Protein dimer interfaces : ligands, sidechains, hot spots and binding

Lei, Li

2006

Lei, L. (2006). Protein dimer interfaces : ligands, sidechains, hot spots and binding.
Master's thesis, Nanyang Technological University, Singapore.

<https://hdl.handle.net/10356/5990>

<https://doi.org/10.32657/10356/5990>

Nanyang Technological University

Downloaded on 26 Apr 2023 10:56:07 SGT

PROTEIN DIMER INTERFACES: LIGANDS, SIDE- CHAINS, HOT SPOTS AND BINDING

LEI LI

School of Mechanical & Aerospace Engineering

A thesis submitted to the Nanyang Technological University
in fulfilment of the requirement for the degree of
Master of Engineering

2006

Acknowledgements

I would like to express my deepest gratitude and appreciation to my supervisor, Professor Jacob Gan for his guidance, encouragement and patient assistance. I wish to acknowledge Professor Pandjassarame Kanguane for his expertise and advice on various aspects of my project.

I also would like to thank Mr. Cui Zhanhua, Mr. Zhao Bin and Ms Yu Yiting for their valuable suggestions.

Last but not least, I would like to thank all the members in my family: my wife Xiangrong Cheng, my dearest parents and my parents-in-law, for their understanding, encouragement, help and support. I sincerely dedicate this thesis to them.

Table of Contents

Acknowledgements	i
Table of Contents	ii
Summary	vi
List of Figures	viii
List of Tables	x
Nomenclatures	xi
Chapter 1 Introduction	1
1.1 Protein-protein interactions	1
1.2 Computational techniques	2
1.2.1 Computational techniques based on atomic structure	4
1.3 Objectives and outline of the research project	5
Chapter 2 Investigating the effect of interface ligands on interface features	7
2.1 Introduction	7
2.2 Materials and methods	9
2.2.1 Dataset creation	9
2.2.2 Determination of LIACV	16
2.3 Results	17
2.3.1 Effect of interface ligands on interface area	17
2.3.2 Effect of interface ligands on interface hydrophobicity	20
2.3.3 Effect of interface ligands on interface residue composition	20

	Summary
2.4 Discussion	21
2.5 Conclusion	25
Chapter 3 Types of inter-atomic interactions across the protein subunit interfaces	27
3.1 Introduction	27
3.2 Methods	28
3.2.1 Dataset Creation	28
3.2.2 Definition of inter-atomic interactions	29
3.2.3 Calculation of inter-atomic interactions	32
3.3 Results & Discussion	32
3.4 Conclusion	36
Chapter 4 Identification of energetically hot spot residues at interface	37
4.1 Introduction	37
4.2 Methods	40
4.2.1 Definition of interface residue	40
4.2.2 Dataset of alanine-mutated interface residues	41
4.2.3 Definition of inter-atomic contacts	41
4.2.4 Classification of inter-atomic contacts	42
4.2.5 Creation of three parameters	45
4.3 Result	58
4.3.1 “NA”	58

	Summary
4.3.2 “NA _I -NA _{II} ”	58
4.3.3 “NC _I -NC _{II} ”	61
4.4 Discussion	61
4.4.1 A ‘hot spot’ prediction approach	61
4.4.2 Comparison with other ‘hot spot’ predication methods	62
4.4.3 Investigation of misidentified hot spots by our method	66
4.5 Conclusion	67
Chapter 5 Distinguishing between specific and non-specific protein-protein interfaces	68
5.1 Introduction	68
5.2 Methods and results	69
5.2.1 Datasets of specific and non-specific interactions	69
5.2.2 Monomer Interface area and monomer surface area	72
5.2.3 Hot-spot interface residues	75
5.3 Discussion	77
5.3.1 Distinguishing between specific and non-specific interfaces	77
5.3.2 Specific interfaces vs non-specific interfaces	77
5.4 Conclusion	83
Chapter 6 Structural difference between 2S (2 state) and 3S (3 state) homodimers	84
6.1 Introduction	84

	Summary
6.2 Methods	86
6.2.1 Dataset creation	86
6.2.2 Analyses of 2S and 3S homodimers	86
6.3 Results	91
6.3.1 Distribution of 2S and 3S on monomer length and subunit interface area	91
6.3.2 Exterior, interior and interface hydrophobicity	94
6.3.3 Factor of interface to interior hydrophobicity (F_{hp}) value	94
6.3.4 Correlation of the three parameters	95
6.4 Discussion	96
6.5 Conclusion	98
Chapter 7 Conclusions	99
List of Publications	100
References	101

Summary

Protein dimer interactions (homodimers and heterodimers) are common in molecular catalysis and regulation. Prediction of interaction sites or interaction partners from sequence is challenging. The significance of interface ligands, sidechain-sidechain prevalence at the interfaces, “hot spot” interface residues (high energy interface residues) and binding are discussed using structural data.

A dataset of 62 identical homodimer pairs (one structure determined with an interface ligand and the other without an interface ligand) suggests that interfaces having ligands are less hydrophobic with small interface area compared to those without ligands and the effect of ligands occupying $\leq 7\%$ interface area is negligible on dimer interactions. In addition, the analysis of homodimers (122 structures), heterodimers (327 structures) and crystal-packing monomers (188 structures) shows preference for sidechain-sidechain interactions. “Hot spot” interface residues are compared with other interface residues using a dataset of 296 alanine-mutated interface residues. It is found that “hot spot” residues have strong positive inter-atomic interactions and a simple effective “hot spot” prediction method is developed which has better accuracy compared with the prevailing ‘hot spot’ prediction methods.

Summary

Some dimers form from unfolded monomers without intermediate (two-state (2S)) and others through the formation of stable intermediate (three-state (3S)). 25 '2S' homodimers and 16 '3S' homodimers is analyzed to estimate their structural differences. The analysis suggests that a combination of the properties such as monomer length, subunit interface area, ratio of interface to interior hydrophobicity can predominately distinguish 2S and 3S dimers.

List of Figures

Figure 1.1	Global network of protein-protein interactions in yeast built from data on two-hybrid screens	3
Figure 1.2	A flowchart describing the project workflow	6
Figure 2.1	Distribution of 1324 homodimers in PDB based on %L	8
Figure 2.2	Flowchart describing the creation of 62 IDPs	10
Figure 2.3	Relationship between ΔB and %L is shown for 62 IDP	18
Figure 2.4	Mean % ΔB and standard deviation for 62 IDP	19
Figure 2.5	Euclidean distance Δf between target and reference dimers for each IDP group	23
Figure 2.6	Difference in interface hydrogen bonds (ΔH) between target and reference structure for each IDP group	26
Figure 3.1	Distribution for the average number of inter-atomic contacts at the homo-, hetero-dimers and crystal-packing monomer interfaces	34
Figure 3.2	Mean and standard deviation distributions for the four types of inter-atomic interactions in homo-, hetero-dimers and crystal-packing monomers	35
Figure 4.1	Flow chart illustrating alanine scanning procedure.	38
Figure 4.2	Illustration of NA , NA_f , NA_{unf} , NC_f and NC_{unf} .	46
Figure 4.3	Distribution for the percentages of the three types of interface residues based on the parameter NA	59

List of figures

Figure 4.4	Percentage distribution of the three residue types in 182 interface residues with NA>0	60
Figure 5.1	Relationship between monomer interface area (MIA) and surface area (MSA)	73
Figure 5.2	Distribution of predicted hot-spot interface residues	76
Figure 5.3	Identification of homodimers and crystal-packing monomers	79
Figure 5.4	Identification of heterodimers and crystal-packing monomers	82
Figure 6.1	Correlation between monomer length (ML) and subunit interface area (IA/2) for three groups of homodimers	93

List of Tables

Table 2.1	Identical homodimer pairs (IDP) of reference and target structures	11
Table 2.2	25 IDPBN set	14
Table 2.3	Non-polar fractions of interface area	22
Table 3.1	Three datasets (homodimers, heterodimers and crystal-packing monomers)	30
Table 4.1	15 complexes with 296 mutations at the interface	43
Table 4.2	Classification of contact between atoms of different classes	44
Table 4.3	Results of analysis on 296 mutated interface residues	47
Table 4.4	Classification of the 296 residues using NA , NA_f , NA_{unf} and NC_f , NC_{unf}	63
Table 4.5	Evaluation of ‘hot spot’ prediction approaches	65
Table 5.1	Datasets representing specific interfaces and non-specific interfaces	70
Table 5.2	Average properties of protein-protein interfaces	74
Table 5.3	Discriminating between homodimers and crystal-packing monomers	80
Table 6.1	Dataset of homodimeric proteins divided into three groups according to their unfolding pathways	88
Table 6.2	Correlation coefficient (R^2) between parameters	95

Nomenclatures

ΔASA	Interface area of a protein-protein complex
ΔG	Change in the standard state free enthalpy of the system on dissociation
$\Delta\Delta G$	Change of ΔG upon mutation
%L	Occupation of ligands to interface area
3D	Three dimension
ASA	Accessible surface area
B	Interface area of protein-protein complexes calculated without ligands
B_L	Interface area of protein-protein complexes calculated with ligands
F_{hp}	Factor of interface to interior hydrophobicity value
H_{ext}	Exterior hydrophobicity
H_{inf}	Interface hydrophobicity
H_{int}	Interior hydrophobicity
HDL^+	Homodimer with interface ligands
HDL^-	Homodimer without interface ligands
IA	Interface area
IDP	Identical homodimer pair
MIA	Monomer interface area
ML	Monomer length (aa)
MSA	Monomer surface area
$N_{\text{hot-spot}}$	Number of hot spot residues

Nomenclatures

NA	Number of atoms in a residue involved in sidechain-sidechain contacts across the interface
NA _f	Number of atoms in a residue involved in energetically favorable sidechain-sidechain contacts across the interface
NA _{unf}	Number of atoms in a residue involved in energetically unfavorable sidechain-sidechain contacts across the interface
NC _f	Number of energetically favorable inter-atomic sidechain-sidechain contacts across the interface
NC _{unf}	Number of energetically unfavorable inter-atomic sidechain-sidechain contacts across the interface
R ²	Coefficient correlation

Chapter 1

Introduction

Proteins are among of the most important constituents of the cells in all living systems. They serve as components for various catalysts, regulators and signal transmitters. Hence, understanding the nature, structure and behavior of proteins is fundamental to the knowledge of biological systems.

1.1 Protein-protein interactions

Protein-protein interaction refers to the process where two or more proteins interact each other physically or chemically. Such interactions may produce other transient compounds. Protein-protein interactions underlie almost all cellular process, including DNA replication, transcription, splicing, translation, cell cycle regulation, signal transduction and cell growth. Therefore, identification and analysis of protein-protein interactions are essential for understanding the molecular mechanisms of cellular processes — a prerequisite for most pharmaceutical applications.

