surface} \]

A propensity of >1.0 denotes that a residue occurs more frequently in the interface than on the protein surface. Propensities were calculated for the permanent and ‘optional’ classes of complexes (Figure 3.6).

With the exception of methionine and phenylalanine, the hydrophobic residues showed a greater preference for the interfaces of the permanent complexes than for those of the ‘optional’ complexes. The lower propensities for hydrophobic residues in the ‘optional’ interfaces were balanced by an increased propensity for the polar residues. The aromatic residue tryptophan had a very high propensity in the ‘optional’ data set. This result should be viewed with caution as it reflects the relatively small number of these residues in the exterior and interface subsets of the proteins of this class.

**Figure 3.6:** Residue interface propensities for permanent and ‘optional’ dimeric complexes.

### 3.4.5 Hydrophobicity Including Hydrogen Bonding

It has often been assumed that proteins will associate through hydrophobic patches on their surfaces (e.g. Korn and Burnett, 1991; Young et al., 1994). However, polar interactions between subunits are also common and, in terms of the driving force for complexation, it is
important to explore the relative contributions of these effects, including reference to the subunit's ability to exist independently.

A mean hydrophobicity value (based on the experimental scale of Fauchère and Pliska (1983)) was calculated for residues defined in the interface, interior, and exterior of each complex.

\[
\text{Mean Hydrophobicity} = \frac{\sum_{i=1}^{N} (HV_{AA})}{N} \tag{3.3}
\]

where \(HV_{AA}\) = value assigned to an amino acid residue in the Fauchère and Pliska (1983) hydrophobicity scale

\(N\) = number of residues in the subset (interface, interior or exterior)

Mean values were calculated for each subset of residues in each category of complex and for the 2 classes of complex (permanent and ‘optional’) (Figure 3.7). In each of the 5 categories of complex, and in the 2 classes, the interfaces had a hydrophobicity value intermediate between that of the interior (hydrophobic) and the exterior (hydrophilic). In addition it was observed that the interface of the ‘optional’ complexes were more closely related to the exterior than those of the permanent complexes.

**Figure 3.7:** Relative hydrophobicity of interior, interface and exterior residues in different types of complex. The hydrophobicity value was calculated using the experimentally derived Fauchère and Pliska (1983) hydrophobicity scale.
Comparison of Protein-Protein Interactions

This difference in hydrophobicity, also observed from the interface residue propensities (section 3.4.4), and previously by Janin and Chothia (1990), can be explained by the roles of the two types of complex. The homo-dimers rarely occur or function as monomers, and hence their hydrophobic surfaces are permanently buried within a protein-protein complex. The 23 'optional' dimer complexes also occur as monomers in solution, and have biological functions in this state. Hence these interfaces cannot be as hydrophobic as those of the homo-dimers, as a large exposed hydrophobic patch on the protein would be energetically unfavourable.

To identify the major polar interactions between the components in the complexes, the mean number of hydrogen bonds per $100\text{Å}^2$ of ΔASA was calculated for each of the 5 categories of complex (Table 3.3), and for the permanent and 'optional' classes (Table 3.4). The homo-dimers had the lowest mean number of hydrogen bonds per $100\text{Å}^2$ of ΔASA (0.70). The enzyme-inhibitor complexes had the highest number of hydrogen bonds per $100\text{Å}^2$ of ΔASA (1.37). When the 23 'optional' dimeric complexes were analysed together they had more intermolecular hydrogen bonds per ΔASA than the permanent dimeric complexes. This distribution was expected from the results of the residue interface propensities, which showed that the 'optional' complexes contained more hydrophilic residues in their interfaces than the permanent complexes. It would appear that electrostatic interactions play a larger role in the 'optional' complexes, compared to the permanent complexes in which hydrophobic interactions play a more important role.

3.4.6 Segmentation

The number of discontinuous segments of the polypeptide chain involved in the interface is important since the ability of peptides or small molecules to mimic one half of the interaction may depend upon it. For example, in an interface which is dominated by one segment, a single peptide would probably be a good mimic. However, the design of molecules to mimic multi-segmented interfaces would certainly be more difficult.

To analyse the discontinuous nature of the interfaces, in terms of the amino acid sequence, the mean number of segments in the interfaces was calculated for each of the 5 categories of complex (Table 3.3), and for the 2 classes (Table 3.4). It was defined that interface residues separated by more than 5 residues were allocated to different segments (Chapter 2, section 2.5.10). In the 63 complexes studied the number of segments varied from 1–11. In fact only 1 complex, relaxin, had 1 segment at the interface. Relaxin is a very small protein derived from a
single chain precursor, with only 24 residues in the A-chain and 29 in the B-chain (Eigenbrot et al., 1991). The inhibitors of the enzymes were unusual in having only 2-5 segments interacting. This class of inhibitors has evolved to mimic an elongated segment of polypeptide chain, in the conformation required for cleavage by the enzyme, and therefore all present a protruding canonical loop structure which dominates the interaction (Huber, 1979; Hubbard et al., 1992). In contrast the other interfaces were highly segmented, especially the long binding site cleft in the proteinases, which had a mean of 7.8 segments.

3.4.7 Conformational Changes on Complex Formation

The packing conformation of the complexes was analysed by calculating the number of interface atoms which overlapped (calculated using van der Waals radii) as 2 complex components were moved apart. The radii used were those assigned by Chothia (1976). One complex component was moved from its partner perpendicular to the line of the best fit plane through the interface atoms of both components. The number of atom overlaps was calculated at 0.1, 0.5, 1, 5, 10, 15, 20, 25, 30 and 40Å distance intervals. One interface atom could overlap more than 1 atom at any 1 distance, and these multiple overlaps were included. The total number of overlaps made by each interface atom from 1 complex component were summed and then divided by the number of residues defined in the interface (see Chapter 2, section 2.4.2). These values were recorded at each distance for the 4 categories of dimeric complexes (Figure 3.8) (Only those proteins with atom overlaps are shown).

The number of overlaps observed between complex components (per number of residues in the interface) can be divided into 3 categories, none, slight and gross. Two of these categories (none and gross overlaps) are shown schematically in Figure 3.9. Overall, the permanent complexes were observed to be those for which gross overlaps were recorded, and the enzyme-inhibitors those for which slight overlaps were recorded. However no overlaps were observed for the antibody-protein complexes.
Figure 3.8: Number of overlapping atoms per number of interface residues in protein-protein complexes, as one component of each complex is moved from the other at set distances (Å).
Figure 3.9: Protein-protein complexes in which one component is moved from its partner perpendicular to the line of the best fit plane through the interface atoms of both components. (1) No overlaps: the interface is interdigitating (a) but the movement of one partner (g) does not result in atom overlaps (b). (2) Gross Overlaps: the interface is intertwined (a) and movement of one partner (y) results in gross atom overlaps (b).

Twelve of the 32 homo-dimers recorded overlaps, and of these, 8 involved gross overlaps, with up to 161 overlaps per interface residue. These 8 structures had the highest rms deviations of interface atoms from a best fit plane (see Chapter 2, section 2.5.3). They included structures in which the 2 subunits were twisted together across the interface (e.g. isocitrate dehydrogenase (3icd)), and structures which had subunits with 'arms' clasping the 2 halves of the structure together (e.g. aspartate aminotransferase (3aat)) (see Chapter 2, Figure 2.7). The inter-locking nature of these interfaces means that the subunits cannot be drawn apart without moving part of one subunit through the other. These results infer that such structures must undergo gross conformational changes on dimerization.

In the 'other' hetero-complex data set of 6 structures, 2 structures, human chorionic gonadotropin (1hrp) and cathepsin D (1iya), showed gross overlaps. Both these structures only have a biological function in the dimeric state. Human chorionic gonadotropin is a hetero-dimer in which the β subunit is wrapped around the α subunit and is covalently joined like a 'seat belt' by a disulphide bond (Lapthorn et al., 1994). Moving the 2 halves of the structure apart causes one subunit to be drawn through the second. As in some of homo-dimers, the results infer that gross conformational changes must occur within this structure upon dimerization.
Comparison of Protein-Protein Interactions

The enzyme-inhibitor complexes revealed very different results. In each of the 9 structures where overlaps occurred, there are far fewer overlaps per interface residue than in the structures discussed above. In addition, the overlaps were only recorded at the 0.5–5Å distance intervals, unlike the homo-dimers and the 2 ‘other’ hetero-complexes in which the overlaps extended up to the 30Å distance interval. The enzyme-inhibitor complexes appear to undergo relatively small conformational changes on binding, which could constitute small re-arrangements of amino acid side chains.

No atom overlaps were recorded for the 6 antibody-protein complexes. This could indicate that the binding of the antibodies with the proteins does not involve any conformational changes at the antibody-protein interfaces, or that conformational changes do occur but are minor, and hence are not detected by the current method. In other studies a wide range of conformational changes have been observed in both the antibody and the protein upon binding (Davies and Padlan, 1992; Rini et al., 1992; Friedman et al., 1994; Wilson and Stanfield, 1994). For example, major antigen-induced domain rearrangements have been observed in an antigen binding fragment (Fab) directed against an HIV-1 peptide (Wilson and Stansfield, 1993). The conformational changes observed in this structure not only affected the antigen binding site but also the interface between the variable light and heavy chains.

3.5 Conclusion

This chapter has highlighted the need to take into account the type of protein-protein complex when characterising the interfaces within them. Complexes can be permanent or ‘optional’, and the constraints imposed upon each of these classes result in different interface characteristics. The requirement for the molecules of ‘optional’ complexes to exist as independent entities results in their interfaces being less hydrophobic than those which exist permanently in a complexed state. This was also reflected in the greater number of inter-molecular hydrogen bonds observed in the ‘optional’ complexes. The ‘optional’ complexes had smaller interfaces, and were less well packed at the interface than their permanent counterparts. In addition differences were observed in interface characteristics between the 5 categories of complex. The selective evolutionary pressure exerted upon different complexes was reflected in the complementarity observed at their interfaces. The antibody-protein complexes, which often constitute structures with a very short evolutionary history, had less complementary interfaces.
than those of the homo-dimers or enzymes which have been exposed to selective pressures over a much longer time period.

The occurrence and extent of conformational changes in proteins upon binding is still unclear (Huber, 1979; Wilson and Stanfield, 1994), and currently there are few proteins which have been structurally determined (by crystallography or nuclear magnetic resonance) before and after complexation. The simple method described in this work distinguished between gross conformational changes, which occurred in some homo-dimers and permanent hetero-dimers, and the much smaller changes (possibly side-chain movements) observed in the enzyme-inhibitor complexes. However, the method failed to detect conformational changes in the antibody-protein complexes which have been observed in other studies (e.g. Herron et al., 1991; Rini et al., 1992; Freidman et al., 1994, Wilson and Stanfield, 1994). Overall, both rigid and flexible docking (the latter involving conformational changes) will most likely occur in different circumstances, but there will always be an energetic price to pay for reducing flexibility.

The comparative analysis described in this chapter has updated the earlier work of Chothia and Janin (1975). The results presented are derived from a relatively small data set of protein complexes which needs to be extended. The comparative analysis was difficult because of the lack of information on the in vivo complex-status in the current (1995) PDB file entries; so, for example, extracting proteins in a specific oligomeric state is a very labour-intensive process. However, as the PDB of protein structures grows (and its format is enhanced) it should be possible to include higher order complexes (e.g. tetramers, hexamers, do-decamers) and analyse such factors as the relationship between interface properties and binding constants. The latter are often difficult to determine experimentally and are almost never deposited in the PDB with the co-ordinates, yet are essential if the kinetics and thermodynamics of complex formation are to be understood.
CHAPTER 4

Patch Analysis of Protein-Protein Interaction Sites

4.1 Introduction
Protein interfaces have been observed to be hydrophobic patches on the surfaces of proteins (e.g. Chothia and Janin, 1975; Janin and Chothia, 1990). Such observations have been made by comparing the interface sites with the interior and the surface of the protein. There have been 2 studies specifically comparing the hydrophobicity of protein interfaces with the remainder of the protein surface. Korn and Burnett (1991) quantified the hydrophobic nature of protein interfaces, calculating ‘hydropathy complementarities’ for contact surfaces (i.e. the interface) and non-contact surfaces on multimeric proteins. This analysis revealed that the contact surfaces were more hydrophobic than the non-contact surfaces. Young et al. (1994) analysed clusters of residues on protein surfaces for hydrophobicity, finding that those clusters which formed interfaces with ligands were, in general, the most hydrophobic of all the surface clusters. Thus, they compared the observed interface with other sites on the surface of the protein, for a single parameter. This chapter addresses the problem of comparing the observed interface with other similar sized sites on the protein surface for multiple parameters.

In preceding chapters the interface regions have been analysed in isolation, but it is also instructive to explore whether these regions are significantly different from the rest of the protein surface in any way. The problem to be addressed is ‘given a protein of known structure (but with no known structure for its complex) is it possible to identify the interface region on its surface?’ This involved 2 stages,

- The identification of properties that differentiate the interface region from the rest of the protein surface
- The application of these properties to the prediction of putative interface sites (this is addressed in Chapter 5)

The analysis discussed in this chapter has been conducted upon a subset of the 32 homo-dimers shown in Table 2.1 in Chapter 2. This subset comprised 29 homo-dimers which were complete proteins. Three structures (1fc1, 2or1 and 2rhe) were excluded from the original data set, as
they represented single domains or fragments of complete proteins. The analysis was also extended to the data set of 10 enzyme/inhibitor complexes shown in Chapter 3, Table 3.1b. It should be noted that all the enzyme-inhibitor complexes involved proteinases; and with the exception of papain and subtilisin, all were related to trypsin, although the corresponding inhibitors are non-homologous. The basis of this work was the definition and comparison of a series of patches on the surfaces of proteins.

**4.2 Definition of a Surface Patch**

Surface residues of a protein were defined as those residues with a relative accessible surface area of >5% (see Chapter 2, section 2.4.1). A patch was defined as a central surface accessible residue and \( n \) nearest surface accessible neighbours, as defined by \( C_\alpha \) positions, where \( n \) was taken as the number of residues observed in the homo-dimer interface (defined in Chapter 2, section 2.4.1). To avoid patches that included residues from opposite 'sides' of a protein, which could have resulted in patches forming rings around the surface, or patches sampling through the centre of a protein, vector constraints were applied. A vector (termed a solvent vector) for each surface residue was defined such that the vector pointed into the solvent. This vector was defined by taking a single central \( C_\alpha \) atom of a surface residue and calculating the centre of gravity of the nearest 10 \( C_\alpha \) atom neighbours. The inverse of the vector from the \( C_\alpha \) of the central surface residue to this centre of gravity was then calculated (Figure 4.1). In this way each surface residue was assigned a vector which pointed into the solvent. A patch was only defined when the angle between the solvent vectors of the central residue and each of the \( n \) nearest surface accessible neighbours was \( \leq 110^\circ \). This procedure defined a number of contiguous, overlapping patches of surface accessible residues. For example, in the HIV protease structure (5hvp) there were 81 such patches defined. The analysis was conducted upon the total number of different patches generated from the surfaces of each subunit. Each patch was analysed for 6 different parameters.
4.3 Calculation of Patch Parameters

4.3.1 Solvation Potentials

Solvation potentials are knowledge-based potentials which give a measure of the propensity of each amino acid type for a certain degree of solvation, approximated by the residue solvent accessible surface area (ASA). These potentials were generated from structures in the PDB, as part of a threading algorithm for the prediction of protein tertiary structure from sequence, using the strategy of recognising known folds (Jones et al., 1992). For each residue type there was a solvation potential (SP) associated with a given ASA, where the ASA was divided into 10% ranges (Figure 4.2). The solvation potentials can be negative or positive. For example in the case of leucine (leu) the solvation potential is large and positive over the ASA range 80-100%, and negative at 0%; hence this residue shows a preference for being buried. Using these values the change in solvation potential between an exposed and buried patch (ΔSP) was calculated for each patch as shown in equation 4.1.
Figure 4.2: Solvation potentials for the 20 amino acids. These solvation potentials have been generated for each 10% of accessible surface area from structures in the PDB as part of a threading algorithm (Jones et al., 1992).
Patch Analysis

\[
\text{Patch } \Delta SP = \frac{\sum_{i=1}^{N_p} \text{SP}(\text{AA})_{\text{ASA}_m} - \text{SP}(\text{AA})_{\text{ASA}_0}}{N_p}
\]  

where \( \text{SP}(\text{AA})_{\text{ASA}_m} \) = Solvation potential of an amino acid residue (AA) with an ASA of \( \text{ASAm} \) in the monomer

\( \text{SP}(\text{AA})_{\text{ASA}_0} \) = Solvation potential of an amino acid residue (AA) with zero accessibility (ASA0)

\( N_p \) = Number of residues in the patch

The \( \Delta SP \) can be negative or positive; the larger and more positive the \( \Delta SP \) value the greater the preference for burial.

4.3.2 Planarity

The planarity of each surface patch was calculated by evaluating the root mean squared (rms) deviation of all the patch atoms from the least squares plane through the atoms, using an algorithm developed by R. Laskowski (see Chapter 2, section 2.5.3).

4.3.3 Residue Interface Propensities

The relative frequency of different amino acid residues in the interfaces of complexes can be used to derive propensities for prediction. Residue interface propensities were calculated for each amino acid as described in Chapter 3 (section 3.4.3; equation 3.2). The natural logarithms \( \ln \) of these propensities were calculated for a data set of 63 protein-protein complexes (listed in Chapter 3, Table 3.1) (Table 4.1, column a). The higher the logarithm of the propensity value the more likely a residue is to occur in a protein-protein interface. These logarithmic propensities were used to calculate a mean propensity value for each surface patch:

\[
\text{Patch Interface Propensity} = \frac{\sum_{i=1}^{N_p} \ln \text{IP}_{\text{AA}}}{N_p}
\]

where \( \text{IP}_{\text{AA}} \) = Interface propensity for an amino acid residue in the patch

\( N_p \) = Number of residues in the patch
Table 4.1: (a) Natural Logarithms ($\ln$) of the interface residue propensities derived from a data set of 63 protein-protein complexes, (b) The Fauchère and Pliska (1983) hydrophobicity scale.

<table>
<thead>
<tr>
<th>Amino Acid</th>
<th>(a) $\ln$ Residue Interface Propensity</th>
<th>(b) Hydrophobicity Scale</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ala</td>
<td>-0.17</td>
<td>0.31</td>
</tr>
<tr>
<td>Arg</td>
<td>0.27</td>
<td>-1.01</td>
</tr>
<tr>
<td>Asn</td>
<td>0.12</td>
<td>-0.60</td>
</tr>
<tr>
<td>Asp</td>
<td>-0.38</td>
<td>0.77</td>
</tr>
<tr>
<td>Cys</td>
<td>0.43</td>
<td>1.54</td>
</tr>
<tr>
<td>Gln</td>
<td>-0.11</td>
<td>-0.22</td>
</tr>
<tr>
<td>Glu</td>
<td>-0.13</td>
<td>-0.64</td>
</tr>
<tr>
<td>Gly</td>
<td>-0.07</td>
<td>0.00</td>
</tr>
<tr>
<td>His</td>
<td>0.41</td>
<td>0.13</td>
</tr>
<tr>
<td>Ile</td>
<td>0.44</td>
<td>1.80</td>
</tr>
<tr>
<td>Leu</td>
<td>0.40</td>
<td>1.70</td>
</tr>
<tr>
<td>Lys</td>
<td>-0.36</td>
<td>-0.99</td>
</tr>
<tr>
<td>Met</td>
<td>0.66</td>
<td>1.23</td>
</tr>
<tr>
<td>Phe</td>
<td>0.82</td>
<td>1.79</td>
</tr>
<tr>
<td>Pro</td>
<td>-0.25</td>
<td>0.72</td>
</tr>
<tr>
<td>Ser</td>
<td>-0.33</td>
<td>-0.04</td>
</tr>
<tr>
<td>Thr</td>
<td>-0.18</td>
<td>0.26</td>
</tr>
<tr>
<td>Trp</td>
<td>0.83</td>
<td>2.25</td>
</tr>
<tr>
<td>Tyr</td>
<td>0.66</td>
<td>0.96</td>
</tr>
<tr>
<td>Val</td>
<td>0.27</td>
<td>1.22</td>
</tr>
</tbody>
</table>

4.3.4 Hydrophobicity

To compare the relative hydrophobicities of surface patches the experimentally derived scale of Fauchère and Pliska (1983) (Table 4.2, column b) was used. This scale is based on the measurement of octanol/water distribution coefficients of the 20 amino acids, and extends the work of Yunger and Cramer (1981) who measured these coefficients for 12 residues. This scale was chosen as it provides hydrophobicity values for all 20 amino acid residues and is the experimental scale used most often (Cornette et al., 1987). The scale was used to assign a mean hydrophobicity value to each surface patch:

$$\text{Patch Hydrophobicity} = \frac{\sum_{i=1}^{N_p} (HV_{AA})}{N_p} \quad \text{[4.3]}$$

where $HV_{AA} =$ value assigned to an amino acid residue in the Fauchère and Pliska (1983) hydrophobicity scale

$N_p =$ number of residues in the patch
4.3.5 Protrusion Index

A protrusion index gives an absolute value for the extent to which a residue protrudes from the surface of a protein, and this measure has been used to locate ‘antigenic’ peptides in proteins (Thornton et al., 1986). The indices are calculated on the assumption that the shape of the protein can be modelled by an ellipsoid (Taylor et al., 1983). An algorithm, implemented by S. Hubbard, was used to calculate the ellipsoid that had the same moments of inertia as the protein, known as the equi-momentell ellipsoid. In the algorithm the ratios of the principal axes of the equi-momentell ellipsoid and their directions are defined. The absolute size of the ellipsoid was then assigned arbitrarily to include a specified percentage of residues. For example a 90% ellipsoid was defined to include 90% of the protein residues within its bounds. Using a series of percentage ellipsoids in 10% steps, residues in a protein were assigned a protrusion index (PI), which specifies the percentage ellipsoid at which the residue first falls outside the area enclosed by the ellipsoid (Thornton et al., 1986). Hence a residue that becomes external at the 80% ellipsoid was assigned a PI of 8, and a residue that becomes external at the 10% ellipsoid was assigned a PI of 1. In the current work, a mean PI was calculated for each surface patch:

\[
\text{Patch PI} = \frac{\sum_{i=1}^{N_p} P_{IA \alpha}}{N_p}
\]  

[4.4]

where \( P_{IA \alpha} \) = Protrusion index of an amino acid residue calculated using the \( C_\alpha \) atom co-ordinates

\( N_p \) = Number of residues in the patch

4.3.6 Accessible Surface Area

The absolute accessible surface area (ASA) of each residue in a patch was calculated using the program ACCESS (Hubbard, 1990), as described in Chapter 2, section 2.4. A mean ASA for each patch was calculated as shown in equation 4.5.
Patch Analysis

\[
\text{Patch ASA (Å}^2\text{)} = \frac{\sum_{i=1}^{N_p} \text{ASA}_{AA}}{N_p} \tag{4.5}
\]

where \( \text{ASA}_{AA} \) = Accessible surface area of an amino acid residue in the patch
\( N_p \) = Number of residues in the patch

In summary, 6 parameters (solvation potential, planarity, residue interface propensity, hydrophobicity, protrusion and accessible surface area) were evaluated for each patch defined on the surface of a protein. The same 6 parameters were also calculated for the experimentally observed interface region, defined at the residue level (see Chapter 2 section 2.4.1).

4.4 Evaluation of Patch Parameters

For each parameter the distribution of values for all the patches on one protein, including the observed interface patch, was plotted (for example HIV protease (5hvp) in Figure 4.3). A ranking of the observed interface, relative to all other possible patches, was then calculated on a scale of 1–10. A rank of 1 indicates that the observed interface scores in the highest range of a parameter distribution of all surface patches. A rank of 10 indicates that the observed interface scores in the lowest range of the parameter distribution for all surface patches. With this approach it becomes possible to plot the rankings of all the observed patches for each protein as a histogram (Figures 4.4 and 4.5), to assess which parameters best differentiate the interface region.

