Protein-protein interaction : heterodimers and homodimers

Cui, Zhanhua

2006

Cui, Z. H. (2006). Protein-protein interaction : heterodimers and homodimers. Doctoral thesis, Nanyang Technological University, Singapore.

<https://hdl.handle.net/10356/35985>

<https://doi.org/10.32657/10356/35985>

Protein-Protein Interaction: Heterodimers and Homodimers

CUI ZHANHUA

School of Mechanical & Aerospace Engineering

**A thesis submitted to the Nanyang Technological University
in fulfillment of the requirements for the degree of
Doctor of Philosophy**

2006

Acknowledgement

Acknowledgement

I would like to express my deepest gratitude to Professor Jacob Gan and Professor Pandjassaram Kanguane for their guidance, encouragement and patient assistance, without whom this work might have not been completed

I would like to thank Li lei for many invaluable discussions.

Finally, special thanks to my family and friends. They encouraged and helped me in all possible ways.

Table of contents

Acknowledgement	i
Table of Contents	ii
Summary	vi
List of Figures	viii
List of Tables	x
Nomenclatures	xi
Chapter 1 Introduction	1
1.1 Background	1
1.2 Objectives	3
1.3 Thesis organization	3
Chapter 2 Literature review	5
2.1 Characteristics of protein-protein interaction	5
2.1.1 Interface size	6
2.1.2 Hydrogen bonds	8
2.1.3 Hydrophobicity	11
2.1.4 Amino acid composition at interfaces	14
2.1.5 Other properties of protein-protein interfaces	18
2.1.5.1 Geometrical shape	19
2.1.5.2 Geometrical complementarity	19
2.1.5.3 Secondary structure state	20
2.1.5.4 Electrostatic complementarity	21
2.1.5.5 Water molecule effects on protein-protein interactions	24
2.1.5.6 Conformational changes in protein-protein interactions	26
2.2 Interaction site prediction models	29
2.2.1 From sequence	30
2.2.2 Correlated mutation	34

Table of Contents

2.2.3 Homology modeling and threading	35
2.2.4 Protein-protein interaction network	37
2.2.5 Other methods	38
2.3 Conclusion	41
Chapter 3 Identification of critical heterodimer protein interface parameters by multi-dimensional scaling in Euclidian space	42
3.1 Introduction	42
3.2 Material and methods	43
3.2.1 Creation of a heterodimer structural dataset	43
3.2.2 Interface parameters	47
3.2.3 Parameter normalization	51
3.2.4 Representation of parameters	51
3.2.5 Calculation of eigenvectors and eigenvalues	52
3.2.6 Selection of interface descriptors	52
3.2.7 Calculation of distances in the parameter space	53
3.2.8 Calculation of correlation coefficient	54
3.3 Results	54
3.3.1 Quantitative descriptors for heterodimer interface parameters in six dimensions	54
3.3.2 Selection of critical interface parameters from highly correlated descriptors	56
3.4 Discussion	58
3.4.1 Interface H-bonds	59
3.4.2 Interface aromatic and aliphatic residues	59
3.4.3 Interface residues and interface hydrophobicity	59
3.4.4 Gap volume	60
3.5 Conclusion	60
Chapter 4 Protein subunit interfaces: heterodimer vs homodimer	61

Table of Contents

4.1	Introduction	61
4.2	Material and methods	64
4.2.1	Creation of heterodimer and homodimer structural dataset	64
4.2.2	Interface parameters	67
4.3	Results	67
4.3.1	Quantitative descriptors for homodimer interface parameters in six dimensions	67
4.3.2	Selection of critical interface parameters from highly correlated descriptors	70
4.4	Discussion	72
4.4.1	Interface residues	72
4.4.2	Interface H-bonds	75
4.4.3	Interface hydrophobicity	77
4.4.4	Interface residue composition	78
4.5	Conclusion	81
Chapter 5	Interface residue prediction using sequence data	82
5.1	Introduction	82
5.2	Material and methods	86
5.2.1	Heterodimer structural dataset	86
5.2.2	Prediction algorithm	87
5.2.3	Definition of window size	90
5.2.4	Accuracy definition	92
5.3	Results and discussion	93
5.3.1	Prediction accuracy	93
5.3.2	Window size effect on accuracy	96
5.3.3	Sequence length effect on accuracy	97
5.3.4	Comparison with other studies	99
5.4	Conclusion	100

Table of Contents

Chapter 6	Conclusion and future work	102
6.1	Conclusion	102
6.2	Future work	103
List of Publications		104
References		105

Summary

Protein-protein interaction is amongst the ubiquitous type of interactions in biological systems and plays a key role in all cellular processes. These interactions are largely protein dimers (two body collision). The dimers are either heterodimers (complexation of non-identical monomers) or homodimers (complexation of identical monomers). Thousands of such dimer structures have been collected and documented at the PDB. However, our knowledge on these interactions is far from satisfactory. Therefore, it is our interest to study dimer interface properties to use them for detecting potential interaction sites. We used a non-redundant dataset of 156 heterodimers and 170 homodimers determined by X-ray crystallography with high resolution (2.5Å) for this purpose.

A large number of structurally important physical-chemical interface parameters characterize dimer interfaces. Yet, a handful of such interface parameters are known to dominate dimer interfaces. Hence, we applied a multi dimensional scaling procedure in Euclidean space to determine the most critical parameters that characterize heterodimer (homodimer) interfaces. The method suggests that the six critical parameters (Heterodimer: interface residues, interface H-bonds, interface hydrophobicity, gap volume, interface aliphatic residues, and interface aromatic residues; Homodimer: interface residues, interface H-bonds, interface hydrophobicity, hydrophobic residues, interface aliphatic residues, and interface aromatic residues) effectively replace an original set of 118 parameters.

Summary

Homodimers and heterodimers have significantly different structure properties at the interface. Nonetheless, these differences were not comprehensively documented. Here we used parameters (interface residues, H-bonds, hydrophobicity and residue composition) to study the difference between 170 homodimers and 156 heterodimers that are mutually independent.

Our ambitious goal is to develop a model to predict interface sites from the sequence. Therefore, we used the structural information gained from our previous analysis to develop a GOR (GOR- Garnier-Osguthorpe-Robson) model for prediction. The GOR method was based on information theory and Bayesian statistics. The model uses three states (interface, interior and surface) identified from interface analysis for model training. Results show a prediction accuracy of 29.5%. This suggests that prediction of interaction sites from sequence data is challenging and non-trivial. Future investigation should help to identify other parameters to strengthen prediction of protein-protein interaction sites.

List of figures

Figure 1.1	Protein structure growth	2
Figure 3.1	Distribution of eigenvalues	53
Figure 3.2	Correlation coefficients for parameter distances	55
Figure 3.3	The distance between six selected descriptors and 118 parameter	56
Figure 3.4	The distance between six selected parameters and 118 parameter	57
Figure 4.1	Distribution of eigenvalues	69
Figure 4.2	Correlation coefficients for parameter distances	69
Figure 4.3	The distance between six selected descriptors and 118 parameter	70
Figure 4.4	The distance between six selected parameters and 118 parameter	71
Figure 4.5	Interface residues in heterodimer	73
Figure 4.6	Interface residues in homodimer	74
Figure 4.7	Interface area in heterodimer interface	74
Figure 4.8	Interface area in homodimer interface	75
Figure 4.9	Hydrogen bonds in heterodimer interface	76
Figure 4.10	Hydrogen bonds in homodimer interface	76
Figure 4.11	Hydrophobic, hydrophilic and charged residue fraction in heterodimers and homodimers	79
Figure 4.12	Propensity difference in heterodimer and homodimer (heterodimer-homodimer)	79
Figure 4.13	Ratio of interface to surface & interface to interior propensity in heterodimer	80
Figure 4.14	Ratio of interface to surface & interface to interior propensity in homodimer	80
Figure 5.1	Information difference in chain A and chain B	89
Figure 5.2	Interface residue length distribution in heterodimer	91
Figure 5.3	Interface residue length distribution in homodimer	91
Figure 5.4	The distribution of accuracy in chain A of heterodimer	94

List of Figures

Figure 5.5	The distribution of accuracy in chain B of heterodimer	95
Figure 5.6	The distribution of accuracy in chain A of homodimer	95
Figure 5.7	Window size effect on prediction accuracy in chain A of heterodimer	96
Figure 5.8	Window size effect on prediction accuracy in chain B of heterodimer	97
Figure 5.9	Window size effect on prediction accuracy in chain A of homodimer	97
Figure 5.10	Prediction accuracy in different length group	99

List of tables

Table 3.1	Heterodimers used in this study	44
Table 3.2	List of 118 heterodimer interface parameters	47
Table 3.3	Critical parameters for heterodimer interfaces	53
Table 4.1	Homodimers used in this study	64
Table 4.2	Critical parameters for homodimer interfaces	71
Table 5.1	Prediction accuracies with window size 11	93
Table 5.2	Observed number and observed accuracy	94
Table 5.3	Groups according to heterodimer and homodimer sequence length	98

Nomenclatures

ΔASA	Interface area of a protein-protein complex
ASA	Accessible surface area
CATH	Class, Architecture, Topology and Homologous Super-family
NMR	Nuclear Magnetic Resonance
PDB	Protein DataBank
DIP	Database of Interacting Proteins
RMSD	Root Mean Square Deviation
SCOP	Structural Classification of Protein
SVM	Support vector machines

Amino acids

A	ALA	Alanine	N	ASN	Asparagine
C	CYS	Cysteine	P	PRO	Proline
D	ASP	Aspartic acid	Q	GLN	Glutamine
E	GLU	Glutamic acid	R	ARG	Arginine
F	PHE	Phenylalanine	S	SER	Serine
G	GLY	Glycine	T	THR	Threonine
H	HIS	Histidine	V	VAL	Valine
I	ILE	Isoleucine	W	TRP	Tryptophan
K	LYS	Lysine	Y	TYR	Tyrosine
L	LEU	Leucine	N	ASN	Asparagine
M	MET	Methionine	P	PRO	Proline

CHAPTER 1

Introduction

1.1 Background

Proteins are worker molecules in living cells. They help to digest food, allow blood to carry oxygen, fight infections, transmit signals, and carry out many other critical functions. Each protein is folded to specifically bind one or several small molecules, nucleic acids, and other proteins. Furthermore, many diseases, including cystic fibrosis and sickle cell disease, are caused by defects in a single protein or only one amino acid in the protein sequence. Some other diseases, such as Alzheimer's disease and "mad cow" disease, result from proteins with incorrect structures. Thus, the analysis of protein-protein interactions is becoming one of the main foci of attention in the post-genomic era.

Protein-protein interaction is probably amongst the most ubiquitous type of interactions in biological systems and plays a key role in all cellular processes. These interactions include chemical transformation, regulation of metabolic pathways, immunologic recognition, protein synthesis, metabolic cycles, DNA transcription and replication, different signaling cascades and many additional processes [Bock & Gough, 2001; Zhou & Shan, 2001; Janin & Seraphin, 2003; Wodak & Mendez, 2004; Martin, et al. 2005].

Up to now, all the researchers working in bioinformatics assume that similar sequences of proteins carry out similar functions while different sequences carry out different functions. There is a close correlation between protein structures and

Chapter 1

functions, for the protein sequences determine three-dimensional structures of proteins. The function of the protein is very often dictated by some minute details at the protein interfaces, such as the grooves, ridges or pockets which could be evaluated by RMSD and gap volume which is a critical interface parameter in our research. Therefore, the determination of the detailed, three-dimensional structures of proteins helps to provide an insight into how proteins function normally and how faulty protein structures cause diseases and to provide pointers to the design and prediction of new proteins or active sites at the known proteins. Moreover, the structures of disease-related proteins can be used to help develop new medicines and diagnostic techniques.

The structures of many protein complexes at atomic level have been determined by X-Ray crystallography and NMR. A large body of chemical and geometrical data (Figure 1.1) is available from PDB [Westbrook et al., 2003]. These data enable us to analyze and understand protein-protein interactions.

Chapter 1

Figure 1.1: Protein structure growth (<http://pd-beta.rcsb.org/pdb>).

There are various techniques, such as classical yeast-two-hybrid method [Fields & Song, 1989], split ubiquitin systems [Stagljar et al., 1998] and biosensor-based assays [Myszka, 1999; Rich & Myszka, 2000], combined to meet the challenge of identifying the relationship between protein structures and functions. Now large sets of data of protein structures are generated by many computational, biochemical or experimental methods and managed by specialized databases that are being developed. Moreover, most of them are free and open to academic use. These databases include PDB, DIP, Swiss-Prot, SCOP and CATH. However, it's a big challenge to interpret these data and to unravel the nature of protein-protein interactions.

Chapter 1

1.2 Objective

Protein-protein interaction is a wide area with challenging difficulties. This investigation is focused on 5 main topics.

1. Creation of high resolution ($\leq 2.5\text{\AA}$) structures of homodimers and heterodimers.
2. Study interface properties in structural data.
3. Apply multi-dimensional scaling procedure for dimension reduction to identify critical interface properties.
4. Differentiate heterodimer and homodimer interfaces using interface properties.
5. Develop a model to predict interaction sites using the sequence data.

1.3 Thesis organization

The structure of the thesis is as follows:

- Chapter 2 will give a review on protein-protein interaction and the development of the prediction of interaction sites;
- Chapter 3 will discuss the identification of critical heterodimer interface parameters;
- Chapter 4 will study the differences between heterodimer and homodimer interfaces using critical interface parameters;
- Chapter 5 will deliver a method to detect potential interface site using the sequence data;
- Chapter 6 will make a summation and present the future work.

CHAPTER 2

Literature review

2.1 Characteristics of protein-protein interactions

Protein-protein interactions occur at protein surfaces and are biophysical phenomena, governed by interface geometrical properties (interface size, planarity, sphericity and complementarity) and chemical properties (the types of chemical groups in amino acids, hydrophobicity, electrostatic interactions and hydrogen bonds). Towards the common goal of understanding how proteins interact, these properties have been studied using different dataset of protein structures by numerous groups [Ippolito, et al., 1990; Janin & Chothia, 1990; Korn & Burnett, 1991; Stickle, et al., 1992; Lawrence & Colman, 1993; Young, et al., 1994; McDonald & Thornton, 1994; Clackson & Wells, 1995; Janin, 1995, 1996; Jones & Thornton, 1997a; Wells, 1996; Lijnzaad & Argos, 1997; Tsai, et al., 1996, 1997ab; Tsai & Nussinov, 1997; Xu, et al., 1997b; Dasgupta, et al., 1997; Bogan & Thorn, 1998; Larsen, et al., 1998; Xu & Regnier, 1998; Lo Conte, et al., 1999; Norel, et al., 1999; Hirakawa, et al., 1999; Eisenhaber, 1999; Jones, et al., 2000; Sheinerman, et al., 2000; Zhou & Shan, 2001; Glaser, et al., 2001; Valdar & Thornton, 2001b; Chakrabarti & Janin, 2002; Sheinerman & Honig, 2002; Brinda, et al., 2002; Ma, et al., 2003; Bahadur, et al., 2003; Nooren & Thornton, 2003b; Caffrey, et al., 2004]. These studies are influenced by dataset size and their characteristics. It should also be noted that these studies are based on limited dataset consisting of heterogeneous (mixture of homodimers and hetero-complexes) data.

Chapter 2

2.1.1 Interface size

Interface size is an important property used to describe protein interfaces and is usually characterized by interface area or the number of interface residues. The number of interface residues is shown to be linearly correlated to interface area (correlation coefficient: $r \geq 0.96$) [Chakrabarti & Janin, 2002; Bahadur, et al., 2003]. Jones & Thornton [1996] showed that homodimer interface area ranges from 368 Å² to 4746 Å² (based on a dataset of 32 dimers) and hetero-complex interface area ranges from 639 Å² to 3228 Å² (based on a dataset of 27 hetero-complexes). They also showed that, in general, interface area is often proportional to the total protein size. Dasgupta et al. [1997] showed that the interface area per subunit ranges from 670 to 5540 Å² based on a sample of 23 oligomeric proteins.

Lo Conte et al. [1999] noted that most of the protein-protein complexes have an interface area in the range of 1200-2000 Å² and defined their interfaces as “standard size.” For all the dataset (used a dataset of 75 hetero-complexes), the interface area varies from 1140 Å² to 4660 Å² with the mean value of 1940 Å².

Jones et al. [2000] studied the differences between protein domain interfaces and oligomeric protein interfaces. The interface area in the 46 monomeric two domain proteins ranges from 260 Å² to 3580 Å² and the interface area derived from 105 oligomeric or protein complexes ranges from 95 Å² to 2813 Å².

Valdar & Thornton [2001b] showed that the residue conservation can help identify biological relevant crystal contacts using a dataset of 53 family of homodimers and

Chapter 2

65 families of monomers. The biological contact is shown to have 53.7 residues on average and account for 25.9% of the protein surface residues. In contrast, the average non-biological contact has only 7.6 residues and covers 4.2% of the surface.

Chakrabarti & Janin [2002] studied 70 hetero-complexes and refined the identity of a typical interaction 'patch' (having an area of at least $800 \text{ \AA}^2$, involving somewhat more than 20 residues and somewhat less than 100 atoms). These patches are composed of a core and a rim. Only the core residues are shown to have a composition distinct from that of the rest of the surface. The rim is interpreted as mainly isolating the core of the patch from the solvent, recalling the 'O-ring' theory [Bogan & Thorn, 1998]. Single patch interface contains 47 ± 11 residues or 23 residues per recognition site. The average interface area per subunit is $1906 \text{ \AA}^2$ and the average number of interface residues is 57.

Nooren & Thornton [2003b] analyzed the characteristics of transient protein-protein interactions using a dataset of 16 homodimers and 23 heterodimers. The interface area in homodimers is shown to range between $478 \text{ \AA}^2$ and $926 \text{ \AA}^2$, which is 7% to 18% of the monomer surface. As compared to homodimers, the heterodimer interface area ranges between $570 \text{ \AA}^2$ and $2213 \text{ \AA}^2$.

Bahadur et al. [2003] showed that the range of interface area extends from $500 \text{ \AA}^2$ to $7000 \text{ \AA}^2$ with the mean value of $1970 \text{ \AA}^2$. Interfaces bury 16% of the subunit surface on average, but this fraction varies from 3% to 44% for all dataset of 122 homodimers. The average interface is shown to contain 52 residues per subunit. Although the smaller proteins obviously cannot form very large interfaces, the

Chapter 2

correlation with size is mediocre.

Caffrey et al. [2004] used a dataset to study protein-protein interface conservation. The dataset consists of 42 chains that form homodimers, 12 chains that form heterodimers, and 10 chains that form transient complexes. The interface area is showed to range from 415 to 3568 Å² for heterodimers, 550 to 4718 Å² for homodimers, and 423 to 2361 Å² for transient complexes. The average number of interface residues is 44.4 (in homodimers) and 42.2 (in heterodimers).

As we have seen from the above mentioned studies, interface sizes obtained can vary and this can be due to different dataset and their features, such as resolutions, sizes or types these dataset.

2.1.2 Hydrogen bonds

A hydrogen bond is a polar interaction between two electronegative atoms, one of which acts as a donor and the other an acceptor [Jones & Thornton, 1995]. A hydrogen atom lies between them and they are covalently bonds to each other. The donor attracts the electron on the hydrogen with the result that the electron's orbit is more towards the donor itself. This leaves a partial positive charge on the hydrogen, which is electro-statically attracted towards the electronegative acceptor. The interaction is energetically favourable in a number of ways, including in terms of the polarization energy and the covalent energy and particularly the electrostatic energy. Generally speaking, there are three main types of hydrogen bonds in protein structures, which are formed between main-chain and main-chain, main-chain and

Chapter 2

side-chain, side-chain and side-chain, apart from the solvent and hetero-atoms mediated hydrogen bonds [Janin & Chothia, 1990; Ippolito, et al., 1990; Stickle, et al., 1992]. Although the energy of an average inter-molecular hydrogen bond is quite small, 20KJ/mol (5Kcal/mol) compared to 200KJ/mol (Kcal/mol) in the case of a covalent bond, a great number of hydrogen bonds play an important role in protein-protein interactions [McDonald & Thornton, 1994]. Studies show that there is a complementarity of hydrogen bond donor/acceptor sites [Janin & Chothia, 1990; Krystek, et al., 1993] and that having two or more intermolecular hydrogen bonds is intrinsic [Meyer, et al., 1996]. The numbers of inter-subunit hydrogen bonds found vary in different studies [Janin & Chothia, 1990; Jones & Thornton, 1995; Xu, et al., 1997b; Lo Conte, et al., 1999; Bahadur, et al., 2003; Nooren & Thornton, 2003b].

Janin & Chothia [1990] showed that there are 8-13 hydrogen bonds and a mean value of 10 hydrogen bonds per complex using 15 protease-inhibitor complexes and 4 antibody-antigen complexes determined by x-ray crystallography. In protease-inhibitor complexes there are two thirds of hydrogen bonds involving main-chain atoms. In contrast, in antibody-antigen complexes most hydrogen bonds involve side-chain atoms.

Jones & Thornton [1995] showed that there are 0-46 hydrogen bonds base on a dataset of 32 homodimers. On average, there is 0.88 hydrogen bond per 100 Å² interface area with a *r* value of 0.77 between hydrogen bonds and interface area.

Chapter 2

Xu et al. [1997b] showed 11 hydrogen bonds per subunit with a r value of 0.89 between hydrogen bonds and interface area based on the studies using a dataset of 319 protein-protein interfaces. They also showed that the estimated number of hydrogen bonds is sensitive to the geometric parameters of the bonds. The geometrical distribution of hydrogen bonds across the interfaces is non-optimal and is different from that of protein interior. The reason for this difference may be that there are more hydrophilic side-chains buried in the binding interfaces than in the protein interior. When folding, proteins are completely free to attain their optimal configurations. But when binding each other, protein molecules are already folded and have limited freedom of six degrees of translation and rotation available to achieve the most favourable configuration [Xu, et al., 1997b; Archakov, et al., 2003].

Lo Conte et al. [1999] showed there are 1 to 34 hydrogen bonds with the mean value of 10.1 hydrogen bonds based on the study using a dataset of 75 hetero-complexes. However, there are 9 hydrogen bonds in standard interfaces (defined by them with interface area range of between $1200 \text{ \AA}^2$ and $2000 \text{ \AA}^2$). They also analyzed effects on the number of hydrogen bonds caused by dataset resolution. Study showed that there is one hydrogen bond per $170 \text{ \AA}^2$ interface area and a r value of 0.84 between hydrogen bonds and interface area for 36 complexes (resolution $\leq 2.4 \text{ \AA}$). For studies using structures with lower resolution, there were fewer hydrogen bonds and the correlation with interface area vanishes, which suggests that errors in atomic coordinates mask existing hydrogen bonds.

Chapter 2

Bahadur et al. [2003] (used a dataset of 122 homodimers) showed an average 9.0 hydrogen bonds per homodimer interface with a r value of 0.75 between hydrogen bonds and interface area. On average, there is one hydrogen bond per 210 Å² and the correlation is better with the polar interface area ($r = 0.83$).

These findings suggest that the number of hydrogen bonds and the correlation with interface area are influenced by dataset size and their characteristics, especially the structure resolution of the dataset used.

2.1.3 Hydrophobicity

It has been demonstrated that the hydrophobic forces play an important role in the protein-protein interaction [Korn & Burnett, 1991; Wells, 1996; Tsai, et al., 1996, 1997a; Lijnzaad & Argos, 1997; Bahadur, et al., 2003]. The average values of contact surface hydrophobicity usually represent the mean of the hydrophobicity of the protein core and its surface [Jones & Thornton, 1996; Eisenhaber, 1999]. Also, the ratio between buried hydrophobic and buried hydrophilic residues is used to measure the hydrophobic effect [Tsai, et al., 1997a]. Studies showed that hydrophobic residues (except ALA) and the charged residue ARG are found to have an increased presence at protein-protein interfaces [Dasgupta, et al., 1997; Lijnzaad & Argos, 1997; Lo Conte, et al., 1999; Zhou & Shan, 2001; Brinda, et al., 2002; Bahadur, et al., 2003; Ma, et al., 2003]. Especially, TYR and TRP are found to have the highest propensity to stay interfaces. It has even been suggested that the binding energy of two proteins derives from the burying of hydrophobic surface areas [Chothia & Janin, 1975].