The current analysis on the proteomes (complete set of proteins) of several organisms reveals the importance of protein-protein interactions. Scientists have realized that living organisms are a lot more than the sum of their component proteomes. In living systems, these components provide various functions under different circumstances, and

affect each other in various, complex ways. This leads to **functional network** paradigm, where an organism is described by its complicated interwoven biochemical pathways (Figure 1.1). Therefore, an understanding of protein-protein interactions is fundamental to fill in the gaps in our knowledge of these functional pathways.

1.2 Computational techniques

The traditional experimental methods for detecting protein-protein interactions are associated with many non-trivial issues. For example, most library-based methods have a high error rate (up to 40%) (Fields and Sternglanz 1994; Phizicky and Fields 1995). Additionally, some methods are typically laborious and time-consuming, in some cases, taking up to years to verify and analyze a single protein-protein interaction.

Computational techniques for predicting protein-protein interactions have been showing great promise (Marcotte et al. 1999; Tong et al. 2002). A statistical or machine-learning technique which detects common features of the interactions may enhance the understanding of the interactions. Moreover, such features may be used to predict interactions which have not yet been experimentally observed. This serves as a valuable filtering process for experimentation.

Figure 1.1 Global network of protein-protein interactions in yeast built from data on two-hybrid screens (Jeong et al. 2001). Nodes are colored according to the phenotypic effect of removing the corresponding protein from the organism (red, lethal; green, non-lethal; orange, slow growth; yellow, unknown).

1.2.1 Computational techniques based on atomic structure

One type of computational techniques for protein-protein interaction is based on protein atomic structures. The structures, accessible through the PDB (Protein Data Bank) (Berman et al. 2000), are determined by X-ray crystallography and NMR (Nuclear Magnetic Resonance). Since they provide a wealth of chemical and geometrical data, they underlie the analysis of interface features for protein-protein interactions.

The interface features studied using structural protein-protein complexes are described below.

- ***Interface area.*** It is defined based on the change in solvent accessible surface area (ASA) of proteins when going from a monomeric to a dimeric state. It provides a crude estimate to the extent of inter-monomeric contact. Since proteins carry out their functions by contacting and forming certain interfaces, it is necessary to analyze the information contained in the interface area.
- ***Chemical composition and polar interactions at protein-protein interfaces.*** It is important to understand what chemical composition at the interfaces result in protein functions and how polar interactions at the interfaces influences protein-protein interactions. Many features have been analyzed, such as interface hydrophobicity, interface hydrogen bonds and interface residue composition (Jones and Thornton 1996; Bahadur et al. 2003).

- *Geometric complementarity and atomic packing between two monomers.*

Geometric complementarity of two protein surfaces in contact is a major element of the recognition process between two proteins. A number of parameters have been developed to measure geometric complementarity, such as gap volume and gap index (Jones and Thornton 1996). Geometric complementarity can also be estimated by analyzing atomic packing between two monomers, for complementary surfaces must form compact interfaces with closed-packed atoms and few cavities (Tsai et al. 1999).

Based on these features and binding energy calculation, a number of computational methods (called docking algorithms) have been developed to predict protein-protein binding sites. The docking algorithms have been assessed using CAPRI (a Critical Assessment of PRedicted Interactions) since 2001. Results show that the algorithms are far from satisfactory, although some are promising (Janin et al. 2003). It indicates that the observed interface features contain limited information of specificity for protein-protein interfaces. Thus, it is crucial to determine several other interface features which could efficiently identify specific protein-protein interfaces.

1.3 Objectives and outline of the research project

The objective of my research project is the analysis of protein-protein interface features. The project is divided into five sections (Figure 1.2).

Figure 1.2 A flowchart describing the project workflow is shown. Three-dimensional protein-protein complexes (3D PPC) are collected from PDB. Then datasets of protein complexes are created for the analysis of protein-protein interfaces. The analysis includes five different parts: (1) the effect of interface ligands on the interfaces; (2) the contributions of different types of inter-atomic interactions across the interfaces (i.e. backbone-backbone/sidechain-sidechain interactions) on protein-protein interactions; (3) identification of energetically hot spot residues ($\Delta\Delta G \geq 1.5 \text{ kcal}\cdot\text{mol}^{-1}$) at the interface; (4) Differentiation between specific and non-specific interfaces; and (5) structural difference between 2S (two-state) and 3S (three-state) homodimers.

CHAPTER 2

Investigating the effect of interface ligands on interface features

2.1 Introduction

Protein-protein interfaces, defined as protein surface involved in monomer contacts, have been widely examined to unravel the principles of protein-protein complexation (Jones and Thornton 1995; 1996; 1997; Brinda et al. 2002; Ofran and Rost 2003). Most studies on interface features deal with protein component of the interfaces but disregard ligand components existing at interface (i.e. non-peptide compounds, containing biological and non-biological components). Thus, the effects of interface ligands on interface features have been neglected.

Recently, the effects of interface ligands have been noticed. Nooren and Thornton studied transient protein-protein interactions using the complexes without interface ligands (Nooren and Thornton 2003). Bahadur and colleagues dissected interfaces of homodimers where ligands occupy less than 5% (the cut-off value chosen was chosen without investigation) of interface area (Bahadur et al. 2003). Although these studies take into account the effects of interface ligands on interfaces, a comprehensive study of the effects is not available. As over half of the homodimers in PDB have ligands at interface (Figure 2.1), it is of our interest to standardize the “ligand interface area cut-off value” (LIACV) for discarding complexes with ligands during interface analysis.

Figure 2.1 Distribution of 1324 homodimers in PDB (July 2004) based on %L (see Formula 2.1), the occupation of ligands to interface area. Over half of the homodimers have interface ligands.

2.2 Materials and methods

2.2.1 Dataset creation

We selected 1324 high-resolution ($\leq 2.5 \text{ \AA}$) homodimers (each monomer > 100 residues) using the information in biological unit record from PDB. The dataset was subject to rigorous analysis for the creation of an IDP (identical homodimer pair) dataset using a methodology (Figure 2.2) described below.

Step 1: Identification of homodimers with and without interface ligands.

Interface area was calculated twice for 1324 homodimer structures using NACCESS (an accessible surface area calculation program) with and without ligands (toggle $-H$ option in NACCESS) (Hubbard 1993). The interface area calculated with ligands is designated as B_L ($\text{\AA}^2$) and the interface area calculated without ligands is designated as B ($\text{\AA}^2$). We defined %L, the occupation of ligands to interface area, using the following equation:

$$L (\%) = (B - B_L)/B \quad (2.1)$$

Based on the value of %L, homodimers were clustered into two categories (< 1 and ≥ 1). Homodimers with $\%L < 1$ are considered as those without interface ligands (the cutoff $\%L = 1$ was selected to discard small molecules which are not interface ligands, such as the precipitant SO_4). Homodimers with $\%L \geq 1$ are considered as those with interface ligands. In this dataset (Figure 2.2), we identified 698 homodimers with $\%L < 1$ (designated thereafter as HDL^-) and 626 homodimers with $\%L \geq 1$ (designated thereafter as HDL^+). Thus, two subsets of HDL^- and HDL^+ were created.

Chapter 2 Interface ligands

Figure 2.2 Flowchart describing the creation of 62 IDPs. 698 homodimers without interface ligands are designated as HDL⁻ group and 626 homodimers with interface ligands are designated as HDL⁺ group. IDPBN refers to identical homodimer pairs as background noise. B_r refers to the interface area of the reference dimer (from HDL⁻ group) and B_t refers to that of the target dimer (from HDL⁺ group). ΔB is the difference of interface areas between reference and target dimers of each IDP (see formula 2.2).

Chapter 2 Interface ligands

Table 2.1 Identical homodimer pairs (IDP) of reference and target structures

Protein name	Reference state		Target state				
	PDB ID	B(Å ²)	H-bonds	Ligand names			
Group 1: 1-2% interface area occupied by ligands (8)							
Autocrine Motility Factor [%]	1IRCD	12884	63	1IRIAB	12952	55	erythrose-4-phosphate
Transthyretin [#]	1F41	1782	16	1E4H	1749	16	glycerol*
Recombination Endonuclease VII	1EN7	6604	33	1E7D	6845	32	zinc ion
Neopullulanase	1J01	5552	24	1J0K	5256	27	isopanose
Triosephosphate Isomerase [%]	1R2RAB	3414	28	1R2RCD	3411	28	2-amino-2-hydroxymethyl-propane-1,3-diol
Fructose 1,6-Bisphosphatase/Inositol Monophosphate	1LBV	2387	7	1LBX	2401	5	d-myo-inositol-1-phosphate
Glucosamine-Phosphate N-Acetyltransferase	1I12BD	5734	36	1I1DBC	5807	38	n-acetyl-d-glucosamine-6-phosphate
Siroheme Synthase	1PJT	12307	46	1PJQ	12337	46	phosphoserine
Group 2: 2-3% interface area occupied by ligands (14)							
Recombination Endonuclease VII	1EN7	6604	33	1E7L	6702	34	zinc ion & sulfide ion*
Triosephosphate Isomerase	1TCD	2967	19	1C11	3036	20	hexane
ADP-Ribose Pyrophosphatase	1G0S	7678	33	1KHZ	7147	27	α-β methylene adp-ribose
Transthyretin (tetramer)	1F41	1782	16	1BM7	1748	16	2-[[3-(trifluoromethyl)phenyl]amino] benzoic acid
L-3-Hydroxyacyl-Coa Dehydrogenase	1F14	3223	9	1F12	3266	5	3-hydroxybutanoyl-coenzyme a
Neopullulanase	1J01	5552	24	1J0J	5235	25	maltotetraose
Quinolinate Acid Phosphoribosyl Transferase	1QPOAB	4843	14	1QPRAB	5146	18	phthalic acid
D-Amino Acid Aminotransferase	2DAB	4617	18	1G2W	4553	18	acetate ion
Hiv-1 Integrase	1BIS	2993	8	1EXQ	3090	13	sulfate ion*
Cyanovirin-N	1L5E	6543	42	1L5B	6306	36	2-[n-cyclohexylamino]ethane sulfonic acid* & sodium ion*
Beta-Lactamase Oxa-10	1E3UBD	2553	11	1E4DAC	2636	11	lysine nz-carboxylic acid & sulfate ion*
Siroheme Synthase	1PJT	12307	46	1PJS	12414	50	nicotinamide-adenine-dinucleotide
Orotidine Monophosphate Decarboxylase	1LP6	1595	8	1LOL	1570	7	xanthosine-5'-monophosphate
Branched-Chain Amino Acid Aminotransferase	1EKV	4679	16	1KT8	4720	19	L-isoleucine
Group 3: 3-4% interface area occupied by ligands (8)							