Rankings of observed interface patches relative to all other surface patches were calculated for 2 data sets
- Protomers from 29 homo-dimers (Figure 4.4)
- Enzyme components of 10 enzyme-inhibitor complexes (Figure 4.5)
Figure 4.3: Distribution of parameters for all patches in HIV protease (5hvp). Distributions are shown for: (a) Solvation potentials, (b) Residue interface propensities, (c) Hydrophobicity (based on the scale of Fauchère and Pliska (1983)), (d) RMS deviation of atoms from the least squares plane through the interface atoms, (e) Protrusion index, (f) Accessible surface area.

On each graph all the surface patches are represented by the grey bars and the observed interface patch is represented by the black bars. Relative rankings (on a scale of 1-10) were calculated from each distribution and are indicated on each graph.
Figure 4.4: Patch analysis distributions for 29 homo-dimers; rank ordering (on a scale of 1-10 (see Figure 4.3)) of observed interface patches relative to other patches on the surface of the protein. The 29 observations (one for each homo-dimer) were combined for each parameter separately. The distributions shown are: (a) Solvation potentials, (b) Residue interface propensities, (c) Hydrophobicity (based on the scale of Fauchère and Pliska (1983)), (d) Rms deviation of atoms from the least squares plane through the interface atoms, (e) Protrusion index, (f) Accessible surface area.
Figure 4.5: Patch analysis distributions for 10 enzymes; rank ordering (on a scale of 1-10 (see Figure 4.3)) of observed interface patches relative to other patches on the surface of the protein. The 10 observations (one for each enzyme) were combined for each parameter separately. The distributions shown are: (a) Solvation potentials, (b) Residue interface propensities, (c) Hydrophobicity (based on the scale of Fauchère and Pliska (1983)) (d) RMS deviation of atoms from the least squares plane through the interface atoms, (e) Protrusion index, (f) Solvent accessible surface area.
4.4.1 Homo-Dimers

It was observed that no single parameter absolutely differentiated the interfaces from all other surface patches. For example with the planarity parameter (Figure 4.4d) 58% of the interfaces were ranked as 10 (i.e. they were the most planar patches), whilst others were relatively non-planar. The most striking correlation was for the accessible surface area (Figure 4.4f). This observation in part reflects the fact that the side-chains from one protomer extend from the surface to interact with the other half of the dimer. Therefore, in isolation, they become highly accessible, and such a strong signal would not be expected for the structure of an isolated molecule prior to complexation, as the side-chains probably change their conformation and 'stretch out' to form the complex. As expected from the accessibility data, the interfaces tended to protrude from the surface (Figure 4.3e) although the signal was weaker, perhaps as a consequence of the requirements for planarity. The solvation potentials (Figure 4.4a), showed some discriminating power, suggesting the parameter carries relevant information. The residue interface propensities (Figure 4.4b) were calculated by 'jack knifing' the propensities. This involved calculating the interface propensities with each of the dimers removed in turn, and then using these new propensities to find the distribution in the one dimer which was excluded from the propensity calculation. The propensities did show some discriminating power, although the trend was not as marked as for some of the other parameters. The weakest correlation was seen for the hydrophobicity measure, derived from the Fauchère and Pliska (1983) hydrophobicity scale (Figure 4.4c). The use of this scale revealed that the observed interface patches tended to be more hydrophobic, but the distribution was widespread with some observed interfaces amongst the least hydrophobic patches.

4.4.2 Enzymes

These complexes represent the combination of large (enzyme) and small (inhibitor) components, and the distinguishing features of the observed interfaces were different to those observed for the homo-dimers. The solvation potential (Figure 4.5a) and the residue interface propensity (Figure 4.5b) (which were 'jack knifed') showed some power in discriminating the interface from the remainder of the protein surface. The hydrophobicity scales (Figures 4.5c) and the planarity parameter (Figure 4.5d) also had a limited power, with the interface showing a tendency to be amongst the most hydrophobic surface patches and amongst the most planar.

For the enzyme data set the protrusion and the accessible surface area parameters (Figure 4.5e and 4.5f) showed inverse distributions to those observed in the homo-dimers. The enzyme
interfaces were amongst the least protruding patches and were patches with low accessibility. This was in contrast to the homo-dimers (Figures 4.4e and 4.4f) in which the interface patches were shown to be amongst the most protruding and the most accessible. These differences can be explained by the relative sizes of the complex components. The enzyme-inhibitor complexes represent large components (enzymes) interacting with small components (inhibitors). Inhibitors often pack tightly into small clefts or pockets on the enzyme surfaces e.g. the complex of human stefin B and papain (Stubbs et al., 1990). It has also been observed that, in general, ligands bind to enzymes in the largest clefts on the enzymes surface (R. Laskowski, manuscript in preparation). Hence the enzyme interfaces tend to be those surface patches that are invaginated or folded into clefts, meaning the residues involved will be amongst the least accessible and the least protruding.

None of the distributions were definitive, in that the observed interface regions were never always at one extreme, but all showed trends for the observed interface to be distinguished from other surface patches. This type of comparative analysis, including many different parameters rather than a single value, can potentially be used to predict the location of putative interface sites on protein surfaces (see Chapter 5).

4.5 Sampling of the Observed Interface

The definition of surface patches was at the residue level and based on Cα atom distances which resulted in patches that were circular and contiguous. As a consequence it was unlikely that for a single protomer any one surface patch would exactly match the residues in the observed interface. To analyse how the patches sampled the residues in the observed interface a percentage overlap was calculated:

\[
\text{Percentage Overlap} = \frac{\text{Number of residues in the } \text{observed interface patch}}{\text{Number of residues in the calculated patch}} \times 100
\]

where

- \( \text{NrO} = \text{Number of residues in the observed interface patch} \)
- \( \text{NrC} = \text{Number of residues in the calculated patch} \)
Protein-Protein Interactions: Chapter 4  Patch Analysis

Figure 4.6: Distributions of percentage overlaps (defined in equation 4.6) in 29 homo-dimers. The maximum percentage overlap is recorded on each graph. The homo-dimers are recorded in increasing size of percentage interface accessible surface area (%ASAi, see Chapter 2, section 2.5.1).
The distribution of overlap percentages for patches in the 29 homo-dimers are shown in Figure 4.6. In addition the maximum percentage overlap in each protomer is indicated on each graph in the figure. These data were generated for each protein with a patch size $n$, equal to the size of the observed interface patch.

It was observed that the maximum overlaps ranged from 54—87%. The differences observed between the percentage overlaps in different protomers can be attributed to the size and shape of the observed interface. In terms of size the ASA of the observed interface compared to the ASA of the protomer as whole (percentage interface ASA ($\%\text{ASA}_i$) see Chapter 2, section 2.5.1) is an important factor. For example, an interface that constitutes more that 25% of the ASA of the protomer will be sampled more often by patches defined on the surface than an interface that contributes only 5% of the ASA. In Figure 4.6 the protomers have been ordered in increasing size of $\%\text{ASA}_i$. The majority of the protomers showed a similar pattern with a high percentage of patches sampling between 1—5% of the observed interface, and a relatively small number of patches sampling a much higher percentage of the observed interface (upto 70% and 80% in some proteins). With a few exceptions these were the protomers in which the $\%\text{ASA}_i$ of the interface was $\leq 20\%$. When the $\%\text{ASA}_i$ was $>20\%$ the distributions of percentage overlaps changed. In these structures (e.g. 2wrp, 5hvp and 1utg), there were relatively few patches which sampled between 1—5% of the observed interface and larger numbers of patches which sampled higher percentages.

### 4.6 Conclusions

The definition of patches of surface residues was based on C$_\alpha$ atom co-ordinates. The definition resulted in approximately circular patches which were contiguous on the surface of the protein. The patches sampled the observed interface patches with varying degrees of success; the maximum percentage overlaps (calculated as overlap/observed interface patch) ranged from 54—87%. These differences observed between proteins were related to the size and the shape of the observed interface patch and the calculated patch.

Using the definition of surface patches, 6 parameters (solvation potential, residue interface propensity, hydrophobicity, planarity, protrusion index and accessible surface area) were
identified to differentiate an observed interface site from all other patches defined on the surface of a protein. In both homo-dimers and enzymes, the interfaces were observed to be, in general, the most hydrophobic and planar patches, and those patches with high solvation potentials and high residue interface propensities. However some parameters had different relative values depending on the specific protein-protein interactions. The homo-dimer interfaces were found to be amongst the most accessible and the most protruding patches, whilst the enzyme interfaces were amongst the least accessible and the least protruding of all the surface patches. These differentiating characteristics can been used as the basis of an algorithm for the prediction of putative interaction sites on the surfaces of proteins; this is discussed in Chapter 5.
CHAPTER 5

Interface Prediction by Patch Analysis

5.1 Introduction
The reliable prediction of protein-protein interaction sites is an important goal in the field of molecular recognition. It is of direct relevance to the design of drugs for blocking or modifying protein-protein interactions. Predictions can be divided into two main areas. The first area is the docking of two proteins of known structure. The second area is the identification of putative interaction sites upon the surface of an isolated protein, known to be involved in protein-protein interactions, but where the structure of the complex is not known.

The docking of one protein with another is a complex process that involves two basic types of complementarity; geometric and chemical. The majority of docking algorithms are based upon methods of evaluating and matching the geometric shape of the potential interacting surfaces. This type of matching requires a molecular surface representation and in many cases this is achieved by depicting the molecular envelope as a Connolly surface (Connolly 1983a; Connolly 1983b). This method represents a molecular surface by a dense population of dots. Many of these can be excluded, leaving only the critical points, which describe local knobs and holes on the protein surface. Connolly (1986) developed this representation of a molecular surface into a docking algorithm, matching quartets of knobs and holes on each surface, and then screening for overlaps between the two molecules. This method proved successful for the docking of α and β subunits of haemoglobin, but failed for the trypsin-trypsin inhibitor complex. Recently Norel et al. (1994) improved this method, reducing the complexity of the docking method, and successfully docked 14 out of 16 protein complexes. Many other algorithms involve similar matching of knobs and holes (Greer and Bush 1978; Kuntz et al., 1982; Lee and Rose, 1985; Jaing and Kim, 1991), and many more match geometric criteria using similar simplified interpretations of the molecular surface (e.g. Wodak and Janin, 1978; Helmer-Citterich and Tramontano, 1994).

Whilst geometric complementarity has been the main focus for automatic predictive algorithms, other computational tools have also been developed, which give an insight into the chemical complementarity of interacting surfaces. Amongst the first to examine electrostatic
complementarity was Salemme (1976), who matched molecular surface electric charges to propose a structure for the redox complex of cytochromes c and b₅. Warwicker (1989) conducted a similar matching technique, calculating and graphically mapping the electrostatic potential energy between molecular surfaces of proteins brought together in different orientations. This technique proved successful for docking the trypsin-trypsin inhibitor complex and the Fab-HyHEL5 antibody-antigen complex.

Hydrogen bonding has also been used for interaction site prediction. HSITE (Danziger and Dean 1989a; Danziger and Dean 1989b) creates a map of hydrogen bonding regions of protein surfaces. The algorithm was designed as an interactive tool to build a knowledge base on protein surfaces, for the subsequent automatic construction of novel ligands to fit specific binding sites. Hydrophobicity has also been used as the basis for prediction. Zielenkiewicz and Rabczenko (1984) developed a method identifying complementary surfaces based on maximising the number of atom coincident contacts. Coincidence was established by dividing atoms into hydrophilic and hydrophobic categories, and a coincident contact defined as one in which the atoms were in the same category. Coincidence was also measured in a second version of the program by the capability of interacting atoms to form hydrogen bonds (Zielenkiewicz and Rabczenko, 1985). The algorithm was used successfully to find the dimer and hexamer complementary surfaces of insulin. In a non-automated method Korn and Burnett (1991) used the presence of hydrophobic areas on the surface of a protomer from a dimeric protein (inorganic pyrophosphatase), to predict the position of the dimer interface. Young et al. (1994) took this approach further and produced an automated predictive algorithm based on the analysis of the hydrophobicity of residue clusters on protein surfaces. This algorithm has successfully been used to predict sites of protein-protein interactions in a number of enzymes, antibody fragments, and a number of other proteins with associated ligands.

Geometric and chemical complementarity are of parallel importance in protein-protein interactions, and some docking algorithms have combined the two. For example, Walls and Sternberg (1992) developed an algorithm to dock molecules based on surface complementarity, which involved both geometric matching and a simple electrostatic model to eliminate unfavourable interactions. More recently, algorithms initially based upon geometric matching have been enhanced to include chemical complementarity. The docking program DOCK (Kuntz et al., 1982), which places ligands in to receptors using local geometric features, has been developed to include a chemical filter. Interacting atoms are assigned to a chemical category and
the docking algorithm restricted to chemically complementary matches (Shoichet and Kuntz, 1993). The addition of chemical matching enhanced the performance of the DOCK algorithm, producing fewer solutions in a faster time, and ranking known inhibitors higher than geometric matching alone. A similar enhancement was made by Vakser and Aflalo (1994) to a geometric docking algorithm based on assessing molecular surface contact and penetration between surface atoms projected onto a 3-dimensional grid system (Katchalski-Katzir et al., 1992). The enhancement involved the classification of surface atoms into hydrophobic and 'non-hydrophobic', and the geometric matching of the molecular surfaces represented by hydrophobic atom groups only.

Most of the protein-protein recognition algorithms, discussed so far, involve the representation of protein molecules as rigid bodies, which undergo little or no change in conformation on complexation. It is not clear to what extent proteins change their conformation on forming a complex (e.g. Huber, 1979), and currently there are few direct structural data for molecules before and after complexation. However, there is evidence that small changes in conformation do occur on complexation of some molecules (e.g. Wilson and Stanfield, 1993). Structural comparisons of antibody-complexed and native lysozyme show that conformational changes of main-chain atoms at the interface regions are of the order of 0.5Å (Davies and Padlan, 1992). In the light of such data new algorithms have been designed to account for conformational changes on complexation (e.g. Jaing and Kim, 1991).

The problem of predicting putative interaction sites on proteins is complex, and one that has been approached in many different ways over the past 20 years. Some algorithms are automatic, and produce a list of possible docking solutions between two proteins of known conformation, which then require additional filtering. Other algorithms, such as HSITE (Danziger and Dean 1989a; Danziger and Dean 1989b), and the hydrophobic evaluation of surface clusters (Young et al., 1994), are designed to be used as tools to identify putative interaction sites when analysing the molecular surface of an isolated protein. The patch analysis of protein surfaces, presented in Chapter 4, is developed in the current chapter to be used as a computational tool for the prediction of putative recognition sites from isolated proteins.

The analysis described in Chapter 4 identified 6 parameters (solvation potential, residue interface propensity, hydrophobicity, planarity, protrusion and accessible surface area) which provided information differentiating protein-protein interaction sites from the remainder of the
protein's surface. Here the parameters have been incorporated into a computational tool and used to predict the interface sites on the surfaces of the protomers from 29 homo-dimers (see Chapter 4, section 4.1). This tool could be used to identify putative interaction sites for a protein that is known to be involved in protein-protein interactions, whose structure has been determined, but where the structure of the complex is not known.

5.2 The Surface Residue Patch

A surface patch was defined in Chapter 4 (section 4.2) as a central surface accessible residue and \( n \) nearest surface accessible neighbours, as defined by \( C_\alpha \) positions, where \( n \) was taken as the number of residues in the observed interface. For the purposes of prediction the same definition applied, except \( n \), the size of the patch in terms of the number of residues, was defined as a variable. The choice of the size of the patch was crucial to the prediction. It was observed (see Chapter 2, section 2.5.1) that the size of an interface region was correlated to the size of the protomer. For prediction, this correlation was calculated in terms of the number of residues in the protomer \( (NR_p) \) and the number of residues in the observed interface region \( (NR_i) \) (Figure 5.1). A regression line, of the form \( y=\alpha x^b \), fitted to the data of 29 protomers gave the equation:

\[
NR_i = 1.92 NR_p^{0.56}
\]  

[5.1]

with a correlation coefficient \( (r) \) of 0.71. In the predictive tool the size of the patch can be set to an arbitrary value, or a size can be estimated using equation 5.1. For evaluation purposes, on the homo-dimer data set, equation 5.1 was used to calculate a patch size.

![Figure 5.1: The relationship between the number of residues in the interface and the number of residues in the protomer. A regression line of \( y = 1.92x^{0.56} \) has been fitted.](image)
5.3 The Prediction Algorithm

For each isolated protein in the data set of 29 homo-dimers (see Chapter 4, section 4.1) all surface patches were generated (as described in Chapter 4, section 4.2), and the 6 parameters (solvation potential, residue interface propensity, hydrophobicity, planarity, protrusion and ASA) calculated for each patch. The solvation potential, residue interface propensity, hydrophobicity, protrusion, and ASA parameters were expected to score highly for the interface sites, whilst the planarity score was expected to be low for such sites as they were usually planar (see Chapter 4, section 4.4.1). The aim of the prediction algorithm, named PATCH, was to identify the surface patch(es) that would most likely form a protein-protein interface.

The predictive procedure involved 3 stages :-

- Scoring of patches for individual parameters
- Calculation of a combined score from multiple parameters
- Calculation of the percentage frequency of surface residues in best patches

5.3.1 Individual Parameter Score

For an individual parameter there was a range of values over all surface patches, and hence the patches could be scored by their relative position to all other surface patches in a single protomer. For each parameter the range was calculated for a given protein, and then divided into 100 separate ranges. Thus each patch parameter value was normalised and assigned a score of 1-100. The lowest parameter value was assigned a score of 1, and the highest parameter value a score of 100. Thus each patch had 6 individual parameter scores assigned; for example a patch could have a score of 60 for solvation potential, 20 for residue interface propensity, 80 for hydrophobicity, 1 for planarity, 100 for protrusion and 90 for ASA. This approach weights all 6 parameters equally, and is relative, rather than absolute.

5.3.2 Combined Parameter Score

A combined score was then calculated which incorporated the individual scores of a patch for all 6 parameters. The combined score gave a probability (on a scale of 1-100) of any one patch (Pj) forming protein-protein interactions, and was calculated as:
Combined Score $P_j = \frac{S_{\text{sp}} + S_{\text{rp}} + S_{\text{hy}} + S_{\text{pi}} + S_{\text{asa}} + (100 - S_{\text{pl}})}{N_{\text{par}}}$ \[5.2\]

where $S_{\text{sp}}$ = score of patch $P_j$ in the solvation potential distribution
$S_{\text{rp}}$ = score of patch $P_j$ in the residue interface propensity distribution
$S_{\text{hy}}$ = score of patch $P_j$ in the hydrophobicity distribution
$S_{\text{pi}}$ = score of patch $P_j$ in protrusion index distribution
$S_{\text{asa}}$ = score of the patch $P_j$ in the accessible surface area distribution
$S_{\text{pl}}$ = score of patch $P_j$ in the planarity distribution
$N_{\text{par}}$ = number of parameters

Thus the combined score was on a scale of 1–100, where 1 denoted a very low (poor) probability of forming a putative interaction site, and 100 a very high (good) probability of forming a putative interaction site. Equation 5.2 was the default definition of the combined score, but the definition was a user defined variable and alternative definitions could be used (see section 5.3.4). In equation 5.2 the score for the planarity parameter was inverted, as interfaces have been observed to be the most planar patches (see Chapter 4, section 4.4.1) and therefore the prediction algorithm scored a planar patch highly.

5.3.3 Residue Frequencies and Patch Profiles

There are 2 ways in which the patches with the highest combined scores (denoted as best patches) can be selected.

- The algorithm places each patch in descending order of its combined score. The first $n$ number of patches can be selected as best patches.
- A cut-off can be defined such that a percentage of patches with the highest combined scores are selected as best patches. For example, if a cut-off of 10% is defined then those patches that have a combined score in the top 10% of the distribution will be selected as best patches.

For the purposes of evaluation the first method was used; the 3 patches with the highest combined scores were selected as best patches. The cut-off method of selection is used in Chapter 6, where the applications of the prediction algorithm are described.
By definition surface patches overlap and residues could occur in the best patches more than once. Hence the frequency with which surface residues occurred in the best patches was calculated in the third and final stage of the algorithm. A graph drawing program, ROMPLOT (Laskowski, 1990), was then used to produce a postscript™ file of patch profiles, which recorded the value of each parameter, and the combined score, against the number of the residue at the centre of each patch (Figure 5.2 a-g). The frequency of each residue in the best patches was also recorded as a histogram (Figure 5.2h).

The PATCH predictive algorithm provides information on the properties of a series of overlapping patches on the surface of a protein. It can be used to extract a prediction of the residues with a high probability of being involved in protein-protein interactions. In addition the algorithm provides information regarding individual properties of each surface patch.

5.3.4 Input Variables

The algorithm was developed as an interactive tool, and a number of user defined variables have been included. The inclusion of variables allows the user to determine, to some degree, how putative interaction sites are selected. The variables are:

- **Patch size**: This can be selected arbitrarily or calculated using equation 5.1 (see section 5.2)

- **% Cut-off when selecting best patches**: This can be set at any level from 1–100%. The higher the value, the greater the number of best patches selected.

- **Definition of combined score**: By default all 6 parameters are used to calculate the combined score. If, however, the user decides that one or more individual parameters are not important for an individual protein, it is possible to eliminate parameters from the combined score calculation.

- **Definition of how the individual parameters were scored**: From the discussion in Chapter 4 (section 4.4) it is known that the distributions of parameters vary between different types of complex. For example the interfaces in the homo-dimers tend to be accessible and protruding, whilst the enzyme interfaces show inverse characteristics. This variation is incorporated into the PATCH algorithm. For example, for the homo-dimers the recommended combined score definition is indicated in equation 5.2. For the enzymes,
in which the interfaces are not protruding or accessible, the preferred combined score definition is:

\[
\text{Combined Score } P_j = \frac{S_{sp} + S_{rp} + S_{oy} + (100 - S_{pi}) + (100 - S_{asa}) + (100 - S_{pl})}{N_{par}}
\]  \[5.3\]

(For description of parameters see equation 5.2).

In this definition of the combined score the scores for the protrusion, accessible surface area and planarity were inverted, as enzyme interfaces were expected to score low values for all 3 parameters.

A flow diagram that outlines the main elements of the predictive algorithm, including the input variables is shown in Figure 5.3.

**5.4 Evaluation of the Prediction Algorithm**

The PATCH prediction algorithm was used to predict the putative interface sites on the surface of protomers from 29 homo-dimers.

In the evaluation process two main areas were addressed:

- The sampling of the observed interface by the surface patches
- The effectiveness of the combined score for selecting putative interface sites

**5.4.1 Sampling of the Observed Interface**

The definition of surface patches was at the residue level and based on C\(\alpha\) atom distances. In addition the size of the patches was predicted using the approximate relationship based on the number of residues in the monomer (equation 5.1). As a consequence it was unlikely that for a single protomer any one surface patch would exactly match the residues in the observed interface. The evaluation of the sampling of the observed interface can be divided into 2 main areas, the size of the calculated patches and the overlap achieved with the observed interface.
Figure 5.2: Patch profile for subunit A of Mannose Binding Protein (1msb). The patch size was 28 residues and the cut-off 15%. The best patches selected from the combined score profile are indicated by vertical dashed lines. The frequency of surface residues in each best patch is indicated in the histogram (h), and the ‘*’ symbols mark the residues in the observed interface.
Protein-Protein Interactions: Chapter 5

Interface Prediction

Figure 5.3: Flow diagram summarising the events in the PATCH algorithm, designed to predict putative interaction sites on the surfaces of isolated proteins.
The accuracy of the prediction of patch size (using equation 5.1) was evaluated by calculating a percentage difference between the predicted patch size and the size of the observed interface:

\[
\text{Percentage Size Difference} = \frac{\text{NrC} - \text{NrO}}{\text{NrO}} \times 100
\]

where \( \text{NrO} \) = Number of residues in the observed patch

\( \text{NrC} \) = Number of residues in the calculated patch

The use of equation 5.1 results in protomers having patch sizes ranging from 47% smaller to 81% larger than the observed interface size (Figure 5.4). The size of the patch used in the prediction will obviously influence how the observed interface is sampled. A patch that is defined to be significantly larger than the observed interface will be capable of sampling more of the observed interface than a patch defined to be significantly smaller. This evaluation revealed how the approximate nature of the relationship between the number of residues in the protomer and those in the interface (equation 5.1) significantly affected the predictions. This inaccuracy was taken into account when the overlap between the best patches and the observed interface was evaluated, by using a relative overlap value (see section 5.4.2).