Chapter 2

Jones & Thornton [1996] also assumed that proteins will associate with each other by hydrophobic patches. He calculated the mean hydrophobic value for all interface residues of each complex. In all of the calculated complexes, the interface hydrophobicity is intermediate between that of the interior and the exterior. Generally speaking, the protein has a hydrophobic core in the interior and a more hydrophilic surface [Hirakawa, et al., 1999]. When comparing the interface hydrophobicity between homodimers and hetero-complexes, it is found that the homodimer interfaces are more hydrophobic. The different hydrophobicity may originate from the roles of two types of complexes [Janin, et al., 1988]. The hetero-complexes often function as monomers in solution. Hence their interfaces cannot be as hydrophobic as homodimer interfaces, because it is energetically unfavourable to have a large exposed hydrophobic area on the proteins. In contrast, the homodimers usually function as the form of dimer, hence they have hydrophobic interfaces permanently buried within the inter-subunits.

Tsai et al. [1997a] studied the hydrophobic effect in protein-protein interactions using a dataset of 362 protein-protein interfaces and 57 symmetry-related oligomeric interfaces. The hydrophobic effect was measured by the buried non-polar surface area or percent burial of residue types. Studies showed that the ratio between buried hydrophobic and buried hydrophilic residues is approximately 1.5. The interior of the interfaces appears to constitute a compromise between the stabilization contributed by the hydrophobic effect on the one hand and avoiding patches on the protein surfaces that are too hydrophobic on the other. The overriding conclusion is that the hydrophobic effect plays a dominant role in protein-protein interfaces [Dill, 1990].

Chapter 2

However, it is not as strong as that observed in protein folding [Tsai, et al., 1997a]. Nevertheless, there are exceptions, where there is no sign of a significant hydrophobic contribution at the interface [Tsai, et al., 1997a].

Most studies tended to analyze the average hydrophobicity over a diverse set of protein-protein interfaces. It suffers the drawback that it blurs information on how individual interfaces are stabilized and cannot show how hydrophobic feature is distributed over the individual interfaces. Larsen et al. [1998] studied the hydrophobic features of each protein-protein interface and highlighted them with images. Although all the interfaces are formed between two globular subunits, the hydrophobic distribution pattern over the interfaces is quite variable: (1) Some interfaces show a recognizable hydrophobic core, with a single large, continuous, hydrophobic patch surrounded by a ring of inter-subunit polar interactions; (2) Some interfaces do not have a single hydrophobic core, instead they have many small hydrophobic patches which consist of 1-3 amino acids. These small hydrophobic patches are not discrete, they are linked with inter-subunit hydrogen bonds and water molecules and all patches are distributed across the interface. Moreover the scattered hydrophobic patches do not dominate the interface character like the single large hydrophobic core; (3) There is still a small portion of the interfaces formed by extensive inter-digitation of the two subunit chains. These interfaces are highly hydrophobic similar to the hydrophobic cores inside the folded domain. They are usually associated with proteins that are quite stable and internally symmetric. From the studies on proteins folding path, they may be formed through two-state folded theory.

Chapter 2

Hydrophobic effect is important to protein-protein interactions. So, understanding it will help unveil the nature of protein-protein interactions, and help build the prediction model of protein interaction sites.

2.1.4 Amino acid composition at interfaces

Many studies analyzed amino acids composition at protein-protein interfaces [Jones & Thornton, 1996; Lijnzaad & Argos, 1997; Dasgupta, et al., 1997; Lo Conte, et al., 1999; Glaser, et al., 2001; Chakrabarti & Janin, 2002; Brinda, et al., 2002; Ma, et al. 2003; Bahadur, et al., 2003 ; Neuvirth, et al., 2004; Rajamani, et al., 2004; Halperin, et al., 2004; Keskin, et al., 2005]. Different studies found that the amino acid distributions differ in different protein complexes. In small globular proteins the interface consists of 57% non-polar residues, 24% neutral polar residues, and 19% charged residues respectively [Miller, et al., 1987]. While in the oligomeric proteins it consists of 65% non-polar residues, 22% neutral polar residues, and 13% charged residues respectively [Janin, et al., 1988]. It indicates that the amino acid composition at the protein-protein interface is more similar to that of the protein surface than protein interior.

Jones & Thornton [1996] studied residue propensities at the interfaces. The results showed that the charged and polar residues, especially ARG and ASP, show an increased affinity for the interface. In addition, the hydrophobic residues MET and PRO show a small increased affinity for the interface. On average, the interface comprises 56% non-polar carbon-containing groups, 29% neutral polar groups, and 15% charged groups [Jones & Thornton, 1996].

Chapter 2

Lijnzaad & Argos [1997] showed that aliphatic and aromatic residues as well as PRO are the largest contributors to interface. LEU, ILE, PHE, VAL and PRO occur progressively more often in the larger patches, whereas the contributions of TRP and TYR roughly depend on the patch size. And, hydrophobic residues were abundant in large interfaces while polar residues were more abundant in small interfaces [Glaser, et al., 2001; Zhou & Shan, 2001]. ALA and MET are intermediate contributors. Charged residues are shown to have less affinity to the interface, especially as the patch size grows.

Dasgupta et al. [1997] showed that hydrophobic interactions at oligomer interfaces favour aromatic amino acids and MET over aliphatic amino acids. However, ARG is the top residues appearing at the oligomeric interfaces. This suggests that ARG might be very helpful to inter-subunit interactions. Musafia et al. [1995] showed that ARG is especially adaptable to the formation of multiple salt bridge interactions, which may explain its structural role at oligomer and crystal interfaces.

Lo Conte et al. [1999] showed that interfaces are much richer in aromatic residues HIS, TYR, PHE and TRP than the average protein surface (21% versus 8 %), and somewhat richer in aliphatic residues LEU, ILE, VAL and MET (17% versus 11 %). They are depleted in the charged residues ASP, GLU and LYS, but not ARG [Zhou & Shan, 2001], which is the only residue that makes the largest overall contribution to interfaces (10 %). Some variations are seen between different types of complexes, but the largest contribution of ARG and depletion in LYS are general conclusions.

Chapter 2

Chakrabarti & Janin [2002] showed using a dataset of 70 protein-protein complexes that TRP and TYR have the highest propensity for the core of recognition sites, and SER and THR have a negative propensity. ARG is the most abundant core residue but also generally abundant on the protein surface.

Brinda et al. [2002] analyzed 20 homodimer interfaces using graph-spectral methods. ARG, HIS, PHE, TYR and GLU are found to be the most preferred residues in the interface clusters. There is also a significant contribution from TRP and MET in the interface clusters when compared with the other amino acids. It is clear that the charged and aromatic residues prefer to stay interface clusters. In contrast, GLY, ALA, VAL and CYS are rarely found in the interface side chain clusters.

Ma et al. [2003] studied structurally conserved residues between binding sites and exposed protein surfaces. The results showed that conservation of TRP on the protein surface indicates a highly potential binding site. To a less extent, conservation of PHE and MET also implies a binding site. For all three residues, there is a significant conservation in binding sites, whereas there is no conservation on the exposed surface.

Bahadur et al. [2003] studied subunit interfaces using a dataset of 122 homodimers. LEU was shown to be the most abundant residue at the homodimer interfaces, which contributed about 10% of the interface area. Other aliphatic residues, ILE, VAL and MET are also main contributors to interfaces. Aromatic residues, PHE, TYR and TRP are a little more abundant at the interface than at the protein surface. Charged residues, ASP, GLU and LYS are depleted at the interface. But ARG is the second largest contributor to homodimer interfaces after LEU.

Chapter 2

Neuvirth et al. [2004] counted TYR, MET, CYS and HIS as the most favoured to be at the interface. While residues, THR, PRO, LYS, GLU, and ALA, were least commonly found at the interfaces. In this study, ARG was not found to be abundant at the interfaces.

Rajamani et al. [2004] showed a particularity of protein interfaces which is the existence of so-called ‘anchor’ residues. These residues can bind to specific pockets in the target protein and create a weak intermediate state that can rapidly convert to the final complex. Molecular dynamics simulations suggest that anchor residues adopt similar conformations in the complex and in the corresponding isolated protein, although this may not be reflected by their conformations in protein crystal structures due to packing or environmental effects.

Another well-characterized property of protein interfaces is the existence of ‘hot-spot’ residues [Ma, et al. 2003; Halperin, et al., 2004; Keskin, et al., 2005]. These residues are found to make a significant thermodynamic contribution to the complex formation in so-called ‘alanine scanning’ experiments. It has been shown that polar residue occur at binding interfaces and are correlated with residue conservation. The number of such residues appears to be proportional to the size of the interface.

In conclusion, hydrophobic residues (except ALA) showed an increased affinity to protein-protein interfaces. Moreover, aromatic residues are a little more abundant at the interfaces than aliphatic residues. TYR and TRP are found to have the highest propensity at the interfaces. Hydrophilic and charged residues (except ARG) are found to be depleted at the interfaces. Most studies showed that ARG is among the

Chapter 2

top contributors to the interfaces.

2.1.5 Other properties of protein-protein interfaces

Besides the properties discussed above, there are also some other features characterizing protein-protein interfaces. These features include: (1) interface shape [Hurley et al., 1989; Smith et al., 1989; Jones & Thornton, 1996; Chakrabarti & Janin, 2002], (2) geometrical complementarity [Lawrence & Colman, 1993; Harpaz, et al., 1994; Jones & Thornton, 1996; Tsai, et al., 1999; Lo conte, et al., 1999], (3) secondary structures of interface residues [Jones & Thornton, 1996; Dasgupta, et al., 1997; Tsai et al., 1997a; Neuvirth, et al., 2004], (4) electrostatic complementarity [Janin & Chothia, 1990; Nicholls, et al., 1991; Roberts, et al., 1991; Novotny & Sharp, 1992; Karshikov, et al., 1992; Demchuck, et al., 1994; Hendsch & Tidor, 1994; Slagle, et al., 1994; Lescar, et al., 1995; Jones & Thornton, 1995; McCoy, et al., 1997; Xu, et al., 1997ab; McCoy, et al., 1997; Camacho, et al., 1999; Stevens, et al., 2000; Sheinerman, et al., 2000; Ivanov, et al., 2001], (5) water molecule effect on protein-protein interactions [Davies & Cohen, 1996; Wells, 1996; Guinto & Di Cera, 1996; Stites, 1997; Xu, et al., 1997b; Dall'Acqua, et al., 1998; Larsen, et al., 1998; Janin, 1999; Lo conte, et al., 1999; Janin, 1999; Archakov, et al., 2003], (6) conformational changes upon protein-protein association [De Vos, et al., 1992; Stanfield, et al., 1993; Wilson & Stanfield, 1993, 1994; Davies & Cohen, 1996; Jones & Thornton, 1996; Xu, et al., 1997a; McCammon, 1998; Dall'Acqua, et al., 1998; Vaughan, et al., 1999; Betts & Sternberg, 1999; Lo Lonte, et al., 1999; Najmanovich, et al., 2000; Sundberg & Mariuzza, 2000; Kimura, et al., 2001; Fradera, et al., 2002; Cole & Warwicker,

Chapter 2

2002; Kallblad & Dean, 2003; Liu, et al., 2003; Emberly, et al., 2003; Richards, et al., 2003]. These features will be discussed in details as the followings.

2.1.5.1 Geometrical shape

Interface shape varies in different types of protein structures. Protein-protein association may form a relatively flat or a twisted interface. Interface planarity (evaluated by RMSD) is used to measure how flat the protein-protein interfaces are. RMSD is a measure of how far the interface residues deviate from a best-fit plane [Jones & Thornton, 1996]. The more flat interfaces have relatively smaller RMSD value. Studies suggest that most interfaces are relatively flat, with a RMSD of less than 6 Å [Jones & Thornton, 1996]. They showed that hetero-complex interfaces are more planar than homodimer interfaces. There are also dimers in which the two subunits are twisted together across the interface, such as Isocitrate dehydrogenase [Hurley et al., 1989], or proteins that have subunits with arms apparently clasping the two halves of the structure together, such as Aspartate aminotransferase [Smith et al., 1989]. Moreover, Chakrabarti & Janin [2002] showed that antibody-antigen interfaces are more planar than protease-inhibitor using a dataset of 70 hetero-complexes. And the mean value of RMSD is 3.0 for the whole dataset.

2.1.5.2 Geometrical complementarity

The shape complementarity of the interface is a major element of the recognition process between two protein molecules. There are several ways to evaluate the protein geometric complementarity: (1) shape correlation index (S_c) [Lawrence & Colman, 1993]; (2) the ratio of the sum V of the Voronoi volumes to the sum V_0 of

Chapter 2

the reference volumes [Harpaz, et al., 1994; Tsai, et al., 1999]; (3) Gap index [Jones & Thornton, 1996].

Generally speaking, the larger the ratio is, the looser the atoms packing is. If the ratio (V/V_0) is equal to 1, the set of atoms should be close-packed. Lo conte et al. [1999] showed that at the protein interfaces the ratio (V/V_0) is in the range of 0.97-1.06. They also showed that the packing of the atoms at the interfaces is very similar to that of the protein interior.

Gap index is used by us, as well as by Jones & Thornton [1995]. They showed that the gap index value is in the range of 0.35-5.2 Å and the overall mean is 2.5 Å using a dataset of 32 homodimers. To some extent, gap index is an indicator of the strength of protein-protein interaction. The more complementary the two protein molecules are, the stronger they associate.

2.1.5.3 Secondary structure states

Studies show that the distribution of secondary structures of the interface residues varies a little in different studies and most interfaces contain more than one type of secondary structures. Jones & Thornton [1995] showed that secondary structures have the following orders: α -helix 53.1%, β -strand 21.8%, coil 12.5%, turn 12.5% using a dataset of 32 homodimers. However, they showed in an extended dataset of 32 homodimers and 27 hetero-complexes that there is an approximately equal proportion of helical, strand, and coil residues involved [Jones & Thornton, 1996]. Dasgupta et al. [1997] used a dataset of 58 oligomeric proteins to show that

Chapter 2

secondary structure has the following order: α -helix 40.4%, β -strand 36.7%, coil 22.9%.

Tsai et al. [1997a] showed that the secondary structure distribution in permanent complexes is similar to the protein cores using a dataset of 362 protein-protein interfaces and 57 oligomeric interfaces. However, this distribution in temporary complexes resembles more as protein surfaces.

Neuvirth et al. [2004] showed a different distribution using a dataset of 67 heterodimers. Results suggest that protein interfaces prefer β -strand and disfavour α -helix. This preference has major implications towards the interface structure, as β -strands form flat areas with three-dimensional connectivity, while α -helices form cylindrical perturbing surface structures with the three-dimensional structure following the one dimensional sequence.

The difference of secondary structures in these studies may result from dataset type. Neuvirth et al. [2004] used a dataset that contained only heterodimers, without homodimers or other oligomers. This may also suggest that secondary structure state at the protein interface is an important feature to characterize protein-protein interfaces.

2.1.5.4 Electrostatic complementarity

Electrostatic forces are shown to be significant forces involved in protein- protein interactions [Xu, et al., 1997a; Stevens, et al., 2000; Sheinerman, et al., 2000; Ivanov, et al., 2001]. It has long been a subject of study to unravel the nature of the

Chapter 2

electrostatic interaction between proteins. Studies showed that the opposite charge pairs have a highly favourable tendency to form salt bridges between protein complexes. There have been some elementary analyses to simply count the number of charged residues and salt bridges at protein-protein interfaces [Janin & Chothia, 1990; Jones & Thornton, 1995; McCoy, et al., 1997]. There is a range of 0-12 charged groups [Xu, et al., 1997b] and 0-5 salt bridges [Jones & Thornton, 1995] per protein-protein interface. But there has been no clear correlation between salt bridges and the interface size. Studies also showed that homo-complexes have less salt bridges and exhibit an extreme general preference for interactions between identical residues [Ofra & Rost, 2003a].

With the development of the continuum electrostatics model for proteins, the electrostatic potential of protein complexes can be calculated and displayed by numerically solving the Poisson-Boltzmann equation of the protein-solvent system, such as the program DelPhi [Nicholls, et al., 1991]. Analyses showed that protein interfaces have a charge complementarity [Roberts, et al., 1991; Novotny & Sharp, 1992] or an electrostatic complementarity [Karshikov, et al., 1992; Demchuck, et al., 1994; Hendsch & Tidor, 1994; Lescar, et al., 1995]. When compared with charge complementarity, electrostatic complementarity is more appropriate to characterize the protein-protein interactions [McCoy, et al., 1997]. Other study also showed that the charge complementarity does not always occur between protein subunits [Xu, et al., 1997b]. The results can be explained by the fact that all protein charge groups, not just those at the interfaces, contribute to the value of electrostatic complementarity at

Chapter 2

the interface. However, charge groups far away from the interfaces contribute less to the electrostatic potential of protein-protein interfaces [McCoy, et al., 1997].

Studies on electrostatics showed that in some cases salt bridges can contribute significantly to the stabilization of the protein subunits, but in other cases they may contribute little or even destabilize some protein complexes. Electrostatic contribution to the protein interaction may come from a generally more favourable, or less unfavourable, electrostatic free energy for transferring a hydrophilic pair from the protein surface in the folded but unbound state to the complex interface in the bound state, as compared to the change from the unfolded state to the folded state [Slagle, et al., 1994; Xu, et al., 1997b]. It appears that the hydrophobic effect and the electrostatic interaction contradict each other, but in fact it is not necessary. On the one hand, if there are more hydrophobic atoms at the interface, there would be less hydrophilic pairs, and vice versa. On the other hand, if the protein interfaces are too hydrophobic, the unbound state would be unstable; if the interfaces are too hydrophilic, there would be more unfavourable interactions. In fact, a favourable electrostatic interaction requires hydrophobic residues to work cooperatively with hydrophilic residues [Xu, et al., 1997b]. Some studies showed that in protein-protein association the electrostatic forces will steer the ligand into protein active pocket, then short-range electrostatics with desolvation further steer the ligand toward the local energy minima [Camacho, et al., 1999]. This is also proved by the experimental results [Schreiber & Fersht, 1996; Vijayakumar, et al., 1998].

Chapter 2

Therefore, electrostatic complementarity is more accurate to characterize protein-protein interaction compared to charge complementarity. To some extent, it can be evaluated by inter-subunit salt bridges. Understanding the nature of electrostatic complementarity and potential at the protein interface will be helpful to solve the problem of the protein docking and the prediction of protein interaction sites.

2.1.5.5 Water molecules effects on protein-protein interactions

Water molecules are frequently present at protein-protein interfaces [Davies & Cohen, 1996; Wells, 1996; Janin, 1999; Archakov, et al., 2003]. Because water molecules are not well determined and reported in medium-resolution or low resolution, to analyze water molecule effect on protein-protein interactions, the selected protein structures must be restricted to high resolution, which was proposed to be 2.4 Å or better [Lo conte, et al., 1999]. Besides resolution, there are many other factors influencing the determination of water molecules between proteins complexes [Larsen, et al., 1998], such as the level and the method of refinement.

The numbers of water molecules found at the interfaces vary in different studies. Lo conte et al. [1999] showed an average of 18 water molecules per subunit with a wide range of 3-50 water molecules. They also showed that there is a mean value of one water molecule per 100 Å². The results suggest that water-mediated polar interactions are more numerous than direct hydrogen bonds. Davies and Cohen [1996] showed that the number of water molecules varies from 1 to 50 at the protein interface. Studies also indicated that water molecules may surround the interfaces or may be buried in them [Davies & Cohen, 1996; Larsen, et al., 1998].

Chapter 2

The number of interactions with water molecules may still be underestimated even at the high resolution. And the larger standard deviation is partly artificial, for there is no established practice among crystallographers to describe the solvent structures. Although the number of water molecules has a weaker correlation with the interface area than hydrogen bonds do, but nearly all reported interface waters molecules are involved in forming hydrogen bonds [Larsen, et al., 1998]. They form hydrogen bonds with protein residues or other water molecules, bridging hydrogen bond network [Xu, et al. 1997b; Dall'Acqua, et al., 1998; Janin, 1999]. They both mediate non-complementary donor-donor or acceptor-acceptor pairs, and connect non-optimally originated donor-acceptor pairs [Xu, et al., 1997b]. A great number of buried water molecules at the interface may be responsible for the large change in heat capacity for the protein-protein association [Guinto & Di Cera, 1996; Stites, 1997].

Therefore, in view of structure and energy, interface water molecules play an important role in stabilizing protein complexes by forming additional hydrogen bonds, interacting with charge groups, and increasing shape and charge complementarity [Stites, 1997; Xu, et al., 1997b; Larsen, et al., 1998; Janin, 1999]. However, the understanding of the nature of the effect caused by water molecules on protein-protein interactions is still far from satisfaction. And most studies on protein-protein interactions neglect their effect. So it results in difficulty in building prediction model of protein interaction sites.

Chapter 2

2.1.5.6 Conformational changes in protein-protein interactions

When protein molecules interact and form complexes, there always exist some conformational changes between subunits [McCammon, 1998; Betts & Sternberg, 1999; Lo Lonte, et al., 1999; Sundberg & Mariuzza, 2000; Kimura, et al., 2001]. However, they have only the limited conformational freedom of six degrees of translation and rotation for optimization of the subunit structures. This limitation will result in cavities, water molecule presence and non-optimal hydrogen bonds at the protein interface [Xu, et al., 1997a; Dall'Acqua, et al., 1998; Vaughan, et al., 1999]. Therefore, studies about side-chain conformational entropy at the interface indicate that the protein interface is less flexible than the rest of the protein surface [Cole & Warwicker, 2002; Liu, et al., 2003]. If the subunit structures are available independently of the complexes, the extent and nature of the conformational changes can be assessed more easily. However, today they are not very clear because the available structures are limited. However, it is possible to distinguish various levels of conformational changes. These levels of conformational changes include no change, side chain movement alone, segment movement involving the main-chain (e.g. hinged loop) and domain movement [Jones & Thornton, 1996; Betts & Sternberg, 1999].

For moderate conformation changes, the extent can be evaluated by RMSD. It is derived by the method of least-squares superposition using equivalent main chain atoms between the bound and unbound subunits. Richards et al. [2003] showed that many protein-protein complexes have a RMSD in the range of 0.5-1 Å. At the interface, the range of the main chain RMSD is typically 1-2 Å and a few side

Chapter 2

chains reorient. In solution, proteins have also been observed to undergo fluctuation of the similar type and amplitude. When comparing the different crystal structures of the same protein, there are local movements of up to 1 Å and side chain rotations [Richards, et al., 2003]. It is found that the ligand bound introduces in general small structural changes at the protein binding site. However, such small structural changes can lead to important alterations in the protein recognition pattern [Fradera, et al., 2002]. Some protein complexes may have large movements and the RMSD is well over 2 Å [Najmanovich, et al., 2000; Richards, et al., 2003]. For example, antigen-antibody recognition is usually accompanied with a wide range of conformational changes [Stanfield, et al., 1993; Wilson & Stanfield, 1993, 1994; Davies & Cohen, 1996].

The conformation change may be located at the interface or affect the whole protein. In some complexes, α -helix may unwind or change its orientation. However, most conformation is due to collective motion of many residues. Although every residue in α -helices and β -strands has a small movement, the cumulative result may play an important role in protein-protein association [Emberly, et al., 2003]. There are also some proteins observed to have quaternary structure changes upon complexation [Stanfield, et al., 1993; Wilson & Stanfield, 1994]. An example is the human growth hormone receptor, which dimerizes on binding the hormone [De Vos, et al., 1992].