Chapter 2 Interface ligands

Glutathione S-Transferase	1EOG	2369	10	1PX6	2417	11	glutathione & 2-(n-morpholino)-ethanesulfonic acid
Fructose 1,6-Bisphosphatase/Inositol Monophosphate	1LBV	2387	7	1LBY	2302	4	fructose-6-phosphate
Quinolate Acid Phosphoribosyl Transferase	1QPOAB	4843	14	1QPREF	5138	19	phthalic Acid
Glucosamine-Phosphate N-Acetyltransferase	1I12BD	5734	36	1I1DAD	5843	35	n-acetyl-d-glucosamine-6-phosphate
Alcohol Dehydrogenase E Chain	1HEU	3306	11	1QV6	3397	13	nicotinamide-adenine-dinucleotide
Phosphotriesterase	1EZ2	2741	25	1I0D	2599	26	ethylene glycol*
Probable Fosfomycin Resistance Protein	1NNR	5177	32	1LQO	5211	36	thallium ion
Nitric Oxide Synthase	1ED6	5595	26	1DMJ	5610	28	5,6-Cyclic-Tetrahydropteridine
<i>Group 4: 4-5% interface area occupied by ligands (9)</i>							
ADP-Ribose Pyrophosphatase	1G0S	7678	33	1G9Q	7600	32	ADP- Ribose
Beta-Lactamase Oxa-10	1E3UBD	2553	11	1E4DBD	2650	12	sulfate ion*
Fructose 1,6-Bisphosphatase/Inositol Monophosphate	1LBV	2387	7	1LBZ	2258	6	fructose-1,6 bisphosphate
Beta-Ketoacyl Synthase	1DD8AB	5784	22	1EK4AB	5788	27	lauric acid
Ribonuclease, Seminal	1R5D	3784	24	1R5C	3827	26	2'-deoxycytidylyl(3',5')-2'-deoxyadenosine
Cyanovirin-N	1L5E	6543	42	1M5J	6341	39	A-d-mannose
Orotidine Monophosphate Decarboxylase	1LP6	5595	8	1LOSAB	5575	10	6-azauridine 5'-monophosphate
Alcohol Dehydrogenase E Chain	1HEU	3306	11	1HSO	3295	14	nicotinamide-adenine-dinucleotide
Nitric Oxide Synthase	1ED6	3161	26	1D1X	3061	26	(6r,1'r,2's)-5,6,7,8 tetrahydrobiopterin
<i>Group 5: 5-6% interface area occupied by ligands (8)</i>							
Beta-Lactamase Oxa-10 ^o	1E3UBD	2553	11	1E3UAC	2597	10	1,2-ethanediol*
Fructose 1,6-Bisphosphatase/Inositol Monophosphate	1LBV	2387	7	1G0I	2290	7	1,2,3,4,5,6-hexahydroxy-cyclohexane
L-3-Hydroxyacyl-CoA Dehydrogenase	1F14	3223	9	1F0Y	3235	5	acetoacetyl-coenzyme a
Cyanovirin-N	1L5E	6543	42	1M5M	6315	36	α -d-mannose & 2-[n-cyclohexylamino]ethane sulfonic acid
Transthyretin	1F41	1782	16	1I1K	1710	16	B-mercaptoethanol*
Glutathione S-Transferase A1	1K3O	2860	3	1K3L	2956	5	s-hexylglutathione
Phosphotriesterase	1EZ2	2741	25	1I0B	2595	27	ethylene glycol & sodium ion
Aspartate 1-Decarboxylase	1PPY	1790	13	1PQF	1739	9	s-oxy cysteine
<i>Group 6: 6-7% interface area occupied by ligands (7)</i>							
Thymidylate Synthase ^o	1F28CD	4104	22	1F28AB	4113	22	2'-deoxyuridine 5'-monophosphate

Chapter 2 Interface ligands

Transthyretin	1F41	1782	16	1111	1713	15	B-mercaptoethanol*
Undecaprenyl Pyrophosphate Synthetase	1V7U	3531	16	1UEH	3529	14	magnesium ion
Glutathione S-Transferase A1	1K3O	2860	3	1ML6	2910	6	mgsta2-2 Toward benzo[A]Pyrene-(7R,8S)-diol-(9S,10R)-epoxide
Glucosamine-Phosphate N-Acetyltransferase	1112AC	5569	33	1121AB	5240	33	selenomethionine*
Protein Maf	1EX2	1078	11	1EXC	1048	9	deoxyuridine-5'-triphosphate
Dtdp-4-Dehydrothamnose 3,5-Epimerase	1DZR	2773	24	1DZT	2797	28	3'-o-acetylthymidine-5'-diphosphate & glycerol*
<i>Group 7: 7-8% interface area occupied by ligands (3)</i>							
Transthyretin	1F41	1782	16	111N	1712	15	B-mercaptoethanol*
Glutathione S-Transferase A1	1K3O	2860	3	1K3Y	2845	5	s-hexylglutathione
Glucosamine-Phosphate N-Acetyltransferase	1112BD	5734	36	1121AB	5240	33	selenomethionine*
<i>Group 8: 8-9% interface area occupied by ligands (3)</i>							
Capsid Protein P40: Assemblin Protease	1IED	3059	12	1WPO	2729	9	selenomethionine*
Aspartate 1-Decarboxylase	1PPY	1790	13	1PT1	1905	14	sulfate anion
Glucosamine-Phosphate N-Acetyltransferase	1112BD	5734	36	1121XY	5335	33	selenomethionine*
<i>Group 9: 9-10% interface area occupied by ligands (2)</i>							
Cytosine Deaminase	1P6O	2944	20	1OX7	2480	18	selenomethionine*
Aspartate 1-Decarboxylase	1PPY	1790	13	1PQH	1810	11	malonic acid & s-oxy cysteine

B = interface area. * indicates non-biological ligand.

Chapter 2 Interface ligands

Table 2.2 25 IDPBN set

PDB entry	Protein	B(Å²)	H-bonds
1CZ3 / 1D1G	Dihydrofolate Reductase	3011 / 3127	26/26
1D8IBC / 1D8HAB	Mrna Triphosphatase Cet1	4089 / 4307	24/27
1DAP / 3DAP	Diaminopimelic Acid Dehydrogenase	5267 / 5319	22/24
1DD8AB / 1DD8CD	β-Ketoacyl [Acyl Carrier Protein] Synthase I	5784 / 5839	16/16
1E3F / 1F4I	Transthyretin	1709 / 1782	23/23
1EBH / 2ONE	Enolase	3575 / 3703	12/12
1I0S / 1I0R	Conserved Hypothetical Protein	4373 / 4376	12/13
1IED / 1IEG	Capsid Protein P40: Assemblin Protease	3059 / 3067	63/69
1IRICD / 1JIQCD	Autocrine Motility Factor	12884 / 12949	7/7
1LBW / 1LBV	Fructose 1,6-Bisphosphatase/Inositol Monophosphatase	2323 / 2387	6/6
1MI3AB / 1K8CAB	Xylose Reductase	2407 / 2419	38/32
1MVP / 2RSP	Rous Sarcoma Virus Protease	2915 / 2982	10/10
1OBU / 1OBQ	Crustacyanin C1 Subunit	2209 / 2241	48/50
1P5R / 1P5H	Formyl-Coenzyme A Transferase	12214 / 12396	5/4
1Q1E / 1Q1BAB	Maltose/Maltodextrin Transport ATP-Binding Protein Malk	2161 / 2290	14/15
1QRK / 1GGY	Coagulation Factor Xiii	4246 / 4423	22/26
1R2SAB / 1R2T	Triosephosphate Isomerase	3404 / 3435	15/13
1P18 / 1P19AB	Hypoxanthine Phosphoribosyltransferase	3300 / 3327	31/30
1SES / 1SET	Seryl-tRNA Synthetase	4466 / 4536	12/9
1SLCAB / 1SLBAB	Galectin-1 (S-Lectin)	1059 / 1111	12/9
1SMN / 1QL0	Nuclease	1735 / 1782	6/6
1UQ5 / 1UQ4	Ricin	1828 / 1892	49/49
1URA / 1ALK	Alkaline Phosphatase	7565 / 7665	35/40
1V26 / 1V25	Long-Chain-Fatty-Acid-CoA Synthetase	5726 / 5728	44/41
2NAD / 2NAC	Nad-Dependent Formate Dehydrogenase	7590 / 7599	15/19

The interface areas for IDPBN pairs are separated by a forward slash.

Step 2: Identification of IDPs

The 1324 homodimer sequences were clustered using HSSP (homology derived secondary structures of proteins) at ≥ 70 cut-off value (Mika and Rost 2003). This procedure created 585 clusters (each cluster containing highly homologous sequences), where 321 clusters contained at least two sequences each. The entries in each of these 321 clusters were compared with HDL⁺ and HDL⁻ group members using PDB ID. Subsequently, the clusters were further grouped into three categories, namely (1) category I (154 clusters consisting of at least one member from each HDL⁺ and HDL⁻), (2) category II (65 clusters consisting of all members from HDL⁻) and (3) category III (102 clusters consisting of all members from HDL⁺). We used category I members (154 clusters) and created 483 identical pairs in all combinations, such that each pair consists of one member of HDL⁺ and HDL⁻ from the same cluster. We further divided these pairs into 9 groups based on the value of L ranging from 1 to 10 with a step size of 1 as shown in Figure 2.2. In this grouping, dimers with %L>10 in HDL⁺ group do not have an identical dimer pair in HDL⁻ group and are therefore discarded. Then the redundant pairs in each group were removed with an HSSP cutoff value of ≤ 5 . We thus created 69 IDPs consisting of one member from HDL⁺ and the other from HDL⁻ (Figure 2.2). The HDL⁺ member in each IDP is thereafter called the target dimer, and the HDL⁻ member in each IDP is thereafter called the reference dimer. Subsequently, we calculated the difference of interface areas (% ΔB) between reference and target dimer of each IDP using the equation:

$$\Delta B (\%) = (B_r - B_t)/B_r \quad (2.2)$$

Chapter 2 Interface ligands

where B_r is the interface area for reference dimer and B_t is that of the target dimer. It is found that most IDP pairs have ΔB values in the range of -6 to +16%. However, 7 pairs have ΔB values in the range of -18% to -200% and these target dimers have significantly larger interface area than reference dimers (due to conformational flexibility, such as orotidine 5'-phosphate decarboxylase (reference 1DQW vs. target 1DQX) (Miller et al. 2000)). Hence, these 7 pairs were eliminated, resulting in 62 IDP for further analysis (Table 2.1).