**Figure 5.4:** Percentage differences in the predicted size of the surface patches and the size of the observed interface, in terms of the number of residues in each (see equation 5.4), for the data set of 29 homo-dimers.
Overlap

To analyse how the patches sampled the residues in the observed interface, 2 percentage values were calculated:

\[
P_1 = \frac{\text{overlap/observed patch}}{\text{NrO}} \times 100 \tag{5.5}
\]

\[
P_2 = \frac{\text{overlap/calculated patch}}{\text{NrC}} \times 100 \tag{5.6}
\]

where \(\text{NrO} = \) Number of residues in the observed patch

\(\text{NrC} = \) Number of residues in the calculated patch

Each overlap value scored the number of residues that occur in both the calculated patch and the observed interface. In P1 this overlap was calculated as a percentage of all the residues in the observed interface. In P2 this overlap was calculated as a percentage of all the residues in the calculated patch. These percentage overlaps (P1 and P2) were equivalent to the measures M1 and M2 used by Young et al. (1994) in the evaluation of their automated predictive algorithm.

The distributions of P1 values (and the maximum P1 (Max. P1) and the maximum P2 (Max. P2) values) for patches in the 29 homo-dimers are shown in Figure 5.5. The distribution observed for 1pyp is typical, with a high percentage of patches sampling between 1–5% of the observed interface, and a relatively small number of patches sampling a much higher percentage of the observed interface (≥ 50%). However there are a small number of protomers that exhibit very different distributions, such as 2wrp. In this protomer every patch samples some of the interface. This results from the large size and convoluted shape of the interface, due to the large degree of inter-digitation occurring between the 2 subunits.

The Max. P1 values ranged from 34–100%, and the Max. P2 values ranged from 47–89%. These values indicate the difference between the predicted size of the patch and the observed interface size. If the predicted size was the same as the observed interface size then Max. P1 = Max. P2. This was true of only one protomer, 2gn5. Where Max. P1 > Max. P2 the predicted patch size was greater than the size of the observed interface. This was true of the only protomer to have a Max. P1 of 100%, 1msb. A value of 100% for Max. P1 indicated that all the residues in the observed interface were included in one single patch defined on the surface of the
protomer. However the Max. P2 value for 1msb was only 68%, indicating that this patch contained additional residues that were not present in the observed interface.

The definition of patches at the C\textsubscript{\alpha} level and the approximate prediction of patch size meant that the observed interface was never sampled accurately in any protomer; i.e. Max. P1 and Max. P2 were never both 100%. Clearly this will influence the quality of the predictions, and was therefore taken into account in assessing the results.

5.4.2 Effectiveness of the Combined Score

To evaluate the effectiveness of the combined scoring system, 2 values were calculated for each of the top 3 patches (i.e. the 3 patches with the highest combined score) in each protomer. The first was the P1 value calculated as indicated in equation 5.5 (Table 5.1, column d). The second was the relative overlap value (Table 5.1, column f) defined as:

$$\text{Relative Overlap} = \frac{P1}{\text{Max. P1}} \times 100$$ [5.7]

where 

P1 = defined in equation 5.5

Max. P1 = the patch with the highest percentage overlap with the observed interface (Table 5.1, column e).

By definition the surface patches were overlapping. To evaluate if the patches selected as the top 3 patches (on the basis of their combined scores) overlapped (e.g. see Figure 5.6), an overlap value between each pair of the top 3 patches in each protomer was calculated. If the overlap between any 2 patches in a set of 3 was $\geq 50\%$ then the 2 patches were scored as 1 patch. The number of different patches (i.e. those that overlapped by $<50\%$) in each set of 3 top patches were recorded (Table 5.1, column g).

In addition the rank order of the patch with the maximum overlap with the observed interface (i.e. the 'correct' patch, Max P1), was calculated relative to the total number of patches on the surface of each protomer (Table 5.1, column h). A rank order of 1 denoted that the patch with the maximum overlap with the observed interface (the 'correct' patch) had the highest combined score of all surface patches, i.e. the best possible prediction.
Figure 5.5: Distributions of percentage overlap (P1) in 29 homo-dimers. The Max. P1 and Max. P2 values are recorded on each graph. The homo-dimers are recorded in increasing order of the ranking of the ‘correct’ patch (i.e. that which overlaps most with the observed interface) (see Table 5.1)
### Table 5.1: Results of the prediction algorithm (PATCH) for protomers from 29 homo-dimers.

- **Column (b).** Total number of patches on the surface of each protomer.
- **Column (c).** The number of residues in a patch.
- **Column (d).** Overlap value P1 for each of the 3 patches with the highest combined score.
- **Column (e).** Overlap value P1 for the patch with the maximum overlap value with the observed interface.
- **Column (f).** Relative overlap value for each of the 3 patches with the highest combined score.
- **Column (g).** Number of ‘different’ patches that the top 3 patches represent (patches were defined as ‘different’ if they had an overlap of <50%).
- **Column (h).** Rank order of the correct patch i.e. that which overlaps most with the observed interface. A rank order of 1 denotes that the patch with the maximum overlap with the observed interface had the highest combined score of all surface patches. The homo-dimers are recorded in increasing order of this ranking.
- **Those protomers with the PDB Code shaded were predicted incorrectly, based on a relative percentage overlap cut-off of 60%.**

<table>
<thead>
<tr>
<th>(a)</th>
<th>(b)</th>
<th>(c)</th>
<th>(d)</th>
<th>(e)</th>
<th>(f)</th>
<th>(g)</th>
<th>(h)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PDB Code</td>
<td>No. Patches</td>
<td>Patch Size</td>
<td>% Overlap P1 of top 3 patches</td>
<td>Max. P1 (%)</td>
<td>Relative % Overlap of top 3 patches</td>
<td>No. Diff Patches</td>
<td>Rank Order Max P1</td>
</tr>
<tr>
<td>1msb</td>
<td>88</td>
<td>28</td>
<td>100</td>
<td>84</td>
<td>74</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>1sdh</td>
<td>121</td>
<td>32</td>
<td>74</td>
<td>26</td>
<td>70</td>
<td>74</td>
<td>100</td>
</tr>
<tr>
<td>1ypi</td>
<td>174</td>
<td>43</td>
<td>85</td>
<td>52</td>
<td>48</td>
<td>85</td>
<td>100</td>
</tr>
<tr>
<td>2cts</td>
<td>304</td>
<td>59</td>
<td>47</td>
<td>46</td>
<td>46</td>
<td>47</td>
<td>100</td>
</tr>
<tr>
<td>2stl</td>
<td>234</td>
<td>46</td>
<td>93</td>
<td>90</td>
<td>55</td>
<td>92</td>
<td>100</td>
</tr>
<tr>
<td>3grs</td>
<td>356</td>
<td>63</td>
<td>57</td>
<td>58</td>
<td>48</td>
<td>58</td>
<td>98</td>
</tr>
<tr>
<td>5adh</td>
<td>257</td>
<td>54</td>
<td>61</td>
<td>78</td>
<td>61</td>
<td>78</td>
<td>78</td>
</tr>
<tr>
<td>1ppp</td>
<td>219</td>
<td>47</td>
<td>46</td>
<td>46</td>
<td>85</td>
<td>85</td>
<td>55</td>
</tr>
<tr>
<td>1utg</td>
<td>67</td>
<td>21</td>
<td>41</td>
<td>38</td>
<td>44</td>
<td>44</td>
<td>94</td>
</tr>
<tr>
<td>2wp</td>
<td>100</td>
<td>27</td>
<td>41</td>
<td>37</td>
<td>47</td>
<td>47</td>
<td>88</td>
</tr>
<tr>
<td>1cdt</td>
<td>57</td>
<td>19</td>
<td>60</td>
<td>60</td>
<td>67</td>
<td>93</td>
<td>64</td>
</tr>
<tr>
<td>4mdh</td>
<td>235</td>
<td>51</td>
<td>59</td>
<td>51</td>
<td>51</td>
<td>62</td>
<td>96</td>
</tr>
<tr>
<td>5hvp</td>
<td>81</td>
<td>25</td>
<td>54</td>
<td>54</td>
<td>54</td>
<td>56</td>
<td>96</td>
</tr>
<tr>
<td>2vc</td>
<td>177</td>
<td>43</td>
<td>70</td>
<td>68</td>
<td>70</td>
<td>87</td>
<td>81</td>
</tr>
<tr>
<td>1ph</td>
<td>276</td>
<td>56</td>
<td>35</td>
<td>35</td>
<td>59</td>
<td>79</td>
<td>44</td>
</tr>
<tr>
<td>2gn5</td>
<td>78</td>
<td>24</td>
<td>62</td>
<td>71</td>
<td>58</td>
<td>83</td>
<td>75</td>
</tr>
<tr>
<td>1ppp2</td>
<td>105</td>
<td>29</td>
<td>32</td>
<td>27</td>
<td>16</td>
<td>60</td>
<td>56</td>
</tr>
<tr>
<td>2sd</td>
<td>110</td>
<td>32</td>
<td>60</td>
<td>50</td>
<td>20</td>
<td>85</td>
<td>70</td>
</tr>
<tr>
<td>3gap</td>
<td>178</td>
<td>39</td>
<td>13</td>
<td>13</td>
<td>10</td>
<td>58</td>
<td>23</td>
</tr>
<tr>
<td>2ssi</td>
<td>93</td>
<td>27</td>
<td>19</td>
<td>19</td>
<td>4</td>
<td>73</td>
<td>26</td>
</tr>
<tr>
<td>2sc</td>
<td>193</td>
<td>58</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>58</td>
<td>0</td>
</tr>
<tr>
<td>2sc</td>
<td>193</td>
<td>58</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>58</td>
<td>0</td>
</tr>
<tr>
<td>3icd</td>
<td>298</td>
<td>58</td>
<td>56</td>
<td>47</td>
<td>52</td>
<td>59</td>
<td>95</td>
</tr>
<tr>
<td>11l8</td>
<td>65</td>
<td>21</td>
<td>13</td>
<td>9</td>
<td>39</td>
<td>61</td>
<td>21</td>
</tr>
<tr>
<td>3sdp</td>
<td>151</td>
<td>37</td>
<td>18</td>
<td>15</td>
<td>0</td>
<td>81</td>
<td>23</td>
</tr>
<tr>
<td>2rux</td>
<td>306</td>
<td>63</td>
<td>47</td>
<td>34</td>
<td>13</td>
<td>58</td>
<td>29</td>
</tr>
<tr>
<td>3enl</td>
<td>273</td>
<td>59</td>
<td>0</td>
<td>0</td>
<td>16</td>
<td>82</td>
<td>0</td>
</tr>
<tr>
<td>1vsg</td>
<td>282</td>
<td>54</td>
<td>33</td>
<td>30</td>
<td>31</td>
<td>34</td>
<td>97</td>
</tr>
<tr>
<td>3aat</td>
<td>286</td>
<td>56</td>
<td>8</td>
<td>8</td>
<td>7</td>
<td>45</td>
<td>18</td>
</tr>
</tbody>
</table>

**Note:** The homo-dimers are recorded in increasing order of this ranking.
Figure 5.6: Sampling the observed interface. The dark grey circle represents the observed interface on the protein (represented by the light grey circle). The open circles represent the calculated patches. The patches with the highest combined ranking could be: (a) 3 patches sampling the observed interface but overlapping each other by <50%. In this case all 3 patches would be counted. (b) 3 patches, 2 of which overlap by ≥50%. In this case only 2 patches would be counted. (c) 3 patches, all of which overlap each other by ≥50%. In this case only 1 patch would be counted. (See Table 5.1, column g.)

If the relative overlap (equation 5.7) was ≥ 60% for any of the top 3 patches of a protomer the prediction was defined as correct. On this criterion 69% (20/29) of the interfaces were predicted correctly. This criterion was defined in terms of the relative overlap but it was also important to consider how they relate to the PI values. The highest PI values of each set of top 3 patches ranged from 34% to 100% for the 20 protomers whose interfaces were correctly predicted (Figure 5.7).

The evaluation of the overlapping nature of the top 3 patches, with the highest combined score, revealed that in 20 of the 29 protomers the top 3 patches overlapped by ≥ 50% and represented only a single patch. Hence, in the majority of cases, the top 3 patches relate to the same area on the surface of the protein; and they do not represent alternative putative interface sites.

The number of patches defined upon the surface of the protomers ranged from 57 to 356, reflecting the overall size of the protomers. The ranking of the patch with the largest overlap with the observed interface (Max. PI, the ‘correct’ patch) revealed that in 66% (19/29) of the protomers this patch ranked in the top 20 surface patches.
Figure 5.7: Percentage overlaps (relative overlap, overlap P1 and overlap P2, (equations 5.7, 5.5 and 5.6 respectively)) for the patch with the highest P1 overlap in the top 3 patches for protomers from 20 homo-dimers with correctly predicted interfaces. The protomers have been placed in order of increasing relative overlap.

It was interesting to observe that 1 protomer, variant surface glycoprotein (1vsg), was defined as correctly predicted, based on the 60% criterion for the relative overlap value of 1 of the top 3 patches, but the ‘correct’ patch only ranked 188th out of a total of 282 surface patches. This inconsistency probably resulted from the uncharacteristic shape of this structure’s interface. The interface is elongated, as is the protomer as a whole, and as was seen in the distribution of overlap values (Figure 5.5), practically all the circular patches defined on the surface sampled the interface. The maximum overlap achieved with the size and shape of the patch used, was the smallest in all 29 homo-dimers, only 34%. Hence the 60% cut-off in this structure was not relevant, as approximately half of the patches achieved this sampling overlap.

In any predictive algorithm of this nature the selection of criteria for the definition of correct and incorrect predictions is somewhat arbitrary. However with this method a correlation between the P1 overlap and the combined score was observed in many of the protomers studied (Figure 5.8), indicating that the combined score does carry useful information for the selection of putative interface sites.
Figure 5.8: The relationships between the percentage overlap with the observed interface (P1) and the combined score for 29 protomers. The black dots represent the calculated patches, and the open triangle the 'correct' patch (i.e. that which overlaps most with the observed interface) (see Table 5.1). The homo-dimers are recorded in increasing order of the ranking of the 'correct' patch.
Of the 29 protomers there are 9 structures with interfaces that were never predicted (those with PDB codes shaded in Table 5.1). The top 3 patches for these 9 structures were examined to find why they were selected in preference to the observed interface region. The size of the patches used in each case could have been a significant factor in the failure of the prediction, hence the predictions were repeated using the size of the observed interface as the patch size. This resulted in 3 structures (1pp2, 2ccy and 2rus) being correctly predicted, but for 6 structures (2ssi, 2tsc, 3aat, 3enl, 3gap, 3sdp) the predictions were still unsuccessful.

It is possible that other interaction sites on the surfaces of these proteins affected the predictions. For example, the subtilisin inhibitor protomer (2ssi) has an enzyme interaction site as well as a site for the dimer interaction. The presence of the active site influences the relative scoring of the observed interface to other surface patches. When the surface patches for 2ssi were scored for each of the 6 parameters, as discussed in Chapter 4 (section 4.4), the observed interface site did not score at the extreme of the distribution for 5 of the 6 parameters. Except for the planarity parameter, the observed interface only ranked 5 or 6 (on a scale of 1-10; where a rank of 1 was expected in all parameter distributions except planarity) in each of 5 distributions. Hence using the predictive algorithm with the combined score based on the premise that the observed interface will rank in the extremes of each of the 6 distributions, would obviously prove unsuccessful for this structure.

5.5 Prediction of Interaction Sites in Enzymes

The PATCH algorithm was used to predict the interaction sites on non-homologous enzymes involved in 3 enzyme-inhibitor complexes (alpha-chymotrypsin/Eglin C (1acb), subtilisin carlsberg/Eglin C (1cse) and papain/inhibitor stefin B mutant (1stf)). The fact that these complexes represent the interaction between a large and a small component, meant that no valid relationship between the number of residues in the enzyme and the number of residues in the interface could be made. Hence for the purposes of prediction the patch size was set to 30 residues for all 3 enzymes. This was the mean number (to the nearest integer) of residues involved in the observed enzyme interfaces in the enzyme-inhibitor data set discussed in Chapter 3 (Table 3.1). It was observed (see Chapter 4, section 4.4) that the enzyme interfaces were less accessible and less protruding, relative to other surface patches, hence the combined score definition used was that shown in equation 5.3.
The percentage size differences (equation 5.4) were relatively small for the enzymes (ranging from 11% smaller to 10% larger than the observed interface), compared to those of the homo-dimers (which ranged from 37% smaller to 81% larger). This was because the patch size used was not predicted using the relationship shown in equation 5.1, but set to a value of 30 residues, based on the size of observed interfaces in other enzyme structures.

The results of the predictions for the 3 enzymes are shown in Table 5.2. The same criterion for a correct prediction was used for the enzymes as for the homo-dimers (if the relative overlap was ≥ 60% for any of the top 3 patches of an enzyme, the prediction was defined as correct). On this criterion all 3 enzyme interfaces were predicted correctly. The rankings of the patch with the largest overlap with the observed interface; i.e. the ‘correct’ patch, revealed that in all 3 enzymes this patch ranked in the top 20 surface patches. In addition a correlation was observed between the percentage overlap P1 and the combined score of the surface patches in each of the enzyme structures (Figure 5.9).

Table 5.2: Results of the prediction algorithm (PATCH) for enzymes from 3 enzyme-inhibitor complexes. For a description of each column see caption to Table 5.1. The enzymes are recorded in order of the ranking in column (g).

<table>
<thead>
<tr>
<th>(a)</th>
<th>(b)</th>
<th>(c)</th>
<th>(d)</th>
<th>(e)</th>
<th>(f)</th>
<th>(g)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PDB Code</td>
<td>N° Patches</td>
<td>% Overlap P1 of top 3 patches</td>
<td>Max P1 %</td>
<td>Relative % Overlap of top 3 patches</td>
<td>N° Different Patches</td>
<td>Rank Order Max P1</td>
</tr>
<tr>
<td>1acb</td>
<td>169</td>
<td>78 52 67</td>
<td>78</td>
<td>100 67 86</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>1cse</td>
<td>173</td>
<td>52 0 22</td>
<td>74</td>
<td>70 0 80</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>1stf</td>
<td>148</td>
<td>47 25 50</td>
<td>69</td>
<td>68 36 72</td>
<td>3</td>
<td>11</td>
</tr>
</tbody>
</table>

Figure 5.9: Relationship between the percentage overlap with the observed interface (P1), and the combined score for 3 enzymes. The black dots represent the calculated patches, and the open triangle the ‘correct’ patch (i.e. that which overlaps most with the observed interface). The enzymes are recorded in increasing order of the ranking of the ‘correct’ patch (Table 5.2, column g).

123
5.6 The PATCH Software

To summarise; the PATCH software has 2 basic functions:

- The analysis of the distribution of 6 parameters (solvation potential, residue interface propensity, hydrophobicity, planarity, protrusion, and accessible surface area) in patches defined on the surface of an isolated protomer
- Prediction of putative interface sites on the surface of isolated protomers using patch analysis.

The PATCH software operates on the 3-dimensional atomic co-ordinates in the current (1995) PDB file format, and an additional input file containing the user specified variables as discussed in section 5.3.4. For the analysis function, PATCH produces histograms (in Postscript™ format) of the 6 parameter distributions for all surface patches. For the prediction function, PATCH produces patch profiles (in Postscript™ format) indicating values of each surface patch for each of the 6 parameters, and for the combined score (e.g. see Figure 5.2).

5.7 Conclusions

The predictive algorithm is based on the definition and comparison of surface patches at the Cα atom level. The simple nature of the definition has advantages and disadvantages. Like all docking and prediction algorithms there is a balance to be reached between the accuracy of the method and the time taken. It is possible that more accurate predictions could be achieved if the patches were defined in terms of surface atoms, and each of the 6 parameters calculated at the atom level. This would be easier for some parameters than others. For example, there have been few atomic hydrophobicity scales calculated, and those, such as the atomic solvation parameter (Eisenberg and McLachlan, 1986), have not been extensively evaluated. However, an alternative measure of atomic hydrophobicity could be formulated by simply calculating the percentage of polar and non-polar atoms.

The definition of the surface patches at the Cα atom level meant that the maximum percentage overlaps with the observed interface (P1 and P2) were never both 100%, i.e. a patch was never defined that contained all the observed interface atoms and no others. To achieve this the patches would need to be defined at the atom level and the definition must permit discontinuous patches. However this would result in a combinatorial explosion, with thousands of patches being defined. Restricting the patch definition to contiguous surface patches defined at the
residue level, reduces the combinatorial problem, but at the expense of the accuracy of the predictions. However, the prediction algorithm is relatively fast; the prediction of the interaction site on HIV protease (PDB code 5hvp) takes 36 seconds on an SGI-Challenge with R4400 CPU running at 150MHz.

By definition the surface patches were overlapping and this caused problems in the evaluation of the predictions. In many homo-dimer structures the 3 patches with the highest combined score overlapped each other by $\geq 50\%$, and represented a single region on the surface of each protomer. One way of overcoming the problem of overlapping patches, would be to assign the combined score for a patch centred on a single residue to that central residue. Hence each residue would be assigned a combined score that describes its local environment upon the surface of the protomer. Putative interface sites could then be selected to comprise those residues with the highest combined scores.

In some respects the method described in this chapter is similar to that used by Young et al (1994) in their analysis of the hydrophobicity of surface residue clusters in proteins. In their method clusters of surface residues were ranked on the basis of the hydrophobicity of the constituent amino acids to identify the interaction sites of proteins (including enzymes and antibody fragments) with their associated ligands. The definition of surface clusters resulted in the clusters being of different sizes, ranging from 3-15 residues. Young et al. (1994) claim that their method had a high success rate, with 25 out of 28 cases having an 'exact' correspondence between the predicted cluster and the ligand binding site. However, if the position of the most hydrophobic cluster coincided with at least 1/3 of the surface buried by the bound ligand, the prediction was classified as 'exact'; hence the criterion for a successful prediction was extremely lenient. When this criterion was used in the evaluation of the PATCH algorithm, the location of 25/29 (86%) interface sites were successfully predicted in the homo-dimers.

The Young et al. (1994) criterion was considered to be too lenient, and the evaluation of the PATCH algorithm, described in this chapter, was undertaken with a more stringent criterion. A prediction was only defined as successful if the relative overlap between the observed interface and the surface patch was $\geq 60\%$ for 1 of the top 3 patches. The current method was more successful than the Young et al. (1994) method, because it used multiple parameters, rather than a single parameter, to rank surface patches. Such an approach is useful for identifying candidate interface residues, which can be mutated experimentally, and tested for their effect on
complex formation. As was observed, the nature of the interface can vary and a 'perfect' prediction would be an unrealistic expectation, unless the structure of the partner is known and full docking can be pursued. The method described here provides a rapid means to identify possible interaction sites as a guide to future experiments. Further examples of the use of this predictive tool are discussed in Chapter 6.
CHAPTER 6

Applications of Patch Analysis in Proteins

6.1 Introduction
The analysis of surface residue patches on proteins, described in Chapters 4 and 5, has a number of practical applications. This chapter discusses how the method can be used to compare surfaces between related structures, and to predict putative interaction sites in a number of different systems. Many related proteins exhibit the same quaternary structure. However, there are instances where related proteins possess different oligomeric states. In this chapter families of proteins in the PDB have been analysed to record the occurrence of this phenomenon, and one example, haemoglobin, is examined in detail. In addition, the method of patch analysis is applied to the prediction of crystal contact sites on the surface of sickle cell haemoglobin. The method is also applied to the prediction of structural epitopes on 2 protein antigens, lysozyme and human chorionic gonadotropin. The computational tool (PROFACE), described in Chapter 2 (section 2.6) for the characterisation of protein-protein interfaces, is used to make preliminary observations upon the specific interfaces in question. The software (PATCH) described in Chapter 5 (section 5.6) is then used to predict interaction sites in different systems. The results are validated from known structures, or from experimental analysis.