In some cases, there are disorder-to-order transitions upon protein-protein association, in which part or all the polypeptide chain in one component is unfolded to the free form and becomes structured again upon complexation. The free TRY loops have the

Chapter 2

disordered polypeptide chain up to about 25%. After association with PTI, they become ordered and take the same conformation as in the mature enzyme [Marquart, et al., 1983]. Studies suggest that essentially all protein complexes with B-factor > 2000 Å² (B-factor is an indicator of thermal motion about atom. In one PDB file, the atom with higher B-factor would have higher oscillation or vibration around the position in the model.) undergo large changes and disordered-to-ordered transitions on association [Richards, et al., 2003].

Besides backbones, conformation changes may result from side-chain flexibility. Only considering the dihedral angles of rotatable side-chain bonds in binding pocket, the flexibility scale of side-chain has the following order: LYS > ARG, GLN, MET > GLU, ILE, LEU > ASN, THR, VAL, TYR, SER, HIS, ASP > CYS, TRP, PHE. LYS side-chains flex 25 times more often than the PHE side-chains do. Side-chains are found to be more rigid in the binding pockets than at the interface. This is perhaps due to functional constraints in ligand recognitions [Betts & Sternberg, 1999; Najmanovich, et al., 2000]. It is also found that there is no correlation between the residue backbone movement and its side-chain flexibility [Najmanovich, et al., 2000].

It is assumed that the interacting proteins should drive the appropriate conformational changes, leading to their complementarity. However, this hypothesis can not explain their fast association rate. Side-chains generally change their conformation on association [Kallblad & Dean, 2003]. Studies about key side-chains in protein-protein recognition may give an explanation that in all likelihood when ligands interact with

Chapter 2

their receptors their key side-chains are their right bound-like conformations [Kimura, et al.,2001].

The plasticity of the protein structures plays a great role in protein-protein interactions. First, it enables protein interfaces to accommodate mutations as they co-evolve; Second, it allows a single molecular surface to interact with numerous, structurally distinct binding partners [Sundberg & Mariuzza, 2000]; Third, it is a need of low-energy conformational changes, which enable hydrogen bond or salt bridge formation and packing the amino acid residues [Betts & Sternberg, 1999]. As more structures of protein complexes and their unbounded components are solved, the extent and nature of conformational changes will be unraveled on protein association [Betts & Sternberg, 1999]

2.2 Interaction site prediction models

With the development of new technologies, the pace of discovery of protein-protein interactions has been accelerated. Progress in the field was stimulated by the completion of dozens of genome sequencing projects, followed by the development of high-throughput experimental methods aimed at functional characterization of the newly discovered genes [Rigaut, 1999; Yates, 2000; Sobott & Robinson, 2002]. The vast amount of data collected so far necessitates the systematic organization of the new information in a form amenable to analysis on the scale of entire genomes. Moreover, there are many discrepancies between the interacting partners identified in high-throughput studies and those identified in small scale experiments. Even when interactions have been validated, one needs to be aware that the majority of the

Chapter 2

experimental evidence cannot distinguish between direct interactions and those mediated by at least one intermediate protein [Edwards, et al., 2002]. Therefore, the development of reliable and efficient methods to detect and characterize protein–protein interactions is central to functional genomics [Janin & Seraphin, 2003]. It is expected that these analyses can reveal the large-scale patterns of protein interactions responsible for higher level properties of organisms, such as adaptation, robustness and error correction, as well as shedding light on the evolutionary history of genomes [Salwinski & Eisneberg, 2003].

Studies have shown that protein-protein interfaces have detectable properties that can distinguish them from the rest of the protein surface [Young, et al., 1994; Clackson & Wells, 1995; Jones & Thornton, 1997a; Bogan & Thorn, 1998; Larsen, et al., 1998; Lo Conte, et al., 1999; Hu, et al., 2000; Jones, et al., 2000; Glaser, et al., 2001; Valdar & Thornton, 2001a ; Kortemme & Baker, 2002; Cole & Warwicker, 2002; Fernandez & Scott, 2003; Nooren & Thornton, 2003ab; Neuvirth, et al., 2004; Halperin, et al., 2004; Bahadur, et al., 2004]. These properties are widely used to detect potential interaction sites or functional sites. Researchers have been trying to accomplish this challenge from different aspects, which will be discussed as the followings.

2.2.1 From sequence

In biology, it is virtually axiomatic that the protein sequence specifies conformation [Anfinsen, 1973]. This suggests an intriguing postulate that knowledge of the protein sequence alone might be sufficient to evaluate protein-protein interactions and derive useful biological function [Bock & Gough, 2001]. This is very useful especially in

Chapter 2

cases in which the protein's three-dimensional structure is unavailable. Usually residues that are conserved for the entire family or a subfamily within the alignment are predicted to be functional [Casari, et al., 1995; Livingstone & Barton, 1996; Caffrey, et al., 2000; Hannenhalli & Russell, 2000]. Studies show that interface residues, particularly those completely buried in the interface, are likely to be more conserved than other surface residues [Livingstone & Barton, 1993; Grishin & Phillips, 1994; Valdar & Thornton, 2001a; Matthews, et al., 2001; Yao, et al., 2003]. It has been shown that protein-protein interfaces are more conserved than expected for a random distribution and that in most cases this conservation is statistically significant at the 5% level [Valdar & Thornton, 2001a]. However, recently it has been argued that the residue conservation is rarely sufficient for complete and accurate prediction of the protein-protein interface [Bradford & Westhead, 2003; Caffrey, et al., 2004].

Several investigators have tried to predict functional site using multiple sequence alignments [Casari, et al., 1995; Andrade, et al., 1997; Hannenhalli & Russell, 2000; Li, et al., 2003]. Casari et al. [1995], for example, applied principal component analysis using a vector representation of protein sequences in a multidimensional "sequence space," to derive subfamily-specific residues involved in protein function. Andrade et al. [1997] proposed a rigorous clustering algorithm based on a self-organizing map to identify protein subfamilies and retrieve characteristic sequence patterns.

Some studies considered methods that map the conservation score to the three-

Chapter 2

dimensional structure. These methods are mainly extensions of Evolutionary Trace approach [Lichtarge, et al., 1996; Lichtarge & Sowa, 2002; Glaser, et al., 2003], Consurf [Armon, et al., 2001], Rate4Site [Pupko, et al., 2002], phylogenetic trees [Sjolander, 1998; Aloy, et al., 2001; Madabushi, et al., 2002; Del Sol Mesa, et al., 2003; La, et al., 2005], and the method of Landgraf and Eisenberg [Landgraf, et al., 2001]. They identify conserved positions and map them onto the three-dimensional structure of the protein (or onto the structure of a representative homolog), looking for spatially contiguous patches of conserved residues. Frequently, structural clusters of Evolutionary Trace positions, which are statistically significant [Madabushi, et al., 2002], are used to the fine-tune functional site prediction [La, et al., 2005]. A potential caveat of these methods is that they require reliable methods to identify orthologs or to define protein families or subfamilies in which the considered interaction or function is conserved. This is a very difficult task, as recently discussed [Abascal & Valencia, 2003]. The availability of a representative three-dimensional structure is also important, because establishing the spatial proximity of the conserved positions is crucial for singling out those that are functionally relevant.

Another set of approaches used neural networks to predict interaction sites [Zhou & Shan, 2001; Fariselli, et al., 2002; Ofra & Rost, 2003ab]. Zhou & Shan [2001] used sequence profiles of neighboring residues and their solvent exposure as input to predict interface residues. For a set of 35 test proteins, the correct fractions of interface residues were 69% for bound structures and 70% for unbound structures. Fariselli et al. [2002] achieved above 70% accuracy for classifying residues in and outside of the interface using two very basic sources of information: evolutionary

Chapter 2

information as accumulated in sequence profiles derived from family alignments and surface patches in protein structures identified as sets of neighbor residues exposed to solvent. Ofra & Rost [2003a & 2003b] showed that protein interfaces have a significantly different amino acid composition from the rest of the protein surface and that the majority of interacting residues are clustered in sequence segments of several contacting residues. Indeed, recent studies indicate that the interfaces of many protein complexes can be subdivided into distinct patches where contacts are often provided by different protein domains [Chakrabarti & Janin, 2002]. Also, interactions with contiguous chain segments are clearly a hallmark of specific recognition by many signaling domains [Pawson & Nash, 2003]. They made use of these two conclusions and neural network to predict protein interaction sites from protein sequences and achieved a surprisingly high level of accuracy [Ofra & Rost, 2003ab].

SVM is also widely used to predict interaction sites [Bock & Gough, 2001; Sprinzak & Margalit, 2001; Koike & Takagi, 2004; Martin, et al. 2005; Bradford & Westhead, 2005]. Koike & Takagi [2004] used the profiles of sequentially/spatially neighboring residues and additional information to predict interface residues. When structural information was used, the predicting performance was improved greatly. Bradford & Westhead [2005] combined the SVM approach with surface patch analysis to predict protein-protein binding sites. Using a leave-one-out cross-validation procedure, they successfully predicted the location of the binding site with an accuracy of 76% of their dataset made up of proteins with both transient and obligate interfaces. SVM approach is also used to predict protein-protein interactions combined with the sequence information and experimental data [Bock & Gough, 2001; Sprinzak &

Chapter 2

Margalit, 2001; Martin, et al., 2005]. There are some differences between these studies. Bock & Gough transformed sequence information into physico-chemical information (charge, hydrophobicity and surface tension). However, the signature descriptor used by Martin and colleagues did not require performing such a transformation. Furthermore, Bock and Gough encoded and compared protein pairs by concatenating normalized versions of the amino acid sequences of each protein, and hence used a global representation of a protein pair. However, Martin and colleagues used explicit pair-oriented, local sequence information (product signature). These amino acid subsequence pairs should occur together when two proteins interact

It is a great challenge for all researchers to detect putative protein interaction sites from the sequence information only. Success in this area will encourage the development of genomics and proteomics. This will also greatly benefit protein design and drug design. Therefore, it is our interest to predict interaction sites from homodimer and heterodimer sequence data by analyzing their interface properties.

2.2.2 Correlated mutation

Through genome-wide analyses and studies of many specific systems, some researchers showed that interacting protein pairs tend to evolve in a correlated manner [Pazos, et al., 1997; Marcotte, et al., 1999; Enright, et al., 1999; Eisenberg, et al., 2000; Huynen, et al., 2000; Pazos & Valencia, 2001, 2002; Tillier & Lui, 2003]. These studies include using genomic context to infer protein-protein interactions between the gene products [Huynen, et al., 2000], and inferring from genome sequences, given observed homologies in other organisms, where interacting proteins

Chapter 2

have fused into a single protein chain [Marcotte, et al., 1999; Enright, et al., 1999].

By analyzing correlated mutations between pairs of residues in multiple sequence alignments of two proteins, several studies showed that positions where changes occur in a correlated fashion in the two interacting protein molecules tend to be close to the protein-protein interfaces [Pazos, et al., 1997; Eisenberg, et al., 2000; Pazos & Valencia, 2001, 2002]. This method can predict interacting pairs of residues from the sequence alone without the knowledge of the interacting protein structures. Although so far only used for the prediction of protein pairs that are likely to interact, it has the potential to identify interaction sites and ultimately interacting residues, provided that the sample of aligned sequences is large enough.

Tillier & Lui [2003] proposed a method that computes the co-variation of two or more residues in a multiple sequence alignment and factors out co-variations resulting from phylogenetic relationships. This approach is proposed for the identification of functional residues and interaction sites, with a very low rate of false positives.

2.2.3 Homology modeling and threading

The prediction of protein-protein interactions from sequence information can go all the way to building a detailed atomic model of the complex. Some studies detect protein interaction sites using homology modeling and threading. Marcotte et al. [1999] proposed a computational method to infer protein interactions from genome sequences on the basis of the observation that some interacting protein pairs have homologs in another organism fused into a single protein chain.

Chapter 2

Some methods used the known protein structure as the template for building a model of the target assembly. In one approach, also available on the web (<http://www.russel.embl.de/interprets>) [Aloy & Russell, 2003], the sequences of the component proteins are assigned to domains annotated in PFAM [Bateman, et al. 2004]. These domains are then matched against a database of known protein structures. Once a match is found, a model is built using similar techniques to those employed in homology modeling.

Cases in which the component proteins cannot be reliably assigned to PFAM domains can, in principle, be rescued by using the threading approach [Lu, et al., 2002, 2003]. This approach first identifies the putative folds of the interacting partners using an established threading method. Once these are matched to a complex of known structure, multi-chain threading is performed to produce sequence/structure alignments from which a detailed model of the complex can be built. So far, however, this method has been evaluated for its capacity to identify interacting partners in fully sequenced genomes and not for inferring information on the actual interaction mode. Indeed, extrapolating information on interactions from the structure of one complex to homologs of its components clearly carries a substantial risk. This risk is analogous to that encountered in other methods for inferring function on the basis of sequence similarity [Abascal & Valencia, 2003]. Assessing this risk in the realm of protein-protein interactions, Aloy & Russell [2003] showed recently that close homologs, displaying at least 30-40% sequence identity, interact in the same way, whereas the interaction mode is rarely conserved among more distantly related proteins.

Chapter 2

2.2.4 Protein-protein interaction network

Currently some studies tried to infer yet unknown functions for the unclassified proteins by exploiting the global protein-protein interaction network structure on the base that the rest of the interaction graph is at least partially known. These methods include the so-called Neighboring Counting Method [Schwikowski, et al., 2000], the χ^2 -method [Hishigaki, et al., 2001], the Bayesian approaches [Deng, et al., 2002; Letovsky & Kasif, 2003], the Redundancy Method [Samanta & Liang, 2003] and a more recent Monte Carlo simulated annealing (SA) approach [Vazquez, et al., 2003]. Letovsky & Kasif successfully predicted protein function based on the local functional density of interacting neighbors using Markov Random Field formalism [Letovsky & Kasif, 2003].

Albert & Albert [2004] showed that the conserved properties of the protein-protein interaction network can be used to identify and validate interaction candidates. They applied a number of machine learning algorithms to the protein connectivity information and achieve a surprisingly good overall performance in predicting interacting proteins. Using a ‘leave-one-out’ approach they find average success rates between 20 and 40% for predicting the correct interaction partner of a protein. The success of these methods is based on the presence of conserved interaction motifs within the network. So it only employed work without requiring additional biological knowledge and used the connectivity data as the only source of implicit information.

Leone & Pagnani [2005] proposed a fast online analysis algorithm to predict protein-protein interactions. The algorithm can easily be generalized into more complex

Chapter 2

graph topologies taking into account hypergraphs, i.e. complexes of more than two interacting proteins. This method adopted a message passing algorithm known as Belief Propagation, which accepts the network of protein physical interactions and a catalog of known protein functions as input, and returns the probabilities for each unclassified protein of having one chosen function.

With development of technologies, the protein-protein interaction network will become more and more complete. Then, more information can be used to predict protein-protein interactions and protein functions.

2.2.5 Other methods

There are many other studies employed that attempted to detect protein-protein interaction sites. Some researchers attempted to solve this problem from protein three-dimensional structures [Jones & Thornton, 1997ab; Neuvirth, et al., 2004]. Jones & Thornton [1997a & 1997b] analyzed protein-protein interface properties using patch analysis. Then, they showed that interfaces can be differentiated from the rest of protein surface by combining six interface properties. Neuvirth et al. [2004] showed that the binding sites have a preference for β -sheets and for relatively long non-structured chains, but not for α -helices.

Gallet et al. [2000] proposed a RBD (receptor-binding domains) method to detect putative interaction sites and promising mutations from sequences. It is based on the analysis of sequence hydrophobicity and principally detects hydrophilic domains. The procedure could be combined with multiple sequence alignments to identify

Chapter 2

homologous binding sites or, in contrast, to elicit the absence of functional interaction domains

Supervised Bayesian Learners have also been used to combine multiple genomic features into reliable predictions of interacting proteins [Jansen, et al., 2003]. Inference rules for new protein functions were formed by combining known protein features with interaction partner information [Oyama, et al., 2002].

Fernandez-Recio et al. [2004] applied protein docking simulations and analysis of the interaction energy landscapes to identify protein-protein interaction sites. They found that highly populated low-energy regions consistently corresponded to actual binding sites

Panchenko et al. [2004] presented a method which is based on the assumption that the structural location of functional sites is conserved between homologous proteins and that functionally important residues tend to cluster together in space, forming three-dimensional residue clusters or surface patches. In this method, each residue is assigned a score which depends on its own conservation in homologs and the conservation of residues in its spatial neighborhood, as judged from the analysis of known structures within a given protein family. They hypothesized that high-scoring sites are more likely to be involved in specific binding or catalysis, and that one may identify functionally important residues even in the absence of structural data on protein-ligand or macromolecular complexes.

Although the interface can be quite large, it was shown in some systems that only a

Chapter 2

small fraction of residues contributes to the majority of the binding energy [Clackson & Wells, 1995]. Furthermore, these so-called hotspots of binding energy tend to have preferred residue types that often have a high degree of burial at the interface [Bogan & Thorn, 1998]. Interestingly, there is evidence (for 11 families) to suggest a relationship between the enrichment of a residue type in a hotspot and the propensity of the corresponding residue to be conserved [Hu, et al., 2000]. Significant progress has been made in predicting the location of hot-spot residues using a free energy formulation based on Lennard-Jones, solvation, hydrogen bonding and electrostatic terms combined with Monte Carlo simulated annealing. Although side-chain repacking could also be taken into account, almost 80% of the hot spots in the trial could be detected without this refinement [Kortemme & Baker, 2002]. These hotspots usually suggest interaction sites.

It is difficult at present to judge the predictive power of the various methods outlined above. Because all of them are based on learning from examples, their performance might vary from case to case, essentially depending on the type of considered protein-protein interaction and on the size of the learning sample. Benchmarks that take these parameters into consideration are still lacking. However, it is true that no single parameter absolutely differentiates protein interaction sites from other surface patches [Jones & Thornton, 1997a], so we should combine more than one of physical-chemical interface properties to detect interaction sites [Jones & Thornton, 1997b; Zhou & Shan, 2001; Fariselli, et al., 2002; Keil, et al., 2004; Neuvirth, et al., 2004].

Chapter 2

2.3 Conclusion

Dimer interactions are characterized by a large combination of physical-chemical parameters. Analysis of dimer structures can provide insight into the principles of protein-protein complexation and help develop models to predict interaction sites.

With the development of computational and experimental technologies, the vast amount of protein-protein interaction data has been collected so far. These data have been accelerating the pace of discovery of protein-protein interactions. However, the knowledge on these interactions is far from satisfactory. Dataset are various in different studies, such as dataset type, dataset size, criteria of selected dataset. Moreover, different interface features are used to describe protein-protein interactions. Therefore, we describe interface features using a more complete parameter set and extract critical ones to reduce the redundancy. Then, we discuss interface differences between heterodimers and homodimers using a selected set of properties.

Although, there have been many methods to analyze and detect protein-protein interaction sites, the results are still far from satisfactory. There is still more room to improve the prediction models. Success in the development of prediction model will encourage the development of genomics and proteomics. This will also greatly benefit protein design and drug design. Therefore, it is our interest to predict interaction sites from homodimer and heterodimer sequence by analyzing their interface properties.

Chapter 3

Identification of critical heterodimer protein interface parameters by multi-dimensional scaling in Euclidian space

3.1 Introduction

Protein-protein interactions play a key role in many biological processes such as signal transduction, gene regulation and antibody-antigen recognition [Loregian, et al., 2002; Shulman-Peleg, et al., 2004]. Therefore, a study on the principles of protein-protein interaction is critical for developing reliable prediction models from sequence data. Current models largely depend on the available knowledge of protein-protein interaction sites [Xu, et al., 1997ab; Halperin, et al., 2002]. However, many model parameters have not been fully captured due to limited structural data and lack of rigorous mathematical formulations.

Studies indicate the presence of charge and electrostatic complementation at the protein-protein interfaces [McCoy, et al., 1997; Hu, et al., 2000]. Formation of hydrogen bonds between subunits plays an important role in the association and stability of protein subunits [Meyer, et al., 1996; Fernandez & Scheraga, 2003]. Residue propensity between interior, exterior and interface regions of oligomeric proteins has been examined [Jones & Thornton, 1995]. They showed the selective occurrence of non-polar residues in the interior and at interface regions of proteins, while polar (or charged) residues prefer solvent exposed exterior regions. Thus, a number of parameters have been known to characterize a protein-protein interface.

Chapter 3

Nonetheless, it is extremely difficult to capture all the non-linear dependencies of such parameters.

A number of methods have been used to identify interface parameters in oligomeric complexes. These methods utilize conserved residues at interface [Ma, et al., 2003], surface patches [Jones & Thornton, 1997b], sequence features [Zhou & Shan; 2001; Ofra & Rost, 2003b; Panchenko, et al., 2004], atomic contact vectors [ACV] [Mintseris & Weng, 2003], topological entities [Chou & Cai, 2004], neural network trained sets [Fariselli, et al., 2002], interface energy landscapes [Recio, et al., 2004], and SVM [Koike & Takagi, 2004]. However, these methods lack sufficient parameters for model development and are often less conclusive in prediction. Here, we analyze 118 interface parameters in 156 heterodimer structures to select the most critical ones in subunit interactions.

3.2 Material and Methods

3.2.1 Creation of a hetero dimer structural dataset

We created a dataset of 156 high resolution ($\leq 2.5\text{\AA}$) heterodimer structures determined by X-ray crystallography for this analysis (Table 3.1). These structural data were obtained from PDB. The dataset was selected such that each polypeptide in these heterodimers is at least 50 residues long.