2.2.2 Determination of LIACV

Interface features could be affected by factors (e.g. surface ligands and conformational flexibility of interacting surfaces) other than interface ligands. The effect of these factors was considered as background noise in the study. We used category II members consisting of 65 HDL⁻ clusters to define IDPBN (IDP for background noise) structures (Figure 2.2). The redundant pairs were removed at a HSSP cutoff value of ≤ 5 . Thus, we created a non-redundant dataset consisting of 25 IDPBN (Table 2.2). The selected pairs consist of one dimer with the smallest interface area and the other with the largest interface area from each cluster.

The effects of interface ligands on interface features were compared with the other factors to determine the LIACV. If the effects of interface ligands are not significantly different from those of the other factors (according to p value=0.01 for one-tail t-test),

Chapter 2 Interface ligands

the effects of interface ligands could be negligible in interface analysis. In the IDP groups having negligible interface ligands, the largest value %L determines the LIACV.

2.3 Results

2.3.1 Effect of interface ligands on interface area

The effect of interface ligands on interface area was studied by measuring %L and ΔB in 62 IDPs (Figure 2.3). Figure 2.3 shows that ΔB range from -6% to 16%. The interface area of reference dimers is predominantly larger than target dimers and their differences (ΔB) increase with ligand occupation (%L) to interface area with a weak correlation co-efficient ($R = 0.36$). The mean value and standard deviation (SD) of ΔB (Figure 2.4AB) are shown for each of the 9 IDP groups clustered based on the value of %L. Figure 2.4A shows that mean ΔB increases with %L and it is significantly large for IDP groups 7-9. The mean for groups 1-6 is less than the mean for IDPBN. SD about mean ΔB is consistently small for IDP groups 1-7 and SD is high for groups 8 and 9 (Figure 2.4B). These SD values are generally larger than the SD in IDPBN. It was found that the means for IDP group 1-6 are not significantly different from the mean for IDPBN set, according to p value = 0.01 for one-tail t-test.

Figure 2.3 The relationship between ΔB and %L is shown for 62 IDP. The correlation co-efficient r is 0.36.

Figure 2.4 A: The mean % ΔB (difference between target and reference structures) for each IDP group is shown. The dash line through 2.3 is a measure of background noise. B: The standard deviation (SD) about the mean ΔB of each IDP group is given. The SD value for IDPBN set is 1.9 (dash line).

2.3.2 Effect of interface ligands on interface hydrophobicity

It was our interest to study the effect of interface ligands on interface hydrophobicity. This was done by calculating the properties of interface chemical groups. These groups were categorized as non-polar (carbon atoms) and polar (nitrogen, oxygen and sulphur atoms) groups. Table 2.3 shows the non-polar fractions of interface area in target and reference dimers for each IDP group. On average, the non-polar fraction is greater in reference structures than target structures. This suggests that hydrophobic interface in target structures are disturbed by interface ligands. The differences between them are 0.7 in groups 1-3, 0.1 in groups 4-6 and above 1.7 in groups 7-9. These differences are generally greater than the differences (0.3) in IDPBN structures. These observations indicate that the dimer interface with ligands is less hydrophobic than that without interface ligands and the distinction increases with ligands occupation. In addition, the mean and SD of the differences of non-polar fraction in IDP groups and in IDPBN set were calculated. It was found that the means for IDP group 1-6 are not significantly different from the mean for IDPBN set, according to p value = 0.01 for one-tail t-test.

2.3.3 Effect of interface ligands on interface residue composition

The effect of interface ligands was studied on interface residue composition. The similarity between the residue composition of target/reference interfaces in each IDP group was quantified by calculating their Euclidean distance Δf (Bahadur et al. 2004):

$$(\Delta f)^2 = 1/19 \sum_{i=1}^{20} (f_i - f'_i)^2 \quad (2.3)$$

Chapter 2 Interface ligands

where f_i and f'_i are the percent contribution of residue type i to the interface area of target and reference dimers, respectively. Figure 2.5 shows that Δf is < 0.2 for groups 1-4, 0.45 for groups 5-6 and > 0.9 for groups 7-9. This suggests that Δf increases with %L. It also shows that the Δf for groups 4-9 are greater than background noise in IDPBN. The mean and SD of the Δf s in IDP groups and in IDPBN set are calculated. It was found that the means for IDP group 1-6 are not significantly different from the mean for IDPBN set, according to p value = 0.01 for one-tail t-test.

2.4 Discussion

Protein homodimers are structural assemblies formed by identical monomers to create active centers for biological reactions. Therefore, it was of great interest to understand the principles of homodimer association. These associations are generally governed by weak interactions. Several efforts have been made to study homodimer interaction using structural data and many structural features are found important (Jones and Thornton 1995; 1996; 1997; Brinda et al. 2002; Nooren and Thornton 2003; Ofra and Rost 2003). PDB also contains many homodimers containing small molecule ligands at the subunit interface (Figure 2.1). The ligands contribute to subunit association by both positive and negative effect. However, the effects of interface ligands on interface features have generally been neglected in interface analysis. Thus, it is our interest to analyze the effect of interface ligands on interface features and standardize the cut-off value of %L, called LIACV, to neglect the ligand effects.

TABLE 2.3 Non-polar fractions of interface area

Group Name	Reference dimers (%)	Target dimers (%)	Difference (%) ^a
IDP group 1	64.6	64.0	0.6
IDP group 2	66.3	65.5	0.8
IDP group 3	65.2	64.6	0.6
IDP group 4	64.9	65.1	-0.2
IDP group 5	64.9	64.8	0.1
IDP group 6	62.0	61.7	0.3
IDP group 7	64.4	62.7	1.7
IDP group 8	65.1	61.5	3.6
IDP group 9	57.5	54.9	2.6
IDPBL set ^b	65.3	65.0	0.3

Fraction of the interface area contributed by non-polar (carbon-containing) class; polar (N, O, and S containing) class contribute the complement to 1.

^a Difference of the non-polar fractions between reference and target dimers in each group.

^b For the IDPBL set, the fraction in the “reference dimer” column is calculated from dimers with large interface area and the fraction in the “target dimers” column is calculated from dimers with small interface area.

Figure 2.5 The Euclidean distance Δf between target and reference dimers for each IDP group is shown. The dash line through 0.2 is a measure of background noise.

Chapter 2 Interface ligands

We created a dataset of 62 identical homodimer pairs, designated as IDP (Table 2.1). Each pair consists of a homodimer with interface ligands and a homodimer without interface ligands. We also created a dataset of homodimer pairs, designated as IDPBN, to determine background noise (Table 2.2). Each pair in this dataset consist of identical homodimers with varying interface conformational flexibility, such that one homodimer has large interface area and other has small interface area. It should also be noted that homodimer pairs in IDPBN do not contain interface ligands, yet exhibit flexible interface conformation. These datasets were created from an initial set of 1324 homodimers using the methodology described in the methods section (Figure 2.1).

These two datasets were independently studied using three parameters: (1) interface area, (2) interface hydrophobicity, and (3) interface residue composition. These interface properties were also studied for increasing values of %L, a measure of ligand occupation at interface area. The mean change in interface area between homodimers with (target dimers) and without (reference dimers) interface ligands (ΔB) is less than the background noise when %L is less than 7% (Figure 2.4A). The SD about the mean difference is also small when %L is less than 7% (Figure 2.4B). It should be noted that the SD about the mean is also large with a larger mean ΔB when %L is $\geq 7\%$. In addition, most IDP groups show a change in interface non-polar fraction larger than the background noise of 0.2 (Table 2.3). The Euclidean distance (Δf) interface residue composition between target and reference structures is also found to be larger than background noise when %L is $\geq 4\%$ (Figure 2.5). Moreover, the effect of interface

Chapter 2 Interface ligands

ligands on hydrogen bonds were analyzed (Figure 2.6). It was found that the difference of hydrogen bond between target and reference structures is not related to the increase in %L.

The effect of interface ligands on interfaces was compared with that of the other factors to determine the LIACV, according to p value=0.01 for one-tail t-test. For the three parameters (interface area, interface hydrophobicity and interface residue composition), the effects of interface ligands in IDP group 1-6 are not significantly different from the effects of the other factors. Therefore, the effect of interface ligand on interface features is negligible when %L is <7.

2.5 Conclusion

The challenge in understanding the principles of protein dimer associations are multifaceted. Dimer associations are either homo/hetero dimer in nature. This association is critical in the formation of active reaction centers, which are important for substrate binding. Therefore, it is essential to study disturbances caused by substrate binding to dimer interfaces. Here, we created a dataset of 62 identical homodimer pairs consisting of one dimer with a bound interface ligand and the other without any interface ligand. Analysis of this dataset suggests that ligand effect on homodimer interface can be neglected if ligands occupy <7% of interface area. It should be noted that this observation is based on a dataset of 62 identical homodimer pairs.

Figure 2.6 The difference in interface hydrogen bonds (ΔH) between target and reference structure is shown for each IDP group. The hydrogen bonding between subunits was calculated using HBPLUS (McDonald and Thornton 1994). The dash line represents ΔH value for IDPBN (0.4). Data suggest no significant correlation between ΔH and %L.

Chapter 3

Types of inter-atomic interactions across the protein subunit interfaces

3.1 Introduction

Proteins interact each other through the recognition of specific protein surfaces, which form protein-protein interface. Thus, elucidating specific interfaces is necessary to understand protein-protein interactions. Many studies investigated residue-residue interactions across the interface (Glaser et al. 2001; Ofra and Rost 2003). It is found that residue pair preferences at the interface are statistically different for homodimers and heterodimers (Ofra and Rost 2003). Also, the preferences could suppress false-positive prediction in protein docking methods (a computational method for predicting protein-protein binding site) (Moont et al. 1999). Residue-based criteria reflect interface nature less accurately than atom-based analysis. However, the atom-based criteria may be not appropriate, as protein-protein binding involves adjustment of the rotameric state of the residues. Thus, a combination of residue-based analysis and atom-based analysis can be more useful.

The residue-residue interactions across the interface can be divided into four types of atom-atom interactions: (1) B₁B₂I, (2) B₁S₂I, (3) S₁B₂I and (4) S₁S₂I, where suffix 1 refers to one monomer and suffix 2 refers to the other monomer, “B” means backbone, “S” means sidechain and “I” means interaction. Jackson analyzed protease-inhibitor and

Chapter 3 Inter-atomic interactions

antigen-antibody complexes and observed that S_1S_2I contributes predominantly to the interaction energy of protease-inhibitors, while B_1B_2I plays a dominant role in antibody-antigen binding (Jackson 1999). Reichmann and coworkers investigated interfaces based on the four types with varying weights (S_1S_2I with weight 2, B_1S_2I and S_1B_2I with 1, and B_1B_2I with 0), as these types show different interaction specificity (Reichmann et al. 2005). However, a comprehensive study on the four types is not available for protein-protein interaction.