6.2 Multiple Multimeric States within Protein Families
Proteins in the PDB have been classified into a series of fold families, by a number of structure comparison methods (e.g. Holme et al., 1992; Orengo, 1992). These fold classifications were exclusively based on the secondary and tertiary structure of proteins. Here the quaternary structure of proteins, classified in the same fold family, were analysed.

6.2.1 Classification of Oligomeric States in Protein Families
The Sequential Structure Alignment Program (SSAP) (Taylor and Orengo, 1989) has been used to classify the PDB into a number of different fold families, in which proteins have a related fold but can vary widely in sequence similarity (Orengo, 1992; 1994). In this method a SSAP score indicates the structural similarity between proteins. SSAP scores ≥ 80 are associated with
highly similar structures, which are usually homologous. In the current analysis of protein quaternary structure, this was the level at which proteins were defined to be in the same family. Classifying the oligomeric states of the structures within these protein families (as described in Chapter 2, section 2.2) revealed 3 families with members that adopt different multimeric states (Table 6.1).

<table>
<thead>
<tr>
<th>Protein Family</th>
<th>Family Members</th>
<th>Source</th>
<th>Multimeric State</th>
</tr>
</thead>
<tbody>
<tr>
<td>PDB Code Name</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Haemoglobin</td>
<td>4sdh Haemoglobin</td>
<td>Arcid clam Human</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>2hhb Haemoglobin</td>
<td></td>
<td>4</td>
</tr>
<tr>
<td>RUBISCO</td>
<td>2rus RUBISCO</td>
<td>Rhodospirillum rubrum expressed in E.coli</td>
<td>2</td>
</tr>
<tr>
<td>(Ribulose-1,5 bisphosphate carboxylase/oxygenase)</td>
<td>8rub RUBISCO</td>
<td>Spinach</td>
<td>16</td>
</tr>
<tr>
<td>Malate Dehydrogenase</td>
<td>4mdh Cytoplasmic malate dehydrogenase</td>
<td>Porcine heart</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>6ldh M4 apo-lactate dehydrogenase</td>
<td>Dogfish muscle</td>
<td>4</td>
</tr>
</tbody>
</table>

Table 6.1: Protein fold families with protein members in different oligomeric states (derived from the July 1993 release of the PDB). The family name is that of a representative protein within the family.

The classification identified the well-documented case of lactate dehydrogenase (LDH) and cytoplasmic malate dehydrogenase (MDH); the former being a tetramer and the latter a dimer. The difference in oligomeric states was attributed to an additional length of polypeptide chain in the tetrameric LDH, for which there is no counterpart in the MDH structure (Matthews and Bernhard, 1973). It was proposed that these residues form an extended arm that interacts with the other subunit, thus stabilising the tetrameric structure (Matthews and Bernhard, 1973). This case highlights the need to examine related proteins to find regions that differ, either structurally or chemically. The method of analysing patches of surface residues on proteins could prove to be potentially useful in understanding how 2 structurally homologous proteins can attain different quaternary states.
6.2.2 Haemoglobin (Hb): Dimeric and Tetrameric

The structure of mammalian haemoglobin, solved by Perutz in 1960, is one of the most well documented proteins. Its classic tetrameric structure, comprising 2 α and 2 β subunits, has been used to study the mechanism of biochemical co-operativity (e.g. Perutz 1970; 1979). Co-operativity in Hb is manifested when the binding of the oxygen ligand increases the affinity for subsequent oxygen binding (Ackers et al., 1992). More recently the structure of haemoglobins in a family of arcid clams (Scapharca inaequivalvis) has also been solved (Royer et al., 1985; Royer et al., 1989; Condon and Royer 1994; Royer, 1994). Of particular importance is the presence of a homo-dimeric Hb (which comprises 2 α subunits), that has a very different method of co-operativity to that seen in the mammalian tetrameric structure. Both the dimeric and tetrameric Hb α subunits exhibit the 'globin fold', comprising a series of 7 α-helices (classified as A, B, C, E, F, G, H in the dimeric Hb (Royer, 1994)) with non-helical regions between the helices and at the amino and carboxyl terminals. The β subunit in the Hb tetramer has 8 helices. In both structures each subunit contains a haem group, at the centre of which is an iron atom to which the oxygen ligand binds. The haem groups are connected to the protein by a covalent bond formed between the central iron atom to a histidine residue, denoted the proximal histidine (Baldwin and Chothia, 1979). The dimeric and tetrameric Hb structures are shown as MOLSCRIPT (Kraulis, 1991) diagrams in Figure 6.1.

The sequences of the α subunits from the 2 Hb structures were aligned, using a structure-based method SSAP (Taylor and Orengo, 1989), and the residues forming the interface contacts located on each sequence (Figure 6.2). From this alignment, it was observed that the position of the residues forming the interface in the dimeric Hb is different to those forming interfaces in the tetrameric Hb. When the dimeric Hb structure was resolved (Royer et al., 1985; Royer, 1994) it was found that the interface regions were located on the E and F helices. The equivalent helices in human tetrameric Hb structure face the solvent, and hence the oligomeric arrangement in dimeric Hb has been termed 'back to front' relative to the human Hb tetramer (Royer et al., 1985; Royer, 1994).
Figure 6.1: MOLSCRIPT (Kraulis, 1991) diagrams of: (a) dimeric deoxy-haemoglobin from *Scapharca* (4sdh) and (b) human tetrameric deoxy-haemoglobin (2hhb).
Figure 6.2: Alignment of the amino acid sequence of one α subunit of haemoglobin from the arcid clam *Scapharca inaequivalvis* (4sdh) and the α subunit from human haemoglobin (2hhb). The interface residues in each subunit are indicated on each sequence.

**Characterisation of Hb Interfaces**

In the current work the interacting surfaces of the human deoxy-Hb tetramer (Fermi *et al.*, 1984) (PDB code 2hhb) and the arcid clam deoxy-Hb dimer (Royer, 1994) (PDB code 4sdh) have been compared. The program PROFACE (Chapter 2, section 2.6) was used to define and analyse the protein-protein interfaces involving the α subunit within each of the Hb structures (Table 6.2). The parameters were initially calculated without the haem groups present in each structure. However the haem groups in the Hb dimer are an integral part of the interface (Royer *et al.*, 1985), and the ΔASA and complementarity were also calculated with the haem groups present in the structure (Table 6.2).

The inclusion of the haem group in the Hb dimer increased the ΔASA by 46 Å² and increased the %ΔASA from 12% to 14%. The inclusion of the haem groups also decreased the gap volume index, indicating an increase in the complementarity of the interface. These changes indicated that the haem groups play a significant role in the dimer interface of this Hb structure. The significance of the haem groups in the dimer interface was first observed when the structure was solved at high resolution (Royer *et al.*, 1990; Royer, 1994). Royer (1994) classified the interface into 4 regions, 3 involved amino acid residues in 3 of the 7 helices (designated “E-E”, “E-F”, “B-F”) and the fourth involved the interaction of the F-helix with the haem propionates of the symmetry related subunit (“Haem-F”). Hydrogen bonds between the F-helix and the haem...
group were also observed, which contribute to the stability of the dimer interaction (Royer, 1994).

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Hb Dimer</th>
<th>Hb Tetramer</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Interface</td>
<td>Interface</td>
</tr>
<tr>
<td></td>
<td>AB</td>
<td>AB</td>
</tr>
</tbody>
</table>
| (i) ΔASA (%ΔASA) | 954 (12%) | 823 (10%)   | 277 (4%)    | 691 (9%)    
|           | (1000 (14%)*) |             |             |             |
| (ii) Shape | 0.54     | 0.55        | 0.36        | 0.58        |
| (iii) Planarity | 1.61     | 1.92        | 1.81        | 1.73        |
| (iv) Segments | 3        | 2           | 2           | 3           |
| (v) Polarity | 48%      | 32%         | 49%         | 40%         |
| (vi) N° Inter-molecular H-Bonds | 7 | 7 | 6 | 7 |
| (vii) Gap Volume Index | 2.3 (1.9*) | 2.9 | 7.9 | 2.6 |

Table 6.2: Results of structural analysis on the protein-protein interfaces within dimeric (4sdh) and tetrameric (2hhb) deoxy-haemoglobinins. Values marked with "*" indicate those calculated with the haem groups in place. Parameters i - iv and vi - vii are described in the legend to Table 3.3 in Chapter 3. Polarity is the percentage of polar atoms in the interface.

The interface in the Hb dimer forms a completely different region of the subunit's surface compared to the AB interface in the Hb tetramer (Figure 6.3), and this was reflected in differences observed when each of the interfaces were analysed (Table 6.2). The dimer Hb interface was larger, with respect to the ΔASA, than any of the 3 interfaces in the Hb tetramer, with and without the inclusion of the haem groups. The dimer interface was also more polar than the AB interface of the tetramer: the dimer had 48% polar atoms and the tetramer AB interface only 32%. However, the AC tetramer interface was also similarly polar, with 49% polar atoms. The AC interface in the tetramer was very small (the %ΔASA was only 4% compared to 10% and 9% for the AB and AD interfaces respectively), and could be considered as a secondary interface that occurs as a consequence of the interaction of the 4 subunits. This interface was also very poorly packed with a high complementarity, compared to the less polar, but more closely packed AB and AD interfaces.

**Patch Analysis and Comparison**

The α subunit of the Hb tetramer has 3 interface regions on its surface (one interacting with subunit B (AB), one with subunit C (AC) and one with subunit D (AD)), and the α subunit of the Hb dimer has 1 interface region. Hence there were 4 comparisons that could be made between 'equivalent' patches (Figure 6.4). The 'equivalent' interfaces were located on each Hb structure using the structural alignment in Figure 6.2, and compared using the PATCH software.
Figure 6.3: CPK model of the Hb dimer and Hb tetramer α subunits, indicating the position of the interfaces on each. In each diagram the non-interface residues of the α subunit are coloured blue and the haem group purple. (1) Hb Dimer: the AB interface is coloured red. (2) Hb Tetramer: the AB interface is coloured red, the AD interface is yellow and the AC interface is green.

In 1(a) and 2(a) the subunits are orientated such that the Hb dimer AB interface is viewed. In 1(b) and 2(b) the subunits are orientated such that the 3 interfaces of the Hb tetramer are viewed.
In Chapter 4, it was discussed how surface patches could be ranked relative to other surface patches (on a scale of 1–10) for 6 parameters. A rank of 1 indicated that the observed interface scored in the highest range of a parameter distribution for all surface patches. A rank of 10 indicated that the observed interface scored in the lowest range of a parameter distribution for all surface patches. In the analysis discussed below, the ranking for the planarity parameter has been inverted, as interfaces rank amongst the most planar patches (see Chapter 4, section 4.4.1). Hence an ‘idealised’ interface would rank 1 for each distribution. The relative ranking of the observed and ‘equivalent’ interfaces on the 2 Hb structures have been calculated (Table 6.3). In addition a mean composite rank has been derived by calculating the mean rank over all 6 parameters (to the nearest integer).

**Table 6.3:** Comparison of relative ranking of the observed and ‘equivalent’ interfaces between the Hb dimer and Hb tetramer.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Relative Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Dimer AB</td>
</tr>
<tr>
<td></td>
<td>Dimer</td>
</tr>
<tr>
<td>Solvation Potential</td>
<td>9</td>
</tr>
<tr>
<td>Residue Propensity</td>
<td>6</td>
</tr>
<tr>
<td>Hydrophobicity</td>
<td>9</td>
</tr>
<tr>
<td>Planarity</td>
<td>1</td>
</tr>
<tr>
<td>Protrusion</td>
<td>2</td>
</tr>
<tr>
<td>ASA</td>
<td>1</td>
</tr>
<tr>
<td>Mean Composite Rank</td>
<td>5</td>
</tr>
</tbody>
</table>

**Figure 6.4:** Schematic to show the equivalent interfaces between the α subunit (represented as a shaded circle) of the Hb dimer and the Hb tetramer. The interfaces and their ‘equivalent’ patches are represented as ellipses; with the native interfaces in each structure coloured the same as the background circle, and the ‘equivalent’ patches shown in a different shade.
It was observed from Table 6.2 that the interface in the Hb dimer was polar and hence the relative rankings for the residue interface propensity and hydrophobicity were high. The Hb dimer interface was observed to be one of the most planar patches, and one of the most accessible and protruding. Very similar relative rankings were observed for the equivalent interface on the Hb tetramer.

The largest difference in relative rankings was observed between the AB interface in the tetramer and the ‘equivalent’ patch in the dimer; the mean composite rank was 3 in the tetramer interface but 6 in the ‘equivalent’ patch on the dimer. The rankings for the solvation potential, residue interface propensity, hydrophobicity and ASA were higher in the ‘equivalent’ patch on the Hb dimer. This indicated that the ‘equivalent’ patch in the tetramer was one of the least hydrophobic patches, that had a low accessibility and a low residue interface propensity. The differences between the relative rankings for the AD and AC interfaces and their ‘equivalent’ patches on the dimer were small. Except for ASA, the relative rankings for the AC interface in the tetramer were very low, which probably resulted from the small and polar nature of this interface (see Table 6.2).

In summary, the AB and AD tetrameric interfaces fit the generalised interface ‘picture’ in which an interface represents a planar, relatively hydrophobic patch on the surface of the molecule, which is protruding and accessible (see Chapter 4, section 4.4.1). The mean composite rank for these 2 interfaces was 3, which is close to the ‘idealised’ interface that would have a mean composite rank of 1. The AC tetramer interface was very small, polar and poorly packed, and could be considered as a secondary interface in the tetramer structure. This interface had a correspondingly higher mean composite rank of 5. The dimeric Hb represented a very different ‘picture’ to the AB and AD Hb tetramer interfaces. The dimer interface was relatively polar and this was reflected in the comparatively high ranks of the hydrophobicity and residue interface propensity parameters for this interface; and the correspondingly high mean composite rank. The ‘equivalent’ patch on the surface of the tetramer had corresponding high rankings, and hence would not be expected to form part of a protein-protein interface. The polarity of the interface in the Hb dimer is linked to the presence of the haem groups buried in the interface. These form the basis of the communication mechanism which occurs between the 2 subunits. This mechanism is an integral part of the dimer’s co-operative oxygen binding function (Royer et al., 1990; Chiancone et al., 1993; Condon and Royer, 1994; Royer 1994).
6.3 Crystal Contacts in Sickle Cell Haemoglobin

6.3.1 Structure of Sickle Cell Haemoglobin: Crystal and Polymer

Sickle cell haemoglobin (HbS) differs from normal haemoglobin (HbA) in a single residue substitution, a valine substituted for the normally occurring glutamate at the β6 position. Since its discovery (Pauling et al., 1949; Ingram, 1959) HbS has been extensively studied to gain an understanding of how a single residue substitution at the surface results in haemoglobin polymerisation (see review by Eaton and Hofrichter (1990)). The polymerisation of deoxy-HbS is thought to be facilitated by the Val-β6 substitution, and by the formation of a hydrophobic pocket between the E and F helices (Wishner et al., 1975). HbS molecules aggregate to form regular fibrous polymers that cause the occlusion of micro-capillaries, resulting in the physiological conditions observed in sickle cell anaemia (Makowski and Magdoff-Fairchild, 1986).

It has been demonstrated that the packing arrangements of HbS in polymers are very similar to the packing observed in the deoxy-HbS single crystal (Edelstein, 1981; Carragher et al., 1988; Eaton and Hofrichter, 1990). In the crystal there are paired linear strands of HbS tetramers, in which the pairs are half-staggered with respect to each other (Padlan and Love, 1985b). The residue at the site of the substitution, Val-β6, is involved in the side-side interaction between paired strands (Figure 6.5). The deoxy-HbS crystal structure has been solved at 3.0Å resolution (Padlan and Love, 1985a; Padlan and Love, 1985b), and its asymmetric unit contains 2 complete tetramers. Hence there are 4 copies of the α subunit and 4 copies of the β subunit. In contrast the deoxy-HbA crystal structure (solved at 1.7Å) has only one HbA molecule in the asymmetric unit, and hence 2 copies of each subunit (Fermi et al., 1984). Although there are differences between HbA and HbS related to the variation in structural resolution (1.7Å and 3.0Å respectively); some comparisons between the two are valid, as it has been shown that the 2 tetramers in the HbS crystal differ more from each other, than either do from the HbA tetramer (Eaton and Hofrichter, 1990).
Figure 6.5: The arrangement of subunits in a polymer of sickle cell haemoglobin (HbS). Each tetramer is drawn end on, and the Val-β6 substitutions are indicated by black dots. It is seen that specific Val-β6 substitutions are involved in side-side interactions between the paired strands in the polymer. (Diagram adapted from Eaton and Hofrichter, 1990).

6.3.2 Characterisation of Crystal Contacts

The interaction between the 2 tetramers is heterologous. Each tetramer in the crystal was treated as one protomer, and designated A and C for the analysis (Figure 6.6). The interfaces on each tetramer were analysed using the program PROFACE (Chapter 2, section 2.6) and the characteristics of each interface are listed in Table 6.5. It was observed that the valine residue (Val-β6) substituted for the normal glutamate residue in the C tetramer contributed 28% of the ΔASA, when going from a single tetramer to the dimer of tetramers.

Figure 6.6: Interaction of tetramers of HbS in the asymmetric unit of the crystal. There are 2 tetramers, designated A and C. The approximate position of the Val-β6 substitutions are indicated by the small shaded circles. The small black circle indicates the position of the Val-β6 substitution that is involved in the interaction between the 2 tetramers in the crystal.
Table 6.4: Results of structural analysis for the interface between the 2 tetramers in the asymmetric unit of sickle cell haemoglobin (PDB code 1hbs). Parameters i - iv and vi - vii are described in the legend to Table 3.3 in Chapter 3. Polarity is the percentage of polar atoms in the interface.

The interaction between the tetramers is a crystal contact, and the \%ΔASA were small, and the complementarity poor, with a high gap volume index (10.7). The comparable complementarity factor observed between homo-dimers was only 2.2 (with a maximum of 4.43) (Chapter 3, Table 3.3). However the shape and planarity parameters, and the number of segments in the crystal contact, were well within the criteria expected for normal dimer contacts (see Chapter 3, Table 3.3). In addition, there were differences observed between the interfaces of each tetramer. The interface of tetramer A was more polar than that of tetramer C, and was composed of 1 segment, compared to the 4 segments of the interface of tetramer C.

6.3.3 Patch Analysis and Prediction of Crystal Contacts

A relative ranking of the observed interface between tetramers was calculated for each tetramer using the PATCH software (Table 6.5). The relative rankings were assigned on a scale 1-10, where an ‘idealised’ interface ranked 1 for each distribution (see section 6.2.2). The observed interface in each tetramer was observed to be amongst the most planar, the most accessible, and most protruding patches. However, the observed interfaces were not recorded at the extremes of the distributions for the other 3 parameters (solvation potential, residue interface propensities, and hydrophobicity (based on the Fauchère and Pliska (1983) scale)).
Table 6.5: Relative rankings of the observed interface in each tetramer of the HbS crystal for 6 parameters. An idealised interface would rank 1 for each parameter.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Tetramer A</th>
<th>Tetramer C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Solvation Potential</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>Residue Propensity</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>Hydrophobicity</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>Planarity</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Protrusion</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>ASA</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Mean Composite Rank</td>
<td>4</td>
<td>4</td>
</tr>
</tbody>
</table>

With this information the PATCH software was used to predict the position of putative interaction sites on the surfaces of the HbS tetramers (A and C), and, for comparison, the HbA tetramer. A patch size of 10 residues was used (selected from a consideration of the number of residues in the observed interfaces), and all 6 parameters were used for the combined score as defined in in Chapter 5 equation 5.2. A 10% cut-off was chosen to select the patches with the top combined score. The frequency of surface residues in the top 10% of patches for the 2 HbS tetramers (A and C) and the HbA tetramer are shown in Figure 6.7.

In the HbS A tetramer the patch with the largest overlap with the observed interface (i.e. the 'correct' patch) was ranked third and in the HbS C tetramer the 'correct' patch was ranked twentieth. Using the definition of a successful prediction defined in Chapter 5 (section 5.4.2) the interface on the HbS A tetramer was correctly predicted but the interface of the HbS C tetramer was not. However the top 10% of surface patches for this latter structure did include one patch in the region of the Val-β6 substitution (residue 434), which was involved in the crystal contact between the 2 tetramers.

There were significant differences observed between each of the profiles (Figure 6.7). In both the HbS A and HbS C profiles residues were selected in the region of the first Val-β substitution at residue 148; but residues were only selected in the region of the second Val-β substitution at residue 434 in the HbS C tetramer. This tetramer is the one in which the region of the 434 Val-β6 residue is involved in the crystal contact. The profile for the HbA tetramer indicates that no residues were selected in either of the regions of the Val-β substitutions which occur in the HbS structure; but segments were selected that were also observed in the residue frequency profiles of the HbS A and HbS C tetramers.
Figure 6.7: Frequency profiles of the HbS and HbA tetramers. In each, the residues have been numbered sequentially from 1–574. The subunits comprising each tetramer are indicated on each profile. The site of the Val-β substitutions in the HbS structure are indicated on the profiles by "*". The Val-β substitution at residue 434 is the one involved in the contact between the 2 tetramers in the HbS crystal.
The results indicate that the Val-β6 substitutions change the nature of the protein surface in 2 regions such that they are more likely to form protein-protein interactions. The substitutions change the surface sufficiently for the PATCH algorithm to select both regions as putative interaction sites; whilst these regions were not selected in the normal haemoglobin tetramer (HbA). However these results must be treated with some caution, as the difference in resolution of the HbS and HbA structures (3.0Å and 1.7Å respectively) obviously causes differences in all regions of the structure (Eaton and Horfrichter, 1990).

### 6.4 Prediction of Protein Antigen Epitopes

The term 'epitope' has been multiply defined as "surface configurations, single determinants, structural themes, immunogenic elements, haptenic groups, antigenic patterns, specific areas" (Jerne, 1960, as cited in Laver et al., 1990). A number of different classes of epitope have since been distinguished, namely *sequential*, *structural*, *energetic* and *functional*. A sequential epitope is one that involves a single continuous length of the polypeptide chain. A structural epitope is one that involves several discrete amino acid sequences widely separated along the protein sequence, which come together on the surface when the polypeptide chain is folded (Sela, 1969, as cited in Laver et al., 1990). An energetic epitope is one that comprises those residues that are the most energetically important in the binding process (Novotny, 1991). A functional epitope comprises those residues defined to be important in the binding process from experimental binding analysis, such as mutagenesis experiments. Hence a functional epitope is defined to be formed of those residues that, when mutated, terminate antigen binding.

The aim of the current work was to predict the location of structural epitopes on protein antigens. Such epitopes have been characterised for a number of antibody-protein complexes solved by X-ray crystallography (e.g. Sheriff et al., 1987; Padlan et al., 1989; Fischmann et al., 1991). They exhibit similar characteristics, being composed of 12-22 residues, involving 2-5 surface loops, and being discontinuous in terms of the protein sequence. (Colman, 1988; Laver et al., 1990). To date there have been many methods of epitope prediction and the majority, such as those of Hopp and Woods (1981; 1983), and Welling et al. (1985), involve the prediction of sequential epitopes. The 2 methods named above, assign a numerical value to each of the 20 amino acids (a hydrophilicity value in the case of Hopp and Woods (1981; 1983), and an antigenicity value in the case of Welling et al. (1985)). This value is repeatedly averaged
along the amino acid sequence. A method has also been devised to locate antigenic peptides from the Cα atom co-ordinates of a protein, based on protrusion from the protein's globular surface (Thornton et al., 1986). A comparison between this method and the use of accessibility and flexibility of amino acid residues found that none were definitive in locating the antigenic peptides, but that the protrusion method was superior to that of Hopp and Woods (1981; 1983) (Thornton et al., 1986). This was to be expected as a 3-dimensional structure will obviously give more information than the amino acid sequence.