Chapter 3

Table 3.1: Heterodimers used in this study

PDB code	Resolution (Å)	Chain one	Name of chain one	Length	Chain two	Name of chain two	Length
1YCS	2.2	B	53BP2	193	A	P53	191
1ABR	2.1	B	Abrin-A	267	A	Carbohydrate	251
1KU6	2.5	A	Acetylcholinesterase	535	B	Fasciculin 2	61
1LFD	2.1	B	Active ras protein	167	A	Ras-interacting domain of ralgs	87
1JIW	1.7	P	Alkaline metalloproteinase	470	I	Proteinase inhibitor	105
1BPL	2.2	B	Alpha-amylase	290	A	Alpha-amylase	179
1KXV	1.6	A	Alpha-amylase	496	C	Camelid VHH domain cab10	119
1TMQ	2.5	A	Alpha-amylase	470	B	Ragi bifunctional inhibitor	117
1BVN	2.5	P	Alpha-amylase	496	T	Tendamistat	71
1ACB	2.0	E	Alpha-chymotrypsin	241	I	Eglin C	63
1CHO	1.8	E	Alpha-chymotrypsin	238	I	Turkey ovomucoid third domain	53
1CGI	2.3	E	Alpha-chymotrypsinogen	245	I	Trypsin inhibitor	56
1SLU	1.8	B	Anionic trypsin	216	A	Ecotin	131
1RE0	2.4	B	ARF guanine-nucleotide exchange factor 1	195	A	ADP-ribosylation factor 1	162
1KSH	1.8	A	ARF-like protein 2	164	B	Cyclic phosphodiesterase delta-subunit	141
1MG9	2.3	B	ATP dependent CLP protease	143	A	Protein YLJA	84
1BRL	2.4	A	Bacterial luciferase	340	B	Bacterial luciferase	319
1AVA	1.9	A	Barley alpha-amylase 2	403	C	Barley alpha-amylase/subtilisin inhibitor	181
1B27	2.1	A	Barnase	110	D	Barstar	90
1LUJ	2.5	A	Beta-catenin	501	B	Beta-catenin-interacting protein ICAT	71
1SOW	2.3	A	Beta-lactamase tem	263	C	Beta-lactamase inhibitory protein	165
1BND	2.3	A	Brain derived neurotrophic factor	109	B	Neurotrophin 3	108
1D4X	1.8	A	C. Elegans actin 1/3	368	G	Gelsolin	124
1G4Y	1.6	R	Calmodulin	147	B	Calcium-activated potassium channel RSK2	81
1DTD	1.7	A	Carboxypeptidase A2	303	B	Metallocoarboxypeptidase inhibitor	61
1NW9	2.4	B	Catalytic domain of caspase-9	238	A	Inhibitor of apoptosis protein 3	91
1OKK	2.1	D	Cell division protein	265	A	Signal recognition particle protein	290
1H1S	2.0	A	Cell division protein kinase 2	296	B	Cyclin A2	258
1OHZ	2.2	A	Cellulosomal scaffolding protein A	140	B	Endo-1,4-beta-xylanase Y	56
1HL6	2.5	A	CG8781 protein	119	B	Mago nashi protein	137
1P5V	1.7	A	Chaperone protein CAF1M	191	B	F1 capsule antigen	136
1PDK	2.4	A	Chaperone protein PAPD	296	B	Protein PAPK	258
1N0L	2.3	A	Chaperone protein PAPD	212	B	Mature fimbrial protein PAPE	116
1FFG	2.1	B	Chemotaxis protein chea	68	A	Chemotaxis protein chey	128
1EAY	2	A	Chey	128	C	Chea	67
1P2M	1.8	A	Chymotrypsinogen A	238	B	Pancreatic trypsin inhibitor	58
1HOG	2.2	A	Coagulation factor	236	B	Coagulation factor	51
1V74	2.0	A	Colicin D	107	B	Colicin D immunity protein	87
1E44	2.4	B	Colicin E3	96	A	Immunity protein	84
1FR2	1.6	B	Colicin E9	131	A	Colicin E9 immunity protein	83
1F5Q	2.5	A	Cyclin dependent kinase 2	296	B	Gamma herpesvirus cyclin	247
1FIN	2.3	A	Cyclin-dependent kinase	298	B	Cyclin A	260
1BLX	1.9	A	Cyclin-dependent kinase 6	305	B	P19ink4D	160
1M9E	1.7	A	Cyclophilin A	164	D	HIV-1 capsid	135
1S6V	1.9	A	Cytochrome C peroxidase	294	B	Cytochrome C	108
1R8S	1.5	E	Cytohesin 2	187	A	ADP-ribosylation factor 1	160
1UJZ	2.1	B	Designed colicin E7 dnase	127	A	Designed colicin E7 immunity protein	87
1NLV	1.8	A	Dictyostelium discoideum actin	364	G	Gelsolin	123
1H31	1.5	A	Diheme cytochrome C	260	B	Cytochrome C	138
1EM8	2.1	A	DNA polymerase III CH1 subunit	147	B	DNA polymerase III PSI subunit	110
1JQL	2.5	A	DNA polymerase III, beta chain	366	B	DNA polymerase III delta subunit	140
1EAI	2.4	A	Elastase	240	C	Chymotrypsin isoinhibitor 1	61
1EFV	2.1	A	Electron transfer flavoprotein alpha chain	312	B	Electron transfer flavoprotein beta chain	252

Chapter 3

1F60	1.7	A	Elongation factor EEF1A	440	B	Elongation factor EEF1BA	90
1TA3	1.7	B	Endo-1,4-beta-xylanase	301	A	Xylanase inhibitor protein I	274
1TE1	2.5	B	Endo-1,4-xylanase	190	A	Xylanase inhibitor protein I	274
3FAP	1.9	A	FK506-binding protein	107	B	FKBP12-rapamycin associated protein	94
1FCD	2.5	A	Flavocytochrome C sulfide dehydrogenase	401	C	Flavocytochrome c sulfide dehydrogenase	174
1NF3	2.1	A	G25k GTP-binding protein	194	C	PAR-6B	123
1NQL	2	B	Galactosyltransferase	272	A	Alpha lactalbumin	123
1WQ1	2.5	G	Gapette	320	R	Harvey-RAS	166
1OR0	2.0	B	Glutaryl acylase beat subunit	510	A	Glutaryl acylase alpha subunit	152
1AXI	2.1	B	Growth hormone receptor	191	A	Growth hormone	175
2NGR	1.9	B	Gtpase activating protein	196	A	GTP binding protein	191
1TX4	1.7	A	Gtpase-activating protein rhogap	196	B	Transforming protein RHOA	174
1AY7	1.7	A	Guanyl-specific ribonuclease SA	96	B	Barstar	89
1HX1	1.9	A	Heat shock cognate 71 KDA	377	B	Bag-family molecular chaperone regulator-1	112
1USU	2.2	A	Heat shock protein HSP82	246	B	AHA1	132
2HBE	2.0	B	Hemoglobin	146	A	Hemoglobin	141
1GPW	2.4	A	Hisf protein	253	B	Amidotransferase HISF	200
1CXZ	2.2	A	His-tagged transforming protein RHOA	182	B	PKN	86
1US7	2.3	B	HSP90 chaperone protein kinase	194	A	Heat shock protein HSP82	207
1KXP	2.1	D	Human vitamin D-binding protein	438	A	Actin, alpha skeletal muscle	349
1H2A	1.8	L	Hydrogenase	534	S	Hydrogenase	267
1KA9	2.3	F	Imidazole glycerol phosphatase synthase	251	H	Imidazole glycerol phosphatase synthase	195
1IBR	2.3	B	Importin beta-1 subunit	458	A	GTP-binding nuclear protein ran	169
1PVH	2.5	A	Interleukin 6 signal transducer	201	B	Leukemia inhibitory factor	169
1IAR	2.3	B	Interleukin-4 receptor alpha chain	188	A	Interleukin	129
1I1R	2.4	A	Interleukin-6 receptor beta chain	301	B	Viral IL-6	167
1O6S	1.8	A	Internalin A	461	B	E-cadherin	105
1KI1	2.3	B	Intersectin long form	342	A	G25k GTP-binding protein	178
2KIN	1.9	A	Kinesin	238	B	Kinesin	100
1PPF	1.8	E	Leukocyte elastase	218	I	Ovomucoid inhibitor	56
1OP9	1.9	B	Lysozyme C	130	A	Hl6 camel VHH fragment	121
1UUZ	1.8	D	Lysozyme C	129	A	Inhibitor of vertebrate lysozyme	130
1OOO	1.9	A	Mago nashi protein	144	B	Drosophila Y14	92
1SVX	2.2	B	Maltose-binding periplasmic protein	369	A	Ankyrin repeat protein OFF7	157
1PQZ	2.1	A	MCMV M144	238	B	Beta-2-microglobulin	99
1MEE	2.0	A	Mesentericopeptidase	275	I	Eglin-C	64
1JW9	1.7	B	Molybdopterin biosynthesis moeb protein	240	D	Molybdopterin converting factor	81
1Q40	2.0	B	Mrna export factor MEX67	180	A	Mrna transport regulator MIR2	163
1SHW	2.2	B	Neural kinase	181	A	Ephrin-A5	138
1QAV	1.9	B	Neuronal nitric oxide synthase	115	A	Alpha-1 syntrophin	90
1E96	2.4	B	Neutrophil cytosol factor 2	185	A	Ras-related C3 botulinum toxin substrate 1	178
1NPE	2.3	A	Nidogen	263	B	Laminin gamma-1 chain	164
1GL4	2.0	A	Nidogen-1	273	B	Proteoglycan core protein	89
1M4U	2.4	A	Noggin	199	L	Osteogenic protein 1	112
1FYH	2.0	A	Nterferon-gamma	242	B	Interferon-gamma receptor alpha chain	201
1STF	2.4	E	Papain	212	I	Stefin B	98
1F34	2.5	A	Pepsin A	325	B	Major pepsin inhibitor PI-3	138
1UBK	1.2	L	Periplasmic hydrogenase large subunit	534	S	Periplasmic hydrogenase small subunit	267
1JLT	1.4	B	Phospholipase A2	122	A	Phospholipase A2 inhibitor	122
1L4Z	2.3	A	Plasminogen	248	B	Streptokinase	125
1DHK	1.9	A	Porcine pancreatic alpha-amylase	495	B	Bean lectin-like inhibitor	195
3YGS	2.5	P	Procaspase 9	97	C	Apoptotic protease activating factor 1	95
1FT1	2.3	B	Protein farnesyltransferase	416	A	Protein farnesyltransferase	315
1G4U	2.3	S	Protein tyrosine phosphatase SPTP	360	R	Ras-related C3 botulinum toxin substrate 1	180
1CT4	1.6	E	Proteinase	185	I	Ovomucoid inhibitor	51
1VG0	2.2	A	Rab escort protein 1	481	B	Ras-related protein rab-7	182
1F2T	1.6	A	Rad50 abc-atpase N-terminal domain	145	B	Rad50 abc-atpase C-terminal domain	143

Chapter 3

1GUA	2.0	A	Rap1A	167	B	C-raf1	76
1HE1	2.0	C	Ras-related C3 botulinum toxin substrate 1	176	A	Exoenzyme S	135
1DS6	2.4	A	Ras-related C3 botulinum toxin substrate 2	181	B	RHO GDP-dissociation inhibitor 2	179
1C1Y	1.9	A	Ras-related protein	167	B	Proto-onkogene serine	77
1DFJ	2.5	E	Ribonuclease A	124	I	Ribonuclease inhibitor	456
1DZB	2.0	A	SCFV fragment 1F9	224	X	Turkey egg-white lysozyme C	129
1H2S	1.9	A	Sensory rhodopsin II	225	B	Sensory rhodopsin II transducer	60
1P57	1.8	B	Serine protease hepsin heavy chain	247	A	Serine protease hepsin light chain	110
4SGB	2.1	E	Serine proteinase B	185	I	Potato inhibitor	51
1SMP	2.3	A	Serratia metallo proteinase	468	I	Erwinia chrysanthemi inhibitor	100
1NRJ	1.7	B	Signal recognition particle receptor	191	A	Docking protein	147
1RJ9	1.9	A	Signal recognition protein	277	B	Signal recognition particle protein	282
1JTP	1.9	A	Single-domain antibody	135	L	Lysozyme C	129
1SGD	1.8	E	Streptogrisin B	185	I	Ovomucoid	51
1LW6	1.5	E	Subtilisin BPN	281	I	Ubtillisin-chymotrypsin inhibitor-2A	63
2SIC	1.8	E	Subtilisin BPN	275	I	Streptomyces subtilisin inhibitor	107
1SPB	2.0	S	Subtilisin BPN	264	P	Subtilisin BPN prosegment	71
1R0R	1.1	E	Subtilisin carlsberg	274	I	Ovomucoid	51
1CSE	1.2	E	Subtilisin carlsberg	274	I	Eglin-C	63
1SCJ	2.0	A	Subtilisin E	275	B	Subtilisin E	71
2SNI	2.1	E	Subtilisin novo	275	I	Chymotrypsin inhibitor 2	64
1EUC	2.1	B	Succinyl-coa synthetase, beta chain	393	A	Succinyl-coa synthetase, alpha chain	306
1ONQ	2.2	A	T-cell surface glycoprotein CD1A	274	B	Beta-2-microglobulin	99
1JTD	2.3	A	Tem-1 beta-lactamase	262	B	Beta-lactamase inhibitor protein II	273
1KTZ	2.2	B	TGF-beta type II receptor	106	A	Transforming growth factor beta 3	82
2TEC	2.0	E	Thermitase	279	I	Eglin-C	63
1JKG	1.9	B	Tip associating protein	180	A	NTF2-related export protein 1	139
1D4V	2.2	B	TNF-related apoptosis inducing ligand	163	A	Death receptor 5	117
1AVW	1.8	A	Trypsin	223	B	Trypsin inhibitor	171
1BRB	2.1	E	Trypsin	223	I	BPTI	51
1F5R	1.7	A	Trypsin II	216	I	Pancreatic trypsin inhibitor	57
1K9O	2.3	E	Trypsin II anionic	223	I	Alaserpin	376
1D6R	2.3	A	Trypsinogen	223	I	Bowman-birk proteinase inhibitor	58
1OPH	2.3	B	Trypsinogen	223	A	Alpha-1 protease inhibitor	375
1P2J	1.4	A	Trypsinogen	220	I	Pancreatic trypsin inhibitor	56
1S1Q	2.0	A	Tumor susceptibility gene 101 protein	137	B	Ubiquitin	71
1ITB	2.5	B	Type 1 interleukin-1 receptor	310	A	Interleukin-1 beta	153
1J7D	1.9	B	Ubiquitin-conjugating enzyme E2-17 KDA	149	A	MMS2	140
1EUV	1.3	A	ULP1 protease	221	B	Ubiquitin-like protein SMT3	79
1UGH	1.9	E	Uracil-dna glycosylase	223	I	Uracil-DNA glycosylase inhibitor	82
1UZX	1.9	A	Vacuolar protein sorting-associated protein	135	B	Ubiquitin	75
1JTT	2.1	A	VH single-domain antibody	133	L	Lysozyme	129
1RKE	2.4	A	Vinculin	262	B	VCL protein	176
1MA9	2.4	A	Vitamin D-binding protein	442	B	Actin, alpha skeletal muscle	356
1YVN	2.1	A	Yeast actin	372	G	Gelsolin	125
1OXB	2.3	A	YDP1P	166	B	Osmolarity two-component system protein	124

Chapter 3

3.2.2 Interface parameters

Each of the 156 heterodimer interfaces was studied using 118 parameters and the corresponding values were determined. The parameter list is given in Table 3.2.

Consequently, a 156 X 118 matrix was generated for the 156 heterodimers.

Table 3.2: List of 118 heterodimer interface parameters

No.	Parameter names
1	Interface area ($\text{\AA}^2$)
2	Interface non-polar area ($\text{\AA}^2$)
3	Interface polar area ($\text{\AA}^2$)
4	RMSD [‡]
5	Interface salt bridge [#]
6	Interface salt bridges per 100 $\text{\AA}^2$ interface area
7	Interface salt bridges per interface residue
8	Gap volume ($\text{\AA}^3$) [‡]
9	Gap index
10	Interface helix residues
11	Interface strand residues
12	Interface loop residues
13	Ratio of interface helix residues to interface residues
14	Ratio of interface strand residues to interface residues
15	Ratio of interface loop residues to interface residues
16	Interface H-bonds
17	Interface H-bonds between main chain – main chain
18	Interface H-bonds between main chain – side chain
19	Interface H-bonds between side chain – main chain
20	Interface H bonds between side chain – side chain
21	Interface H-bonds per 100 $\text{\AA}^2$ interface area
22	Interface H-bonds between main chain - main chain per 100 $\text{\AA}^2$ interface area
23	Interface H-bonds between main chain - side chain per 100 $\text{\AA}^2$ interface area
24	Interface H-bonds between side chain - main chain per 100 $\text{\AA}^2$ interface area
25	Interface H-bonds between side chain - side chain per 100 $\text{\AA}^2$ interface area
26	Interface H-bonds per interface residue
27	Interface H-bonds between main chain – main chain per interface residue
28	Interface H-bonds between main chain – side chain per interface residue
29	Interface H-bonds between side chain – main chain per interface residue
30	Interface H-bonds between side chain – side chain per interface residue
31	Interface hydrophobicity
32	Interface residues
33	Interface hydrophobic residues
34	Interface polar residues

Chapter 3

35	Interface sulphur containing residues
36	Interface charged residues
37	Interface negative charge residues
38	Interface positive charge residues
39	Interface ALA
40	Interface ARG
41	Interface ASN
42	Interface ASP
43	Interface CYS
44	Interface GLN
45	Interface GLU
46	Interface GLY
47	Interface HIS
48	Interface ILE
49	Interface LEU
50	Interface LYS
51	Interface MET
52	Interface PHE
53	Interface PRO
54	Interface SER
55	Interface THR
56	Interface TRP
57	Interface TYR
58	Interface VAL
59	Hydrophobic residues per 100 interface residues
60	Polar residues per 100 interface residues
61	S-containing residues per 100 interface residues
62	Charged residues per 100 interface residues
63	Negative charge residues per 100 interface residues
64	Positive charge residues per 100 interface residues
65	Interface ALA per 100 interface residues
66	Interface ARG per 100 interface residues
67	Interface ASN per 100 interface residues
68	Interface ASP per 100 interface residues
69	Interface CYS per 100 interface residues
70	Interface GLN per 100 interface residues
71	Interface GLU per 100 interface residues
72	Interface GLY per 100 interface residues
73	Interface HIS per 100 interface residues
74	Interface ILE per 100 interface residues
75	Interface LEU per 100 interface residues
76	Interface LYS per 100 interface residues
77	Interface MET per 100 interface residues
78	Interface PHE per 100 interface residues
79	Interface PRO per 100 interface residues

Chapter 3

80	Interface SER per 100 interface residues
81	Interface THR per 100 interface residues
82	Interface TRP per 100 interface residues
83	Interface TYR per 100 interface residues
84	Interface VAL per 100 interface residues
85	Interface residues per 100 complex residues
86	Interface hydrophobic residues per 100 complex residues
87	Interface polar residues per 100 complex residues
88	Interface S-containing residues per 100 complex residues
89	Interface charged residues per 100 complex residues
90	Interface negative charge residues per 100 complex residues
91	Interface positive charge residues per 100 complex residues
92	Interface ALA per 100 complex residues
93	Interface ARG per 100 complex residues
94	Interface ASN per 100 complex residues
95	Interface ASP per 100 enzyme residues
96	Interface CYS per 100 complex residues
97	Interface GLN per 100 complex residues
98	Interface GLU per 100 complex residues
99	Interface GLY per 100 complex residues
100	Interface HIS per 100 complex residues
101	Interface ILE per 100 complex residues
102	Interface LEU per 100 complex residues
103	Interface LYS per 100 complex residues
104	Interface MET per 100 complex residues
105	Interface PHE per 100 complex residues
106	Interface PRO per 100 complex residues
107	Interface SER per 100 complex residues
108	Interface THR per 100 complex residues
109	Interface TRP per 100 complex residues
110	Interface TYR per 100 complex residues
111	Interface VAL per 100 complex residues
112	Interface aromatic residue
113	Interface aliphatic residue
114	Interface aromatic per 100 interface residues
115	Interface aliphatic per 100 interface residues
116	Interface aromatic per 100 complex residues
117	Interface aliphatic per 100 complex residues
118	Residue interface propensities

‡ Gap volume between subunits and RMSD of interface are calculated SURFBET [McDonald & Thornton, 1994].

The number of salt bridges formed between subunits is counted by WHATIF [Laskowski, 1995].

Chapter 3

Interface area

ASA (accessible surface area) was calculated using NACCESS [Hubbard, 1993] with a probe radius of 1.4 Å and interface area is defined by ΔASA (change in ASA upon complexation from monomer to dimer state).

Inter-subunit H-bonds

A hydrogen bond is a polar interaction between two electronegative atoms, where a donor and an acceptor participate. The number of H-bonds formed between subunits was calculated using the program HBPLUS [Rodriguez, et al., 1998].

Hydrophobicity

Interface hydrophobicity was estimated using the equation $\sum_{i=1}^N (HV) / N$ [Jones & Thornton, 1997a], where N is the number of interface residues, and HV is the hydrophobicity scale for each residue [Radzicka & Wolfenden, 1988].

Interface residues propensity

Interface residues show a ΔASA (change in accessibility) of $\geq 5\%$ upon complexation. Interface residue propensities were calculated using the percentage frequencies of 20 residues using the following functions:

$$P_{IS}(i) = f_{interface}(i) / f_{surface}(i) \quad (2-1)$$

$$P_{II}(i) = f_{interface}(i) / f_{interior}(i) \quad (2-2)$$

where $P_{IS}(i)$ is residue interface propensity compared to protein surface, $P_{II}(i)$ is residue interface propensity compared to protein interior, $f_{interface}(i)$ is residue

Chapter 3

frequency at the protein interface, $f_{surface}(i)$ is residue frequency at the protein surface, $f_{interior}(i)$ is residue frequency at the protein interior.

3.2.3 Parameter normalization

Each parameter value was normalized such that standard deviation is equal to one and the average is equal to zero. The standard deviation was calculated using the STDEVP function in Microsoft Excel. The normalization ensures that all parameters are expressed as dimensionless numbers. The normalized parameter value is represented by S using α (parameter index whose range is from 1 to 118), i (heterodimer interface index whose range is from 1 to 156), P (parameter value), n (number of heterodimers i.e. 156 in number), $\overline{P_\alpha}$ (parameter mean) and σ_{P_α} (standard deviation). By definition, S is given as

$$s_\alpha(i) = \frac{(P_\alpha(i) - \overline{P_\alpha})}{\sigma_{P_\alpha}} \quad (2-3)$$

where,

$$\sigma_{P_\alpha} = \sqrt{\frac{n \sum_{i=1}^n P_\alpha(i)^2 - (\sum_{i=1}^n P_\alpha(i))^2}{n^2}} = \sqrt{\frac{156 \sum_{i=1}^{156} P_\alpha(i)^2 - (\sum_{i=1}^{156} P_\alpha(i))^2}{156^2}} \quad (2-4)$$

This procedure generated a 156 x 118 matrix containing normalized parameter values.

Chapter 3

3.2.4 Representation of parameters

There are 156 heterodimer structures used in this analysis and each dimer interface (i) is represented as a vector $\overrightarrow{S(i)}$ in 118 dimensional ‘continuous space’, where the components $S_\alpha(i)$ are the normalized parameter values. The scalar product Q_{ij} between two vectors $\overrightarrow{S(i)}$ and $\overrightarrow{S(j)}$, where j is another index for a heterodimer interface, is given by

$$Q_{ij} = \overrightarrow{S(i)} \cdot \overrightarrow{S(j)} = \sum_{\alpha=1}^{118} s_\alpha(i) \cdot s_\alpha(j) \quad (2-5)$$

The 156 x 156 matrix Q is positive symmetric and consists of the scalar products of the parameter vectors $\overrightarrow{S(i)}$ and $\overrightarrow{S(j)}$, where $i = 1$ to 156 and $j = 1$ to 156.

3.2.5 Calculation of eigenvectors and eigenvalues

The symbolic eigenvectors (E) and eigenvalues (λ) of a square matrix Q are computed using MATLAB. The eigenvalues of Q are the zeros of the characteristic polynomial of Q . As Q is of order 156, we will have 156 eigenvectors and eigenvalues and the smallest eigenvalue, λ_{156} , is near zero due to normalization of the parameters. The eigenvalues and their corresponding eigenvectors are indexed in decreasing order of eigenvalues.

3.2.6 Selection of interface descriptors

The distribution of the eigenvalues of the Q matrix (Figure 3.1), containing the scalar products between all pairs of the 118 dimensional heterodimer vectors, rapidly decreases from the largest value λ_1 to λ_{156} . The rapid decrease of the eigenvalues derived from the 118 physical – chemical parameters shows a large anisotropy of the

Chapter 3

distribution of the parameter values. This anisotropy is a consequence of the large redundancy in the sets of parameter values. This suggests that the number of parameters can be reduced while retaining approximately the same distribution of heterodimers in the property space. We found that the eigenvalues rapidly decrease within the first six largest eigenvalues.

Figure 3.1: Distribution of eigenvalues

3.2.7 Calculation of distances in the parameter space

If μ represents the index of eigenvalue and eigenvector, each heterodimer can be represented as a vector in a six-dimensional Euclidean space with each dimension perpendicular to each other. The co-ordinates of the i^{th} heterodimer can be written as:

$$\sqrt{\lambda_{\mu}} E_i^{\mu} \quad (2-6)$$

where μ varies from 1 to 6.

The distance between the i^{th} and j^{th} heterodimer interface is given by

Chapter 3

$$d_{ij} = \sqrt{\sum_{\mu=1}^n (\sqrt{\lambda_{\mu}} E_i^{\mu} - \sqrt{\lambda_{\mu}} E_j^{\mu})^2} \quad (2-7)$$

where n is 6.

Distances computed between heterodimers in the six dimensional Eigen sub-space constitute the parameter distance matrix (PDM). Small distances values between two heterodimers indicate that they are similar in all of the 118 physical and chemical parameters.

3.2.8 Calculation of correlation coefficient

Pearson's correlation coefficient between pairs of parameter values (x_i, y_i) is calculated using the correlation function (CORRCOEF) in MATLAB.