It is of our interest to study the prevalence of these inter-atomic types across the interfaces of homodimers, heterodimers and crystal-packing monomers using 3D structures. The three different datasets represent different kinds of protein-protein interfaces. Homodimers are composed of two identical monomers, while heterodimers are composed of two different monomers. Crystal-packing monomers are crystal artifacts which are formed by two monomers in crystallized conditions and thus their interactions are not biologically relevant. Here, we show the distribution of B_1B_2I , B_1S_2I , S_1B_2I and S_1S_2I to the interaction of homodimers, heterodimers and crystal-packing monomers.

3.2 Methods

3.2.1 Dataset Creation

We used three datasets: (1) the homodimer dataset, consisting of 122 complexes derived from (Bahadur et al. 2003); (2) the heterodimer dataset, containing 327 complexes

Chapter 3 Inter-atomic interactions

derived from (Mintseris and Weng 2003); and (3) the dataset of crystal-packing monomers, containing 188 structures with buried surface area $> 800\text{\AA}$ derived from (Bahadur et al. 2004). The crystal-packing monomers are undistinguishable from real dimers on the basis of interface size. The three datasets are given in Table 3.1. The homodimer and heterodimer datasets do not contain interface ligands with $\%L > 7$, where $\%L$ is the percentage of interface area occupied by ligands (See chapter 2).

3.2.2 Definition of inter-atomic interactions

Inter-monomeric residue-residue interactions were classified into four types of atom-atom interactions: (1) B_1B_2I , (2) B_1S_2I , (3) S_1B_2I and (4) S_1S_2I , where suffix 1 and 2 refers to different monomers, “B” means backbone, “S” means sidechain and “I” means interaction. For homodimers, two monomers were randomly designated as monomer 1 and monomer 2, since they are two-fold symmetry. For heterodimers, large monomer (e.g. enzymes and antibodies) was designated as monomer 1 and small monomer (e.g. inhibitors and antigens) was designated as monomer 2. For crystal-packing monomers, the monomer which buries large surface area was designated as monomer 1 and the monomer burying small surface area was designated as monomer 2. Monomer buried surface area was defined as the change in solvent accessible surface area (ASA) of the monomer when going from a monomeric to a dimeric state, calculated using the program NACCESS (Hubbard 1993).

Chapter 3 Inter-atomic interactions

Table 3.1. Three datasets (homodimers, heterodimers and crystal-packing monomers)

PDB entries for homodimers (122) ¹									
12as	1a3c	1a4i	1a4u	1aa7	1ad3	1ade	1af5	1afw	1ajs
1alo	1amk	1aor	1aq6	1auo	1b3a	1b5e	1b67	1b8a	1b8j
1bam	1bbh	1bd0	1bif	1biq	1bis	1bjw	1bkp	1bmd	1brw
1bsl	1bsr	1buo	1bxg	1bxk	1cdc	1cg2	1chm	1cmb	1cnz
1coz	1csh	1ctt	1cvu	1czj	1daa	1dor	1dpg	1dqs	1dxg
1e98	1ebh	1fl3	1fip	1fro	1gvp	1hhp	1hjr	1hss	1hxp
1icw	1imb	1isa	1ivy	1jhg	1jsg	1kba	1kpf	1lyn	1m6p
1mkb	1mor	1nox	1nse	1nsy	1oac	1opy	1pgt	1pre	1qfh
1qhi	1qr2	1r2f	1reg	1rfb	1rpo	1ses	1slt	1smn	1smt
1sox	1tc1	1tox	1trk	1uby	1utg	1vfr	1vok	1wtl	1xso
2arc	2ccy	2hdh	2ilk	2lig	2mcg	2nac	2ohx	2spc	2sqc
2tct	2tgi	3dap	3grs	3sdh	3ssi	4cha	4kbp	5csm	5rub
8prk	9wga								
PDB entries for heterodimers (327) ²									
1e50	1e96	1a14	1a2k	1a6d	1acb	1agr	1ahw	1ak4	1akj
1ao7	1ar1	1aro	1atn	1aui	1ava	1avg	1avx	1avz	1awc
1ay7	1azz	1b4u	1b6c	1b7y	1b8m	1b9y	1bdj	1be3	1bgx
1bi8	1bj1	1bkd	1bml	1bqh	1buh	1buv	1bvn	1bzq	1c1y
1c3o	1c4z	1ccw	1cgi	1clv	1cmx	1cpc	1cs4	1cxz	1d2z
1d4x	1d5x	1dce	1de4	1dee	1dev	1df9	1dfj	1dhk	1dii
1dj7	1dkf	1dkg	1dm0	1dn1	1doa	1dpj	1dtd	1dtw	1du3
1dx5	1dxt	1e6e	1e6j	1e6v	1e8o	1e9z	1eai	1eay	1ebd
1ebp	1eer	1eex_1	1eex_2	1efu	1efv	1efx	1eg9	1eja	1ep3
1es7	1euv	1evt	1exb	1ezv_1	1ezv_2	1ezv_3	1ezx	1f02	1f34
1f3u	1f3v	1f51	1f60	1f80	1f83	1f93	1fak	1fbi	1fbv
1fc2	1fcd	1ffu	1ffv	1fg9	1fin	1fle	1flt	1fm0	1fns
1fq1	1fqj	1fqv	1fs0	1fsk	1fxw	1g0y	1g4y	1g73	1g8k
1gaq	1gc1	1gcq	1gh6	1ghq	1gka	1gl1	1gl4	1gla	1go3
1go4	1gp2	1gpw	1grn	1gvn	1gxd	1gzs	1h2k	1h2r	1h2v
1h32	1h4i	1h59	1h8e	1hcn	1he1	1hez	1hfe	1hr6	1hsa
1hwg	1hx1	1hxm	1hzz	1i1a	1i2m	1i3o	1i4d	1i4e	1i7w
1i85	1i9r	1ib1	1ibr	1icf	1ihf	1iis	1ijk	1im3	1iod
1iqd	1ir1	1is8	1itb	1jb0_1	1jb0_2	1jb0_3	1jb0_4	1jb0_5	1jb7
1jch	1jiw	1jk0	1jk8	1jkj	1jma	1jmx	1jnz	1jnr	1jro
1jsu	1jtd	1jtg	1jv2	1jw9	1jwh	1k28	1k3u	1k3z	1k4c
1k5d_1	1k5d_2	1k8k_1	1k8k_2	1k8k_3	1k8k_4	1k8k_5	1k8k_6	1k90	1kac
1keg	1kfu	1kgy	1ki1	1kkl	1kmi	1kqf_1	1kqf_2	1ktd	1kxp
1kxt	1kyo	1kzy	1l0o	1l6x	1l7v	1l9j	1lb1	1ld8	1ldj

Chapter 3 Inter-atomic interactions

1li1	1lk3	1lpb	1lti	1luc	1m10	1m1e	1m2o	1m2v	1m4u
1mah	1mbu	1mjg	1ml0	1mr1	1mro_1	1mro_2	1mro_3	1n2c	1n98
1nbf	1nbw	1nf5	1noc	1nsn	1o6s	1o94	1osp	1poi	1pre
1qav	1qdl	1qfu	1qfw	1qgw	1qkz	1qla	1qlb	1qo0	1raf
1req	1rlb	1sbb	1sgf	1smf	1spp	1stf	1t7p	1tab	1tbq
1tmq	1toc	1ugh	1vcb	1vkx	1wej	1wq1	1www	1xd1	1yca
1ytf	1zbd	2ahj	2btc	2btf	2hmi	2jel	2kau_1	2kau_2	2min
2mta_1	2mta_2	2mta_3	2pcc	2prg	2ptc	2sic	2tec	3gtu	3pce
3sgb	3ygs	4cpa	4htc	4rub	4sgb	7cei			
PDB entries for crystal-packing monomers (188) ³									
13pk	1a7v_1	1a7v_2	1ad5	1afk	1ag9	1ah7_1	1ako_1	1ako_2	1amu_1
1atl	1aw7	1ayl	1blj_1	1blj_2	1b3j	1bc2	1bea	1bin	1bkz
1bs2_1	1byo	1c02	1caq	1ck7_1	1ck7_2	1cki_1	1clu	1cqx	1dsu
1dys	1e0s_1	1e0s_2	1ehy	1epa	1ewf_1	1feh	1fgk	1fjm	1fkf
1fmt_1	1fmt_2	1g2a	1gar	1gjm	1hf8	1hfv_1	1ilr_1	1kpt	1kwa_1
1mpg_1	1mpg_2	1mss	1naw_1	1naw_2	1np4	1pbg_1	1pbg_2	1pda_1	1ppo_1
1ppo_2	1ppo_3	1qaz	1qci	1qdm	1qha	1qip	1qme_1	1qpa	1qtq_1
1qtq_2	1rb3	1rhs_1	1rne	1shk	1the_1	1the_2	1tht	1toa	1ton
1urp_1	1urp_2	1vbt	1xgs	256b	256l	2acy	2atj_1	2bc2	2bls_1
2erc	2g3p	2ihl	2mbr	2scp	2shp	2tps	2ugi	3mht	3pmg
5tss	830c_1	8pti	2rn2	3cms	3ng1_1	3ng1_2	830c_2	2end	2nap
1a12_1	1a12_2	1a39	1a3y_1	1a3y_2	1a6q	1a7t	1ae9	1af7	1ah7_2
1ajk	1akz	1amu_2	1aq0	1aqz	1ayi	1b24	1b80	1b8x	1be0
1bf6	1bg0	1bgc	1bs2_2	1cfy	1cki_2	1dxm	1dz4_1	1dz4_2	1elp
1ewf_2	1ewf_3	1fmt_3	1fvk_1	1fvk_2	1hfv_2	1hfv_3	1hrn_1	1hrn_2	1hsl
1i4g_1	1i4g_2	1ihb	1ilr_2	1inp_1	1inp_2	1jfr	1kfs_1	1kfs_2	1kwa_2
1mbl	1ml1	1mwc	1nmt	1nuc	1ome	1ovw	1pda_2	1pmi	1pva
1qme_2	1qnt_1	1qnt_2	1qs8	1qsn	1rge	1rhs_2	1sw6_1	1sw6_2	1trn
1vjw	1vlz_1	1vlz_2	1xca	1ygh	1zin	2atj_2	2bls_2		

¹ The dataset contains 122 homodimers with resolution $\leq 3.0\text{\AA}$. The sequence identity between any two proteins is $< 25\%$ (Bahadur et al. 2003). ² The dataset contains 327 non-redundant heterodimers (each monomer > 25 residues) with resolution ($\leq 3.25\text{\AA}$) (Mintseris and Weng 2003). ³ The dataset contains 188 crystal-packing monomers with buried surface area $> 800\text{\AA}^2$ (Bahadur et al. 2004).