6.4.1 Characterisation of Structural Epitopes

The characterisation of structural epitopes from 6 antibody-protein (antigen) complexes (Chapter 3, Table 3.1c) was conducted using the PATCH software. The residue interface propensities were those based on 63 protein-protein complexes (analysed in Chapter 3), derived as described in Chapter 4 (section 4.3.3). The 6 parameters (solvation potential, residue interface propensities, hydrophobicity, planarity, protrusion, and accessible surface area) were calculated for each surface patch, and for the observed structural epitopes on each antigen. The protrusion index was the same as that used in the prediction method of Thornton et al. (1986). A ranking (on a scale of 1–10 (see Chapter 4, section 4.4)) of the observed structural epitope site relative to other surface patches was represented as a histogram for each antigen (Figure 6.8).

It was observed that no single parameter absolutely differentiated the observed structural epitopes from all the other patches on the surface of the proteins. The most striking correlation was seen for the protrusion index and the accessible surface area. The observed interface was more accessible and more protruding than other surface patches. In addition the distribution observed for the planarity parameter showed a clear trend for an epitope site to be more planar than other surface patches. The solvation potential and residue interface propensities showed little correlation. The hydrophobicity parameter showed that the observed epitopes on the protein antigens tended to be less hydrophobic than other surface patches, although values were not recorded at the extreme of the distribution (i.e. no observed interfaces were ranked as 10). This infers that the observed epitopes show a preference for charged and polar residues, unlike other protein-protein interaction sites that have been observed to be, in general, hydrophobic (Chapter 3, section 3.4). The hydrophilic nature of antigenic sites has been observed previously; and was used by Hopp and Woods (1981) in their prediction of sequential epitopes, based on hydrophilicity values.
Figure 6.8: Patch analysis distributions for 6 protein antigens. Rank ordering (on a scale of 1-10 (see Chapter 4, section 4.4)) of observed interface patches relative to other patches on the surface of the protein. The 6 observations (one for each antigen) were combined for each parameter separately. The distributions shown are: (a) Solvation potentials, (b) Interface residue propensities, (c) Hydrophobicity (based on the scale of Fauchère and Pliska (1983)) (d) RMS deviation of atoms from the least squares plane through the interface atoms, (e) Protrusion index, (f) Accessible surface area.
6.4.2 Prediction of Structural Epitopes

As discussed in Chapter 4 (section 4.5.2) a number of input variables are required in the PATCH algorithm. For the epitope predictions a patch size of 15 residues was selected, as studies have shown that structural epitopes range in size from 12 to 22 residues (e.g. Padlan et al., 1989; Fischmann et al., 1991). This value is significantly larger than that used in the prediction of sequential epitopes (Hopp and Woods 1981; Hopp and Woods, 1983; Welling et al., 1985) in which windows of 6 and 7 residues were used to average numerical values.

For epitope predictions the combined score was defined as:-

\[
\text{Combined Score } P_j = \frac{(100 - S_{hy}) + S_{pi} + S_{asa}}{N_{par}} + (100 - S_{pl})
\]  

[6.1]

where 

- \( S_{hy} \) = rank order of patch \( P_j \) in the hydrophobicity distribution
- \( S_{pi} \) = rank order of patch \( P_j \) in protrusion index distribution
- \( S_{asa} \) = rank order of the patch \( P_j \) in the accessible surface area distribution
- \( S_{pl} \) = rank order of patch \( P_j \) in the planarity distribution
- \( N_{par} \) = number of parameters

In this definition the planarity and hydrophobicity parameters were inverted as epitopes were observed to be the most planar patches and amongst the most hydrophobic patches (see section 6.4.1); and therefore the prediction scored a planar and hydrophobic patch highly. The cut-off for selecting the patches with the highest combined rank was 10%. The positions of structural epitope sites were predicted in two systems; lysozyme and human chorionic gonadotropin (hCG).

Lysozyme

The interactions between lysozyme and antibodies have been extensively documented (see review by Benjamin et al. (1984)) and there are multiple entries of such structures in the PDB. The aim of the current work was to predict the structural epitopes of an unbound structure of lysozyme. The structure used was hen egg white lysozyme (PDB code 132L) (Rypniewski, 1993). The PATCH software produced a list of surface residues which constituted putative interaction sites (i.e. those with a combined score in the top 10% of the distribution and defined as best patches). These are indicated in Figure 6.9, alongside the residues in the structural...
epitopes derived from three different antibody-lysozyme complexes, D1.3 Fab/Hen Egg White Lysozyme (1fdl), HYHEL-5 Fab/Chicken Lysozyme (2hfl), and HYHEL-10 Fab/Chicken Lysozyme (3hfm) (defined using the interface definition described in Chapter 2, section 2.4.1).

Figure 6.9: Prediction of structural epitope sites on lysozyme using the PATCH software.

Good correspondence was observed between the position of the observed and the predicted epitope residues. The predicted residues form 5 segments, each corresponding to different segments of the 3 known structural epitopes. It was difficult to quantify the correspondence in terms of the overlap values (P1 and P2) calculated in Chapter 4 (section 4.5.3), as there were multiple epitopes on the surface of the lysozyme. However, 2 percentage overlaps, TP1 and TP2, between all the residues in the 3 observed epitopes and all the residues involved in the predicted epitopes were calculated i.e.

\[
TP1 = \frac{\text{TNrO} \cap \text{TNrC}}{\text{TNrO}} \times 100 \quad [6.2]
\]

\[
TP2 = \frac{\text{TNrO} \cap \text{TNrC}}{\text{TNrC}} \times 100 \quad [6.3]
\]

where \( \text{TNrO} \) = total number of residues in 3 observed epitopes
\( \text{TNrC} \) = total number of residues in the calculated epitopes

TP1 was calculated as 51% and TP2 as 83%, indicating that of the residues involved in the 3 observed epitope sites 51% were represented in the top 10% of patches (the best patches), and 83% of those residues selected for the putative epitopes occured in one of the 3 observed epitopes.
The percentage overlaps between each of the 5 patches with the highest combined score (which constituted the top 10% of patches) and each of the observed epitopes (overlap/correlated patch) were also calculated (Table 6.6).

<table>
<thead>
<tr>
<th>Rank</th>
<th>Order</th>
<th>Central Residue</th>
<th>% Overlap with Observed Epitopes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st</td>
<td>21</td>
<td>0</td>
<td>93</td>
</tr>
<tr>
<td>2nd</td>
<td>68</td>
<td>0</td>
<td>73</td>
</tr>
<tr>
<td>3rd</td>
<td>69</td>
<td>0</td>
<td>80</td>
</tr>
<tr>
<td>4th</td>
<td>70</td>
<td>0</td>
<td>73</td>
</tr>
<tr>
<td>5th</td>
<td>104</td>
<td>33</td>
<td>53</td>
</tr>
</tbody>
</table>

The results indicate a good correspondence between the predicted and the observed epitopes. The epitopes on the 2hfl and the 3hfm structures were predicted more successfully than that on the 1fdl structure. One segment (residues 117-121 and 124-126) of the 1fdl epitope was never predicted, as patches in this region were relatively hydrophobic and non-planar.

The two algorithms designed to predict sequential epitopes (Hopp and Woods, 1981; Hopp and Woods 1983; Welling et al., 1985) were only partially successful in predicting epitopes. In an evaluation of these predictive methods Getzoff et al. (1988) concluded that the single numerical parameters, upon which the predictions were based, were insufficient to accurately predict the location of sequential epitopes. The method described in the current work was more successful in identifying residues involved in structural epitopes by combining a number of characteristics of the structural epitopes.

**Human Chorionic Gonadotropin**

Human chorionic gonadotropin (hCG) is a hetero-dimer comprised of \(\alpha\) and \(\beta\) subunits, each containing 3 disulphide bonds that form a cysteine knot (Lapthorn et al., 1994). The dimer is stabilised by a segment of the \(\beta\) subunit which wraps around the \(\alpha\) subunit, covalently linking the two by a disulphide bond (Figure 6.10). The hCG protein is a glycoprotein hormone responsible for the maintenance of the early stages of pregnancy in humans (Pierce and Parsons, 1981). Its functional significance in the human reproductive cycle means that agonists are of potential importance for infertility treatments, and antagonists are candidates for contraceptive vaccines (Wu et al., 1994). The protein is a member of a family of hetero-dimeric glycoprotein hormones, which includes luteinizing hormone (LH), follicle-stimulating hormone (FSH), and thyroid-stimulating hormone (TSH). These hormones function by binding to specific cell surface receptors, and hCG and LH bind to the same cell surface receptor, but their secretion...
occurs under different conditions (Wu et al., 1994). Both the $\alpha$ and $\beta$ subunits are required in each glycoprotein for receptor binding, but the $\beta$ subunit is the one that determines the specific activity of the hormone (Lapthorn et al., 1994). There is a significant amino acid sequence similarity between the $\beta$ subunits of the glycoproteins; specifically the $\beta$ subunit of hCG and LH have an 80% sequence identity (Wu et al., 1994).

![Subunit A ($\alpha$) and Subunit B ($\beta$)](image)

**Figure 6.10:** MOLSCRIPT (Kraulis, 1991) diagram of hetero-dimeric human chorionic gonadotropin (hCG).

The sequence and structural similarity between hCG and LH pose a problem in the search for an antagonist of hCG. If an antagonist is isolated to act as a contraceptive vaccine, it is likely that it will also act as an agonist against LH. A specific antagonist needs to be found which will inactivate hCG but leave LH to function normally. Work is currently being conducted into the production of a highly specific immunological contraceptive vaccine (Jackson et al., 1995). This involves the initial identification of epitopes on the surface of the hCG $\beta$ subunit using a combination of site-directed and random mutagenesis (Jackson et al.; 1995; T. Lund, personal communication). The aim is to find and remove epitopes on the surface of hCG that cross-react with those on LH. This experimental work prompted the application of the PATCH software (Chapter 5, section 5.6) to the problem. The hCG system provided an opportunity to compare a theoretical prediction of structural epitopes with experimentally determined functional epitopes derived from mutagenesis studies.
As stated previously hCG is a hetero-dimer, and hence if the 3-dimensional co-ordinates of the β subunit (PDB entry 1hrp, (Lapthorn et al., 1994)) were used in isolation, as the basis for a prediction, the surface of the subunit would contain a dimer interface region as well as structural epitopes. To avoid confusion between the two types of protein-protein interaction, calculations have been made on the dimer structure.

The PATCH software was used to analyse the surface of the hCG dimer (α and β subunits) using a patch size of 15 residues. The combined score was calculated as indicated in equation 6.1 and a 10% cut-off was used for the selection of best patches with the highest combined score. The profiles of the patches on the surface of the β subunit (residues 1-112) are shown in Figure 6.11. There were 7 patches selected as best patches from the combined score, and these were centred on residues 5, 23, 72, 73, 74, 75 and 86. Some of these patches were overlapping, and the last profile (Figure 6.11h) shows the frequency of surface residues in these best patches. This highlights 2 main segments as putative epitope sites; residues 19-27 and 70-78. In addition there are 4 other segments (2-7, 9-14; 60-64; 84, 86-89; 103, 105-106) and 1 isolated residue (32) which also occurred within the best patches.

Panels of monoclonal antibodies (Mabs) have been used to identify epitope sites on the surface of the β subunit in hCG (Jackson et al., 1995). Specific site mutations in βhCG have been found to affect the binding of some monoclonal antibodies (Jackson et al., 1995; T. Lund, personal communication); and these include the 3 substitutions shown in Table 6.7.

<table>
<thead>
<tr>
<th>Mutant</th>
<th>Amino Acid Substitution</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Lys → Asn</td>
</tr>
<tr>
<td></td>
<td>Glu → Arg</td>
</tr>
<tr>
<td></td>
<td>Gly → Glu</td>
</tr>
<tr>
<td></td>
<td>Pro → His</td>
</tr>
<tr>
<td></td>
<td>Val → Tyr</td>
</tr>
<tr>
<td></td>
<td>Arg → Glu</td>
</tr>
<tr>
<td></td>
<td>Arg → Ser</td>
</tr>
<tr>
<td></td>
<td>Gly → His</td>
</tr>
<tr>
<td></td>
<td>Val → His</td>
</tr>
</tbody>
</table>

Table 6.7: Substitutions made in the β subunit of hCG and their position (T. Lund, personal communication)
was 15 residues and the cut-off for the definition of best patches was 10%.

**Figure 6.1:** Patch profiles for human chorionic gonadotropin (hCG). The PATCH software.
With a panel of 13 Mabs, mutant-3 terminated binding in 9, mutant-2 terminated binding in 3 and reduced binding by 75% in another, and mutant-1 terminated binding in 1 and reduced binding by 40% in another (T. Lund, unpublished results). It was observed that the 2 main segments identified as putative epitopes from the prediction, overlap with the epitope sites identified from the mutagenesis studies. However the predictive tool also identified additional sites that were not substantiated by the experimental work.

As stated previously, the aim of the experimental work was to identify cross-reactive epitopes in hCG and LH. The LH structure has not been solved but its 3-dimensional structure was modelled on the hCG structure by A. Lapthom and co-workers using SCULPT (Surles et al., 1994). A prediction was conducted to find ‘equivalent patches’ on the surfaces of hCG and LH, compare the characteristics of each, and predict putative epitopes. A patch was defined on hCG and the 6 parameters (solvation potential, residue interface propensity, hydrophobicity, planarity, accessible surface area, and protrusion) calculated. An ‘equivalent’ patch, was then defined in the LH model by selecting the aligned residues from the sequence alignment in Figure 6.12. Patch profiles were then produced showing the patch parameters, including the combined score, and the best patches selected for both structures, on one profile (Figure 6.13). This highlighted those patches selected as putative epitopes in each individual protein, and those in both structures (possibly indicating cross reactive epitopes).

![Sequence alignment between hCG and LH](image)

**Figure 6.12:** Sequence alignment between hCG and LH. Only those amino acid residues that differ between the 2 sequences are indicated for the LH sequence (T. Lund, personal communication).
Figure 6.13: Comparative patch profiles between hCG and LH β subunit structures. On the combined score profile (g) the short dashed lines (---) indicate a best patch selected on hCG, the long dashed lines (-----) a best patch on LH, and a solid line (—) a best patch on both structures. The frequency profile (h) indicates the residues involved in the hCG best patches (black bars) and the LH best patches (grey bars).
As shown previously, the best patches on the hCG structure are centred at residues 5, 23, 72, 73, 74, 75 and 86. Using equivalent patches on the LH structure the 3 best patches on LH (which constitute the 10% of patches with the highest combined score) were centred at residues 22, 73 and 75. There was a large overlap between the patches centred on residues 22 and 23, and those centred on residue 72, 73, 74, 75. These results suggest that the putative epitopes centred at approximately 23 and at 73 might be cross reactive between hCG and LH, and that the putative epitope centred at residue 5 (which includes residues 2-7, 9-14, 103, 105-106) and that centred at residue 86 (which includes residues 10-13, 32, 60-64, 84, 86-89) could be specific to hCG. The experimental work using panels of Mabs is still continuing, and hence it will remain to be seen if this comparative prediction is correct.

6.5 Conclusions

Applications of analysing and comparing patches of residues on the surfaces of proteins have been described in this chapter. The method has been applied to the analysis of related proteins in different oligomeric states, crystal contacts, and the prediction of structural epitopes on antigenic antigens.

Haemoglobin (Hb) was examined as an example of a protein family that has members in more than one oligomeric state. It was found that the interface within the arcid clam Hb dimer is more hydrophilic than the AB and AD interfaces in the human Hb tetramer, which were relatively hydrophobic. The haem groups are a significant part of the Hb dimer interface, and their role in the co-operative mechanism of oxygen binding could explain the hydrophilic nature of this interface.

The structure of sickle cell haemoglobin (HbS) was used to analyse crystal contacts between protein molecules. The contacts within the deoxy-HbS single crystal are of particular relevance as it has been proposed that the packing arrangement of HbS in polymers is similar to the packing observed in the crystal (Carragher et al., 1988; Eaton and Hofrichter, 1990). The interface area between the tetramers was small and the complementarity factor large compared to interfaces in the functional oligomers analysed in Chapter 3 (section 3.4). The predictive tool identified the Val-β6 substitution at the centre of one putative interface site but also included many additional sites on the surface of the tetramer. The issue of crystal contacts is important, and one that needs to be addressed further. In some instances a structure is solved that includes
more that one molecule in the asymmetric unit, but for which the functionally active state is unknown. In such structures it is important to define the inter-molecular interactions as crystal contacts or protein-protein interactions within a functional oligomer. The characterisation that has been conducted on various types of protein-protein interaction (Chapter 3) revealed a range of values for multiple characteristics observed for protein-protein contacts that occur within functional protein complexes. With such data it should be possible to derive criteria that could be used to distinguish between crystal contacts and interactions within functional oligomers.

The characterisation of a series of protein antigen interfaces provided the basis for the prediction of structural epitopes on the surface of 2 antigenic proteins. The predictions proved successful for the lysozyme structure, with the residues in putative epitopes corresponding with those in epitopes from known structures. The work on lysozyme represented the testing of the prediction method with evidence from crystal structures. The prediction of epitopes of hCG represented the comparison of predicted structural epitopes with functional epitopes derived from mutagenesis studies. Again, there was considerable correspondence between residues in predicted epitopes on the hCG β subunit and residues constituting functional epitopes. This latter system was also used to compared surface patches between related proteins. The analysis revealed putative epitopes on hCG that were also present on the LH molecule, which could represent cross-reactive structural epitopes. In addition 2 sites (patches centred on residue β5 and β86) were only predicted on the hCG molecule, and hence they could represent epitopes specific to hCG. The possibilities of producing a contraceptive vaccine from this system are significant, and theoretical predictions that can identify specific areas to concentrate experimental studies could reduce the time taken to produce such a vaccine.
Conclusions

Protein-protein interactions are fundamental to so many biological processes that efforts to understand their complexities have been intense. This work was conducted to update the current knowledge of the characteristics of protein-protein interactions. With an increased knowledge of the chemical and physical nature of these interactions a new method of protein interface prediction is presented.

The characterisation of protein interfaces within homo-dimeric proteins consolidated the results previously presented by Argos (1988), and Janin et al. (1988). In the current work the interfaces were shown to be hydrophobic patches on the surfaces of proteins, generally globular and planar in shape, which represent discontinuous segments of the polypeptide chain. Whilst the hydrophobic nature of the interactions was apparent, significant numbers of inter-subunit hydrogen bonds were also observed. This is consistent with the hypothesis that hydrophobic interactions are the major component of the energy of association (Chothia and Janin, 1975; Janin and Chothia, 1990), whilst hydrogen bonds confer specificity (Fersht, 1984).

The characterisation of homo-dimer interfaces presents just part of the picture of protein-protein interactions. The analysis of interfaces from distinct types of protein-protein complexes (e.g. homo-dimers, homo-trimers, enzyme-inhibitors and antibody-protein complexes) is also presented. These complexes have been classified as permanent or ‘optional’; and the constraints imposed upon each result in different interface characteristics. The requirement for the molecules of ‘optional’ complexes to exist as independent entities results in less hydrophobic interfaces, compared to those which exist permanently in a complexed state. This is also reflected in the greater number of inter-molecular hydrogen bonds observed in such complexes. The ‘optional’ complexes also have smaller and less well packed interfaces. The selective evolutionary pressure exerted upon different complexes was reflected in the complementarity observed at their interfaces. The antigen-protein complexes, which constitute structures with a very short evolutionary history, have less complementary interfaces than those of the homo-
proteins or enzymes, which have been exposed to selective pressures over a much longer time period. These results are consistent with those of Lawrence and Colman (1993).

Conformational change upon binding was a further aspect addressed in the comparative analysis of different types of protein-protein complex. The occurrence and extent of such changes are still unclear (Huber, 1979; Wilson and Stanfield, 1994), and currently there are few proteins that have been structurally determined (by X-ray crystallography or nuclear magnetic resonance) before and after complexation. The simple method presented here distinguished between gross conformational changes, which occurred in some homo-dimers and permanent hetero-dimers, and the much smaller changes (possibly side-chain movements) observed in the enzyme-inhibitor complexes. The future understanding of the occurrence and significance of conformational changes on binding relies upon the determination of many more protein structures before and after complex formation.

The results presented on the comparative characterisation of both permanent and ‘optional’ complexes were derived from a relatively small data set of proteins that needs to be extended. However, at present, such extensions are hampered by the lack of information on the in vivo complex-status in the current PDB file entries; so for example, extracting proteins in a specific oligomeric state is a very labour-intensive process. However, as the number of protein structures deposited in the PDB grows (and the PDB’s text file format is enhanced for computer readability) it will be possible to include higher order complexes such as tetramers, pentamers and hexamers. The inclusion of such complexes will make it possible to analyse differences in the way proteins organise themselves into higher-order associations; and to determine if interface characteristics are specific to different types of association (e.g. isologous or heterologous). Computational methods of symmetry evaluation should also be derived for such oligomers, as the presence of symmetry imposes additional constraints upon the structure and function of protein associations.

Empirical calculations have been used to quantify the contributions of hydrophobic and electrostatic interactions to the energy of association (e.g. Chothia, 1974; Chothia and Janin, 1975; Janin and Chothia, 1990). More recently a continuum energy model for protein-protein interactions has been proposed which combines factors for electrostatic and hydrophobic components and the loss of conformational energy on binding (Jackson and Sternberg, 1995). The problem of quantifying the components of the energy of association is complex, and one
made more so by the fact that binding constants for associations are difficult to determine experimentally, and are almost never deposited in the PDB with the co-ordinates. The derivation of binding constants, and the analysis of their relationship with interface characteristics, is essential if the kinetics and the thermodynamics of complex formation are to be understood.

A new software tool, PROFACE, has been developed to characterise protein-protein interactions. The program, which operates on the 3-dimensional co-ordinates of protein complexes, will prove a useful method of evaluating new complexes, providing information on the size, shape and chemical nature of the interactions. Such approaches will increase the knowledge of the rules that govern molecular recognition, and improve the ability to design small novel molecules to disrupt protein associations. In addition, it may allow directed mutations of specific residues on the surfaces of proteins to be used for the design of novel complexes.

The prediction of protein-protein interaction sites on protein surfaces can be divided into 2 main areas; the docking of 2 proteins of known structure; and the identification of putative interaction sites upon the surface of an isolated protein, known to be involved in protein-protein interactions, but where the structure of the complex is unknown. The current work has concentrated upon the latter aspect of the prediction problem, introducing a method to define patches of surface residues on a given isolated protein. Six chemical and physical parameters have been identified, which differentiate a known interface site from all other surface patches. The parameters, solvation potential, residue interface propensity, planarity, hydrophobicity, accessible surface area and protrusion, have different relative values dependent upon the type of protein-protein complex. The interfaces in the homo-dimers and the inhibitors were found to be amongst the most accessible and the most protruding, whilst the interfaces in the enzymes were amongst the least accessible and the least protruding of all the surface patches. These differentiating characteristics were used as the basis of a new predictive algorithm, PATCH. The algorithm was used successfully to predict putative interaction sites on the surfaces of isolated protomers from 20 homo-dimers, and enzymes from 3 enzyme-inhibitor complexes.

The PATCH algorithm is based on the definition and comparison of surface patches at the Cα residue level. In all docking and prediction methods there is an equilibrium to be reached between the accuracy of the method and the time taken. With the current method more accurate predictions could be achieved if the patches were re-defined in terms of surface atoms; and each
of the 6 differentiating parameters calculated at the atom level. In addition, a definition of surface patches that allowed discontinuous patches could increase prediction accuracy. However this would result in a combinatorial explosion, with hundreds of thousands of patches being defined. Restricting the patch definition to contiguous surface patches defined at the residue level, as at present, reduces the combinatorial problem, but at the expense of the accuracy of the predictions. The predictive method, presented here, is an extension to the method of Young et al., (1994), in that it uses multiple parameters to rank surface patches. Such an approach is useful for identifying candidate interface residues, which can be mutated experimentally, and tested for the effect on complex formation.