3.3 Results

3.3.1 Quantitative descriptors for heterodimer interface parameters in six dimensions

We used 156 high resolution heterodimer structures (Table 3.1) to derive a comprehensive list of 118 physical/chemical parameters for heterodimer interfaces (Table 3.2). Each heterodimer was represented as a vector in the 118-dimensional space of normalized parameters with mean value of zero and standard deviation 1. Our multi-dimensional scaling approach reveals the high redundancy of the parameter values. The computational approach and justification for reduction to a lower dimensional space follows closely the practice of embedding in distance geometry and it is easy to eliminate redundant variables when describing complex phenomenon

Chapter 3

in molecular recognition. The distribution of eigenvalues decreases rapidly (Figure 3.1). This is due to large redundancy in the parameter set. This suggests that the number of parameters can be reduced while retaining approximately the same distribution of heterodimers in the parameter space. The eigenvalues rapidly decrease within the first six largest eigenvalues. We compared distances in the original parameter space with those regenerated from a subset of n eigenvectors, varying n systematically from 2 to 156 (Figure 3.2). The correlation coefficient between the original and regenerated distances was more than 95% for $n = 6$, and approaches 1 very rapidly. We therefore chose the first six eigenvalues and eigenvectors to calculate the six dimensional descriptors of the heterodimer interfaces. The individual distances in the original parameter space and in the sub-space using the first six eigenvectors were highly correlated (Figure 3.3). The correlation coefficient between the distances was 0.95.

Figure 3.2: Correlation coefficients for parameter distances

Chapter 3

Figure 3.3: The distance between six selected descriptors and 118 parameter space.
The linear correlation coefficient is 0.95.

3.3.2 Selection of critical interface parameters from highly correlated descriptors

We used the first six highly correlated descriptors (dimension 156 X 6) and normalized parameter values (dimension 156 X 118) to calculate the correlation coefficients between the selected descriptors (E_1 to E_6) and the original normalized parameter values. This operation generated a matrix (dimension 6 X 118) containing correlations between the six highly correlated descriptors and the normalized parameter values. We then used this matrix to select the most significant interface parameters using the calculated correlation coefficients (Table 3.3). This further enabled us to select the most significant parameters that best describe a heterodimer protein interface (Table 3.3).

Chapter 3

Table 3.3: Critical parameters for heterodimer interfaces

Top six eigenvectors	Parameter name	Correlation coefficients between the top descriptors (E1 to E6) and the original normalized parameters	Correlation coefficients between the distances of selected parameters and the original normalized parameters
E ₁	Interface residues	-0.94	0.70
E ₂	Interface hydrophobicity	-0.66	0.28
E ₃	Gap volume	-0.52	0.51
E ₄	Interface aliphatic per 100 interface residues	-0.52	0.28
E ₅	Interface aromatic per 100 interface residues	-0.35	0.15
E ₆	Interface H-bonds per 100 Å ² interface area	-0.84	0.13

Figure 3.4: The distance between six selected parameters and 118 parameter space.
The linear correlation coefficient is 0.72.

We then used the parameter values for these six parameters to calculate its distances from the rest of 118 original parameter values. The distances were then used to calculate the correlation coefficients. These values suggest that these six parameters have different weights in heterodimer subunit interactions. Data shows that the H-

Chapter 3

bonds have the highest weight among the six parameters listed in Table 3.3. We also calculated the individual distances between the original parameter values and the six selected parameter values (Figure 3.4). The correlation coefficient between these distances was found to be 0.72.

3.4 Discussion

Heterodimer protein interaction is a common phenomenon in cellular regulation and signaling. This occurs by a huge combination of physical-chemical parameters that characterize their interacting surfaces. The multi dimensional scaling method applied in this study helps to reduce a large pool of interface parameters to a small set of six quantitative descriptors of heterodimer interfaces. Here, we show that the six parameters (Figure 3.3) were sufficient to reproduce the distances in the complete parameter space (Figure 3.4). The most significant parameters that are found to reproduce the original parameter set are given in Table 3.3. They are (1) interface residues, (2) interface hydrophobicity, (3) interface H-bonds, (4) interface aromatic residues, (5) interface aliphatic residues and (6) gap volume. It should be noted that several linear combinations of parameter values represent a descriptors and it is often difficult to further refine or simplify such non-linearity. The goal here is to identify the most critical parameters that represent heterodimer interfaces. In general it is difficult to decide *a priori* which of the many parameters dominate at the interface. Our quantitative descriptors represent a precise spatial relation of all heterodimers with respect to the 118 physical-chemical parameters. This enabled us to identify the most critical parameters and these parameters are further discussed below.

Chapter 3

3.4.1 Interface H-bonds

Intermolecular hydrogen bonds between subunits are important in the association and stability of heterodimers [Lo Conte, et al., 1999]. This analysis suggests that interface H-bonds have a good correlation co-efficient of $r = -0.84$ with the distances of other interface parameters.

3.4.2 Interface aromatic and aliphatic residues

Aromatic and aliphatic residues have greater propensity at the protein-protein interfaces [Bahadur, et al., 2003, 2004]. As shown in Table 3.3, the correlation coefficients of interface aromatic and aliphatic residues with the numerical descriptors (E_2 and E_6) are -0.52 and -0.35. Therefore, aromatic and aliphatic residues were chosen as prominent parameters because of their high correlation coefficients compared to other members of aliphatic or aromatic residue groups.

3.4.3 Interface residues and interface hydrophobicity

Hydrophobicity plays an important role in protein association [Larsen, et al., 1998; Archakov, et al., 2003]. Thus, interface hydrophobicity was among the prominent parameters for heterodimer interaction. The number of interface residues relates to interface area. Stronger protein subunit associations were generally associated with larger interface areas [Jones & Thornton, 1995]. These parameters were already shown to be used in the prediction of heterodimer interaction sites by surface patch analysis [Jones & Thornton, 1997b]. The method detects the most possible interaction sites by the incorporation of this parameter.

Chapter 3

3.4.4 Gap volume

Gap volume, to some extent, can reflect the average strength of the interactions between subunits. Heterodimer interfaces are average less complimentary homodimer interfaces. This can be supported with respect to their binding constants and evolutionary history. For example, the antibody–antigen recognition is a ‘happenstance’ not subject to evolutionary optimization over more than a few days so interfaces between them must be less complimentary [Jones & Thornton, 1995; Nooren & Thornton, 2003a]. So gap volume is more critical in heterodimer structures than in homodimer structures.

3.5 Conclusion

A large number of structurally important physical - chemical parameters characterize heterodimer interfaces and each of these parameters contributes differently to the stability of a heterodimer interface. A weighted value was assigned to each parameter to indicate the differential contribution. Here, we apply a mathematical procedure to determine the most critical parameters that describe a heterodimer interface. The six critical interface parameters discussed here are based on the selected 156 heterodimer structures. The multi dimensional scaling procedure suggests that the six critical parameters effectively replace the original 118 parameter set. This finding is of critical importance in the understanding and development of prediction models for heterodimer interfaces.

Chapter 4

Protein subunit interfaces: heterodimers vs homodimers

4.1 Introduction

Protein subunit interaction (either homodimer or heterodimer) is an important phenomenon in cell regulation and catalysis. Thousands of such interactions are theoretically possible in a combinatorial manner between proteins in cells. The task of documenting each of these interactions is laborious. Therefore, prediction of subunit interaction sites either from folded structures or from primary sequences is required. However, this objective is currently ambitious due to the limited knowledge on the principles of protein subunit interactions using structural data. Therefore, it is our interest to study the principles of subunit interactions. Jones & Thornton (used 59 protein complexes) [1996], Xu et al. (used 319 protein-protein interfaces) [1997b], Tsai et al. (used 362 protein-protein interfaces) [1997a], Lo Conte et al. (used 75 hetero-complexes) [1999], Chakrabart & Janin (used 70 hetero-complexes) [2002], Brinda et al. (used 20 homodimers) [2002], Bahadur et al. (used 122 homodimers)[2003], Nooren & Thornton (used 39 protein dimers) [2003b], Caffrey et al. (used 64 protein-protein interfaces) [2004], Zhanhua et al. (used 65 heterodimers) [2005] utilized a dataset of protein complexes determined by X-ray crystallography to examine the properties of the subunit interactions. Protein subunit interfaces in these studies have been characterized using geometrical properties (interface size, planarity, sphericity and complementarity) and chemical properties (the types of amino acid chemical groups, hydrophobicity, electrostatic interactions and H-bonds). These studies are influenced by dataset size and their characteristics.

Chapter 4

The analyses were based on a limited dataset consisting of heterogeneous (uneven mixture of homodimers and heterodimers) data.

The analyses report on the role of inter-subunit H-bonds in protein subunit association. The number of H-bonds varied in different studies [Jones & Thornton, 1995; Xu, et al., 1997b; Lo Conte, et al., 1999; Bahadur, et al., 2003; Nooren & Thornton, 2003b]. On average, Bahadur & colleagues showed 9.0 H-bonds per homodimer interface with a r value of 0.75 between H-bonds and interface area [Bahadur, et al., 2003]. Jones & Thornton (used 32 homodimers) showed 0.88 H-bonds per 100 Å² interface area with a r value of 0.77 between H-bonds and interface area [Jones & Thornton, 1995]. Lo Conte et al. showed an average of 10.1 H-bonds with one H-bond per 170 Å² interface area and a r value of 0.84 between H-bonds and interface area [Lo Conte, et al., 1999]. Xu & colleagues also showed 11 H-bonds per subunit with a r value of 0.89 between H-bonds and interface area [Xu, et al., 1997b]. The r value between H-bonds and interface area in these studies vary from 0.75 to 0.89. This variation is influenced by dataset size and nature.

Previous studies also show that hydrophobic effect plays an important role in protein association [Lijnzaad & Argos, 1997; Tsai, et al., 1997a; Bahadur, et al., 2003], yet not as much as in protein folding [Tsai, et al., 1997a]. These studies showed that protein interfaces are more hydrophobic than surfaces, but less than interior. Hydrophobic effect was measured by the buried non-polar surface area (or percent burial) of residue types [Tsai, et al., 1997a]. The study showed that the ratio between buried hydrophobic and buried hydrophilic residues is approximately 1.5 [Tsai, et al.,

Chapter 4

1997a]. Hydrophobic residues (except ALA) and the charged residue ARG are predominantly present at protein-protein interfaces with TYR and TRP having highest propensity [Lijnzaad & Argos, 1997; Dasgupta, et al., 1997; Lo Conte, et al., 1999; Brinda, et al., 2002; Bahadur, et al., 2003].

Interface size is yet another important property widely used to describe protein-protein interfaces and it is usually characterized by interface area. The number of interface residues is linearly correlated to interface area ($r \geq 0.96$) in several studies [Chakrabarti & Janin, 2002; Bahadur, et al., 2003]. However, the mean number of interface residues varies between these studies. It is shown that the mean value is 52 [Bahadur, et al., 2003], 57 [Chakrabarti & Janin, 2002], 53.7 [Valdar & Thornton, 2001b], 44.4 (for homodimers) and 42.2 (for heterodimers) [Caffrey, et al., 2004]. Thus, the number of interface residues vary in a narrow range of 42 and 57 in these studies.

Here, we created an extended dataset of exclusively homodimers and heterodimers. We believe that these exclusive datasets can reduce data bias to differentiate heterodimer and homodimer interfaces. The analysis of interface H-bonds, interface size and hydrophobicity are interesting and provide insight in the understanding of protein subunit interactions. These findings suggest that the extracted interface properties are influenced by dataset size and their characteristics. It should also be noted that the variations among these studies are mainly due to their features.

Chapter 4

4.2 Material and Methods

4.2.1 Creation of heterodimer and homodimer structural dataset

The non-redundant dataset of 156 heterodimers is listed in Table 3.1. In order to analyze the differences between heterodimer and homodimer, we then created a non-redundant dataset of 170 homodimers (Table 4.1) such that they satisfy the following conditions. These rules include: (1) each chain ≥ 50 residues; (2) structures determined by x-ray crystallography; (3) resolution ≤ 2.5 Å; (4) the structure with the highest resolution was selected where more than one structure was available; (5) redundant entries were removed at a sequence similarity cut-off of $\geq 30\%$ [Li, et al., 2005].

Table 4.1: Homodimers used in this study

PDB	Resolution (Å)	Name	Source	Chain one	Length	Chain two	Length
1CNZ	1.8	3-isopropylmalate dehydrogenase	Salmonella typhimurium	A	363	B	363
1AFW	1.8	3-ketoacetyl-coa thiolase	Saccharomyces cerevisiae	A	390	B	393
1M4I	1.5	Acetyltransferase	Escherichia coli	A	181	B	176
1LQ9	1.3	Actva-orf6 monooxygenase	Streptomyces coelicolor	A	112	B	112
1ADE	2	Adenylosuccinate synthetase	Escherichia coli	A	431	B	431
1M7H	2	Adenylylsulfate kinase	Penicillium chrysogenum	A	203	B	200
1NA8	2.3	ADP-ribosylation binding protein	Homo sapiens	A	151	B	145
1OR4	2.2	Aerotactic transducer hemat	Bacillus subtilis	A	169	B	158
1BD0	1.6	Alanine racemase	Bacillus stearothermophilus	A	381	B	380
1A4U	1.9	Alcohol dehydrogenase	Drosophila lebanonensis	A	254	B	254
1ALK	2	Alkaline phosphatase	Escherichia coli	A	449	B	449
1LK9	1.5	Alliin lyase	Allium sativum	A	425	B	427
1HSS	2.1	Alpha-amylase inhibitor	Triticum aestivum	A	111	B	111
1S2Q	2.1	Amine oxidase B	Homo sapiens	A	499	B	494
1EKP	2.5	Amino acid aminotransferase	Homo sapiens	A	365	B	365
2GSA	2.4	Aminotransferase	Synechococcus SP	A	427	B	427
1DQT	2	Antigen	Mus musculus	A	117	B	117
1BJW	1.8	Aspartate aminotransferase	Thermus thermophilus	A	381	B	381
1JFL	1.9	Aspartate racemase	Escherichia coli	A	228	B	228
1MJH	1.7	Atp-binding protein	Methanococcus jannaschii	A	143	B	144
1IRI	2.4	Autocrine motility factor	Homo sapiens	A	557	B	557
1LR5	1.9	Auxin binding protein	Zea mays	A	160	B	160
1N80	2.5	Baseplate structural protein	Bacteriophage T4	A	328	B	328
1EWZ	2.4	Beta lactamase oxa-10	Pseudomonas aeruginosa	A	243	C	243
1EBL	1.8	Beta-ketoacyl- <i>acp</i> Synthase III	Escherichia coli	A	309	B	309
1N1B	2	Bornyl diphosphate synthase	Salvia officinalis	A	516	B	519
1KSO	1.7	Calcium-binding protein A3	Homo sapiens	A	93	B	93

Chapter 4

1JD0	1.5	Carbonic anhydrase	Homo sapiens	A	260	B	259
1AUO	1.8	Carboxylesterase	Pseudomonas fluorescens	A	218	B	218
1CDC	2	CD2	Rattus norvegicus	A	96	B	96
1F13	2.1	Cellular coagulation factor	Homo sapiens	A	722	B	719
1NW1	2	Choline kinase	Caenorhabditis elegans	A	365	B	357
1R5P	2.2	Circadian oscillation regulator	Anabaena SP	A	90	B	93
1G64	2.1	Cob(I) alamin adenosyltransferase	Salmonella typhimurium	A	169	B	190
1OTV	2.1	Coenzyme pqq synthesis protein C	Klebsiella pneumoniae	A	254	B	254
1I0R	1.5	Conserved hypothetical protein	Archaeoglobus fulgidus	A	161	B	168
1OAC	2	Copper amine oxidase	Escherichia coli	A	719	B	722
1EAJ	1.4	Coxsackie virus	Homo sapiens	A	124	B	120
1CHM	1.9	Creatinase	Pseudomonas putida	A	401	B	401
1S44	1.6	Crustacyanin A1 subunit	Homarus gammarus	A	180	B	180
1GD7	2	CSAA protein	Thermus thermophilus	A	109	B	109
1L5B	2	Cyanovirin-N	Nostoc ellipsosporum	A	101	B	101
1SO2	2.4	Cyclic Phosphodiesterase B	Homo sapiens	A	363	B	363
1P3W	2.1	Cysteine desulfurase	Escherichia coli	A	385	B	385
1COZ	2	Cytidyltransferase	Bacillus subtilis	A	126	B	126
1P6O	1.1	Cytosine deaminase	Saccharomyces cerevisiae	A	156	B	161
2DAB	2	D-amino acid aminotransferase	Thermophilic bacillus	A	280	B	282
1F17	2.3	Dehydrogenase	Homo sapiens	A	293	B	291
2NAC	1.8	Dehydrogenase	Methylophilic bacterium pseudomonas	A	374	B	374
1NFZ	2	Delta-isomerase	Escherichia coli	A	176	B	180
1D1G	2.1	Dihydrofolate reductase	Thermotoga maritima	A	164	B	164
1DOR	2	Dihydroorotate dehydrogenase A	Lactococcus lactis	A	311	B	311
1AD1	2.2	Dihydropteroate synthetase	Staphylococcus aureus	A	264	B	251
1NU6	2.1	Dipeptidyl peptidase	Homo sapiens	A	728	B	728
1PE0	1.7	DJ-1	Homo sapiens	A	187	B	187
1G1A	2.5	DTDP-D-glucose 4,6-Dehydratase	Salmonella enterica	A	352	B	352
1BBH	1.8	Electron transport	Chromatium vinosum	A	131	B	131
1Q8R	1.9	Endodeoxyribonuclease rusa	Escherichia coli	A	118	B	109
1RVE	2.5	Endonuclease	Escherichia coli	A	244	B	244
1M9K	2	Endothelial nitric-oxide synthase	Homo sapiens	A	400	B	401
1P43	1.8	Enolase 1	Saccharomyces cerevisiae	A	436	B	436
1JR8	1.5	Erv2 protein mitochondrial	Saccharomyces cerevisiae	A	105	B	105
1V26	2.5	Fatty-acid-coa synthetase	Thermus thermophilus	A	489	B	510
1LBQ	2.4	Ferrochelataase	Saccharomyces cerevisiae	A	356	B	354
1RYA	1.3	Gdp-mannose mannosyl hydrolase	Escherichia coli	A	160	B	160
1QFH	2.2	Gelation factor	Dictyostelium discoideum	A	212	B	212
1JV3	2.2	Glcnaclp uridyltransferase	Homo sapiens	A	490	B	484
1DPG	2	Glucose 6-phosphate dehydrogenase	Leuconostoc mesenteroides	A	485	B	485
1QXR	1.7	Glucose-6-phosphate isomerase	Pyrococcus furiosus	A	187	B	187
1EOG	2.1	Glutathione S-transferase	Escherichia coli	A	208	B	208
1N2A	1.9	Glutathione S-transferase	Escherichia coli	A	201	B	187
1M0W	1.8	Glutathione synthetase	Saccharomyces cerevisiae	A	481	B	479
1R9C	1.8	Glutathione transferase	Mesorhizobium loti	A	125	B	118
1F4Q	1.9	Grancalcin	Homo sapiens	A	161	B	165
1DQP	1.8	Guanine phosphoribosyltransferase	Giardia lamblia	A	230	B	230
3SDH	1.4	Hemoglobin	Scapharca inaequivalvis	A	145	B	145
1IPI	2.2	Holliday junction resolvase	Pyrococcus furiosus	A	114	B	114
1FWL	2.3	Homoserine kinase	Methanococcus jannaschii	A	296	B	296
2HHM	2.1	Hydrolase	Homo sapiens	A	272	B	272
1PP2	2.5	Hydrolase	Crotalus atrox	R	122	L	122
1FJH	1.7	Hydroxysteroid dehydrogenase	Comamonas testosteroni	A	236	B	236
1G0S	1.9	Hypothetical Protein	Escherichia coli	A	201	B	202
1JOG	2.4	Hypothetical protein	Haemophilus influenzae	A	129	B	129
1PT5	2	Hypothetical protein	Escherichia coli	A	415	B	415

Chapter 4

1QYA	2	Hypothetical Protein	Escherichia coli	A	293	B	307
1FUX	1.8	Hypothetical protein	Escherichia coli	A	164	B	163
1J30	1.7	Hypothetical rubrerythrin	Sulfolobus tokodaii	A	141	B	137
1LHZ	2.3	Immunoglobulin lambda	Homo sapiens	A	213	B	213
1AA7	2.1	Influenza virus matrix mrotein	Influenza virus	A	158	B	157
8PRK	1.9	Inorganic pyrophosphatase	Saccharomyces cerevisiae	A	282	B	282
1R8J	2	Kaia	Synechococcus elongatus	A	272	B	264
1CQS	1.9	Ketosteroid isomerase	Pseudomonas putida	A	124	B	124
1AQ6	2	L-2-haloacid dehalogenase	Xanthobacter autotrophicus	A	245	B	245
1I2W	1.7	Lactamase	Bacillus licheniformis	A	255	B	256
1BH5	2.2	Lactoylglutathione lyase	Homo sapiens	A	177	B	182
1QMJ	2.2	Lectin	Gallus gallus	A	132	B	132
1K75	1.8	L-histidinol dehydrogenase	Escherichia coli	A	425	B	425
1EHI	2.4	Ligase	Leuconostoc mesenteroides	A	360	B	347
1NWW	1.2	Limonene-1,2-epoxide hydrolase	Rhodococcus erythropolis	A	145	B	146
1UC8	2	Lysine biosynthesis enzyme	Thermus thermophilus	A	240	B	239
1EN5	2.3	Manganese superoxide dismutase	Escherichia coli	A	205	B	205
1A4I	1.5	Methylenetetrahydrofolate	Homo sapiens	A	285	B	295
1FC5	2.2	Molybdopterin biosynthesis	Escherichia coli	A	397	B	396
1JYS	1.9	Mta/sah nucleosidase	Escherichia coli	A	226	B	226
1LNW	2.1	Multidrug resistance operon repressor	Pseudomonas aeruginosa	A	137	B	135
1FP3	2	N-acyl-d-glucosamine	Sus scrofa	A	402	B	402
1FYD	2.3	NAD(+) Synthetase	Bacillus subtilis	A	271	B	246
1HJ3	1.6	Nitrite reductase	Paracoccus pantotrophus	A	544	B	542
1G1M	2.3	Nitrogenase iron protein	Azotobacter vinelandii	A	287	B	289
1G8T	1.1	Nuclease SM2 isoform	Serratia marcescens	A	241	B	241
1EYV	1.6	N-utilizing substance protein	Mycobacterium tuberculosis	A	131	B	133
1M98	2.1	Orange carotenoid protein	Arthrospira maxima	A	316	B	314
1ORO	2.4	Orotate phosphoribosyltransferase	Escherichia coli	A	213	B	206
1DVJ	1.5	Orotidine 5'-phosphate decarboxylase	Methanobacterium thermoautotrophicum	A	239	B	211
1GGQ	2.5	Outer surface protein C	Borrelia burgdorferi	A	162	B	162
1AOR	2.3	Oxidoreductase	Pyrococcus furiosus	A	605	B	605
1BMD	1.9	Oxidoreductase	Thermus flavus	A	327	B	327
1HDY	2.5	Oxidoreductase	Homo sapiens	A	374	B	374
1N2O	2.1	Pantothenate synthetase	Mycobacterium tuberculosis	A	279	B	279
1RN5	2.2	Peptide deformylase	Leptospira interrogans	A	177	B	177
1PN2	2	Peroxisomal hydratase	Candida tropicalis	A	269	B	267
1PN0	1.7	Phenol 2-monooxygenase	Trichosporon cutaneum	A	652	C	656
1BXG	2.3	Phenylalanine dehydrogenase	Rhodococcus SP	A	349	B	347
1M6P	1.8	Phosphate receptor	Bos Taurus	A	146	B	146
1RQL	2.4	Phosphonoacetaldehyde hydrolase	Bacillus cereus	A	257	B	257
1O4U	2.5	Phosphoribosyltransferase	Thermotoga maritima	A	265	B	266
1EZ2	1.9	Phosphotriesterase	Pseudomonas diminuta	A	328	B	328
1EXQ	1.6	Pol polyprotein	Escherichia coli	A	147	B	145
1MNA	1.8	Polyketide synthase	Streptomyces venezuelae	A	276	B	278
1C6X	2.5	Protease	Escherichia coli	A	99	B	99
1FL1	2.2	Protease	Escherichia coli	A	192	B	207
1F89	2.4	Protein YLC351C	Saccharomyces cerevisiae	A	271	B	271
1LHP	2.1	Pyridoxal kinase	Ovis aries	A	306	B	309
1CBK	2	Pyrophosphokinase	Haemophilus influenzae	A	160	B	160
1QR2	2.1	Quinone reductase type 2	Homo sapiens	A	230	B	230
1EN7	2.4	Recombination endonuclease	Bacteriophage T4	A	157	B	157
1EV7	2.4	Restriction endonuclease naei	Nocardia aerocolonigenes	A	295	B	293
1H8X	2	Ribonuclease	Homo sapiens	A	125	B	125
1I4S	2.2	Ribonuclease III	Aquifex aeolicus	A	147	B	147
1KGN	1.9	Ribonucleotide reductase protein	Corynebacterium ammoniagenes	A	296	B	296
1TLU	1.6	S-adenosylmethionine decarboxylase	Thermotoga maritima	A	117	B	117