3.2.3 Calculation of inter-atomic interactions

We analyzed protein-protein interactions for homodimers, heterodimers and crystal-packing monomers by calculating inter-atomic interactions. For a complex, two atoms from different monomers were considered to interact if the distance between their centers is less than the sum of their van der Waals (vdW) radii plus x (Å). We used the vdW radius published elsewhere (Chothia 1976). The value of x was varied from -0.5Å to 4.0Å at increments of $0.1(\text{Å})$. The total number of inter-atomic interactions for each complex was calculated at every value of x . They were grouped into four inter-atomic types (B_1B_2I , B_1S_2I , S_1B_2I and S_1S_2I). We calculated the mean of inter-atomic interactions for each dataset for all values of x (Figure 3.1). We also calculated the mean and standard deviation (SD) for the percentage of the four types (Figure 3.2).

3.3 Results & Discussion

Figure 3.1 shows the mean distribution of inter-atomic contacts in homodimers, heterodimers and crystal-packing monomers. The mean number of inter-atomic contacts at homodimer interfaces is largest for increasing value of x , while crystal-packing monomer interfaces have the least number. This observation is similar to the high interface compactness in homodimers and heterodimers reported by Bahadur and coworkers (Bahadur et al. 2004). The large inter-atomic interactions at the interfaces of homodimers and heterodimers provide insight to the binding mechanisms common in homodimers and heterodimers. Interestingly, crystal-packing monomer interfaces has

Chapter 3 Inter-atomic interactions

the least number of inter-atomic interactions. This suggests non-specific interactions in this type of associations.

Figure 3.2 shows the distribution of the four types of inter-atomic interactions at the interfaces of homodimers, heterodimers and crystal-packing monomers. S_1S_2I (sidechain-sidechain interactions) are predominant in all the three types of dimer interfaces. However, B_1B_2I (backbone-backbone interactions) are the least in all the three types of dimer interfaces. B_1S_2I (backbone-sidechain interactions) and S_1B_2I (sidechain-backbone interactions) are equally distributed at the interfaces of homodimers and heterodimers. This explains the previous observation that interface hydrogen bonds are predominantly contributed by sidechain-sidechain donor-acceptor pairs and the contribution of backbone-backbone interactions is the least (Xu et al. 1997). Figure 3.2 gives the %SD about the mean for all the three types of interfaces. Large %SD are observed for all the four types of inter-atomic interactions at interfaces of homodimers, heterodimers and crystal-packing monomers for x ranging from $-0.5\text{\AA}$ – $0.5\text{\AA}$. It is interesting to note that the four types in homodimers, heterodimers and crystal-packing monomers do not show significant difference on mean percentage and %SD distributions, although interface formations of the three types of complex are different. Nevertheless, the dominance of side chain-sidechain interactions at the interfaces of homodimers and heterodimers imply that sidechain-sidechain stacking plays an important role in dimer association and stability.

Figure 3.1. The distribution for the average number of inter-atomic contacts at the homodimer, heterodimer and crystal-packing monomer interfaces.

Chapter 3 Inter-atomic interactions

Figure 3.2. Mean and standard deviation (SD) distributions for the four types of inter-atomic interactions ($B_1B_2I/B_1S_2I/S_1B_2I/S_1S_2I$) in homodimers, heterodimers and crystal-packing monomers.

3.4 Conclusion

We reported a comprehensive analysis of the four types of inter-atomic interactions at the interfaces of homodimers, heterodimers and crystal-packing monomers. The analysis suggests that homodimer interfaces have larger inter-atomic contact than heterodimers. Both homodimers and heterodimers have predominant sidechain-sidechain stacking. These data are important in the development of protein-protein docking algorithms.

Chapter 4

Identification of energetically hot spot residues at interface

4.1 Introduction

Protein-protein interactions are determined by the residues at the interface. Thus, it is necessary to understand the contribution of individual residues to the free energy of binding. The most popular approach is to mutate the target residue to alanine (i.e. the target side-chain removed) and measure the binding free energies before and after mutation (Wells 1991). The change of binding free energy ($\Delta\Delta G$) indicates the consequence of the residue's sidechain for binding affinity (Figure 4.1). It is observed that despite the large size of protein-protein interface, binding free energy is dominantly contributed by a few interface residues ($\Delta\Delta G \geq 1.5 \text{ kcal}\cdot\text{mol}^{-1}$), called “hot spot” residues (Clackson and Wells 1995; Bogan and Thorn 1998). Besides, it is observed by molecular kinetics that “hot spot” residues have a marked impact on both protein-protein recognition and binding stability, while other interface residues are related with binding stability only. For example, at the BPTI-trypsin interface, hot spot Lys15->Ala mutation ($\Delta\Delta G = 10 \text{ kcal}\cdot\text{mol}^{-1}$) leads to a 200-fold decrease in association rate, while low energy residue Arg17->Ala ($\Delta\Delta G < 0.5 \text{ kcal}\cdot\text{mol}^{-1}$) has little effect on association rate (Castro and Anderson 1996). Therefore, interface specificity is mainly determined by hot spots.

Figure 4.1 Flow chart illustrating alanine scanning procedure.

Chapter 4 Hot spot interface residues

Because hot spots are a good indicator of interface specificity, their characteristics have been widely investigated (Bogan and Thorn 1998; Ma et al. 2003; Halperin et al. 2004; Li et al. 2004; Brinda and Vishveshwara 2005; del Sol et al. 2005; del Sol and O'Meara 2005; Haliloglu et al. 2005). Hot spots are enriched in Trp, Tyr and Arg, and are often surrounded by hydrophobic rings to occlude bulk solvent (Bogan and Thorn 1998). In addition, hot spots statistically correlate with structurally conserved residues in ten protein families (Ma et al. 2003). Moreover, hot spots from different monomers prefer to interact and their couplings are structurally conserved (Halperin et al. 2004). It has also been found that hot spots are related to central interface residues using the small-world network approach (proteins represented as networks, residues as nodes and interactions as edges) (del Sol et al. 2005; del Sol and O'Meara 2005).

In recent years a number of computational methods have been developed to predict hot spots. These methods are classified into two types: (1) energy-based and (2) structure-based. In the energy-based methods, functions are developed to calculate a residue's $\Delta\Delta G$ by simulating residue mutation to Alanine (Massova and Kollman 1999; Guerois et al. 2002; Kortemme and Baker 2002; Verkhivker et al. 2002; Kortemme et al. 2004). These methods give good qualitative prediction results. However, high computational cost and the difficulty in operation (e.g. data processing) make them unsuitable for easy implementation. A good example of structure-based methods is the one described by Gao and colleagues (Gao et al. 2004). In this method, interface residues are covered by a grid box and the contribution by each residue to binding affinity is estimated by rolling different kinds of probes (representing hydrophobic

Chapter 4 Hot spot interface residues

group, hydrogen bonds) over the grids close to the residue. The residues having high energy contribution are predicted as hot spots. This method is subject to complex structural analysis and comparison. From the above analyses, we can see that a simple “hot spot” prediction method has yet been developed.

It has recently been documented that among interface residues, some are involved in strong positive residue-residue interactions, while others have weak interactions (del Sol and O'Meara 2005; Reichmann et al. 2005). The residues with the strong interactions are expected to have more contribution to binding. Thus, it is of interest to investigate whether “hot spot” residues possess strong interactions. We collected 296 alanine-mutated interface residues from 15 complexes and grouped them into three classes (1) hot-spot residues with $\Delta\Delta G \geq 1.5 \text{ Kcal}\cdot\text{mol}^{-1}$, (2) warm residues with $\Delta\Delta G = 0.5\text{--}1.5 \text{ Kcal}\cdot\text{mol}^{-1}$ and (3) unimportant residues with $\Delta\Delta G < 0.5 \text{ Kcal}\cdot\text{mol}^{-1}$. We analyzed each residue in the dataset based on inter-atomic interactions across the interface.

4.2 Methods

4.2.1 Definition of interface residue

Solvent-accessible surface area (ASA) of a residue was calculated using the program NACCESS (Hubbard 1993). A residue with interface area (ΔASA) $> 1\text{\AA}^2$ is defined as an interface residue, where ΔASA is the change in ASA of the residue upon protein dimer formation from monomer state.

4.2.2 Dataset of alanine-mutated interface residues

We created a dataset consisting of 296 alanine-mutated interface residues from 15 complexes (Table 4.1), derived from the database ASEdb (Alanine Scanning Energetics database) (Thorn and Bogan 2001). These residues have $\Delta\Delta G$ in the range $-0.9\sim 10$ kcal $\cdot$ mol $^{-1}$. According to the classification of Gao *et al.* (Gao et al. 2004), we classified the residues into three types: (1) hot-spot ($\Delta\Delta G \geq 1.5$ Kcal $\cdot$ mol $^{-1}$), (2) warm ($0.5 - 1.5$ Kcal $\cdot$ mol $^{-1}$) and (3) unimportant (< 0.5 Kcal $\cdot$ mol $^{-1}$) residues. Thus, unimportant residues are dominant (45%) in the dataset and warm residues (27%) are similar to hot-spot residues (28%). Hot-spot and warm residues were considered as key residues.

4.2.3 Definition of inter-atomic contacts

Protein-protein complexation is determined by inter-atomic interactions between monomers. Hence, we attempted to investigate the three groups of residues (hot spots, warm and unimportant residues) in terms of their contribution to the inter-atomic interactions. The inter-atomic interactions are composed of four categories, namely S_1S_2I , S_1B_2I , B_1S_2I and B_1B_2I (S: sidechain atom, B: backbone atom, subscript 1 and 2 refer to different monomers). The prevalence of these four inter-atomic interactions at the interface of protein-protein complexes, 70 non-redundant complexes (Chakrabarti and Janin 2002), was examined by calculating their means and standard deviations at varying inter-atomic distances (see Figure 3.2). S_1S_2I dominates at the interface and hence we exclusively selected S_1S_2I for studying hot spots, warm and unimportant residues. By definition, two sidechain atoms from different monomers were considered

Chapter 4 Hot spot interface residues

to interact if the distance between their centers is less than the sum of their van der Waals (vdW) radii plus a cutoff distance of 0.5Å, at which cutoff the mean of S_1S_2I is maximum and the standard deviation is minimum (See Figure 3.2).