Applications of defining and comparing patches of residues on the surfaces of proteins were also presented. A small number of protein families (defined using the method of Orengo (1992; 1994)) have been identified that contain proteins in more than one oligomeric state. The dimeric and tetrameric structures of haemoglobin (Hb) were examined as an example of this phenomenon, using the PATCH software. It was found that the interface within the arcid clam Hb dimer was more hydrophilic than 2 of the 3 interfaces in the human Hb tetramer, which were relatively hydrophobic. The haem groups form an integral part of the interface in the Hb dimer and its hydrophilic nature is linked to the way in which the dimer achieves its co-operative oxygen binding function. The characteristics of crystal contacts in sickle cell haemoglobin (Hbs) have also been analysed. The contacts within the deoxy-Hbs single crystal, which contains 2 tetramers, are of particular relevance, as it has been proposed that the packing arrangement of Hbs in polymers in vivo is very similar to the packing observed in the crystal (Carragher et al., 1988; Eaton and Hofrichter, 1990). The interface area between the tetramers was small and the complementarity factor large, compared to interfaces observed in functional oligomers and complexes. The PATCH predictive algorithm, identified the Val-β6 substitution as a component of one putative interface, but also identified many additional sites on the surface of the tetramer.

Crystal contacts represent an important subset of protein-protein interactions, and it is expected that further quantitative analysis of such contacts will reveal their characteristics to be at the extremes of those observed for interactions in functional complexes. The ability to distinguish consistently between crystallographic contacts and those occurring within biologically functional complexes would be advantageous. In some instances a structure is solved which includes more than one molecule in the asymmetric unit, but for which the functionally active state is
unknown. The characterisation conducted here, on various categories of protein-protein interaction, revealed a range of values for multiple parameters observed for protein-protein contacts occurring within functional protein complexes. It should be possible to derive criteria that can be used to distinguish between crystal contacts and those within functional complexes.

The characterisation of a series of protein antigen interfaces provided the basis for the prediction of structural epitopes on the surface of lysozyme and the β subunit of human chorionic gonadotropin. Using the PATCH algorithm the predictions proved successful for lysozyme, with good correspondence observed between predicted epitopes and those from known crystal structures. In the βhCG structure, good correspondence was achieved between residues predicted in structural epitopes, and those in functional epitopes derived from the mutagenesis studies of Lund (unpublished results). Surface patches have also been compared between β subunit of hCG and the β subunit of luteinizing hormone, with which it shares both sequence and structural homology. The comparison revealed putative cross-reactive epitopes, and 2 specific to the hCG structure. The possibilities of developing a contraceptive vaccine from this system are significant and theoretical predictions identifying specific sites for mutagenesis studies, could aid the development process.

The prediction algorithm discussed in this work could be improved to include a patch definition and comparisons at the atom level. Alternatively the combined score attributed, to a patch, could be assigned to the central residue of that patch, and a putative interaction site selected to include those residues with the highest combined scores. The conformational changes on binding observed to occur in some complexes (e.g. see review by Wilson and Stanfield (1994)) make prediction of interaction sites on isolated structures in the bound conformation, to some extent, academic. Many more structures need to be solved to provide additional information on structures before and after complex formation, so that theoretical procedures, such as those described here, can be rigorously tested over many structures.

The protein associations analysed and predicted in the current work represent relatively simple systems. There are a vast number of highly complex protein associations that have still to be solved. One process that involves such complex protein associations is one that is vital to the existence of proteins themselves, the process of protein synthesis and folding. Current interest is centred on the ribosome and molecular chaperones, both structures essential for the production of correctly folded proteins. The ribosome of *E.coli* comprises 55 proteins and 3 ribosomal
RNA's packed into 30S and 50S subunits (Zimmermann, 1995). Recently the use of electron microscopy has revealed a 25Å resolution structure of the 30S and 50S complex in the *E.coli* ribosome (Frank *et al.*, 1995). This remarkable view of the molecule revealed an irregular structure with lobes, channels, bridges and tunnels, which all potentially have functional significance. For example, the 2 ribosomal subunits that are joined by 5 bridges, enclose a space that could accommodate messenger RNA and transfer RNA during protein synthesis (Frank *et al.*, 1995). Molecular chaperones have been defined as proteins that assist in the correct non-covalent assembly of other polypeptides containing structures *in vivo*, but are not themselves components of these assembled structures when they perform their normal biological function (Ellis, 1993). Chaperones effectively protect newly synthesised proteins during folding and transport (see review by Martinus *et al.*, 1995). Proteins are transported across membranes in a largely unfolded state, and this unfolding and refolding, is thought to be mediated by molecular chaperones (Stuart *et al.*, 1994).

The ribosome and molecular chaperones are integral parts of protein synthesis and the folding process, and hence the interactions that occur within them are of particular relevance. The components of the ribosome must provide highly specific binding sites for the growing peptide chain (Zimmermann, 1995). Molecular chaperones, such as GroEL (see Hartle *et al.*, 1994), not only exhibit complex protein-protein interactions within their oligomeric conformations, but must also provide binding sites for partially folded proteins. These proteins, and the interactions they mediate, are of fundamental importance; and their structure and mechanisms are only now beginning to emerge.

The role of protein-protein interactions in a disease process was discussed in Chapter 5, in the analysis of sickle cell haemoglobin. In sickle cell anaemia, the physiological symptoms are caused by the aggregation of the sickle cell haemoglobin molecules into long fibres. The aggregation of proteins is also a feature of another disease process, Alzheimer's Disease (AD), in which there has recently been renewed interest. Alzheimer's is characterised by an abnormally high level of brain lesions, termed senile plaques, which are composed of associations of β-amyloid proteins (Sisodia and Price, 1995). The events that lead to the proliferation and subsequent aggregation of these proteins are, as yet, unclear. However it is known that β amyloid is derived from an amyloid precursor protein, and enzymes that are involved in the secretion of the β-amyloid protein from the precursor are the target for the rational design of anti-AD drugs (Ashall and Goate, 1994).
The recent discoveries of complex protein associations, and the many examples of protein-protein complexes discussed throughout this thesis, give a clear indication of their importance in biological processes. Computational tools designed for interface characterisation and prediction will aid the understanding of the recognition process as new protein-protein associations emerge.

In nature many of the most important biological functions involve huge multi-component complexes, such as the ribosome, and the first steps are only just being taken to understand the principles of molecular recognition in simple systems. However the implications of a better understanding of protein-protein interactions for the design of new therapeutics and environmental products are apparent to all.
REFERENCES


Protein-Protein Interactions

References


Protein-Protein Interactions

References


Protein-Protein Interactions

References


Protein-Protein Interactions

References


175
Protein-Protein Interactions

References


Protein-Protein Interactions

References


A1. Chapter 1: Section 1.3.2 Protein Stoichiometry

**Isologous and Heterologous Associations**

The definition of isologous and heterologous associations proposed by Monod *et al.*, (1965) have been addressed in Chapter 1. In isologous associations the binding set of each protomer is ‘covered’ by the equivalent binding set on the other protomer, hence these associations tend to lead to finite closed structures. Heterologous associations can involve multiple binding sets on a single protomer which can lead to infinite open structures (Matthews and Bernhard, 1973). If an oligomer has an odd number of equivalent protomers, then the associations between them must be heterologous. However if an oligomer has an even number of equivalent protomers, the associations can be either isologous, heterologous or a mixture of the two (Matthews and Bernhard, 1973).

A2. Chapter 1: Section 1.5.3 Shape Complementarity

**Packing Density**

Protein interiors have been found to be closely packed (Richards, 1974; Richards, 1977; Richards and Lim, 1993). However the packing is not perfect and internal cavities have been detected in protein interiors (Rashin *et al.*, 1986; Williams *et al.*, 1994).

A3. Chapter 2: Section 2.3. Non-homologous Data Set of Protein Dimers

**Secondary Structure Alignment Program**

In the secondary structure alignment program (SSAP) of Taylor and Orengo (1989), used in Chapter 2 to calculate a non-homologous data set, a local structural environment is defined for each residue in a protein. Residues in two proteins are then matched by comparing their structural environments and an alignment of the proteins generated by dynamic programming methods. The SSAP algorithm assigns scores of between 0 and 100 to proteins on the basis of how well they align structurally. The higher the score the greater the similarity in terms of the number of residues equivalent and the correspondence of local structural environments. A score of 100 is assigned if one protein is structurally aligned with itself. Proteins are grouped
into structural families at two levels; the 80 structure level (SSAP score $\geq 80$), at which proteins have significant structural similarity and the 70 structure level (SSAP score between 70 and 80) at which proteins show a greater variation in loop structures and the orientation of secondary structures. There is a clear correlation between the SSAP scores and the rms deviation of the aligned structures; the higher the SSAP score the lower the rms value (Orengo et al., 1992). A SSAP score of 80 correlates approximately to an rms deviation of 3Å.

The non-homologous data sets used in this thesis comprised structures with a sequence identity of $<35\%$ and a SSAP score of $\leq 80$, which correlates approximately to an rms deviation, obtained from structural superposition, of $>3Å$.

A4. Chapter 2: Section 2.5.5 Atom Distributions and Hydrophobicity

Further Analysis of Atom Distributions

The ASA contributed by polar (oxygen and nitrogen) and non-polar (carbon and sulphur) atoms has been calculated for each subset of protein atoms (interior, interface and exterior).

<table>
<thead>
<tr>
<th>Atom Sets</th>
<th>% Contribution to ASA of Non-Polar Atoms</th>
<th>Polar Atoms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interior</td>
<td>78.17 (68.94)</td>
<td>21.83 (33.06)</td>
</tr>
<tr>
<td>Interface</td>
<td>65.06</td>
<td>34.94</td>
</tr>
<tr>
<td>Exterior</td>
<td>57.96</td>
<td>42.04</td>
</tr>
</tbody>
</table>

Table A1: ASA contributions by polar and non-polar atoms in the three atom sets. The figures in brackets for the interior subset indicate the results when the interior subset is redefined as in the method of Janin et al. (1988) (see text).

The results (Table A1) clearly indicate that the interface atoms have properties that are intermediate between the interior and the exterior; containing 65.06% non-polar atoms compared to 78.17% and 57.96% in the interior and exterior respectively. This is in contrast to the results of Janin et al (1988), who found that the interface atoms were more non-polar (65%) than both the protein exterior (57%) and interior (58%). The values for the interface and exterior sets are comparable to those shown in Table A1, but the value for the interior set is much lower (58% (Janin et al., 1988) compared to 78% (Table A1)).
Janin et al. (1988) defined the interior set as those residues buried within the protein subunits and those buried between subunits. The percentage ASA contributed by non-polar and polar atoms for an interior set redefined on the basis of that used by Janin et al. (1988) has also been calculated in the current work, and the results shown as bracketed values in Table A1. Whilst the redefined interior set has a much lower percentage of non-polar atoms (68.94%) it is still greater than the 58% calculated by Janin et al., (1988). Even using this redefined value for the protein interior of 68.94%, the interface still has a percentage value (65.06%) which is lower than this.

Hence these additional results confirm that the interface has hydrophobic properties intermediate between the interior and exterior; and are not more hydrophobic than the protein interior as indicated by Janin et al., (1989). The results presented here confirm those of Argos (1988) and Korn and Burnett (1991), who also found the hydrophobicity of the interface to be intermediate between the protein interior and exterior.

A5. Additional References


PROTEIN–PROTEIN INTERACTIONS: A REVIEW OF PROTEIN DIMER STRUCTURES

SUSAN JONES and JANET M. THORNTON

Biomolecular Structure and Modelling Unit, Department of Biochemistry and Molecular Biology,
University College, Gower Street, London WC1E 6BT, U.K.

CONTENTS

I. INTRODUCTION 31

II. CLASSIFICATION OF MULTIMERIC PROTEINS 33

III. THE NON-HOMOLOGOUS DATA SET OF PROTEIN DIMERS 33
   1. Generation of Co-ordinates 33
   2. Validation of Co-ordinates 34

IV. DEFINITIONS OF PROTEIN–PROTEIN INTERFACES 35
   1. The Residue Level Definition 35
   2. The Atom Level Definition 36

V. CHARACTERIZATION OF THE DIMER INTERFACES 36
   1. Accessible Surface Area 36
   2. Sphericity of the Dimeric Proteins 36
   3. Shape of the Dimer Interfaces 42
   4. Amino Acid Composition 43
   5. Distribution of Atom Types 45
      (a) Atom distributions and hydrophobicity 46
   6. Secondary Structural States 47
      (a) Interface classification by secondary structure 49
   7. Hydrogen Bonding 50
   8. Salt Bridges 50
   9. Disulphide Bonding 50
   10. Interface Segmentation 52
   11. Gap Volumes and Surface Complementarity 53
   12. Protrusion of Interface Residues from the Molecular Surface 57
   13. Flexibility of Residues and Atoms in the Dimer Interface 58
      (a) Comparisons of atomic temperature factors within dimeric proteins 61

VI. A TOOL FOR THE ANALYSIS OF INTERFACES IN MULTIMERIC PROTEINS 62

VII. PROTEIN–PROTEIN INTERACTIONS IN DIMERS: A SUMMARY 63

VIII. DISCUSSION 63

REFERENCES 64

I. INTRODUCTION

Protein–protein interactions form the basis of the quaternary structure of multimeric proteins, and represent one of the most complex levels of structural organization in biological molecules. Multimeric proteins are found in every cellular location, including the cell organelles, the cytosol and the cell membranes (Hardy et al., 1988). They are of great importance biologically as they mediate biochemical phenomena such as enzyme cooperativity and signal transduction. Their importance in biological systems is reflected in the volume of research conducted in the field of protein associations.

It was the pioneering dissociation experiments of Svedberg (1926, 1929) which established that some proteins exist as complex associations. Combinations of association and hybridization techniques have led to the discovery of large numbers of proteins possessing...
quaternary structures. The early analysis of multimeric proteins concentrated upon two aspects: stoichiometry and stereochemistry. The spatial arrangements of the subunits within complexes are inseparable from discussions of symmetry in oligomeric structures. An initial understanding of the importance of symmetry was derived principally from the comparative studies of myoglobin (Kendrew et al., 1960) and haemoglobin (Perutz and Rossman, 1960). The importance of symmetry, in terms of protein structures, was also introduced in the theoretical model of Monod et al. (1965), on the allosteric effects of enzymes. From these studies, and the increasing number of protein structures solved by X-ray crystallography, it has been found that many subunits of oligomeric proteins are organized into stable arrays with high symmetry (Gust and Dirks, 1981).

The study of multimeric proteins has developed through a combination of experimental analysis and theoretical models of subunit–subunit associations. Current work in this area concerns the application of computer algorithms to structural analysis. The deposition of the three-dimensional co-ordinates of protein structures (solved mainly by X-ray crystallography, as well as nuclear magnetic resonance) in the Brookhaven Protein Data Bank (Bernstein et al., 1977), has permitted the analysis of relatively large numbers of multimeric proteins for the purpose of structural characterization of protein–protein interactions.

Protein–protein association involves the specific complementary recognition of two macromolecules to form a stable assembly (Duquerroy et al., 1991). Fundamental to the stabilization of protein association is the hydrophobic interaction (Chothia and Janin, 1975). The term hydrophobic interaction is used to describe the gain in free energy which occurs when non-polar residues of proteins associate in an aqueous environment (Kauzmann, 1959). The process of folding and protein–protein aggregation reduces the surface area in contact with water. This is the structural basis of the hydrophobic effect. The folding of polypeptide chains and aggregation of subunits buries the hydrophobic residues of the proteins, and hence minimizes the number of thermodynamically unfavourable solute–solvent interactions. The quantitative evaluation of exactly how much hydrophobic interactions contribute to the stabilization of protein–protein associations is controversial (Privalov and Gill, 1988; Lessor and Rose, 1990).

The second fundamental aspect of protein–protein interactions is complementarity. This applies not only to the shape but also the charge of groups on the surface. The hydrogen bond (a polar interaction between donor and acceptor electronegative atoms) is an intrinsic component of protein–protein interactions. It has been suggested that whilst hydrophobic forces drive protein–protein interactions, hydrogen bonds and salt bridges provide the specificity (Fersht, 1984). The structural shape of the macromolecular interfaces is an additional component of the complementarity factor. A theorem for self-complementarity of surfaces has been defined (Morgan et al., 1979) and more recently, a new shape correlation statistic ($S_c$) has been defined to measure packing and shape complementation in protein–protein complexes (Lawrence and Colman, 1993).

This brief introduction has emphasized that protein–protein interactions are complex, with many fundamental factors contributing to a stable association. The work that will be described in this review concentrates on protein–protein interactions in just one structural system: dimeric proteins. The study has been conducted with the aim of characterizing protein–protein interfaces; with the long term view of using the structural characteristics to predict potential interface sites. This article is divided into several sections, beginning with analysis of the numbers of biological multimeric structures recorded in the Brookhaven Protein Data Bank (PDB). It is followed by a discussion of the methods used to generate a non-homologous data set and to define the protein–protein interfaces. The main section describes the characterization of the interfaces, including analysis of the residue and atom composition, the secondary structure, segmentation, planarity and ionic interactions. The underlying theme has been the differentiation of the interface from the remainder of the protein surface. To this end the characteristics of the interface have been compared with those of the protein surface and the protein interior, where applicable. The article concludes with a discussion of the implications of this work and how it might be applied in the prediction of protein–protein interfaces.
II. CLASSIFICATION OF MULTIMERIC PROTEINS

This current study is concerned with the classification of proteins into their biological multimeric states. This process of classification was conducted by consulting the individual entry for a protein in the PDB. If no information could be found there, the original journal articles in which the structure was published were consulted. This study is based on the July 1993 version of the PDB that contains 974 protein structures, determined experimentally. These proteins have been classified with reference to the total number of subunits in their biological multimeric state (Fig. 1). Four proteins were unclassified as they had no information pertaining to their biological multimeric state in their PDB entries, and the references cited for each are, at the time of writing, unpublished.

The distribution of biological multimeric states in the PDB shows the prevalence of dimers and tetramers, and the relatively small number of odd numbered multimers. Such a distribution is consistent with previous observations on cumulative crystallographic data of oligomeric proteins (Matthews and Bernhard, 1973; Klotz et al., 1975).

III. THE NON-HOMOLOGOUS DATA SET OF PROTEIN DIMERS

The protein dimers in the PDB were chosen as the starting point for the analysis of protein–protein interfaces in multimeric proteins. They constituted the largest set of proteins in one biological multimeric state, with 148 entries in the July 1993 release of the PDB. These entries represent 81 different protein structures, as 30 dimers had multiple file entries in the PDB. These multiple entries record proteins crystallized in different solutions, in different crystal classes, or with different inhibitors bound.

In this analysis only non-homologous proteins were selected for the data set. To define non-homologous proteins both the sequence relationship and the structural similarity have been considered, as often homology can only be detected from the three-dimensional structure. Proteins were selected for inclusion on the basis that they had a sequence identity of <35% and were structurally different. The structural similarity of the proteins was measured using a method of direct structural alignment (SSAP, Taylor and Orengo, 1989). Proteins were selected for the data if they had a SSAP score of ≤80. In the process of selection, only dimers with homologous subunits were considered. This selection resulted in a non-homologous data set of 32 protein dimers, each belonging to a different homologous protein family (e.g. only one immunoglobulin was included) (Table 1).

1. Generation of Co-ordinates

Of the 31 dimers solved by crystallography, the asymmetric units of 14 protein crystals contained only one subunit. For these structures the rotation and translation matrices
Table 1. The Non-homologous Data Set of 32 Protein Dimers

<table>
<thead>
<tr>
<th>Code</th>
<th>Protein</th>
<th>Source</th>
<th>Resolution</th>
</tr>
</thead>
<tbody>
<tr>
<td>lct</td>
<td>Cardiotoxin</td>
<td>Naja mossambica mossambica</td>
<td>2.5</td>
</tr>
<tr>
<td>fc1</td>
<td>FC fragment (immunglobulin)</td>
<td>Human pooled serum</td>
<td>2.9</td>
</tr>
<tr>
<td>il8</td>
<td>Interleukin</td>
<td>Human neutrophil recombinant in E. coli</td>
<td>NMR</td>
</tr>
<tr>
<td>lmsb</td>
<td>Mannose binding protein</td>
<td>Rat expressed in E. coli</td>
<td>2.3</td>
</tr>
<tr>
<td>1phh</td>
<td>P-hydroxybenzoate dehydrogenase</td>
<td>Pseudomonas fluorescens</td>
<td>2.3</td>
</tr>
<tr>
<td>1pp2</td>
<td>Phospholipase</td>
<td>Western diamond black rattlesnake</td>
<td>2.5</td>
</tr>
<tr>
<td>1ppp</td>
<td>Inorganic pyrophosphatase</td>
<td>Baker's yeast</td>
<td>3.0</td>
</tr>
<tr>
<td>lsdh</td>
<td>Haemoglobin (clam)</td>
<td>Arcid clam</td>
<td>2.4</td>
</tr>
<tr>
<td>lugu</td>
<td>Uteroglobin</td>
<td>Rat female genital tract</td>
<td>1.34</td>
</tr>
<tr>
<td>lvsg</td>
<td>Variant surface glycoprotein</td>
<td>Trypanosoma brucei</td>
<td>2.9</td>
</tr>
<tr>
<td>lypi</td>
<td>Triose phosphate isomerase</td>
<td>Yeast</td>
<td>1.9</td>
</tr>
<tr>
<td>2cyc</td>
<td>Citrate synthase C</td>
<td>Rhodospiillum molischianum</td>
<td>1.67</td>
</tr>
<tr>
<td>2ct</td>
<td>Citrate synthase C</td>
<td>Phage 434</td>
<td>2.5</td>
</tr>
<tr>
<td>2gn5</td>
<td>Gene 5 DNA binding protein</td>
<td>Filamentous bacteriophage FD(M13)</td>
<td>2.3</td>
</tr>
<tr>
<td>2orl</td>
<td>434 Repressor</td>
<td>Human myeloma patient RHE urine</td>
<td>1.6</td>
</tr>
<tr>
<td>2hre</td>
<td>Bence-Jones protein</td>
<td>Rhodospiillum rubrum expressed in E. coli</td>
<td>2.3</td>
</tr>
<tr>
<td>2rue</td>
<td>ECO RV endonuclease</td>
<td>E. coli recombinant form expressed in E. coli</td>
<td>3.0</td>
</tr>
<tr>
<td>2sod</td>
<td>Superoxide dismutase</td>
<td>Bovine erythrocyte</td>
<td>2.0</td>
</tr>
<tr>
<td>2ssi</td>
<td>Subtilisin inhibitor (Streptomyces)</td>
<td>Streptomyces alborescens S-3253</td>
<td>2.6</td>
</tr>
<tr>
<td>2ssl</td>
<td>Tyrosyl transferase RNA synthase</td>
<td>Bacillus stearotherophilus NCA 1503</td>
<td>2.3</td>
</tr>
<tr>
<td>2tsc</td>
<td>Thymidylate synthase</td>
<td>E. coli</td>
<td>1.97</td>
</tr>
<tr>
<td>2trp</td>
<td>Trp repressor</td>
<td>E. coli</td>
<td>1.65</td>
</tr>
<tr>
<td>3aat</td>
<td>Aspartate aminotransferase</td>
<td>E. coli</td>
<td>2.8</td>
</tr>
<tr>
<td>3enl</td>
<td>Enolase</td>
<td>Baker's yeast</td>
<td>2.25</td>
</tr>
<tr>
<td>3gvp</td>
<td>Catabolite gene activator protein</td>
<td>E. coli</td>
<td>2.5</td>
</tr>
<tr>
<td>3gts</td>
<td>Glutathione reductase</td>
<td>Human erythrocyte</td>
<td>1.54</td>
</tr>
<tr>
<td>3isc</td>
<td>Isoeucinate dehydrogenase</td>
<td>E. coli</td>
<td>2.5</td>
</tr>
<tr>
<td>3isp</td>
<td>Iron superoxidase</td>
<td>Pseudomonas ovalis wild type</td>
<td>2.1</td>
</tr>
<tr>
<td>4mdh</td>
<td>Cytoplasmic malate dehydrogenase</td>
<td>Horse heart</td>
<td>2.9</td>
</tr>
<tr>
<td>5adh</td>
<td>Alcohol dehydrogenase</td>
<td>NTH strain of human immunodeficiency</td>
<td>2.0</td>
</tr>
<tr>
<td>5hvp</td>
<td>HIV protease</td>
<td>virus type I in E. coli</td>
<td>2.0</td>
</tr>
</tbody>
</table>

Each protein belongs to a different homologous protein family and within each dimer the two subunits are homologous.

provided in the PDB files were applied to produce the complete molecule, using an algorithm implemented by T. Flores.