Chapter 4

1K6Z	2	Secretion chaperone syce	Yersinia pestis	A	120	B	119
1K3S	1.9	Sige	Salmonella enterica	A	106	B	104
1PJQ	2.2	Siroheme synthase	Salmonella typhimurium	A	447	B	454
1HJR	2.5	Site-specific recombinase	Escherichia coli	A	158	C	158
3LYN	1.7	Sperm lysine	Haliotis fulgens	A	122	B	124
2SQC	2	Squalene-hopene Cyclase	Alicyclobacillus acidocaldarius	A	623	B	623
1SCF	2.2	Stem cell factor	Homo sapiens	A	116	B	118
1OX8	2.2	Stringent starvation protein B	Escherichia coli	A	105	B	105
1M3E	2.5	Succinyl-coa	Sus scrofa	A	459	B	460
1R7A	1.8	Sucrose phosphorylase	Bifidobacterium adolescentis	A	503	B	503
1SOX	1.9	Sulfite oxidase	Gallus gallus	A	463	B	458
1L5X	2	Survival protein E	Pyrobaculum aerophilum	A	270	B	272
1REG	1.9	T4 rega	Bacteriophage T4	X	122	Y	120
1MKB	2	Thiol ester dehydrase	Escherichia coli	A	171	B	171
1QHI	1.9	Thymidine kinase	Herpes simplex virus	A	304	B	308
1HSJ	2.3	Transcription/sugar binding protein	Escherichia coli	A	487	B	487
1NY5	2.4	Transcriptional regulator	Aquifex aeolicus	A	384	B	385
1ON2	1.6	Transcriptional regulator	Bacillus subtilis	A	135	B	135
1SMT	2.2	Transcriptional repressor	Synechococcus	A	98	B	101
1TRK	2	Transferase	Saccharomyces cerevisiae	A	678	B	678
7AAT	1.9	Transferase	Gallus gallus	A	401	B	401
1KIY	2.4	Trichodiene synthase	Fusarium sporotrichioides	A	354	B	354
118T	2.4	Udp-galactopyranose mutase	Escherichia coli	A	367	B	367
1F6D	2.5	Udp-n-acetylglucosamine	Escherichia coli	A	366	B	363
1JP3	1.8	Undecaprenyl pyrophosphate synthase	Escherichia coli	A	210	B	207
1JMV	1.9	Universal stress protein A	Haemophilus influenzae	A	140	B	137
1HQO	2.3	URE2 protein	Saccharomyces cerevisiae	A	221	B	217
9WGA	1.8	Wheat germ agglutinin	Triticum vulgaris	A	170	B	170
1MI3	1.8	Xylose reductase	Candida tenuis	A	319	B	319

4.2.2 Interface parameters

First, we use 118 (Table 3.2) to represent homodimer interfaces. Then, we extract critical interface parameters for these 170 homodimers using methods described in chapter 3 (section 3.2).

4.3 Results

4.3.1 Quantitative descriptors for homodimer interface parameters in six dimensions

We derived a comprehensive list of 118 physical/chemical parameters for each homodimer interface. This parameter list was also used to analyze heterodimers in Chapter 3. Each homodimer was represented as a vector in the 118-dimensional space

Chapter 4

of normalized parameters with mean value of zero and standard deviation 1. Our multi-dimensional scaling approach also reveals the high redundancy of the parameter values for homodimer structures. The computational approach and justification for reduction to a lower dimensional space follows closely the practice of embedding in distance geometry and it is easy to eliminate redundant variables when describing complex phenomenon in molecular recognition. The distribution of eigenvalues decreases rapidly for homodimers (Figure 4.1). This is due to large redundancy in the parameter set. This suggests that the number of parameters can be reduced while retaining approximately the same distribution of homodimers in the parameter space. The eigenvalues rapidly decrease within the first six largest eigenvalues. We compared distances in the original parameter space with those regenerated from a subset of n eigenvectors, varying n systematically from 2 to 170 for homodimers (Figure 4.2). The correlation coefficient between the original and regenerated distances was more than 0.93 for homodimers for $n = 6$, and approach 1 very rapidly. We therefore chose the first six eigenvalues and eigenvectors to calculate the six dimensional descriptors of the homodimer interfaces. The individual distances in the original parameter space and in the sub-space using the first six eigenvectors were highly correlated (Figure 4.3). The correlation coefficient between the distances was 0.95 for homodimers.

Chapter 4

Figure 4.1: Distribution of eigenvalues

Figure 4.2: Correlation coefficients for parameter distances

Chapter 4

Figure 4.3: The distance between six selected descriptors and 118 parameter space is given. The linear correlation coefficient is 0.95.

4.3.2 Selection of critical interface parameters from highly correlated descriptors

We used the first six highly correlated descriptors (dimension $n \times n$) and normalized parameter values (dimension $n \times 118$) to calculate the correlation coefficients between the selected descriptors (E_1 to E_6) and the original normalized parameter values (n is the number of homodimers, i.e. 170). This operation generated a matrix (dimension 6×118) containing correlations between the six highly correlated descriptors and the normalized parameter values for homodimer interfaces. We then used this matrix to select the most significant interface parameters using the calculated correlation coefficients (Table 4.2). This further enabled us to select the most significant parameters that best describe a homo dimer interface (Table 4.2). We then used the parameter values for these six parameters to calculate its distances from the rest of 118 original parameter values. The distances were then used to calculate

Chapter 4

the correlation coefficients. These values suggest that these six parameters have different weights in homodimer subunit interactions. Data shows that the number of interface residues, which means that interface size has important affect on protein subunit interactions, has the highest weight among the six parameters for homodimers listed in Table 4.2. We also calculated the individual distances between the original parameter values and the six selected parameter values for homodimers (Figure 4.4). The correlation coefficient between these distances was found to be 0.72 for homodimer structures.

Table 4.2: Critical parameters for homodimer interfaces

Top six eigenvectors		Parameter name	Correlation coefficients between the top descriptors (E1 to E6) and the original normalized parameters	Correlation coefficients between the distances of selected parameters and the original normalized parameters
Homodimer	E ₁	Interface residues	-0.93	0.65
	E ₂	Interface aliphatic per 100 interface residues	0.73	0.57
	E ₃	Interface hydrophobicity	0.49	0.13
	E ₄	Hydrophobic residues per 100 interface residues	0.50	0.14
	E ₅	Interface H-bonds per 100 Å ² interface area	0.64	0.15
	E ₆	Interface aromatic per 100 interface residues	0.62	0.18

Figure 4.4: The distance between six selected parameters and 118 parameter space is given. The linear correlation coefficient is 0.72

Chapter 4

4.4 Discussion

Dimer interactions are characterized by a large combination of physical-chemical parameters. Analysis of dimer structures can provide insight into the principles of protein-protein complexation and help develop models to predict interaction sites. The multi dimensional scaling method applied in a recent study reduced a large pool of interface parameters to a small set of six critical properties for heterodimers [Zhanhua, et al., 2005]. We have showed that the six selected parameters were sufficient to describe subunit interfaces instead of the complete parameter space. Here, we discuss the interface differences between 156 heterodimers and 170 homodimers using a selected set of properties. The properties used in this study are (1) interface residues, (2) interface H-bonds, (3) interface hydrophobicity, (4) interface residue composition.

4.4.1 Interface residues

The number of interface residues is proportional to interface area [Chakrabarti & Janin, 2002; Bahadur, et al., 2003]. Stronger protein subunit associations were generally associated with larger interface areas [Jones & Thornton, 1995]. In our study, the range of heterodimer interface residues varies from 18 to 162 with a mean value of 51. While, the range of homodimer interface residues extends from 15 to 308 with a mean value of 81. Like H-bonds, interface residues also varied with different studies and are affected by dataset size and type [Valdar & Thornton, 2001b; Chakrabarti & Janin, 2002; Bahadur, et al., 2003; Caffrey, et al., 2004]. Hence, we created mutually exclusive datasets of homodimers and heterodimers for this analysis to reduce bias due to data type heterogeneity. Thus, we show that the amount of

Chapter 4

interface residues is significantly different for homodimers and heterodimers. The results also suggest that the previous studies are based on datasets biased with heterodimers. The relation between number of interface residues and monomer length is shown in Figure 4.5 and 4.6. They show that interface residues increase with both heterodimer and homodimer monomer length. However, the relation is causal. Figure 4.7 and 4.8 show a causal relationship between interface area and monomer length for both homodimers and heterodimers. The mean interface residues are larger in homodimers than heterodimers. This is consistent with previous studies [Bahadur, et al., 2003; Caffrey, et al., 2004].

Figure 4.5: Interface residues in heterodimers

Chapter 4

Figure 4.6: Interface residues in homodimers

Figure 4.7: Interface area in heterodimer interface

Chapter 4

Figure 4.8: Interface area in homodimer interface

4.4.2 Interface H-bonds

Intermolecular hydrogen bonds between subunits are important in protein-protein association and stability [Tsai, et al., 1997a; Lo Conte, et al., 1999]. H-bonds in homodimers (range 0 - 51) and heterodimers (range 0 - 98) are different. The mean H-bonds are larger for homodimers (mean = 18) than heterodimers (mean = 12). Figure 4.9 and Figure 4.10 show that there is a high correlation between H-bonds and interface residues. The correlation coefficient is 0.83 in heterodimers and 0.85 in homodimers. This is similar to the previous reports in the range of 0.75 and 0.89 [Jones & Thornton, 1995; Xu, et al., 1997b; Lo Conte, et al., 1999; Bahadur, et al., 2003; Nooren & Thornton, 2003b].

Chapter 4

Figure 4.9: Hydrogen bonds in heterodimer interface

Figure 4.10: Hydrogen bonds in homodimer interface

Chapter 4

However, there is a subtle difference with the previous studies and the variation is affected by structure resolution, dataset size and data type. The dataset used in this study contains structures with resolution $\leq 2.5\text{\AA}$ and the data is either exclusively homodimer or heterodimer. However, previous datasets contain structures with resolution $\leq 3.0\text{\AA}$ and the data is a mixture of heterodimers, homodimers and other oligomers. At low resolution there are fewer H-bonds and the correlation with interface area decreases [Lo Conte, et al., 1999]. Here, we show that the relation between H-bonds and interface residues is highly correlated for both heterodimers and homodimers. This is useful to evaluate inter-subunit H-bonds prediction and their involvement in interface stability. On average there are 0.24 H-bonds per interface residue in heterodimers and 0.22 H-bonds per interface residue in homodimer. The maximum number of H-bonds per interface residue is 0.65 in heterodimers and 0.44 in homodimers. Although there are more intermolecular H-bonds in homodimers, the density of H-bonds per interface residue is lower in homodimers than in heterodimers [Bahadur, et al., 2003].

4.4.3 Interface hydrophobicity

Hydrophobic effect plays an important role in the dimer association [Lijnzaad & Argos, 1997; Tsai, et al., 1997a; Bahadur, et al., 2003]. Here, we used the experimentally derived hydrophobic scale to assign a mean value to each dimer interface [Jones & Thornton, 1997a; Li, et al., 2005]. The mean interface hydrophobicity is -1.00 in heterodimers and -0.61 in homodimers. This suggests that homodimer interfaces are more hydrophobic than heterodimers [Jones & Thornton, 1996; Bahadur, et al., 2003]. The ratio between buried hydrophobic and buried

Chapter 4

hydrophilic residues is 2.46 for heterodimers and 2.14 for homodimers. The results show that hydrophobic residues dominate dimer interfaces.

4.4.4 Interface residue composition

Several studies show the prevalence of certain types of residues at the dimer interfaces [Lijnzaad & Argos, 1997; Dasgupta, et al., 1997; Lo Conte, et al., 1999; Brinda, et al., 2002; Bahadur, et al., 2003]. However, the significance of hydrophobic, hydrophilic, and charged residues at the interface of homodimers and heterodimers is not well documented. Figure 4.11 show the fractional distribution of hydrophobic, hydrophilic and charged residues in homodimer and heterodimer interfaces. Hydrophobic residues (M, F, P, A, B, L), except for I and G are dominant in homodimer interfaces. However, hydrophilic residues (W, C, H, Q, N, Y, S), except for T, are dominant in heterodimer interfaces. This observation is interesting and not surprising because homodimers being made of identical monomer subunits tend to associate by hydrophobic interactions. This is in contrast to the observation in heterodimer interfaces being made of non-identical monomer subunits, associating generally by hydrophilic interactions.

Figure 4.12, shows the ration of interface/surface and interface/interior residue propensity difference between heterodimers and homodimers. Interestingly, the ratio of interface to interior charged residues (D, E, K, R) is significantly larger in heterodimers compared to homodimers (Figures 4.12, 4.13 and 4.14). On the other hand, the ratio of interface to interior hydrophobic residues (A, V, L, M, I, F) are prevalent in homodimers than in heterodimers (Figures 4.12, 4.13 and 4.14).

Chapter 4

Similarly, hydrophilic residues (N, Q, H, Y, S, T) are prevalent in heterodimer interfaces (Figures 4.12, 4.13 and 4.14). However, the propensity difference in the ratio of interface to surface hydrophobic/hydrophilic/charged for homodimers and heterodimers is almost zero (Figure 4.13).

Figure 4.11: Hydrophobic, hydrophilic and charged residue fraction in heterodimers and homodimers

Figure 4.12: Propensity difference in heterodimers and homodimers (heterodimers-homodimers)

Chapter 4

Figure 4.13: Ratio of interface to surface & interface to interior propensity in heterodimers

Figure 4.14: Ratio of interface to surface & interface to interior propensity in homodimers

Chapter 4

4.5 Conclusion

We performed a comprehensive analysis on the differences between 156 heterodimers and 170 homodimers. The homodimer and heterodimer datasets are mutually exclusive and is one of the unique features of the analysis. The analysis documents the differences between homodimer and heterodimer interfaces for the first time in a comprehensive manner. Homodimer interfaces have greater number of interface residues and H-bonds on average. However, the density of H-bonds per residue is greater for heterodimer interfaces. The study also shows that charged residues (D, E, K, R) and hydrophilic residues (N, T, S, Q, H, W, Y) are dominant at the heterodimer interfaces. Nonetheless, hydrophobic residues (A, V, L, M, I, F) are predominant at the homodimer interfaces. These data find utility in the development of independent models for the prediction of homodimer and heterodimer interaction sites.

Chapter 5

Interface residue prediction using sequence data

5.1 Introduction

With the development of computational and experimental technologies, the vast amount of protein-protein interaction data has been collected so far. These data have been accelerating the pace of discovery of protein-protein interactions. However, the knowledge on these interactions is far from satisfactory. Therefore, the development of reliable and efficient methods to detect and characterize protein-protein interactions is a necessary and challenging task to all bioinformatics researchers.

Studies have shown that protein-protein interfaces have detectable properties that can distinguish them from the rest of the protein surface [Young, et al., 1994; Jones & Thornton, 1996, 1997a; Lo Conte, et al., 1999; Hu, et al., 2000; Nooren & Thornton, 2003b; Neuvirth, et al., 2004; Bahadur, et al., 2004]. These properties include geometrical properties (interface size, planarity, sphericity and complementarity) and chemical properties (the types of amino acid chemical groups, hydrophobicity, electrostatic interactions and H-bonds). They are widely used to detect potential interaction sites or functional sites. There have already been some studies to detect protein-protein interaction sites using protein three-dimension structures [Jones & Thornton, 1997ab; Neuvirth, et al., 2004], protein docking simulations and analysis of the interaction energy landscapes [Fernandez-Recio, et al., 2004], correlated mutations [Pazos, et al., 1997; Marcotte, et al., 1999; Enright, et al., 1999; Eisenberg, et al., 2000; Huynen, et al., 2000; Pazos & Valencia, 2001, 2002; Tillier & Lui, 2003] and homology modeling and threading [Marcotte, et al., 1999]. The global protein-

Chapter 5

protein interaction network has also been used to infer unknown functions for the unclassified proteins. These methods include the so-called Neighboring Counting Method [Schwikowski, et al., 2000], the χ^2 -method [Hishigaki, et al., 2001], the Bayesian approaches [Deng, et al., 2002; Letovsky & Kasif, 2003], the Redundancy Method [Samanta & Liang, 2003] and a more recent Monte Carlo simulated annealing (SA) approach [Vazquez, et al., 2003]. In another study, Letovsky & Kasif successfully predicted protein function based on the local functional density of interacting neighbors using Markov Random Field formalism [Letovsky & Kasif, 2003].

In biology, it is virtually axiomatic that the protein sequence specifies its conformation [Anfinsen, 1973]. This suggests an intriguing postulate that knowledge of the protein sequence alone might be sufficient to evaluate protein-protein interactions and derive useful biological function [Bock & Gough, 2001]. This is very useful especially in cases in which the protein three-dimensional structures are unavailable. Studies show that interface residues, particularly those completely buried in the interface, are likely to be more conserved than other surface residues [Livingstone & Barton, 1993; Grishin & Phillips, 1994; Valdar & Thornton, 2001a; Matthews, et al., 2001; Yao, et al., 2003]. The property is used in methods that map the conservation score to the three-dimensional structure. These methods are mainly extensions of Evolutionary Trace approach [Lichtarge, et al., 1996; Lichtarge & Sowa, 2002; Glaser, et al., 2003], Consurf [Armon, et al., 2001], Rate4Site [Pupko, et al., 2002], and phylogenetic trees [Sjolander, 1998; Aloy, et al., 2001; Madabushi, et al., 2002; Del Sol Mesa, et al., 2003; La, et al., 2005]. They identify conserved positions

Chapter 5

and map them onto the three-dimensional structure of the protein (or onto the structure of a representative homolog), looking for spatially contiguous patches of conserved residues.

Another set of approaches used neural networks to predict interaction sites [Zhou & Shan, 2001; Fariselli, et al., 2002; Ofra & Rost, 2003ab]. Zhou & Shan [2001] used sequence profiles of neighboring residues and their solvent exposure as input to predict interface residues. For a set of 35 test proteins, the correct fractions of interface residues were 69% for bound structures and 70% for unbound structures. Fariselli et al. [Fariselli, et al., 2002] achieved above 70% accuracy for classifying residues in and outside of the interface using two very basic sources of information: evolutionary information as accumulated in sequence profiles derived from family alignments and surface patches in protein structures identified as sets of neighbor residues exposed to solvent. Ofra & Rost [2003a & 2003b] showed that protein interfaces have a significantly different amino acid composition from the rest of the protein surface and that the majority of interacting residues are clustered in sequence segments of several contacting residues. Indeed, recent studies indicate that the interfaces of many protein complexes can be subdivided into distinct patches where contacts are often provided by different protein domains [Chakrabarti & Janin, 2002]. Also, interactions with consecutive chain segments are clearly a hallmark of specific recognition by many signaling domains [Pawson & Nash, 2003]. They made use of these two conclusions and neural network to predict protein interaction sites from protein sequences and achieved a surprisingly high level of accuracy [Ofra & Rost, 2003ab].

Chapter 5

SVM is also widely used to predict interaction sites [Bock & Gough, 2001; Sprinzak & Margalit, 2001; Koike & Takagi, 2004; Martin, et al. 2005; Bradford & Westhead, 2005]. Koike & Takagi [2004] used the profiles of sequentially/spatially neighboring residues and additional information to predict interface residues. When structural information was used, the predicting performance was improved greatly. Bradford & Westhead [2005] combined the SVM approach with surface patch analysis to predict protein-protein binding sites. Using a leave-one-out cross-validation procedure, they successfully predicted the location of the binding site with an accuracy of 76% of their dataset made up of proteins with both transient and obligate interfaces. SVM approach is also used to predict protein-protein interactions combined with the sequence information and experimental data [Bock & Gough, 2001; Sprinzak & Margalit, 2001; Martin, et al., 2005]. However, there are some differences between these studies. Bock & Gough transformed sequence information into physico-chemical information (charge, hydrophobicity and surface tension). However, the signature descriptor used by Martin and colleagues did not require performing such a transformation. Furthermore, Bock and Gough encoded and compared protein pairs by concatenating normalized versions of the amino acid sequences of each protein, and hence used a global representation of a protein pair. However, Martin and colleagues used explicit pair-oriented, local sequence information (product signature). These amino acid subsequence pairs should occur together when two proteins interact.

Although, there have been many methods to analyze and detect protein-protein interaction sites, the results are still far from satisfactory. There is still more space to improve the prediction models. Success in the development of prediction model will

Chapter 5

encourage the development of genomics and proteomics. This will also greatly benefit protein design and drug design. Therefore, it is our interest to predict interaction sites from homodimer and heterodimer sequence by analyzing their interface properties.

5.2 Material and Methods

The prediction model is derived from GOR (Garnier-Osguthorpe-Robson) methods which have been successfully used to detect protein secondary structures [Garnier, et al., 1978; Gibrat, et al., 1987; Garnier, et al., 1996]. The method is based on information theory and Bayesian statistics. GOR stands for the first letter of the names of the authors of the original publication [Garnier, et al., 1978]. In this method, protein secondary structures include three states, which are α -helix, β -strand and coil residues. These three states are usually represented by H (α -helix), E (β -strand) and C (coil). In our study, protein sequences are transformed into three state sequences and these three states include interface, interior and surface residues. They are represented by Y (interface residue), N (interior residue) and S (surface residue) respectively. Therefore, from the view of this, the prediction of interaction sites is very similar to the prediction of protein secondary structures. This encourages us to use GOR methods to detect dimer interface residues.

5.2.1 Heterodimer structural dataset

The non-redundant dataset of 156 heterodimers (Table 3.1) and 170 homodimers (Table 4.1) is used in this chapter. These data were used in previous chapters by us.

Chapter 5

5.2.2 Prediction algorithm

The basic idea of the method is to treat the monomer amino acid sequence and the transformed three state sequences as two messages related by a translation process. Also, we consider the information contained in the other sequence that interact with the processed sequence. Information theory with simple Bayesian statistics is used to map this translation process [Garnier, et al., 1978; Gibrat, et al., 1987; Garnier, et al., 1996].

Let us use the following notations:

$\mathbf{r} = (r_1, r_2, \dots, r_n)$ is the sequence of n amino acid residues, where r_i denotes the i th residue. $\mathbf{T} = (T_1, T_2, \dots, T_n)$ is the transformed three state sequence corresponding to respective residues, where $T_i \in \{Y, N, S\} \quad \forall i = 1, 2, \dots, n$.