4.2.4 Classification of inter-atomic contacts

We classified inter-atomic sidechain-sidechain contacts into two groups (energetically favorable and unfavorable contacts) using the scheme described by Sobolev and colleagues (Sobolev et al. 1996). Table 4.2 lists the scheme, where residue atoms are grouped into seven classes: (I) hydrophilic, (II) Acceptor, (III) Donor, (IV) Hydrophobic, (V) Aromatic, (VI) Neutral, (VII) Neutral-donor and (VIII) Neutral-acceptor. Therein, carbon atoms in aromatic rings are considered as favorably contacting with polar groups, in agreement with energy estimations for such contacts. Carbon atoms bonded to polar oxygen or nitrogen atoms are not treated as hydrophobic atoms. It is because when polar oxygen and nitrogen atoms form hydrogen bonds, carbon atoms bonded to the donor/acceptor may be in contact with the acceptor/donor. To prevent such forced contacts from affecting the estimation of inter-atomic contacts, this type of carbon atoms is classified as neutral atoms instead of hydrophobic atoms.

Chapter 4 Hot spot interface residues

Table 4.1 15 complexes with 296 mutations at the interface

PDB ID	Partner A	Partner B	No. mutations in A	No. mutations in B
1a4y	Angiogenin	Ribonuclease inhibitor	11	12
1ahw	Tissue Factor	Immunoglobulin Fab 5G9	7	-
1brs	Barnase	Barstar	7	5
1bxi	Colicin E9			
	immunity protein	Colicin E9 Dnase domain	17	-
1cbw	Basic Pancreatic			
	Trypsin Inhibitor	Chymotrypsin	7	-
	Blood coagulation			
1dan	factor Viia	Soluble Tissue Factor	24	34
1dfj	Ribonuclease A	Ribonuclease Inhibitor	-	14
1dvf	Fv D1.3	Fv E5.2	16	7
1fc2	Protein A	IgG Fc fragment	3	-
1gc1	Cd4	Envelope Protein Gp120	17	-
1jck	Staphylococcal			
	enterotoxin C3	T cell antigen receptor V _β	10	-
	Monoclonal			
1vfb	antibody D1.3	Egg lysozyme	13	11
	Basic pancreatic			
2ptc	trypsin inhibitor	Trypsin	1	-
3hfm	Lysozyme	IgG1 Fab (HyHEL-10)	13	2
	Human growth			
3hhr	hormone	hGH receptor	39	26

Chapter 4 Hot spot interface residues

Table 4.2 Classification of contact between atoms of different classes

Atom class*	favorable (+) or unfavorable (-) contact							
	I	II	III	IV	V	VI	VII	VIII
I Hydrophilic	+	+	+	-	+	+	+	+
II Acceptor	+	-	+	-	+	+	+	-
III Donor	+	+	-	-	+	+	-	+
IV Hydrophobic	-	-	-	+	+	+	+	+
V Aromatic	+	+	+	+	+	+	+	+
VI Neutral	+	+	+	+	+	+	+	+
VII Neutral-donor	+	+	-	+	+	+	-	+
VIII Neutral-acceptor	+	-	+	+	+	+	+	-

*I: Hydrophilic = nitrogen or oxygen atoms that can donate and accept hydrogen bonds (e.g., oxygen of hydroxyl group of Ser or Thr). II: Acceptor = nitrogen or oxygen atoms that can only accept a hydrogen bond. III: Donor = nitrogen that can only donate a hydrogen bond. IV: Hydrophobic = carbon atoms that are not in aromatic rings and do not have a covalent bond to a hydrophilic atom. V: Aromatic = carbon atoms in aromatic rings. VI: Neutral = carbon atoms that have a covalent bond to at least one atom of class I or two or more atoms from class II or III; nitrogen atoms if it has covalent bonds with 3 carbons; sulfur atoms in all cases. VII: Neutral-donor = carbon atoms that has a covalent bond with only one atom of class III. VIII. Neutral-acceptor = carbon atoms that has covalent bond with only one atom of class II. The classification is derived from (Sobolev et al. 1996).

4.2.5 Definition of NA, $NA_f - NA_{unf}$ and $NC_f - NC_{unf}$

We investigated each residue in our dataset using (1) NA, (2) $NA_f - NA_{unf}$ and (3) $NC_f - NC_{unf}$ (Table 4.3), illustrated in Figure 4.2. NA is the number of atoms of a residue participating in sidechain-sidechain contacts. It could examine the involvement for an interface residue in protein-protein interaction. $NA_f - NA_{unf}$ is the difference between the number of atoms in favorable contacts (NA_f) and unfavorable contacts (NA_{unf}). It was employed to explore energetic contribution for a residue to protein-protein interface in terms of atoms. $NC_f - NC_{unf}$ is the difference between favorable contacts (NC_f) and unfavorable contacts (NC_{unf}). It was used to explore energetic contribution for a residue to protein-protein interface in terms of inter-atomic contacts.

Figure 4.2. Illustration of NA , NA_f , NA_{unf} , NC_f and NC_{unf} . The interaction of K15 (BPTI, Chain I) to S190, S195 and V213 (trypsin, Chain E) is shown (PDB ID: 2PTC). K15 has three interacting sidechain atoms (CB, CD and NZ) and therefore the NA value is 3. Therein, the three atoms are all involved in favorable contacts (green line) and only CB participating in unfavorable contacts (red line). Thus, the NA_f value is 3 and NA_{unf} is 1. In addition, K15 has three favorable contacts and one unfavorable contact; hence NC_f is 3 and NC_{unf} is 1. Carbon atom: white; oxygen atom: red; Nitrogen atom: blue.

Table 4.3. Results of analysis on 296 mutated interface residues

PDB ID	Residue	$\Delta\Delta G$ (Kcal/Mol)	ΔASA ($\text{\AA}^2$)	NA	NA _P	NA _{unf}	NC _P	NC _{unf}	Alanine Scanning (Kcal/Mol)	FOLDEF (Kcal/Mol)	PP SITE
1a4yA	R5	2.3	100.1	3		-1		-1	2.54	2.09	H
1a4yA	H8	0.9	24.1	2		2		2	0.84	-0.37	U
1a4yA	Q12	0.3	22.8	2		2		3	0.66	0.4	U
1a4yA	H13	-0.3	4.7	0		0		0	0.02	0.19	U
1a4yA	R31	0.2	119.9	3		2		8	2.76	-0.83	W
1a4yA	R32	0.9	62.2	3		3		9	0.18	0.06	U
1a4yA	N68	0.2	11.1	0		0		0	0.52	0.05	U
1a4yA	H84	0.2	64	3		3		5	1.05	0.25	W
1a4yA	W89	0.2	84.8	5		5		13	2.71	0.88	H
1a4yA	E108	-0.3	43.9	0		0		0	1.58	-0.15	H
1a4yA	H114	0.65	62.6	5		3		3	1.69	0.49	W
1a4yB	W261	0.1	32	0		0		0	1.06	0.6	H
1a4yB	W263	1.2	66	2		2		4	2.27	2.2	H
1a4yB	W318	1.5	42.5	7		7		12	2.2	2.7	H
1a4yB	K320	-0.3	39.5	1		1		1	-0.21	-0.75	U
1a4yB	E344	0.2	16	3		3		6	1.37	-0.44	U
1a4yB	W375	1	56.6	7		5		12	2.83	2.51	H
1a4yB	E401	0.9	19.6	1		1		1	0.02	-0.2	W
1a4yB	Y434	3.3	98.5	6		6		8	2.95	1.93	H
1a4yB	D435	3.5	80.3	3		1		6	0.57	3.2	H
1a4yB	Y437	0.8	143.4	5		4		7	2.98	1.07	H
1a4yB	R457	-0.2	2.8	0		0		0	-0.04	0.32	-
1a4yB	I459	0.7	23.9	1		1		1	0.62	0.41	W
1ahwA	Y156	>4	65.9	3		2		6	4.6	2.00	W
1ahwA	T167	0	41.5	0		0		0	-0.03	0.05	U
1ahwA	T170	1	11.3	0		0		0	-0.06	0.45	U

Chapter 4 Hot spot interface residues

lahwA	D178	-0.5	14.3	0	0	0	0	-0.09	0.15	U
lahwA	T197	1.3	8.9	0	0	0	0	-0.02	0.01	U
lahwA	V198	-0.3	7.2	0	0	0	0	-0.01	-0.03	-
lahwA	N199	1.1	0.7	0	0	0	0	-0.01	0.03	-
lbrsA	K27	5.4	54.6	3	3	5	5	1.88	2.58	H
lbrsA	N58	3.1	2.6	0	0	0	0	-0.03	0.73	U
lbrsA	R59	5.2	144.2	7	1	10	10	3.01	0.83	H
lbrsA	E60	-0.2	65.4	3	0	-2	-2	1.4	0.43	W
lbrsA	E73	2.8	9.5	0	0	0	0	-0.2	-0.59	W
lbrsA	R87	5.5	5.2	1	1	2	2	4.44	-0.12	W
lbrsA	H102	6	95.3	6	4	10	10	5.08	1.81	W
lbrsB	Y29	3.4	86.2	4	4	13	13	3.13	1.86	H
lbrsB	D35	4.5	111	4	1	7	7	1.42	0.54	H
lbrsB	D39	7.7	80.9	4	3	13	13	9.4	3.07	H
lbrsB	T42	1.8	16	2	2	5	5	1.66	2.01	W
lbrsB	E76	1.3	21.7	3	3	5	5	1.54	0.12	U
lboxiA	C23	0.92	10.6	0	0	0	0	-0.08	-0.10	U
lboxiA	N24	0.14	6.4	0	0	0	0	0.00	-0.05	U
lboxiA	T27	0.73	5.3	1	1	3	3	0.59	0.90	U
lboxiA	S28	0.17	0	0	0	0	0	0.00	0.04	U
lboxiA	S29	0.96	6.6	0	0	0	0	0.35	-0.04	U
lboxiA	E30	1.41	100.5	5	3	11	11	2.97	0.94	H
lboxiA	L33	3.42	33.4	2	1	4	4	1.02	1.07	W
lboxiA	V34	2.58	61.5	1	1	2	2	0.98	0.91	H
lboxiA	V37	1.66	16.2	3	3	4	4	0.5	0.44	W
lboxiA	T38	0.9	19.2	1	0	0	0	1.35	1.00	W
lboxiA	E41	2.08	32.8	2	2	3	3	-0.08	0.58	H
lboxiA	S48	0.01	2.1	0	0	0	0	-0.01	-0.05	U
lboxiA	G49	1.49	8.6	0	0	0	0	-	-0.43	U
lboxiA	S50	2.19	33	1	1	2	2	5.4	1.39	W