2. Validation of Co-ordinates

The structure of the 14 dimers generated in the method described in Section III.1 were inspected visually using the graphics software Quanta (Quanta 4.0: Molecular Simulations Inc.). This ensured that the structures produced were valid structures: i.e. that the dimers generated had an area of interface between their subunits.

In addition to visual inspection, the generated structures were validated systematically by checking for the presence of atoms with overlapping van der Waals radii (vdr) (the radii used were those assigned by Chothia (1976)). The distance (d) between each atom in one subunit (p) and each atom in the second (generated) subunit (g) were calculated. If the equation

\[ d < vdr_p + vdr_g + 1 \, \text{Å}^2 \]  

were true then the two atoms were considered as overlapping. The 1 Å error margin was incorporated to account for small errors in the crystallographic co-ordinates. The number of overlapping pairs was calculated in each generated dimer and in each dimer in which both subunits were contained in the asymmetric unit. The number of overlapping atoms was calculated as a percentage of the number of atoms in each dimer (Fig. 2). The mean percentage number of overlapping atoms in the 14 generated dimers was 0.04% and in the 18 dimers that were complete in the asymmetric unit, 0.01%. The higher percentage number of overlapping atoms in the generated dimers may indicate inaccuracies derived from the refinement stage of the crystallographic process. In this study, it was considered that dimers
IV. DEFINITIONS OF PROTEIN-PROTEIN INTERFACES

For this review protein-protein interfaces have been defined at two levels, at the residue level and at the atom level. At both levels the definitions are based on the concept of solvent accessible surface area (asa). The native structure of proteins exists in the presence of water and the asa describes the extent to which a protein can form contacts with water. The asas of the dimers were calculated using an implementation of the Lee and Richards (1971) algorithm developed by Hubbard (1992a). With a probe sphere, of radius 1.4 Å, the accessible surface was defined as the surface mapped out by the centre of the probe as if it were rolled around the van der Waals surface of the protein. Both atomic and residue accessibilities for a given protein were evaluated. In addition, relative asas for each residue were calculated by summing atomic asas in a residue and comparing the totals to the asa that each of the 20 residues possess in a standard extended state. The states used were those with the residue X in a tripeptide ALA–X–ALA, where φ and ψ values of all residues are set to 140° and 135°, and ω = 180° (except for proline) (Hubbard, 1992b). The program was used to calculate the asas of atoms and residues in each dimer, and each subunit as a discrete structure.

1. The Residue Level Definition

On the basis of asa calculations the interface residues were defined as those residues with side chains possessing an asa that decreased by >1 Å² on dimerization. The 1 Å² error factor was used to account for small errors in the crystallographic co-ordinates and computational inaccuracies in the calculation of the asas.

The exterior and interior residue definitions were based on the relative asa of each residue, which range from 0% for residues with no atom contact with the solvent, to 100% for fully accessible residues. On this basis the exterior residues were defined as having relative accessibilities >5% and interior residues as those with relative accessibilities ≤5%. This 5% cut-off was devised and optimized by Miller et al. (1987), who used it to define residues buried in monomeric proteins.
The definition of the exterior/interior residues is based on relative accessibilities and the definition of the interface residues is based on absolute asa. Thus all interface residues were originally defined as interior or exterior residues. At this stage the subset of interface residues was excluded from the subsets of interior and exterior residues. This resulted in three discrete sets of residues for each of the dimers in the non-homologous data set.

2. The Atom Level Definition

The atom level definition is also based on Aasa. The interface was defined as those atoms that showed a decrease of 0.01 Å² in asa on dimerization. The interior was defined as those atoms with an atomic asa of zero, and the exterior as those atoms with an atomic asa of greater than zero. As with the residue level definition, the interface atoms were removed from the exterior and interior subsets, to produce three discrete sets of atoms. The interface atoms of 18 of the 32 dimers in the data set are shown as CPK models in Fig. 3.

V. CHARACTERIZATION OF THE DIMER INTERFACES

The interfaces of the 32 dimers have been analysed with respect to a number of characteristics identified as being potentially important for molecular recognition in multimeric proteins. The analyses use the interface definitions at both the atom and the residue level (Section IV).

1. Accessible Surface Area

The asa for the interface (A_i) of one subunit in each dimer was calculated as

\[ A_i = \frac{(A_{sa} + A_{sb}) - A_{sab}}{2} \]  

where \( A_{sa} \) and \( A_{sb} \) denote the asa of the 2 subunits dissociated and \( A_{sab} \) denotes the asa of the subunits associated in the dimer. The percentage asa buried, per subunit, on dimerization was calculated as

\[ \%A_i = \frac{A_i}{A_{sa}}. \]

The asa buried, per subunit, on dimerization ranged from 368.1 Å² in the 434 repressor (2 or 1) to 4746.1 Å² in citrate synthase (2cts). The percentage asa buried, per subunit, upon dimerization ranged from 6.5% in inorganic pyrophosphatase (Ipyp), to 29.4% in the trp repressor (2wpr).

There is a positive linear relationship between the interface asa (asa_i) and the molecular weight (M_r) of the protomer (Fig. 4). A regression line of the form \( y = ax \) fitted to the data gives the equation:

\[ A_{sa} = 0.065 M_r. \]

The approximate nature of the relationship is indicated by a very high estimated variance of 6.97 x 10^5 and a correlation coefficient (r) of data to model of 0.69.

The atoms in the interface were classified into non-polar and polar groups and the contribution of each to the asa of the interface calculated. This revealed that non-polar atoms contributed 68.1% of the asa buried and polar atoms contributed 31.9%. This is very similar to a previous study on αβ-horse oxyhemoglobin dimer (Chothia and Janin, 1975). In this protein it was found that non-polar atoms contributed 68% of the interface asa and polar and charged atoms contributed 32%.

2. Sphericity of the Dimeric Proteins

The folding of polypeptide chains and aggregation of subunits buries the hydrophobic residues of a protein and hence minimizes the number of thermodynamically unfavourable solute–solvent interactions. In terms of physical shape, a reduction of a protein's surface exposed to solvent is achieved by the protein adopting a structure that is spherical. A sphere has the lowest surface area to volume ratio, hence the globular nature of proteins.

In this work the globular (spherical) nature of the dimer is compared with that of the protomer. An approximation of the surface area to volume ratio was calculated for each
Fig. 3 (i).
Fig. 3 (iii).

Fig. 3. CPK models of 18 of the protein dimers in the non-homologous data set. One subunit of each dimer is shown and the interface atoms are coloured red and the remainder of the subunit atoms are coloured in blue. Each structure is labelled using its PDB code, for the full name of the structure refer to Table 1. Each subunit is viewed looking face onto the interface surface. The scales between the different structures are not comparable.
Fig. 7. CPK models of non-planar interfaces in protein dimers. In pictures 1 (a, b) and 2 (a, b) two subunits are shown: one subunit is coloured blue and one red. The interface atoms in each subunit are coloured differently, the atoms coloured green are the interface atoms in the 'blue' subunit and those coloured yellow are the interface atoms in the 'red' subunit. In pictures 1c and 2c only the interface atoms of the two structures are shown. (1) Aspartate aminotransferase mutant (3aat). (a) Dimer viewed along the subunit interface with the arm of the A subunit shown stretched around the B subunit. (b) Dimer viewed along the subunit interface with the arm of the B subunit shown stretched around the A subunit. (c) Dimer interface only shown, viewed along the interface. (2) Isocitrate dehydrogenase (3icd). (a) Dimer viewed along the subunit interface. (b) Dimer viewed looking down the subunit interface showing the two subunits twisted together at the top. (c) Dimer interface only shown, viewed along the interface.
Fig. 4. The relationship between the molecular weight of the protomer and the interface asa (per subunit). The solid line is the fitted line of the equation $y = 0.06x$ and the dashed lines are the 95% confidence limits of the individual $y$ values. The correlation coefficient ($r$) is 0.69.

Fig. 5. The accessible surface area/molecular weight ratio of 32 protein dimers and their protomers. The solid line traces the asa/mwt ratio of 1:1. All the data points fall below this line indicating that on changing from monomer to dimer the structures effectively reduce the amount of surface area exposed to solvent.

protein (as a dimer and as a protomer), using the ratio of asa to molecular weight. In all the proteins this ratio in the dimer was less than that of the protomer (Fig. 5). Thus in forming the dimer the protein achieves a more globular state than if the protein remained as a monomeric structure.
3. Shape of the Dimer Interfaces

The interfaces form a surface of contact between the two protomers of a protein. The shape (the relative length and breadth) and the planarity of this surface were analysed by calculating a plane of best fit through the three-dimensional co-ordinates of the atoms in the interface.

A program (implemented by R. Laskowski) was used to calculate the best fit plane by principal component analysis. The program calculates the equation for the best fit plane and the root mean square (RMS) of the plane (Fig. 6). In the data set 84.3% of the interfaces were relatively flat, with a RMS of ≤ 6 Å. This is consistent with the observations of Argos (1988) who found that 83% of the protein–protein interfaces he studied were flat. Argos (1988) considered that an overall flatness would be expected when considering the symmetry involved in the associations. Whilst the presence of symmetry does not dictate that the interface should be flat, it does require that any 'protrusion' is complemented by a 'hollow'. This requirement suggests that the gross surface of the interface can be approximated by a plane. However, in the current study, five proteins had comparatively high RMS values (> 6 Å) when planes were fitted. These proved to be proteins in which the two subunits were twisted together across the interface (e.g. 3icd), or proteins that had subunits with 'arms' apparently clasping the two halves of the structure together (e.g. 3aat); two examples can be seen in Fig. 7.

To give a measure of the shape of the interfaces, the standard deviations of the interface atom distributions from the centre of gravity of the best fit plane were calculated. The standard deviations in the x and y dimensions were compared (Fig. 8). In the data set 31 of the 32 dimers had interfaces that were in general circular: having standard deviations of the atom distributions in the x dimension (sdx) very close to those in the y direction (sdy). The ratios of sdy/sdx for 31 proteins lie between 0.48 and 1.00. The only exception is the variant surface glycoprotein (1vsg) which has an elongated interface, with the sdy being 0.25 times the sdx. The elongated interface of the variant surface glycoprotein (1vsg), reflects the elongated nature of the protein as a whole (the protomer has a sdy/sdx ratio of 0.33). Variant surface glycoprotein of Trypanosoma brucei forms a coat on the surface of the parasite (Freymann et al., 1990) and thus its shape may reflect its function. It was also found that, for the 32 dimers, the ratio of sdy/sdx is independent of the size of the protein (correlation coefficient (r) = 0.13) and is very poorly correlated to the shape of the protomer as a whole (r = 0.45). The shapes of 18 of the 32 interfaces can be seen as CPK models in Fig. 3.
4. Amino Acid Composition

The total number of each of the 20 amino acids present in the interface was calculated and expressed as a percentage of the total number of interface residues. The percentage frequencies were also calculated for the interior and the exterior of the proteins for comparison (Fig. 9). These results, and the percentage frequency ratios calculated in Table 2, clearly indicate that the interfaces show a greater similarity to the exterior of the protein than the interior, in terms of their amino acid composition. This characteristic is exemplified in the charged residues, of which the interface has 3.54 times the percentage frequency compared with that of the interior, but only 0.74 times that of the exterior.

These results are in contrast to the work of Janin et al. (1988) who concluded that the...
Table 2. Percentage Frequencies of Hydrophobic, Polar and Charged Residues and the Ratios of Interface/Interior and Interface/Exterior for Each Set of Residues

<table>
<thead>
<tr>
<th>Amino acids</th>
<th>Percentage frequencies</th>
<th>Percentage frequency ratios</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Exterior</td>
<td>Interface</td>
</tr>
<tr>
<td>Hydrophobic</td>
<td>40.90</td>
<td>46.68</td>
</tr>
<tr>
<td>Polar</td>
<td>29.48</td>
<td>31.38</td>
</tr>
<tr>
<td>Charged</td>
<td>29.53</td>
<td>21.94</td>
</tr>
</tbody>
</table>

In this study Ala, Gly, Ile, Leu, Met, Phe, Pro, Val were classed as hydrophobic, Asp, Arg, Lys, Glu as charged and Asn, Cys, Gin, His, Ser, Thr, Trp, Tyr as polar.

interface amino acid composition was more closely related to the interior of the protein. However, Korn and Burnett (1991) found that the hydrophathy of the average interface of multimeric protein lies between that of the exterior and the interior. Argos (1988) also made similar conclusions, stating that the subunit interfaces display characteristics that lie between those of the hydrophobic nature of the protein interior and the hydrophilic properties of the protein exterior.

Propensities have also been calculated from the percentage frequencies of the 20 amino acid residues in the interface. Two propensities for each amino acid (\( j \)) were calculated as follows:

\[
\text{Interface/Protein (IP) Propensity } AA_j = \frac{\% \text{Freq } AA_j(\text{interface})}{\% \text{Freq } AA_j(\text{protein})}, \quad (5)
\]

\[
\text{Interface/Surface (IS) Propensity } AA_j = \frac{\% \text{Freq } AA_j(\text{interface})}{\% \text{Freq } AA_j(\text{surface})}. \quad (6)
\]

The two types of propensities calculated give the 'relative probability' of an amino acid being present in the interface with respect to the probability of it being present in the protein as a whole (eqn 5) and with respect to the probability of it being present in the protein surface (interface and exterior residues) (eqn 6). A propensity of > 1 denotes that a residue occurs more frequently in the interface than in the protein as a whole (or than in the protein surface). A propensity of < 1 denotes that a residue occurs less frequently.

The IP propensities (Fig. 10) reveal that the charged and polar residues, especially arginine and asparagine, show an increased affinity for the interface. In addition, the hydrophobic residues methionine and proline, show a small increased affinity for the interface. This supports the results observed by Janin et al. (1988) and Argos (1988) who found that arginine and methionine have a high affinity for the interface. The two polar residues showing less affinity for the interface are cysteine and tryptophan. It is possible that the double aromatic ring in the side chain of tryptophan is too bulky to be accommodated in the interface of a protein–protein interaction. However in general the aromatics have an affinity for the interface; two out of the three single aromatic residues show an increased affinity. This affinity was also found by Argos (1988) who suggested that aromatics make particularly good 'glue' for sticking protein subunits together.

The IS propensities (Fig. 11) confirm the results of the IP propensities. The three single aromatic residues and arginine all show an increased affinity for the interface relative to the protein surface. In addition the propensities clearly show that the interface contains a greater percentage of hydrophobic residues than the protein surface as a whole. This characteristic is further illustrated in Fig. 12 where the percentage frequency of residues in the interface and exterior are shown with the residues in order of increasing hydrophobicity, based on the Wolfenden Hydrophobicity Scale (Wolfenden et al., 1981). This hydrophobicity scale is the one most closely linked to experiment, with the equilibria of distribution of amino acid side chains, between their dilute aqueous solutions and the vapour phase, being determined by dynamic vapour pressure measurements. The scale thus gives a direct measure of the free energy for transfer from water to the hydrophobic phase of amino acid side chains.

The analysis of the percentage frequencies of amino acid residues relative to their hydrophobicities reveals that the interfaces represent hydrophobic patches on the surfaces of proteins. This equates with the theory that the hydrophobic interaction is fundamental to the stabilization of protein associations (Chothia and Janin, 1975).
Protein-protein interactions

Fig. 10. Interface/protein (IP) propensities. The propensities give the 'relative probability' of an amino acid being present in the interface with respect to the probability of it being present in the protein as a whole.

Fig. 11. Interface/surface (IS) propensities. The propensities give the 'relative probability' of an amino acid being present in the interface with respect to the probability of it being present in the surface of the protein.

5. Distribution of Atom Types

The percentage frequencies of 19 atom types have been calculated. The atom types are those defined by Warme and Morgan (1978), in their studies of long range atomic interaction in proteins. The 19 atom types defined include 15 different side chain atoms and four backbone atoms. The distribution is shown in Fig. 13, in which the atom types have been ordered according to their percentage frequency in the interior subset. This clearly shows that the distribution of the 19 atom types in the interface closely resembles that in the exterior. There is a relatively large percentage frequency of polar CH₂ atoms in the exterior and interface of proteins (polar CH₂ atom types are those occurring in charged amino acid side chains). This shows the increased affinity for polar atoms on the protein surface.
Fig. 12. The percentage frequencies of amino acid residues in the interface and exterior of 32 protein dimers. The amino acids are ordered with respect to increasing hydrophobicity according to the Wolfenden Hydrophobicity Scale (Wolfenden et al., 1981). Proline is not shown as the Wolfenden Scale has no hydrophobicity value for this residue.

Fig. 13. Percentage frequency of 19 atom types in the interior, interface and exterior of 32 protein dimers. The atom types are those defined by Warme and Morgan (1978).

(a) Atom distributions and hydrophobicity

The percentage numbers of polar and non-polar atoms in the interface and the exterior have been calculated for the 32 protein dimers. It was found that the interfaces were slightly more hydrophobic than the exterior: the interface comprising 32.06% polar and 67.94% non-polar atoms and the exterior 36.70% polar and 63.30% non-polar. In all, 88% of the dimers displayed this distribution. This gives additional support to the results of the residue distributions (Section V.4), indicating that the interfaces represent hydrophobic patches on the surfaces of proteins.

To further investigate the nature of the hydrophobicity of the interfaces each interface was
Fig. 14. Method of defining three zones within a dimer interface. The dots represent the interface atom positions as transformed into two dimensions centred at the origin. The four black dots, shown with their co-ordinates, are the maximum and minimum x- and y-co-ordinates used to define the outer boundary of zone 1 (outer). The outer boundary of zone 2 (middle) is defined by taking 25% from each of the maximum and minimum x- and y-co-ordinates. The outer boundary of zone 3 (inner) is defined by taking 50% from each of the maximum and minimum x- and y-co-ordinates. The atoms are coloured in three shades to show that they fall into different zones.

The percentage number of polar and non-polar atoms in each zone was calculated for each dimer and a mean percentage number calculated for the whole data set (Fig. 15). From the mean percentages it can be seen that the interfaces are areas in which the non-polar atoms are concentrated in the centre, with the distribution of these atoms decreasing as the outer parts of the interface are reached. This could be described as a round ‘sticky plaster’ arrangement, with the interface representing the plaster with the ‘sticky’ (non-polar) atoms in the centre protected from the outside solvent by the polar interactions. However it should be noted that this distribution is not marked, the outer zone has 61.81% non-polar atoms and the inner zone 72.31%. In addition when considering the distributions in individual dimers only 41% show this ‘sticky plaster’ arrangement of polar and non-polar atoms.

6. Secondary Structural States

The secondary structural states of those residues in the interface of the dimers have been analysed. The secondary structure definitions (coil, helix, sheet and turn) were based on
Fig. 15. The mean distribution of polar and non-polar atoms within three defined zones from 32 dimer interfaces. The zones are defined as described in the legend to Fig. 14.

Fig. 16. Percentage frequency of secondary structure motifs in interface residues in 32 protein dimers. The secondary structure motifs are those defined by Kabsch and Sander (1983). Those of Kabsch and Sander (1983). The number of interface residues in each type of secondary structural conformation was calculated as a percentage of the total number of interface residues. The percentage distributions were also calculated for the interior and the exterior residues, for comparison (Fig. 16).

All four secondary structural states were represented in the interfaces; with the helical secondary structural configuration being the most common. In the coil, sheet and turn regions, the secondary structure percentages are more closely correlated with those of the exterior residues. These findings are comparable to analysis by Argos (1988) who calculated
percentage distributions of secondary structural states (helix, strand, turn and coil) of interfaces according to their surface contribution.

(a) **Interface classification by secondary structure**

The interfaces in the set of non-homologous dimers were classified into one of four types, α, β, α/β and coil (Fig. 17). The classification is based on the percentage frequency of α and β secondary structure in the interface residues (Table 3). In the data set 53.1% of the protein interfaces were classified as α, 21.8% as β, 12.5% as α/β and 12.5% as coil.

The seven β interface proteins were further classified into three groups (extension, stacked and complex) according to the arrangement of the interface sheets between the two subunits in the protein. The classification was conducted by visually inspecting the structures using the graphics software Quanta (Quanta 4.0: Molecular Simulations Inc.). The interfaces were classified as **extended** if the sheets in the two interfaces formed a continuous sheet between the two subunits. Interfaces where the sheets in one subunit stacked on top of the sheets in the second subunit were classified as **stacked**. Interfaces where the sheets did not clearly stack or form extensions were classified as **complex**. The results of the classifications are shown in

![Fig. 17. The classification of dimer interfaces by secondary structure. The solid lines indicate the 20% cut-offs used in the classification (see Table 4).](image)

**Table 3. The Definitions used in the Classification of 32 Dimer Interfaces by Secondary Structure**

<table>
<thead>
<tr>
<th>Classification</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>α</td>
<td>α &gt; 20% and β &lt; 20%</td>
</tr>
<tr>
<td>β</td>
<td>α &lt; 20% and β &gt; 20%</td>
</tr>
<tr>
<td>α/β</td>
<td>α &gt; 20% and β &gt; 20%</td>
</tr>
<tr>
<td>Coil</td>
<td>α ≤ 20% and β ≤ 20%</td>
</tr>
</tbody>
</table>

Each dimer interface, defined at the residue level was classified according to their percentage content of α and β secondary structure content.
Table 4. The Classification of Seven β Interfaces into Three Types: Extended, Stacked and Complex

<table>
<thead>
<tr>
<th>Extended</th>
<th>Stacked</th>
<th>Complex</th>
</tr>
</thead>
<tbody>
<tr>
<td>1odt</td>
<td>2ssi</td>
<td>1fcl</td>
</tr>
<tr>
<td>5hvp</td>
<td>2rhe</td>
<td>2sod</td>
</tr>
<tr>
<td></td>
<td>2tsc</td>
<td></td>
</tr>
</tbody>
</table>

Table 4 and two examples of the classifications are shown as MOLSCRIPT (Kraulis, 1991) diagrams in Fig. 18. Of the seven β interfaces, two were defined as extended, three as stacked and two as complex. Superoxide dismutase (2sod) was one of the structures classified as complex, having the first and last strands in a flattened β barrel in each subunit, interacting across the dimer interface.

7. Hydrogen Bonding

A hydrogen bond is a polar interaction between two electronegative atoms, a donor and acceptor. To identify these polar interactions between the subunits in the dimers, hydrogen bonds were calculated using a program, HBPLUS (McDonald and Thornton, 1994). The program involves two stages; the generation of a set of possible positions for a hydrogen (H) attached to a donor and secondly a search between all atoms to find donor (D) and acceptor (A) pairs that fit specified geometric criteria. The criteria used for the generation of inter-subunit hydrogen bonds were as follows:

- D–A distance < 3.9 Å
- D–H–A angle > 90°
- H–A distance < 2.5 Å
- D–A–AA angle > 90°
- H–H–AA angle > 90°

where AA is the atom attached to the acceptor. The mean and percentage number of inter-subunit hydrogen bonds in each protein are shown in Table 5; with those for intra-subunit hydrogen bonds for comparison.