The information function, I , measures the mutual information contained between the monomer sequence and the transformed three state sequence, which can be defined as

$$I(T_i, r_i) = \ln \left(\frac{P(T_i / r_i)}{P(T_i)} \right) \quad (4-1)$$

where $P(T_i)$ is the prior probability of having a transformed state T_i and $P(T_i / r_i)$ is the conditional probability for the presence of a transformed state T_i at the location of a residue r_i . From the Bayes rule, these probabilities can be expressed as $P(T_i / r_i) = f(T_i, r_i) / f(r_i)$ and $P(T_i) = f(T_i) / M$, where $f(T_i, r_i)$ is the frequency that r_i and T_i jointly occur, $f(T_i)$ is the frequency that T_i occur, and M is the total number of residues, in a given set of sequences. So in a given window ($w = 2s + 1$), the

Chapter 5

information difference for each of the transformed states for residue i is given as follows:

$$I(\Delta T_i, \mathbf{r}_i) = I(T_i, \mathbf{r}_i) - I(\overline{T}_i, \mathbf{r}_i) \quad (4-2)$$

where $\overline{T}_i$ is the complement of a transformed state T_i , and $\mathbf{r}_i = (r_{i-s}, \dots, r_{i+s})$ represents residues within the window around a residue r_i .

In our research, the information difference comes from two parts. It has the following form:

$$I(\Delta T_i, \mathbf{r}_i) = I_1 + I_2 \quad (4-3)$$

I_1 is derived from GOR III (Gibrat et al. 1987), which follows the approximation:

$$I_1 \approx \sum_{j=1}^l \ln \left(\frac{f(T_i, r_j, r_i)}{f(\overline{T}_i, r_j, r_i)} \right) + l \ln \left(\frac{f(\overline{T}_i, r_i)}{f(T_i, r_i)} \right) \quad (4-4)$$

where $f(T_i, r_j, r_i)$ is the frequency that T_i , r_j and r_i jointly occur. T_i and r_i come from the observed sequence and r_j comes from its interacting partner sequence (Figure 5.1). In this approximation, the assumption is made that the pair-wise combination of the central residue and a residue in the other sequence has an influence that the central residue stays interface, interior or surface. The interacting partner of one protein must have information determines whether a residue stays at interface, interior or surface since it can interact with many different partners in different environment [Jeong, et al., 2001].

Chapter 5

Figure 5.1: Information difference in chan A and Chain B

In this model the interaction is illustrated. Residues are denoted by circles. The sliding window size ($ws=2s+1$) is used to identify interface residue in chain A with reference to chain B of length l . By analysis, the optimized window size is 11 ($ws=11$, $s=6$).

$\mathbf{r}=(r_1, r_2, \dots, r_n)$ is the sequence of n amino acid residues, where r_i denotes the i th residue in chain A, which is the middle residue in the sliding window; r_{i-s} and r_{i+s} denote the residue at the beginning and the end of the sliding window, respectively; And r_j denotes any residue from chain A's interacting partner (chain B).

$\mathbf{T}=(T_1, T_2, \dots, T_n)$ is the transformed three state sequence corresponding to respective residues, where $T_i \in \{Y, N, S\} \quad \forall i=1, 2, \dots, n$, where Y represents interface residue, N represents interior residue and S represents surface residue.

I_2 is derived from GOR IV (Garnier et al. 1996). Another approximation is used to take into account all possible pairs formed by each residue in the given window. The assumption used in this approximation is that the pair-wise combinations of residues in the window influence the central residue to stay interface, interior or surface:

Chapter 5

$$I_2 \approx \frac{2}{w} \sum_{j=-s}^{s-1} \sum_{k=j+1}^s \ln \left(\frac{f(T_i, r_{i+j}, r_{i+k})}{f(\bar{T}_i, r_{i+j}, r_{i+k})} \right) - \frac{(w-2)}{w} \sum_{j=-s}^s \ln \left(\frac{f(T_i, r_{i+j})}{f(\bar{T}_i, r_{i+j})} \right) + \ln \left(\frac{f(\bar{T}_i)}{f(T_i)} \right) \quad (4-5)$$

The above approximations for the mutual information between the amino acid sequences and its transformed three states sequence (Y, N and S) provide means to formulate the influence of local sequence upon a given residue to stay at interface, interior or surface. Using the information value, we can compute the probability that the residue r_i belongs to interface, interior or surface (T_i). The probability is given as follows:

$$P(T_i / \mathbf{r}_i) = \frac{1}{1 + \frac{f(\bar{T}_i)}{f(T_i)} e^{-I(\Delta T_i, \mathbf{r}_i)}} \quad (4-6)$$

This conditional probability can be used to determine whether a given residue stays at interface, interior or surface (T_i). The prediction of each residue is finished by finding the giving maximum probability:

$$T_i = \begin{cases} Y & \text{if } P(Y / \mathbf{r}_Y) \text{ is maximum} \\ N & \text{if } P(N / \mathbf{r}_N) \text{ is maximum} \\ S & \text{if } P(S / \mathbf{r}_S) \text{ is maximum} \end{cases} \quad (4-7)$$

5.2.3 Definition of window size

The sequence contains the information for placing the residues at interface, interior or surface upon folding and binding. However in practice, it is only possible to use a window based approach to consider the local sequence around the residue of interest. As shown in Figure 5.2 and Figure 5.3, most of the consecutive interface residue

Chapter 5

segments are shorter than or equal to 6. For heterodimers, 97.8% and 94.4% of consecutive segments are shorter than 6 residues in chain A and chain B respectively. For homodimers, the percentage is 96.3% in chain A and 96.3% in chain B (we will consider only one chain because the two chains are identical in homodimers). So we choose 11 as the sliding window size for both heterodimers and homodimers analysis.

Figure 5.2: Interface residue length distribution in heterodimers

Figure 5.3: Interface residue length distribution in homodimers

Chapter 5

5.2.4 Accuracy definition

We perform leave-one-out cross-validation to evaluate the prediction method. This method involves removing one protein from the training set, performing training on the remaining proteins and then predicting the position of the interface, interior and surface residues on the selected protein. This process is repeated until all proteins have been left out

We adopt the overall three-state (interface, interior and surface residues) prediction percentage Q_3 defined as the ratio of correctly predicted residues to the total number of residues in the database under consideration.

$$Q_3(\%) = \sum_{t \in \{Y, N, S\}} \frac{a_t}{b_t} \times 100 \quad (4-8)$$

where a_t is the number of correctly predicted residues in state t , and b_t is the number of residues observed in state t .

A measure of the performance on interface, interior and surface state $t \in \{Y, N, S\}$ is the percentage Q_t of correctly predicted residues observed in class t

$$Q_t(\%) = \frac{a_t}{b_t} \times 100 \quad (4-9)$$

These measures can be helpful to gain information on prediction accuracy of each structure state (interface, interior and surface residues).

Chapter 5

5.3 Results and discussion

5.3.1 Prediction accuracy

Table 5.1 shows prediction accuracies with window size 11. Also, the observed accuracy and observed number of interface, interior and surface residues are listed in Table 5.2. For both heterodimers and homodimers the accuracy (Q_n) of interior residues is all greater than observed accuracy (Figure 5.4, 5.5, and 5.6). The accuracy (Q_y) of interface residues of chain A in heterodimers is 0.295. For all these chains, 90% have prediction accuracy greater than observed accuracy (Figure 5.4). At the same time, the accuracy (Q_y) of interface residues of chain B in heterodimers is 0.378, 85% of which have the accuracy greater than observed accuracy (Figure 5.5). The overall prediction accuracy (Q_3) are 0.489, 0.402, and 0.464 for chain A and chain B of heterodimers and chain A of homodimers. They are in a narrow range of 0.4-0.5 for both heterodimers and homodimers.

Table 5.1: Prediction accuracies with window size 11

	Heterodimer		Homodimer
	Chain A	Chain B	Chain A
Accuracy for interface residue (Q_y)	0.295	0.378	0.099
Accuracy for interior residue (Q_n)	0.774	0.746	0.878
Accuracy for surface residue (Q_s)	0.412	0.319	0.418
Accuracy for above three type of residues (Q_3)	0.489	0.402	0.464

Table 5.2: Observed number and observed accuracy

			Observed number	Observed accuracy
Heterodimer	Chain A	Interface residue	4170	0.106
		Interior residue	10543	0.268
		Surface residue	24631	0.626
		Total	39344	1.000
	Chain B	Interface residue	3761	0.172
		Interior residue	4525	0.207
		Surface residue	13522	0.621
		Total	21808	1.000
Homodimer	Chain A	Interface residue	6901	0.149
		Interior residue	12077	0.262
		Surface residue	27247	0.589
		Total	46225	1.000

Observe number is the number of each type of residues in the data set. *Observed accuracy* is the ratio of number of each type of residue to the total sum of three types of residues (interface, interior and surface residues).

Figure 5.4: The distribution of accuracy in chain A of heterodimers

Chapter 5

Figure 5.5: The distribution of accuracy in chain B of heterodimers

Figure 5.6: The distribution of accuracy in chain A of homodimers

5.3.2 Window size effect on accuracy

Here, we show the effect on prediction accuracy caused by different window sizes. As shown in Figures 5.7, 5.8 and 5.9, the window size has little effect on prediction accuracy in both heterodimers and homodimers, especially in homodimers (Figure 5.9). The overall accuracy (Q_3) increases with the increase of window size. However, interface residue prediction accuracy (Q_y) in heterodimers decreases with the increase of window size (Figures 5.7 and 5.8). The accuracy achieves the maximum at the window size 11 in chain A of heterodimer.

Figure 5.7: Window size effect on prediction accuracy in chain A of heterodimer

Chapter 5

Figure 5.8: Window size effect on prediction accuracy in chain B of heterodimer

Figure 5.9: Window size effect on prediction accuracy in chain A of homodimer

5.3.3 Sequence length effect on accuracy

Previous studies showed the different characteristics of interaction sites depending on the interaction surface area [Lo Conte, et al., 1999; Glaser, et al., 2001]. We also know that the number of interface residues is proportional to interface area

Chapter 5

[Chakrabarti & Janin, 2002; Bahadur, et al., 2003]. Stronger protein subunit associations were generally associated with larger interface areas [Jones & Thornton, 1995]. Therefore, we analyze the effect on prediction accuracy caused by monomer sequence length. Heterodimers and homodimers are grouped into several groups according their sequence length (Table 5.3). Prediction accuracy (Q_y) of interface residues increases with the increase of sequence length, except for group L100 in chain A of heterodimers and homodimers (Figure 5.10). We do not consider their results because there are only 4 and 5 sequences in each group respectively (Table 5.3). However, the overall prediction accuracy (Q_3) decreases with the sequence length increase.

Table 5.3: Groups according to heterodimer and homodimer sequence length

Sequence length (residue number)	Heterodimer		Homodimer
	Chain A	Chain B	Chain A
Between 50 and 100	4	59	5
Between 100 and 200	56	74	61
Between 200 and 300	58	15	45
Between 200 and 400	18	8	28
Between 400 and 500	20		20
Greater than 600			11
Total	156	156	170

Figure 5.10: Prediction accuracy in different length groups

5.3.4 Comparison with other studies

There have been several studies on the prediction of the potential interface site residues [Zhou & Shan, 2001; Fariselli, et al., 2002; Ofra & Rost, 2003; Koike & Takagi, 2004; Bradford & Westhead, 2005]. Zhou & Shan [2001] adopted the neural network method with the input of sequence profiles and ASA. They showed a prediction accuracy of 0.51.

Fariselli and colleagues showed using SVM method that the precision of interface residues and surface residues are 0.56 and 0.85 respectively [Fariselli, et al., 2002]. In their dataset the protease complex is eliminated. Moreover, they incorporate two information sources which are evolutionary information as accumulated in sequence profiles derived from family alignments and surface patches in protein structures

Chapter 5

identified as sets of neighbor residues exposed to solvent. Their results are better than ours. But we only use sequence data information.

Ofran & Rost [2003] predicted interaction sites using sequentially neighboring profiles and neural network. They achieved a high precision. Another study to predict interaction sites used local sequence information [Bradford & Westhead, 2005]. These two studies considered the whole interaction site, not the all interface residues. So we can not compare our results with their.

Koike & Takagi [2004] also used SVM to predict interface residues from sequence data. They achieved the accuracy of 0.264 for interface residues when only adopting sequence data information and window size of 11. This is a little lower than our results (Q_y) of heterodimers (0.295 and 0.387 for chain A and chain B of heterodimers respectively).

5.4 Conclusion

We developed a method to predict interface residues from sequence. Our method not only considers the predicted sequence information but also its interacting partners. For chain A and chain B of heterodimers, the prediction accuracy (Q_y) of interface residues are 0.295 and 0.387. Among them, 90% and 85% have the accuracy greater than observed accuracy (R_y) which are 0.106 and 0.172, respectively. The predication accuracy (Q_n) for interior residues is greater than observed accuracy. The overall accuracy (Q_3) are 0.489, 0.402, and 0.464 for chain A (heterodimer) and chain B (heterodimer) and chain A (homodimer), respectively. Our method still has

Chapter 5

much room for improvement. However, it is a challenging task to predict interaction sites from sequence.

Chapter 6

Conclusion and future work

6.1 Conclusion

A large number of structurally important physical - chemical parameters characterize heterodimer interfaces and each of these parameters contributes differently to the stability of a heterodimer interface. A weighted value was assigned to each parameter to indicate the differential contribution. Here, we apply a mathematical procedure to determine the most critical parameters that describe a heterodimer interface. The six critical interface parameters discussed here are based on the selected 156 heterodimer structures. The multi dimensional scaling procedure suggests that the six critical parameters effectively replace the original 118 parameter set. This finding is of critical importance in the understanding and development of prediction models for heterodimer interfaces.

We performed a comprehensive analysis on the differences between 156 heterodimers and 170 homodimers. The homodimer and heterodimer datasets are mutually exclusive and is one of the unique features of the analysis. The analysis documents the differences between homodimer and heterodimer interfaces for the first time in a comprehensive manner. Homodimer interfaces have greater number of interface residues and H-bonds on average. However, the density of H-bonds per residue is greater for heterodimer interfaces. The study also shows that charged residues (D, E, K, R) and hydrophilic residues (N, T, S, Q, H, W, Y) are dominant at the heterodimer interfaces. Nonetheless, hydrophobic residues (A, V, L, M, I, F) are predominant at the homodimer interfaces. These data find utility in the development

Chapter 6

of independent models for the prediction of homodimer and heterodimer interaction sites.

We developed a method to predict interface residues from sequence. Our method not only considers the predicted sequence information but also its interacting partners. For chain A and chain B of heterodimers, the prediction accuracy (Q_y) of interface residues are 0.295 and 0.387. Among them, 90% and 85% have the accuracy greater than observed accuracy (R_y) which are 0.106 and 0.172, respectively. The predication accuracy (Q_n) for interior residues is very high and greater than the observed accuracy. The overall accuracy (Q_3) are 0.489, 0.402, and 0.464 for chain A (heterodimer) and chain B (heterodimer) and chain A (homodimer), respectively. Our method still has much room for improvement. However, it is a challenging task to predict interaction sites from sequence.

6.2 Future work

Future work mainly focuses on improving prediction accuracy. Firstly, we will try other methods or algorithms to improve accuracy. Secondly, we will try to determine new interface features to describe protein-protein interface accurately.

List of Publication

List of Publications

1. Cui, ZH., Gan, GK., Li, L., Sakharkar, MK., Kanguane, P. 2005. Identification of critical heterodimer protein interface parameters by multi-dimensional scaling in Euclidian space. *Front Biosci* 10: 844-852.
2. Cui, ZH., Gan, GK., Li, L., Sakharkar, MK., Kanguane, P. 2005. Protein subunit interfaces: heterodimers versus homodimers. *Bioinformation* 1(2): 28-39.
3. Li, L., Gan, GK., Cui, ZH., Sakharkar, MK., Kanguane, P. 2005. Generation of a dataset for studying ligand effect on homodimer interface. *Front Biosci* 10: 1977-1984.
4. Li, L., Gunasekaran, K., Gan, GK., Cui, ZH., Sakharkar, MK., Kanguane, P. 2005. Structural features differentiate the mechanisms between 2S (2 state) and 3S (3 state) folding of homodimers. *Bioinformation* 1(2): 42-49.

References

Abascal F & Valencia A. Automatic annotation of protein function based on family identification. *Proteins*, 53, 683-692 (2003).

Albert I & Albert R. Conserved network motifs allow protein-protein interaction prediction. *Bioinformatics*, 20(18), 3346-3352 (2004).

Aloy P, Querol E, Aviles FX & Sternberg MJ. Automated structure-based prediction of functional sites in proteins: Applications to assessing the validity of inheriting protein function from homology in genome annotation and to protein docking. *J. Mol. Biol.* 311, 395-408 (2001).

Aloy P & Russell RB. InterPreTS: protein interaction prediction through tertiary structure. *Bioinformatics*, 19, 161-162 (2003).

Andrade MA, Casari G, Sander C & Valencia A. Classification of protein families and detection of the determinant residues with an improved self-organizing map. *Biol. Cybern.* 76, 441-450 (1997).

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

Archakov AI, Govorun VM, Dubanov AV, Ivanov YD, Veselovsky AV, Lewi P & Janssen P. Protein-protein interactions as a target for drugs in proteomics. *Proteomics*, 3, 380-391 (2003).

References

Armon A, Graur D, & Ben-Tal N. ConSurf: An algorithmic tool for the identification of functional regions in proteins by surface mapping of phylogenetic information. *J. Mol. Biol.*, 307, 447-463 (2001).

Bahadur RP, Chakrabarti P, Rodier F & Janin J. Dissecting subunit interfaces in homodimeric proteins. *Proteins*, 53, 708-719 (2003)

Bahadur RP, Chakrabarti P, Rodier F & Janin J. A dissection of specific and non-specific protein-protein interfaces. *J. Mol. Biol.*, 336, 943-955 (2004).

Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer EL, Studholme DJ, Yeats C and Eddy SR. The Pfam protein families database. *Nucleic Acids Res*, 32, D138-D141 (2004).

Betts MJ & Sternberg MJ. An analysis of conformational changes on protein-protein association: implications for predictive docking. *Protein Eng.*, 12, 271-283 (1999).

Bogan AA & Thorn KS. Anatomy of hot spots in protein interfaces. *J. Mol. Biol.*, 280, 1-9 (1998).

Bock JR & Gough DA. Predicting protein-protein interactions from primary structure. *Bioinformatics*, 17(5), 455-60 (2001).

Bradford JR & Westhead DR. Asymmetric mutation rates at enzyme-inhibitor interfaces: implications for the protein-protein docking problem. *Protein Sci.*, 12, 2099-2103 (2003).

References

Bradford JR & Westhead DR. Improved prediction of protein-protein binding sites using a support vector machines approach. *Bioinformatics*, 21(8), 1487-94 (2005).

Brinda KV, Kannan N & Vishveshwara S. Analysis of homodimeric protein interfaces by graph-spectral methods. *Protein Eng.*, 15(4), 265-77 (2002).

Caffrey DR, O'Neill LA & Shields DC. A method to predict residues conferring functional differences between related proteins: Application to MAP kinase pathways. *Protein Sci.*, 9, 655-670 (2000).

Caffrey DR, Somaroo S, Hughes JD, Mintseris J & Huang ES. Are protein-protein interfaces more conserved in sequence than the rest of the protein surface? *Protein Sci.*, 13, 190-202 (2004).

Camacho CJ, Weng ZP, Vajda S & Delisi C. Free energy landscapes of encounter complexes in protein-protein association. *Biophys. J.*, 76, 1166-1178 (1999).

Casari G, Sander C & Valencia A. A method to predict functional residues in proteins. *Nat. Struct. Biol.*, 2, 171-178 (1995).

Chakrabarti P & Janin J. Dissecting protein-protein recognition sites. *Proteins*, 47, 334-343 (2002).

Chothia C & Janin J. Principles of protein-protein recognition. *Nature*, 256, 705-708 (1975).

Chou KC & Cai YD. A novel approach to predict active sites of enzyme molecules. *Proteins*, 55, 77-82 (2004)

References

Clackson T & Wells JA. A hot spot of binding energy in a hormone receptor interface. *Science*, 267, 383-386 (1995).

Cole C & Warwicker J. Side-chain conformational entropy at protein-protein interfaces. *Protein Sci.*, 11, 2860-2870 (2002).

Dall'Acqua W, Goldman ER, Lin W, Teng C, Tsuchiya D, Li H, Ysern X, Braden BC, Li Y, Smith-Gill SJ & Mariuzza RA. A mutational analysis of binding interactions in an antigen-antibody protein-protein complex. *Biochemistry*, 37, 7981-7991 (1998).

Dasgupta S, Iyer GH, Bryant SH, Lawrence CE & Bell JA. Extent and nature of contacts between protein molecules in crystal lattices and between subunits of protein oligomers. *Proteins*, 28(4), 494-514 (1997).

Davies DR & Cohen GH. Interactions of protein antigens with antibodies. *Proc. Natl. Acad. Sci. USA*, 93, 7-12 (1996).

De Vos AM, Ultsch M & Kossiakoff AA. Human growth hormone and extracellular domain of its receptor: crystal structure of the complex. *Science*, 255, 306-312 (1992).

Del Sol Mesa A, Pazos F & Valencia A. Automatic methods for predicting functionally important residues. *J. Mol. Biol.*, 326, 1289-1302 (2003).

Demchuck E, Mueller T, Oschkinat H, Sebald W & Wade RC. Receptor binding properties of four-helix bundle growth factors deduced from electrostatic analysis. *Protein Sci.*, 3, 920-935 (1994).

References

Deng M, Zhang K, Mehta S, Chen T & Sun F. Prediction of protein function using protein–protein interaction data. In *Proceedings of the IEEE Computer Society Bioinformatics Conference (CSB'02)*, Stanford, CA, August 14-16, 197-207 (2002).

Dill KA. Dominant forces in protein folding. *Biochemistry*, 31, 7134-7155 (1990).

Eisenberg D, Marcotte EM, Xenarios I & Yeates TO. Protein function in the post-genomic era. *Nature*, 405, 823-826 (2000).

Eisenhaber F. Hydrophobic regions on protein surface. *Drug. Discovery and Design*, 17, 27-42 (1999).

Emberly EG, Mukhopadhyay R, Wingreen NS & Tang C. Flexibility of α -helices: results of a statistical analysis of database protein structures. *J.Mol. Biol.*, 327, 229-237 (2003).

Enright AJ, Iliopoulos I, Kyrpides NC & Ouzounis CA. Protein interaction maps for complete genomes based on gene fusion events. *Nature*, 402, 86-90 (1999).

Fariselli P, Pazos F, Valencia A & Casadio R. Prediction of protein–protein interaction sites in heterocomplexes with neural networks. *Eur. J. Biochem.*, 269, 1356-1361 (2002).

Fernandez A & Scheraga HA. Insufficiently dehydrated hydrogen bonds as determinants of protein interactions. *PNAS*, 100(1), 113-118 (2003)

Fernandez A & Scott R. Dehydron: a structurally encoded signal for protein interaction. *Biophys J.*, 85, 1914-1928 (2003).

References

- Fernandez-Recio J, Totrov M & Abagyan R. Identification of protein-protein interaction sites from docking energy landscapes. *J. Mol. Biol.*, 335(3), 843-65 (2004).
- Fields S & Song O. A novel genetic system to detect protein-protein interactions. *Nature*, 340, 245-246 (1989).
- Fradera X, De La Cruz X, Silva CH, Gelpi JL, Luque FJ & Orozco M. Ligand-induced changes in the binding sites of proteins. *Bioinformatics*, 18(7), 939-48 (2002).
- Gallet X, Charlotiaux B, Thomas A & Brasseur R. A fast method to predict protein interaction sites from sequences. *J. Mol. Biol.*, 302(4), 917-26 (2000).
- Garnier J, Osguthorpe DJ, & Robson B. Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *J. Mol. Biol.*, 120, 97-120 (1978).
- Garnier J, Gibrat JF & Robson B. GOR method for predicting protein secondary structure from amino acid sequence. *Methods Enzymol*, 266, 541-553 (1996).
- Gibrat JF, Garnier J & Robson B. Further developments of protein secondary structure prediction using information theory. *J. Mol. Biol.*, 198, 425-443 (1987).
- Glaser F, Pupko T, Paz I, Bell RE, Bechor-Shental D, Martz E & Ben-Tal N. ConSurf: identification of functional regions in proteins by surface-mapping of phylogenetic information. *Bioinformatics*, 19, 163-164 (2003).