The number of inter-subunit hydrogen bonds ranged from 0 to 46. Phospholipase (1phh), gene 5/DNA binding protein (2gn5), uteroglobin (1utg) and inorganic pyrophosphatase (1pyp) were found, in the current study, to have no inter-subunit hydrogen bonds. The percentage of each type of hydrogen bond is very different between inter- and intra-subunit bonds. Side chains were involved in 76.4% of the inter-subunit interactions, but only 32.6% of the intra-subunit interactions. There is a clear positive correlation between the size of the interface and the number of hydrogen bonds present (correlation coefficient (r) = 0.77) (Fig. 19). Interfaces covering ≥ 1500 Å² (per subunit) (13 of the 32 dimers) have on average 0.88 (± 0.40) hydrogen bonds per 100 Å² buried (per subunit).

8. Salt Bridges

In protein structures there are positively (N-terminal amide, histidine, arginine and lysine side chains) and negatively charged (C-terminal carboxyl, aspartic acid and glutamine side chains) groups. Within a protein oppositely charged atoms in close proximity are defined to form a salt bridge if they are ≤ 4.0 Å apart (Barlow and Thornton, 1983). The numbers of inter-subunit salt bridges were calculated for the dimers, and they were found to range from 0 to 5. However there was no clear correlation with the size of the interface, with 56% of the proteins having no salt bridges between their subunits.

9. Disulphide Bonding

The number of disulphide bonds between each of the subunits was calculated using the program, HBPLUS (McDonald and Thornton, 1994). The program treats cysteines specifically; if any two cysteines were found with their sulphur atoms within 3 Å they were defined as cystines. Disulphide bonds are formed in reducing environments, and therefore occur in extracellular proteins where they contribute to the structural properties of the
Protein-protein interactions

Stacked Beta Interface
Streptomyces Subtilisin Inhibitor (2ssi)

Subunit B
Subunit A

(i) The Dimer

Extended Beta Interface
Cardiotoxin (1cdt)

Subunit B
Subunit A

(i) The Dimer

(ii) The Interface Strands (ii) Strands in the sheet extended across the interface

Fig. 18. MOLSCRIPT diagrams of extended and stacked β sheets. One subunit of each dimer is shaded dark and the other light and the pictures show the whole dimer (i) and just the strands forming the interface (ii).

Table 5. The Mean and Percentage Number of Inter- and Intra-Subunit Hydrogen Bonds in 32 Protein Dimers

<table>
<thead>
<tr>
<th>Type of hydrogen bond</th>
<th>Inter-subunit H-bonds</th>
<th>Intra-subunit* H-bonds</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>%</td>
</tr>
<tr>
<td>Main chain–Main chain</td>
<td>3.1</td>
<td>23.6</td>
</tr>
<tr>
<td>Side chain–Side chain</td>
<td>5.0</td>
<td>38.2</td>
</tr>
<tr>
<td>Main chain–Side chain</td>
<td>5.0</td>
<td>38.2</td>
</tr>
<tr>
<td>All</td>
<td>13.1</td>
<td>—</td>
</tr>
</tbody>
</table>

*The numbers for the intra-subunit hydrogen bonds are for a single subunit.
protein. In the data set 71.8% of the proteins are intracellular and hence a very small number of inter-subunit disulphide bonds were expected. In fact only two proteins with inter-subunit disulphide bonding were found, uteroglobin (lutg), which had two disulphide bonds (cys3:cys69' and cys69:cys3') and glutathione reductase (3grs), which had a single bond (cys58:cys63'). The small number of intersubunit disulphides was expected as such bonds are rare (Thornton, 1981). However when inter-subunit disulphides do occur in proteins they often play an important role in structural stabilization. Recent protein engineering experiments on two structures, platelet-derived growth factor-B (Prestrelski et al., 1994) and thymidylate synthase (Gokhale et al., 1994) have both shown that the introduction of inter-subunit disulphides into multimeric proteins increases the stability of the proteins.

Uteroglobin (lutg) is an extracellular protein, in which the two disulphide bonds hold the two protomers of the structure together (Morize et al., 1987). Glutathione reductase (3grs) is one of the few intracellular proteins that contain a disulphide bond. In general disulphides in intracellular proteins have functions other than structure stabilization, and in glutathione reductase the disulphide bond has an active catalytic function (Schulz et al., 1982).

10. Interface Segmentation

To analyse the discontinuous nature of the interfaces, in terms of the amino acid sequence, the numbers of segments in the interfaces were calculated. It was defined that interface residues separated by more than five residues were allocated to different segments. For example in the following set of interface residues four segments are defined:

|----------------|-----------------------------------------------|

<table>
<thead>
<tr>
<th>Segment</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
</table>

The number of segments ranges from 2 to 15 and there is a very weak correlation with the size of the interface (correlation coefficient (r) is 0.59) (Fig. 20). Alcohol dehydrogenase (Sadh) has a comparatively large number of segments for the size of its interface, that is probably reflective of the coiled nature of the interface in this structure (Section V.6.a).
The contribution of each segment to the asa of the interface (Fig. 21) and the numbers of hydrogen bonds and salt bridges in each segment of a protein were also calculated (Fig. 22). All the dimers have particular segments that dominate in terms of their asa contribution. For example, those proteins with more than 10 segments generally have 1–4 dominating segments, e.g. alcohol dehydrogenase (5adh), which has 15 segments, has a single predominate segment that contributes 54% of the asa of the protein's interface. As expected, the distribution of hydrogen bonds and salt bridges in the segments correlates closely with the asa contributions of the segments. The larger the asa contribution of the segments, the greater the number of hydrogen bonds and salt bridges within that segment. These predominant segments could represent exploitable 'hot spots' in the interface, which could prove important targets for mutagenesis.

11. Gap Volumes and Surface Complementarity

The complementarity of the interacting surfaces has been evaluated by quantifying the volume of the gaps existing between the subunits in each of the dimers, using a program SURFNET (Laskowski, 1991). This program considers each pair of subunit atoms in turn (a pair consists of one atom from the first subunit and one from the second), placing a sphere (maximum radius 5.0 Å) half-way between the surfaces of the two atoms, such that its surface just touches the surfaces of the atoms in the pair. Checks are then made to test if any other atoms intercept this sphere and each time an intercept is found, the size of the sphere is reduced accordingly. If at any time the size of the sphere falls below a minimum (minimum radius 1.0 Å) the sphere was discarded. If the sphere remained after all the checks its size was recorded. The sizes of all the allowable gap-spheres were then used to calculate the gap volume between the two subunits. It was found that there was a positive correlation between the size of the interface (per dimer) and the gap volume (Fig. 23).

To compare the surface complementarity between different types of protein–protein complex a 'gap volume index' has been defined:

\[
\text{Gap Volume Index} = \frac{\text{Gap Volume (Å}^3\text{)}}{\text{Interface asa (Å}^2\text{) (per complex)}}. \quad (7)
\]

The 'interface asa' parameter in eqn (7) is the asa of interface atoms defined on the basis of
Fig. 21. Interface $asa$ contributions of interface segments. The graphs indicate the percentage $asa$ contributed by each segment defined in each dimer interface.

changes in $asa$ upon dimerization (see Section IV.2). However, the interface atom pairs used in the gap volume calculations were any pairs (one atom from each subunit) whose surfaces could be touched by a sphere of maximum radius 5.0 Å. The interface defined on this basis will have a much larger $asa$ than that defined on the basis of changes in $asa$ on dimerization. Hence, it should be noted that the data shown should not be used to directly derive a value of gap volume per $asa$ of interface. The gap volume indices were calculated as a means of comparing the gap volumes between complexes of different types.
Fig. 22. Bonding in dimer interface segments. The graphs show the numbers of hydrogen bonds (dark blocks) and salt bridges (light blocks) occurring in each segment defined in each dimer interface. Those dimers with no inter-subunit bonds are not shown.

Gap volume indexes were calculated for the 32 non-homologous dimers, and four protein–inhibitor and three antibody-antigen complexes (listed in Table 6), which are those used by Lawrence and Colman (1993). The mean gap volume index for each of the three types of complexes, with the standard deviation of the distributions, are shown in Table 7.

The gap volume indexes indicate that the interacting surfaces in the dimers are the most complementary, whilst the antibody-antigen complexes are the least complementary (although all three distributions do overlap). This data supports the conclusions drawn by
Fig. 23. Gap volumes and accessible surface areas of dimer interfaces. The gaps between the interacting subunits of each dimer were calculated using SURFNET (Laskowski, 1991). The continuous line is the fitted line of the equation \( y = 1.75x \). The correlation coefficient \((r)\) is 0.82.

Table 6. The Protein–Inhibitor and Antibody–Antigen Complexes used in the Comparison of Interface Gap Volumes

<table>
<thead>
<tr>
<th>Protein/inhibitor complexes</th>
<th>Antibody/antigen complexes</th>
</tr>
</thead>
<tbody>
<tr>
<td>α-Chymotrypsin/ovomucoid (1cho)</td>
<td>Lysozyme/D1.3 Fab (1fdl)</td>
</tr>
<tr>
<td>Bovine pancreatic β-trypsin inhibitor (2ptc)</td>
<td>Lysozyme/HYHEL-5 Fab (2hfl)</td>
</tr>
<tr>
<td>Subtilisin/N-acetyl eglin-C (2sec)</td>
<td>Lysozyme/HYHEL-10 Fab (3hfm)</td>
</tr>
<tr>
<td>Subtilisin novo/chymotrypsin inhibitor (2snl)</td>
<td></td>
</tr>
</tbody>
</table>

Table 7. The Gap Volume Indexes for Three Different Types of Protein–Protein Complexes

<table>
<thead>
<tr>
<th>Type of complex</th>
<th>Dimer</th>
<th>Protein–Inhibitor</th>
<th>Antibody–Antigen</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gap Volume Index</td>
<td>2.19 (±0.83)</td>
<td>2.34 (±0.14)</td>
<td>2.57 (±0.50)</td>
</tr>
</tbody>
</table>

The Gap Volume Indexes are mean indexes for each data set of complexes and the values in the brackets are the standard deviations from the means.

Lawrence and Colman (1993) using their new shape complementarity statistic \((Sc)\). The relative size of the gap volume indexes, in this current study, can be explained if the complexes are considered with respect to their binding constants and evolutionary history. The interactions between antibodies and antigens are the molecular equivalent of a 'first encounter'. Most antibody–antigen interactions initially have a binding constant of around \(10^{-9} \text{ mol}^{-1}\) but in subsequent immune responses this may increase substantially as somatic mutations improve recognition and the strength of binding. In contrast both protein–inhibitor and dimeric interactions have been subject to selective evolutionary pressures. Many dimeric interactions, which can have binding constants as high as \(10^{-16} \text{ mol}^{-1}\), are so strong that the monomers have to be denatured to separate the subunits. In the protein–inhibitor interactions, both partners must also exist independently and be soluble in water. Many of these interactions have binding constants ranging from \(10^{-7} \text{ mol}^{-1}\) to
10^{-13} \text{ mol}^{-1}, \text{e.g. the protein inhibitors of proteinases (Laskowski and Kato, 1981). The gap volume indices, calculated in this current work, reflect the average strength of the interactions between three distinct types of dimers, although their ranges do overlap considerably.}

12. Protrusion of Interface Residues from the Molecular Surface

Residues involved in protein–protein interactions could be ‘preformed’, in that key surface residues (the interface residues) could be specifically exposed on the protomer surface, possibly anchored by intra-subunit bonding, offering favourable interaction sites for a second protomer. To investigate this hypothesis residue accessibilities were used as a measure of the ‘protrusion’ of residues from the surface of the dimer. The distributions of relative accessibilities of the 20 amino acids were compared between the interface and the exterior in a single subunit separated from its partner (Fig. 24). The distributions clearly show that the residues in the interface are more accessible than those in the exterior (in an isolated subunit) and the mean of each distribution (Table 8) supports this result. Their increased accessibility implies that the interface residues have fewer intramolecular contacts and hence are more flexible. This implication of a flexible state could be viewed to indicate that the interfaces are not preformed, that the residues are not fixed into a specific conformation until the actual dimerization takes place.

To further investigate the presence of ‘protruding’ residues in the interface, dominant interface residues were defined based on the criteria that they have an absolute asa of more than two standard deviations above the mean asa of all the residues in the dimer. Using this definition, the 32 dimers were found to have between zero and seven dominant residues. As was expected from their definition, the larger residues, such as the aromatics, were more

Table 8. Mean Relative Accessibilities for Each of the Twenty Amino Acids Compared in the Exterior and the Interface

<table>
<thead>
<tr>
<th>Residue</th>
<th>Exterior</th>
<th>Interface</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALA</td>
<td>39.11</td>
<td>39.60</td>
</tr>
<tr>
<td>ASN</td>
<td>45.89</td>
<td>50.46</td>
</tr>
<tr>
<td>ARG</td>
<td>39.79</td>
<td>51.31</td>
</tr>
<tr>
<td>ASP</td>
<td>48.35</td>
<td>55.18</td>
</tr>
<tr>
<td>CYS</td>
<td>19.86</td>
<td>28.28</td>
</tr>
<tr>
<td>GLN</td>
<td>43.46</td>
<td>46.22</td>
</tr>
<tr>
<td>GLU</td>
<td>48.27</td>
<td>50.87</td>
</tr>
<tr>
<td>GLY</td>
<td>44.27</td>
<td>48.30</td>
</tr>
<tr>
<td>HIS</td>
<td>35.73</td>
<td>36.74</td>
</tr>
<tr>
<td>ILE</td>
<td>26.37</td>
<td>36.01</td>
</tr>
<tr>
<td>LEU</td>
<td>25.35</td>
<td>32.59</td>
</tr>
<tr>
<td>LYS</td>
<td>51.09</td>
<td>54.58</td>
</tr>
<tr>
<td>MET</td>
<td>31.99</td>
<td>44.88</td>
</tr>
<tr>
<td>PRO</td>
<td>46.20</td>
<td>49.70</td>
</tr>
<tr>
<td>PHE</td>
<td>20.06</td>
<td>43.09</td>
</tr>
<tr>
<td>SER</td>
<td>44.63</td>
<td>43.54</td>
</tr>
<tr>
<td>TRP</td>
<td>19.35</td>
<td>42.63</td>
</tr>
<tr>
<td>TYR</td>
<td>25.44</td>
<td>37.72</td>
</tr>
<tr>
<td>THR</td>
<td>38.90</td>
<td>44.01</td>
</tr>
<tr>
<td>VAL</td>
<td>27.66</td>
<td>35.51</td>
</tr>
<tr>
<td>Mean total</td>
<td>36.09</td>
<td>43.56</td>
</tr>
</tbody>
</table>

The relative accessibilities were calculated by summing atomic accessible surface areas in each residue and comparing the totals to the accessible surface area that each of the 20 residues possess in the standard extended state, using the program ACCESS (Hubbard, 1992b).
highly represented than the smaller linear residues. Visual inspection of the dominant residues revealed that they protruded from one subunit into the other and in some interfaces were interdigitating (Fig. 25). The overall impression was that each dimer had a small number of residues that protruded from one subunit to the other, often at dispersed sites over the interface, thus physically anchoring the two subunits together.

13. Flexibility of Residues and Atoms in the Dimer Interface

Crystallographic temperature factors (B-factors) have been used to analyse the flexibility of atoms and residues in dimer interfaces. $B$, termed the isotropic temperature factor, or the Debye–Waller factor, is related to the mean square displacement of the atom from its mean position by the expression
Fig. 25. Liquorice bond diagrams of dominant residues in two dimer interfaces. In each diagram one subunit is coloured blue and one red, the dominant residues are displayed in contrasting colours, the green residues are in the 'blue' subunit and the yellow residues are in the 'red' subunit.
The $B$-factor, calculated for each atom as the result of a restrained least squares refinement, can be viewed as an attempt to fit a Gaussian to the spread of the electron density about the average position of that atom. Any factor, including static disorder in the crystal and atomic motion, that contributes to the spreading of this electron density will contribute to the $B$-factor.

The distributions of $B$-factors in the 23 protein dimers (single subunits only) (Fig. 26) reveal their diverse nature, both within single protein structures and between different proteins. Nine proteins (1cdt, 1l18, 1pyp, 2orl, 2ssi, 3aat, 3gap, 3sdp and 5adh) do not have valid $B$-factors assigned, and in the subsequent analysis a subset of 23 protein dimers has therefore been used.

(a) **Comparisons of atomic temperature factors within dimeric proteins**

The percentage frequencies of $B$-factors in the interior, interface and exterior atoms of the proteins...
23 dimers and ratios of these frequencies for interface/interior, interface/exterior and interior/exterior have been calculated (Fig. 27).

In the data set all of the proteins have an exterior/interior ratio of >1.0 (Fig. 27(c)). This indicates that the interior residues have atoms that are less flexible than those on the exterior of the protein. This is expected as protein interiors are less solvent accessible and have more steric hindrance, due to the closely packed nature of atoms, than those in the exterior.

Fifty-eight per cent of the proteins have an interface/interior ratio of >1.0 (Fig. 27(a)), indicating that the interface comprises residues with atoms slightly more flexible than those in the interior. In the comparison of interface to exterior, 100% of proteins have an interface/exterior ratio of <1.0 (Fig. 27(b)). A ratio of <1.0 indicates that the interface residues have atoms that are less flexible than those in the exterior of the structure. This would be expected as the interface residues are involved in protein-protein interactions.

VI. A TOOL FOR THE ANALYSIS OF INTERFACES IN MULTIMERIC PROTEINS

At present, work is in progress to develop a computational tool to analyse the interfaces of multimeric proteins. The program is designed to automatically generate data including the
size of the interface, planarity, polarity, inter-subunit bonding, segmentation and amino acid composition. The program operates on multimeric structures recorded in the current PDB file format, and the user is only required to enter the subunit identifiers of the interface to be analysed. The program produces output as a series of PostScript (Adobe Systems Inc., 1985) tables and graphs, as well as a number of data files. This tool will make it possible to compare interfaces in dimers and in higher multimeric structures relatively quickly.

VII. PROTEIN–PROTEIN INTERACTIONS IN DIMERS: A SUMMARY

It can be concluded that the structural properties of dimer protein–protein interfaces are between those of the exterior and the interior but are more closely correlated to the exterior surface than the interior. Specifically the following characteristics describe our current observations of dimer protein–protein interfaces:

- The $asa$ is approximately linearly related to molecular weight of the protomer ($r = 0.69$).
- The $asa$ buried (per subunit) in the interface ranges from 368.1 Å$^2$ to 4746.1 Å$^2$.
- The $asa/M_r$ ratio decreases from protomer to dimer.
- The interface is more hydrophobic than the exterior but less hydrophobic than the interior:
  - $atoms$: 68.1% non-polar, 31.9% polar.
  - $residues$: 46.7% hydrophobic, 31.4% polar, 21.9% charged.
- The interface has an affinity for hydrophobic residues, as well as arginine and single aromatic residues (histidine, tyrosine, phenylalanine).
- On average there are 0.88 hydrogen bonds per 100 Å$^2$ of $asa$ buried (for interfaces covering $>1500$ Å$^2$ per subunit).
- Interfaces occur between helix, sheet and coil motifs, with both like and non-like interactions across the interface.
- Interfaces are discontinuous, segmented surfaces, with between 2 and 15 segments and a mean of 5.5.
- The gap volume between dimer subunits is approximately proportional to the $asa$ of the interface (per dimer).
- In general dimer interfaces have better shape complementarity than antibody/antigen complexes.

VIII. DISCUSSION

This study of protein–protein interactions in dimers has consolidated the results previously presented by Argos (1988) and Janin et al. (1988). It has also provided additional data on the surfaces involved in protein–protein interactions in multimeric proteins. The interfaces have been shown to be in general circular, planar, hydrophobic patches on the surface of proteins that show good shape complementarity. This work has also identified the dominant segments of the polypeptide chain, which could prove to be exploitable 'hot spots' for mutagenesis.

Multimeric proteins represent just one system in which protein–protein interactions are important and interactions have been studied in many protein–inhibitor, protein–ligand and antibody–antigen complexes (e.g. Janin and Chothia, 1990). These different systems represent different levels of interaction. The interactions between subunits in dimeric proteins are amongst the strongest and most extensive ($asa$ up to 4768 Å$^2$ per subunit). Dimer interactions are long-lived with isolated oligomer subunits rarely achieving their biological function in the monomeric state. These features distinguish dimer interactions from other protein–protein interactions, which, in comparison, are often weaker temporary contacts. It would appear that dimer interactions are amongst the strongest interactions in nature.

The evolutionary pathway followed to achieve such effective and optimal interactions has been sought, and a number of possible modes of oligomerization evolution have been considered (Riddihough, 1994). The most recent hypothesis, termed 'domain swapping' (Bennet et al., 1994), has been observed between the monomeric $\gamma B$ and the homodimeric
βB2 crystallins. Structural analysis has shown that whilst the N- and C-terminal domains interact within the same molecule in the γB-crystallin structure, the corresponding interaction in βB2 is between domains from different subunits (Slingsby et al., 1991). Hence in the β- and γ-crystallin family pairs of domains associate either intramolecularly to form monomeric proteins, or intermolecularly to form oligomers (Slingsby et al., 1992). In effect interfaces that have evolved to be optimal in the monomer are ‘hijacked’ to stabilize the oligomeric structure. However, it should be considered that many dimers are not multidomain proteins and solid evidence for domain swapping is rare, so although it may occur occasionally, it is unlikely to be the dominant evolutionary pathway for forming stable dimers.

Whatever the mode of evolution, protein oligomers have evolved to fulfil a wide variety of biological functions, controlling many cellular processes including signal transduction and enzyme co-operativity. Protein oligomers are observed as allosteric enzymes and multienzyme complexes, and in many such structures the catalytic or functional units are found repeated in each subunit. In some structures the binding of substrates occurs at the protein–protein interface; for example, the trimeric aspartate transcarbamylase from E. coli has an active site that lies at the interface between adjacent subunits. The functional role of oligomer interfaces is also seen in the dimeric DNA binding proteins, such as gene 5 DNA binding protein (Brayer and McPherson, 1983), which has symmetry related DNA binding sites that include interactions with both subunits of the structure.

An understanding of protein–protein interactions, and their relationship to protein functions, is a prerequisite for drug design and for the optimization of drug therapies already in use. In this respect, Wlodawer et al. (1989) suggested a molecule that disrupted the dimerization of the HIV protease, could act as a blueprint for a drug designed to combat the HIV virus. The characterization and understanding of protein–protein interactions would be a preliminary step in such a drug design regime. Such a step would also prove important to current work concerned with the development of dimerization inhibitors as anti-parasitic drugs (Nordhoff et al., 1993).

The importance of protein–protein interactions in biological systems, and their potential use in drug design regimes, has lead to the theoretical prediction of interacting protein surfaces through the development of predictive algorithms (e.g. Wodak and Janin, 1978; Walls and Sternberg, 1992; Zielenkiewicz and Rabczewko, 1984, 1985; Helmer-Citterich and Tramontano, 1994). The majority of such algorithms are primarily based on calculations of shape complementarity. The use of other characteristics of protein–protein interfaces, such as hydrophobicity, as the primary basis for prediction, has been less well explored. However, Korn and Burnett (1991) have used hydrophathy analysis to predict the position of the interface in a dimeric protein using a non-automated method. More recently Young et al. (1994) produced an automated predictive algorithm based on the analysis of the hydrophobicity of clusters of residues in a protein. This algorithm has successfully been used to predict sites of protein–protein interactions in a number of enzymes, antibody fragments and a number of other proteins with associated ligands.

This current review of dimer–dimer interfaces has effectively produced a rule base of interface characteristics including size, residue composition, hydrophobicity, planarity, segmentation and complementarity. With this information it should be possible to formulate a new knowledge-based predictive algorithm for the identification of interface forming regions on the surfaces of promoters. By learning the rules that govern molecular recognition, we can hope to improve our ability to design small novel molecules that will disrupt dimerization, and could thereby interfere in many of the molecular signals used in pathways to control cell development.

REFERENCES