References

Glaser F, Steinberg DM, Vakser IA & Ben-Tal N. Residue frequencies and pairing preferences at protein-protein interfaces. *Proteins*, 43, 89-102 (2001).

Grishin NV & Phillips MA. The subunit interfaces of oligomeric enzymes are conserved to a similar extent to the overall protein sequences. *Protein Sci.*, 3, 2455-2458 (1994).

Guinto ER & Di Cera E. Large Heat Capacity Change in a Protein-Monovalent Cation Interaction. *Biochemistry*, 35, 8800-8804 (1996).

Hannenhalli SS & Russell RB. Analysis and prediction of functional sub-types from protein sequence alignments. *J. Mol. Biol.*, 303, 61-76 (2000).

Halperin I, Ma B, Wolfson HJ & Nussinov R. Principles of docking: An overview of search algorithms and a guide to scoring functions. *Proteins*, 47, 409-443 (2002)

Halperin I, Wolfson H & Nussinov R. Protein-protein interactions; coupling of structurally conserved residues and of hot spots across interfaces. Implications for docking. *Structure (Camb)*, 12, 1027-1038 (2004).

Harpaz Y, Gerstein M & Chothia C. Volume changes on protein folding. *Structure*, 2(7), 641-649 (1994).

Hendsch ZS & Tidor B. Do salt bridges stabilize proteins? A continuum electrostatics analysis. *Protein Sci.*, 3, 211-226 (1994).

Hirakawa H, Muta S & Kuhara S. The hydrophobic cores of proteins predicted by wavelet analysis. *Bioinformatics*, 15 (2), 141-148 (1999).

References

Hishigaki H, Nakai K, Ono T, Tanigami A & Tagaki T. Assessment of prediction accuracy of protein function from protein–protein interaction data. *Yeast*, 18, 523-531 (2001).

Hubbard SJ. NACCESS, computer program. Department of Biochemistry and Molecular Biology, University College, London (1993).

Hurley JH, Thorness PE, Ramalingam V & Helmers NH. Structure of a bacterial enzyme regulated by phosphorylation, isocitrate dehydrogenase. *Proc. Natl. Acad. Sci. USA* 86, 8635-8639 (1989).

Huynen M, Snel B, Lathe W & Bork P. Predicting protein function by genomic context: quantitative evaluation and qualitative inferences. *Genome Res.*, 10, 1204-1210 (2000).

Hu Z, Ma B, Wolfson H & Nussinov R. Conservation of polar residues as hot spots at protein interfaces. *Proteins*, 39, 331-342 (2000).

Ippolito JA, Alexander RS & Christianson DW. Hydrogen bond stereochemistry in protein structure and function. *J. Mol. Biol.*, 215, 457-471 (1990)

Ivanov YD, Kanaeva IP, Karuzina II, Archakov AI, Hoa GH & Sligar SG. Molecular recognition in the p450cam monooxygenase system: direct monitoring of protein-protein interactions by using optical biosensor. *Arch. Biochem. Biophys.*, 391, 255-264 (2001)

Janin J. Principles of protein–protein recognition from structure to thermodynamics. *Biochimie*, 77, 497-505 (1995).

References

- Janin J. Protein-protein recognition. *Prog. Biophys. Mol. Biol.*, 64, 145-166 (1996).
- Janin J. Wet and dry interfaces: the role of solvent in protein-protein and protein-DNA recognition. *Structure Fold. Des.*, 7, 277-279 (1999).
- Janin J & Chothia C. The structure of protein-protein recognition sites. *J. Biol. Chem.*, 265, 16027-16030 (1990).
- Janin J, Miller S & Chothia C. Surface, subunit interfaces and interior of oligomeric proteins. *J. Mol. Biol.*, 204, 155-164 (1988).
- Janin J & Seraphin B. Genome-wide studies of protein-protein interaction. *Curr. Opin. Struct. Biol.*, 13, 383-388 (2003)
- Jansen R, Yu H, Greenbaum D, Kluger Y, Krogan NJ, Chung S, Emily A, Snyder M, Greenblatt JF & Gerstein M. A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science*, 302, 449-453 (2003).
- Jeong H, Mason SP, Barabasi AL & Oltvai ZN. Lethality and centrality in protein networks. *Nature*, 411, 41-42 (2001).
- Jones S, Marin A & Thornton JM. Protein domain interfaces: characterization and comparison with oligomeric protein interfaces. *Protein Eng.*, 13, 77-82 (2000).
- Jones S & Thornton JM. Protein-protein interactions: a review of protein dimer structures. *Prog. Biophys. Molec. Biol.*, 63, 31-65 (1995).
- Jones S & Thornton JM. Principle of protein-protein interactions. *Proc. Natl. Acad. Sci. USA*, 93, 13-20 (1996).

References

Jones S & Thornton JM. Analysis of protein-protein interaction sites using surface patches. *J. Mol. Biol.*, 272, 121-132 (1997a).

Jones S & Thornton JM. Prediction of protein-protein interaction sites using patch analysis, *J. Mol. Biol.*, 272, 133-143 (1997b).

Kallblad P & Dean PM. Efficient conformational sampling of local side-chain flexibility. *J. Mol. Biol.*, 326, 1651-1665 (2003).

Karshikov A, Bode W, Tulinsky A & Stone SR. Electrostatic interactions in the association of proteins: analysis of the thrombin-hirudin complex. *Protein Sci.*, 1, 727-735 (1992).

Keil M, Exner TE & Brickmann J. Pattern recognition strategies for molecular surfaces: III. binding site prediction with a neural network. *J. Comput. Chem.*, 25, 779-789 (2004).

Keskin O, Ma B & Nussinov R. Hot regions in protein-protein interactions: the organization and contribution of structurally conserved hot spot residues. *J. Mol. Biol.*, 345, 1281-1294 (2005).

Kimura SR, Brower RC, Vajda S & Camacho CJ. Dynamical view of the positions of key side chains in protein-protein recognition. *Biophys. J.*, 2001, 80, 635-642 (2001).

Koike A & Takagi T. Prediction of protein-protein interaction sites using support vector machines. *Protein Eng. Des. Sel.*, 17, 165-173 (2004).

Korn AP & Burnett RM. Distribution and complementarity of hydrophathy in

References

multisubunit proteins. *Proteins*, 9, 37-55 (1991).

Kortemme T & Baker D. A simple physical model for binding energy hot spots in protein-protein complexes. *Proc. Natl. Acad. Sci. USA*, 99, 14116-14121 (2002).

Krystek S, Stouch T & Novotny J. Affinity and specificity of serine endopeptidase-protein inhibitor interactions. Empirical free energy calculations based on X-ray crystallographic structures. *J. Mol. Biol.*, 234, 661-679 (1993)

La D, Sutch B & Livesay DR. Prediction protein functional sites with phylogenetic motifs. *Proteins*, 58, 309-320 (2005).

Landgraf R, Xenarios I, & Eisenberg D. Three-dimensional cluster analysis identifies interfaces and functional residue clusters in proteins. *J.Mol. Biol.*, 307, 1487-1502 (2001).

Larsen TA, Olson AJ & Goodsell DS. Morphology of protein-protein interfaces. *Structure*, 6, 421-427 (1998).

Laskowski RA. SURFNET: a program for visualizing molecular surfaces, cavities, interfaces. *J. Mol. Graph.*, 13(5), 323-330 (1995)

Lawrence MC & Colman PM. Shape complementarity at protein-protein interfaces. *J. Mol. Biol.*, 234, 946-950 (1993).

Leone M & Pagnani A. predicting protein functions with message passing algorithms. *Bioinformatics*, 21(2), 239-247 (2005).

References

- Lescar J, Pellegrini M, Souchon H, Tello D, Poljak R, Peterson N, Greene M & Alzari P. Crystal structure of a cross reaction complex between Fab F9.13.7 and guinea-fowl lysozyme. *J. Biol. Chem.*, 270, 18067-18076 (1995).
- Letovsky S & Kasif S. Predicting protein function from protein/protein interaction data: a probabilistic approach. *Bioinformatics*, 19, 197-204 (2003).
- Li L, Shakhnovich EI & Mirny LA. Amino acids determining enzyme-substrate specificity in prokaryotic and eukaryotic protein kinases. *Proc. Natl. Acad. Sci.*, 100, 4463-4468 (2003).
- Li L, Gan JG, Zhanhua C, Sakharkar MK & Kanguane P. Generation of a dataset for studying ligand effect on homodimer interface. *Frontiers in Bioscience*, 10, 1977 (2005).
- Lichtarge O, Bourne HR & Cohen FE. An evolutionary trace method defines binding surfaces common to protein families. *J. Mol. Biol.*, 257, 342-345 (1996).
- Lichtarge O & Sowa ME. Evolutionary predictions of binding surfaces and interactions. *Curr. Opin. Struct. Biol.*, 12, 21-27 (2002).
- Lijnzaad P & Argos P. Hydrophobic patches on protein subunit interfaces: characteristics and prediction. *Proteins*, 28, 333-343 (1997).
- Liu W, Crocker E, Siminovich DJ & Smith SO. Role of side-chain conformational entropy in transmembrane helix dimerization of glycophorin A. *Biophysical journal*, 84, 1263-1271 (2003).

References

Livingstone CD & Barton GJ. Protein sequence alignments-a strategy for the hierarchical analysis of residue conservation. *Comput. Appl. Biosci.*, 9, 745-756 (1993).

Livingstone CD & Barton GJ. Identification of functional residues and secondary structure from protein multiple sequence alignment. *Methods Enzymol.*, 266, 497-512 (1996).

Lo Conte L, Chothia C & Janin J. The atomic structure of Protein-protein recognition sites. *J. Mol. Biol.*, 285, 2177-2198 (1999).

Loregian A, Mardsen HS & Palu G. Protein-protein interactions as targets for antiviral chemotherapy. *Rev. Med. Virol.*, 12, 239-262 (2002)

Lu L, Arakaki AK, Lu H, Skolnick J: Multimeric threading-based prediction of protein-protein interactions on a genomic scale: application to the *Saccharomyces cerevisiae* proteome. *Genome Res.*, 13, 1146-1154 (2003).

Lu L, Lu H & Skolnick J. MULTIPROSPECTOR: an algorithm for the prediction of protein-protein interactions by multimeric threading. *Proteins*, 49, 350-364 (2002).

Ma B, Elkayam T, Wolfson H & Nussinov R. Protein-protein interactions: structurally conserved residues distinguish between binding sites and exposed protein surfaces. *Proc. Natl. Acad. Sci. USA*, 100, 5772-5777 (2003).

Madabushi S, Yao H, Marsh M, Kristensen DM, Philippi A, Sowa ME & Lichtarge O. Structural clusters of evolutionary trace residues are statistically significant and common in proteins. *J. Mol. Biol.*, 316, 139-154 (2002).

References

- Marcotte E, Pellegrini M, Ng H-L, Rice DW, Yeates TO & Eisenberg D. Detecting protein function and protein-protein interactions from genome sequences. *Science*, 285, 751-753 (1999).
- Marquart M, Walter J, Deisenhofer J, Bode W & Huber R. The geometry of the reactive site and the peptide groups in trypsin, trypsinogen and its complexes with inhibitors. *Acta Cryst.*, B39, 480-490 (1983).
- Martin S, Roe D & Faulon JL. Predicting protein-protein interactions using signature products. *Bioinformatics*, 21(2), 218-26 (2005).
- Matthews B. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta*, 405, 442-451 (1975).
- Matthews LR, Vaglio P, Reboul J, Ge H, Davis BP, Garrels J, Vincent S & Vidal M. Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or 'interologs'. *Genome Res.*, 11, 2120-2126 (2001)
- McCammon JA. Theory of biomolecular recognition. *Curr. Opin. Struct. Biol.*, 8, 245-249 (1998).
- McCoy AJ, Chandana Epa V & Colman PM. Electrostatic complementarity at protein/protein interfaces. *J. Mol. Biol.*, 268, 570-584 (1997).
- McDonald IK & Thornton JM. Satisfying hydrogen bonding potential in proteins *J. Mol. Biol.*, 238, 777-793 (1994).

References

Meyer M, Wilson P & Schomburg D. Hydrogen bonding and molecular surface shape complementarity as a basis for protein docking. *J. Mol. Biol.*, 264, 199-200 (1996).

Miller S, Lesk AM, Jani J & Chothia C. The accessible surface area and stability of oligomeric proteins. *Nature*, 328, 834-836 (1987).

Mintseris J & Weng ZP. Atomic contact vectors in protein-protein recognition. *Proteins*, 53, 629-639 (2003)

Musafia B, Buchner V & Arad D. Complex salt bridges in proteins: Statistical analysis of structure and function. *J. Mol. Biol.*, 254, 761-770, 1995

Myszka DG. Survey of the 1998 optical biosensor literature. *J. Mol. Recognit.*, 12, 390-408 (1999)

Najmanovich R, Kuttner J, Sobolev V & Edelman M. Side-chain flexibility in proteins upon ligand binding. *Protein*, 39, 261-268 (2000).

Neuvirth H, Raz R & Schreiber G. ProMate: a structure based prediction program to identify the location of protein-protein binding sites. *J. Mol. Biol.*, 338, 181-199 (2004).

Nicholls A, Sharp K & Honig B. Protein folding and association: insights from the interfacial and thermodynamic properties of hydrocarbons. *Proteins*, 11, 282-290 (1991).

Nooren IM & Thornton JM. Diversity of protein-protein Interactions. *EMBO J.*, 22, 3486-3492 (2003a).

References

- Nooren IM & Thornton JM. Structural characterisation and functional significance of transient protein-protein interactions. *J. Mol. Biol.*, 325, 991-1018 (2003b).
- Norel R, Perrey D, Wolfson HJ & Nussinov R. Examination of shape complementarity in docking of unbound proteins. *Proteins*, 36, 307-317 (1999).
- Novotny J & Sharp K. Electrostatic fields in antibodies and antibody/antigen complexes. *Prog. Biophys. Mol. Biol.*, 58, 203-224 (1992).
- Ofran Y & Rost B. Analysing six types of protein-protein interfaces. *J. Mol. Biol.*, 325, 377-387 (2003a).
- Ofran Y & Rost B. Predicted protein-protein interaction sites from local sequence information. *FEBS Lett*, 544, 236-239 (2003b).
- Oyama T, Kitano K, Satou K & Ito T. Extraction of knowledge on protein-protein interaction by association rule discovery. *Bioinformatics*, 18, 705-714 (2002).
- Panchenko AR, Kondrashov F & Bryant S. Prediction of functional sites by analysis of sequence and structure conservation. *Protein Sci.*, 13, 884-892 (2004)
- Pawson T & Nash P. Assembly of cell regulatory systems through protein interaction domains. *Science*, 300, 445-452 (2003).
- Pazos F, Helmer-Citterich M, Ausiello G & Valencia A. Correlated mutations contain information about protein-protein interaction. *J. Mol. Biol.*, 271, 511-523 (1997).
- Pazos F & Valencia A. In silico two-hybrid system for the selection of physically interacting protein pairs. *Proteins*, 47, 219-227 (2002).

References

Pazos F & Valencia A. Similarity of phylogenetic trees as indicator of protein-protein interaction. *Protein Eng.*, 14, 609-614 (2001).

Pupko T, Bell RE, Mayrose I, Glaser F & Ben-Tal N. Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. *Bioinformatics*, 18(suppl 1), S71-S77 (2002).

Radzicka A & Wolfenden R. Comparing the polarities of the amino acids: side-chain distribution coefficients between the vapor phase, cyclohexane, 1-octanol, and neutral aqueous solution. *Biochemistry*, 27, 1664-1670 (1988)

Rajamani D, Thiel S, Vajda S & Camacho CJ. Anchor residues in protein-protein interactions. *Proc. Natl. Acad. Sci. USA*, 101, 11287-11292 (2004).

Recio JF, Totrov M & Abagyan R. Identification of Protein-Protein Interaction Sites from Docking Energy Landscapes. *J. Mol. Biol.*, 335, 843-865 (2004).

Rich RL & Maszka DG. Survey of the 1999 surface plasmon resonance biosensor literature. *J. Mol. Recognit.*, 13, 388-407 (2000).

Richards FM, Eisenberg DS & Kuriyan J. *Advances in protein chemistry*. Academic press, (2003).

Roberts VA, Freeman HC, Olson AJ, Tainer JA & Getzoff ED. Electrostatic orientation of the electron-transfer complex between plastocyanin and cytochrome c. *J. Biol. Chem.*, 266, 13431-13441 (1991).

References

- Rodriguez R, Chinae G, Lopez TP & Vriend G. Homology modeling, model and software evaluation: three related resources. *Bioinformatics*, 14, 523-528 (1998)
- Rost B & Sander C. Prediction of protein secondary structure at better than 70% accuracy. *J. Mol. Biol.*, 232, 584-599 (1993).
- Salwinski L & Eisenberg D. Computational methods of analysis of protein-protein interactions. *Curr. Opin. Struct. Biol.*, 13, 377-382 (2003)
- Samanta MP & Liang S. Predicting protein functions from redundancies in large-scale protein interaction networks. *Proc. Natl. Acad. Sci. USA*, 100, 12579-12583 (2003).
- Schreiber G & Fersht AR. Rapid electrostatically assisted association of proteins. *Nature Struct. Biol.*, 3, 427-431 (1996).
- Schwikowski B, Uetz P & Fields S. A network of protein-protein interactions in yeast. *Nat. Biotechnol.*, 18, 1257-1261 (2000).
- Sheinerman FB & Honig B. On the role of electrostatic interactions in the design of protein-protein interfaces. *J. Mol. Biol.*, 318, 161-177 (2002).
- Sheinerman FB, Norel R & Honig B. Electrostatic aspects of protein-protein interactions. *Curr. Opin. Struct. Biol.*, 10, 153-159 (2000).
- Shulman-Peleg A, Nussinov R & Wolfson HJ. Recognition of Functional sites in proteins structures. *J. Mol. Biol.*, 339, 607-633 (2004).
- Sjolander, K. Phylogenetic inference in protein superfamilies: Analysis of SH2

References

domains. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, 6, 165-174 (1998).

Smith DL, Almo SC, Toney MD & Ringe D. 2.8 Å Resolution Crystal Structure of an Active Site Mutant of Aspartate Aminotransferase from *E. coli*. *Biochemistry*, 28, 8161-8167 (1989).

Sprinzak E & Margalit H. Correlated sequence- signatures as markers of protein-protein interaction. *J. Mol. Biol.*, 311, 681–692 (2001).

Stagljär I, Korostensky C, Johnsson N & Heesen S. A genetic system based on split-ubiquitin for the analysis of interactions between membrane proteins in vivo. *Proc. Natl. Acad. Sci. USA* 95(5), 5187-5192 (1998).

Stanfield RL, Takimoto-Kamimura M, Rini JM, Profy AT & Wilson IA. Major antigen-induced domain rearrangements in an antibody. *Structure*, 1, 83-93 (1993).

Stevens JM, Armstrong RN & Dirr HW. Electrostatic interactions affecting the active site of class sigma glutathione S-transferase. *Biochem. J.*, 347, 193-197 (2000).

Stickle DF, Presta LG, Dill KA & Rose GD. Hydrogen bonding in globular proteins. *J. Mol. Biol.*, 226, 1143-1159 (1992)

Stites WE. Protein-protein interactions: interface structure, binding thermodynamics and mutational analysis. *Chem. Rev.*, 97, 1233-1250 (1997).

Sundberg EJ & Mariuzza RA. Luxury accommodations: the expanding role of structural plasticity in protein–protein interactions. *Structure Fold. Des.* 2000, 8, R137-R142 (2000).

References

Tillier ER & Lui TW. Using multiple interdependency to separate functional from phylogenetic correlations in protein alignments. *Bioinformatics*, 19, 750-755 (2003).

Tsai CJ & Nussinov R. Hydrophobic folding units at protein-protein interfaces: Implications to protein folding and to protein-protein association. *Protein Sci.*, 6, 1426-1437 (1997).

Tsai CJ, Lin SL, Wolfson HJ, Nissinov R. Protein-protein interfaces: architectures and interactions in protein-protein interfaces and in protein cores. Their similarities and differences. *Crit. Rev. Biochem. Mol. Biol.*, 31, 127-152 (1996).

Tsai CJ, Lin, SL, Wolfson HJ, & Nissinov R. Studies of protein-protein interfaces: a statistical analysis of the hydrophobic effect. *Protein Sci.*, 6, 53-64 (1997a).

Tsai CJ, Xu D & Nissinov R. Structural motifs at protein-protein interfaces: protein cores versus two-state and three state model complexes. *Protein Sci.*, 6, 1797-1809 (1997b).

Tsai J, Taylor R, Chothia C & Gerstien M. The packing density in proteins: standard radii and volumes. *J. Mol. Biol.*, 290, 253-266 (1999).

Valdar WS & Thornton JM. Protein-protein interfaces: Analysis of amino acid conservation in homodimers. *Proteins*, 42, 108-124 (2001a).

Valdar WS & Thornton JM. Conservation helps to identify biologically relevant crystal contacts. *J. Mol. Biol.*, 313(2), 399-416 (2001b).

References

Vaughan CK, Buckle AM & Fersht AR. Structural response to mutation at a protein-protein interface. *J. Mol. Biol.*, 286, 1487-1506 (1999).

Vazquez A, Flammini A, Maritan A & Vespignani A. Global protein function prediction in protein-protein interaction networks. *Nat. Biotechnol.*, 21, 697-700 (2003).

Vijayakumar M, Wong KY, Schreiber G, Fersht AR, Szabo A & Zhou HX. Electrostatic enhancement of diffusion-controlled protein-protein association: comparison of theory and experiment on barnase and barstar. *J. Mol. Biol.*, 278, 1015-1024 (1998).

Wells, J. A., Binding in the growth hormone receptor complex. *Proc. Natl. Acad. Sci. USA*, 93, 1-6 (1996).

Westbrook J, Feng ZK, Chen L, Yang HW & Berman HM. The Protein Data Bank and structural genomics. *Nucleic Acids Res.*, 31, 489-491 (2003).

Wilson IA & Stanfield, RL. Antibody-antigen interactions. *Curr. Opin. Struct. Biol.*, 3, 113-118 (1993).

Wilson IA & Stanfield RL. Antibody-antigen interactions: new structures and new conformational changes. *Curr. Opin. Struct. Biol.*, 4, 857-867 (1994).

Wodak SJ & Mendez R. Prediction of protein-protein interactions: the CAPRI experiment, its evaluation and implications. *Curr. Opin. Struct. Biol.*, 14(2), 242-9 (2004).

References

Xu D, Lin SL & Nussinov R. Protein binding *versus* protein folding: the role of hydrophilic bridges in protein associations. *J. Mol. Biol.*, 265, 68-84 (1997a).

Xu D, Tsai, CJ & Nussinov R. Hydrogen bonds and salt bridges across protein-protein interfaces. *Protein Eng.*, 10, 999-1012 (1997b).

Xu W & Regnier FE. Protein-protein interactions on weak-cationexchange sorbent surfaces during chromatographic separations. *J. Chromatogr. A* 828, 357-364 (1998).

Yao H, Kristensen DM, Mihalek I, Sowa ME, Shaw C, Kimmel M, Kavraki L & Lichtarge O. An accurate, sensitive, and scalable method to identify functional sites in protein structures. *J. Mol. Biol.*, 326, 255-261 (2003).

Young L, Jernigan RL & Covell DG. A role for surface hydrophobicity in protein-protein recognition. *Protein Sci.*, 3, 717-729 (1994).

Zhanhua C, Gan JG, Li L, Mathura VS, Sakharkar MK & Kanguane P. Identification of critical heterodimer protein interface parameters by multidimensional scaling in Euclidian space. *Frontiers in Bioscience*, 10, 844 (2005)

Zhou HX & Shan Y. Prediction of protein interaction sites from sequence profile and residue neighbor list. *Proteins*, 44, 336-343 (2001).