Chapter 4 Hot spot interface residues

1bxiA	D51	5.92	36.6	0	0	0	0	0	0.82	1.07	W
1bxiA	Y55	4.63	90.6	4	4	4	6	6	2.8	3.31	H
1bxiA	P56	1.24	5.4	0	0	0	0	0	-	0.14	U
1cbwA	T11	0.2	37.9	0	0	0	0	0	-0.18	0.06	U
1cbwA	P13	-0.1	46.3	1	1	1	1	1	-	0.75	W
1cbwA	K15	2	147.6	2	-1	-1	-1	-1	-0.03	-0.74	H
1cbwA	R17	0.5	151.1	6	3	3	9	9	1.28	1.22	H
1cbwA	I19	0.1	38.1	0	0	0	0	0	0.68	0.40	W
1cbwA	V34	0	14.1	0	0	0	0	0	0.3	0.40	W
1cbwA	R39	0.2	64.8	1	1	1	1	1	1.61	1.17	W
1danA	L39	0	103.2	3	3	3	7	7	1.24	1.10	-
1danA	I42	0	3.5	0	0	0	0	0	0.51	0.23	-
1danA	K62	0	72	2	0	0	2	2	-0.27	-0.45	-
1danA	Q64	0.8	60.5	1	1	1	1	1	2.98	0.30	-
1danA	I69	1.9	74.7	2	2	2	4	4	1.57	1.66	-
1danA	F71	1.2	100.2	5	5	5	13	13	3.07	2.58	-
1danA	E77	0	60.3	2	0	0	0	0	0.36	-1.19	-
1danA	R79	1.2	110.4	4	2	2	8	8	2.45	0.34	-
1danA	Q88	0	29.4	3	3	3	5	5	0.35	-0.05	-
1danA	V92	0	65.6	1	1	1	4	4	1.05	0.97	-
1danA	N93	0	29.3	3	3	3	6	6	0.12	0.07	-
1danA	E94	0	38.3	0	0	0	0	0	-	-0.13	-
1danA	R271	0	1.3	0	0	0	0	0	-	0.03	-
1danA	F275	0	71.2	7	7	7	9	9	1.66	1.32	-
1danA	R277	0.51	120.6	1	0	0	3	3	3.09	0.95	-
1danA	F278	0	25.8	0	0	0	0	0	0.53	0.27	-
1danA	R304	0.65	2.9	0	0	0	0	0	-	0.88	-
1danA	M306	0.5	110.8	4	0	0	11	11	1.03	1.08	-
1danA	T307	0	6	0	0	0	0	0	0.27	-1.30	-
1danA	Q308	0	72.8	3	2	2	5	5	2.6	0.66	-

Chapter 4 Hot spot interface residues

1danA	D309	0.41	42.5	3	3	4	1.34	1.52	-
1danA	Q312	0	5	0	0	0	-	-0.14	-
1danA	E325	0	3.9	0	0	0	-	0.04	-
1danA	R379	0.51	30.9	3	2	5	1.33	-0.93	-
1danB	K15	-0.4	3.3	0	0	0	-0.02	-0.08	-
1danB	T17	0.1	25.7	2	2	2	0.12	0.22	U
1danB	N18	0.2	11.3	0	0	0	0.04	0.19	U
1danB	K20	2.6	75.9	2	2	2	1.5	0.16	H
1danB	I22	0.7	27.7	2	0	0	0.65	0.64	W
1danB	E24	0.7	14.7	2	2	2	0.64	-0.44	U
1danB	Q37	0.55	27.3	1	1	6	0.75	1.37	W
1danB	K41	0.35	3.2	0	0	0	-0.04	-0.07	U
1danB	S42	-0.1	18.1	0	0	0	-0.05	-0.08	U
1danB	D44	0.7	51.2	0	0	0	0.9	0.89	W
1danB	W45	1.6	49.1	4	2	2	1.05	0.97	H
1danB	K46	0.25	39.1	0	0	0	0.10	0.22	W
1danB	S47	0.05	22.9	0	0	0	-0.05	-0.19	W
1danB	K48	0.4	22.8	2	2	2	0.43	0.04	W
1danB	F50	0.4	101.2	7	7	17	2.61	3.09	H
1danB	Y51	-0.1	34.1	1	1	1	-	0.42	H
1danB	D58	2.18	43.2	0	0	0	1.09	1.13	W
1danB	D61	0.24	50.3	1	0	1	0.01	0.27	W
1danB	E62	0	1.1	0	0	0	-	0.32	-
1danB	L72	-0.06	0.2	0	0	0	-	-0.12	-
1danB	F76	1.2	16	2	2	3	0.61	1.23	W
1danB	Y78	0.7	0.3	0	0	0	-	-0.01	W
1danB	P92	-0.2	7	0	0	0	-	-0.33	U
1danB	Q110	1.4	58.5	2	1	1	1.69	0.02	U
1danB	E128	0.1	14.5	0	0	0	-0.11	-0.13	U
1danB	R131	0	44.4	2	2	6	0.29	0.37	U

Chapter 4 Hot spot interface residues

1danB	L133	0	41.6	2	2	5	1.62	1.70	H
1danB	R135	0.55	68.8	0	0	0	0.94	0.23	W
1danB	N138	0	15.8	0	0	0	-	0.02	-
1danB	F140	1.5	13.3	3	3	3	1.54	0.74	H
1danB	S163	0	24.6	2	2	2	0.42	-0.15	U
1danB	T203	0.1	25.6	3	1	1	0.22	0.10	U
1danB	V207	-0.2	55	2	2	3	1.12	1.40	H
1danB	E208	0	25.8	1	1	1	0.28	0.15	U
1dfjB	E206	1	24.1	1	1	2	0.77	-0.97	W
1dfjB	W261	1.3	39.2	2	2	2	1.58	1.39	H
1dfjB	W263	2.2	43.5	3	3	3	3.09	2.53	H
1dfjB	E287	1.3	23.9	1	1	1	-0.1	-0.08	U
1dfjB	S289	0.8	2.2	0	0	0	0	-0.02	U
1dfjB	W318	1	54.7	2	2	3	1.27	1.24	H
1dfjB	K320	1.3	42.2	1	1	1	0.38	-0.07	W
1dfjB	E344	1.6	5.1	0	0	0	-	0.30	-
1dfjB	E401	1.3	26.3	2	2	6	0.42	0.21	W
1dfjB	Y434	5.9	107.7	7	7	14	2.74	4.23	H
1dfjB	D435	3.6	62.7	2	2	2	0.16	1.03	W
1dfjB	Y437	2.6	158.3	7	5	9	3.62	2.62	H
1dfjB	R457	0.8	28.6	2	2	3	0.21	-0.11	U
1dfjB	I459	0.3	35.3	1	1	1	0.5	0.52	W
1dvfA	T30	0.9	18	0	0	0	0.01	0.18	U
1dvfA	Y32	1.8	22.1	0	0	0	0.11	0.19	W
1dvfA	W52	4.2	65.5	7	5	6	3.26	3.52	H
1dvfA	D54	4.3	65.2	3	3	6	0.72	0.86	W
1dvfA	N56	1.2	53.2	3	1	3	1.07	-0.68	U
1dvfA	D58	1.6	14.1	3	3	3	0.71	0.28	W
1dvfA	E98	4.2	7.4	3	2	5	0.05	1.76	U
1dvfA	R99	1.9	60	0	0	0	0.22	0.94	U

Chapter 4 Hot spot interface residues

ldvfA	D100	2.8	47	3	2	2	0.43	0.69	W
ldvfA	Y101	4	68	4	2	2	2.13	1.62	H
ldvfA	H1030	1.7	18	1	1	1	0.26	0.20	W
ldvfA	Y1032	2	44.3	2	2	2	0.47	0.24	H
ldvfA	Y1049	1.7	23.1	3	3	3	0.87	0.43	W
ldvfA	Y1050	0.7	53	3	3	3	0.55	-0.11	H
ldvfA	W1092	0.3	55.8	4	4	6	1.56	0.68	H
ldvfA	S1093	1.2	11.4	0	0	0	-0.02	0.01	U
ldvfB	H33	1.9	27	2	2	3	0.79	0.46	W
ldvfB	D52	1.7	15.8	0	0	0	-0.18	-0.29	W
ldvfB	N55	1.9	46.2	1	1	3	2.13	-0.53	W
ldvfB	I100	2.7	49.5	1	1	1	1.07	1.29	H
ldvfB	Y101	4.7	165.7	4	3	11	4.77	5.42	W
ldvfB	Q103	1.6	90.6	5	2	9	1.42	1.91	W
ldvfB	R105	4.1	94.4	5	3	6	1.5	1.44	H
lfc2A	N28	0.6	46	3	0	0	0.29	0.92	-
lfc2A	I31	2.2	25.5	2	1	1	0.81	0.32	-
lfc2A	K35	1.2	59.7	2	0	0	0.18	-0.13	-
lgc1A	S23	0.29	4.2	0	0	0	0.27	-0.15	W
lgc1A	Q25	0.03	20.1	1	0	1	0.42	2.82	W
lgc1A	H27	0.28	38.6	5	1	1	0.8	0.44	W
lgc1A	K29	0.59	24.9	2	2	5	2.46	-0.07	W
lgc1A	N32	0.18	3	0	0	0	0	0.02	U
lgc1A	Q33	0.1	17.3	0	0	0	0.03	-0.05	W
lgc1A	K35	0.32	83.9	4	-1	-1	0.55	-1.19	U
lgc1A	Q40	-0.41	79	0	0	0	1.66	4.17	W
lgc1A	S42	0	67.5	0	0	0	0.03	-1.06	W
lgc1A	L44	1.04	7.8	0	0	0	0.07	0.43	U
lgc1A	T45	-0.15	23.7	1	1	1	0.32	0.65	U
lgc1A	N52	0.7	23.3	0	0	0	1.01	-0.04	U

Chapter 4 Hot spot interface residues

lgc1A	R59	1.16	47.2	4	4	4	9	1.02	0.45	W
lgc1A	S60	-0.09	39.8	0	0	0	0	0.14	-0.36	U
lgc1A	D63	-0.32	37.2	0	0	0	0	-0.05	0.49	W
lgc1A	Q64	0.44	29.6	3	1	1	1	1.02	-0.47	W
lgc1A	E85	1.31	1.8	0	0	0	0	2.1	1.02	U
ljckA	T20	1.4	59.7	0	0	0	0	1.26	0.33	W
ljckA	N23	2.5	49	0	0	0	0	1.96	0.27	W
ljckA	Y26	1.7	38.7	0	0	0	0	0.92	0.40	H
ljckA	N60	1.3	42.3	0	0	0	0	0.82	0.40	W
ljckA	Y90	2.5	31.7	2	2	2	2	1.0	0.79	H
ljckA	V91	2.1	71.7	2	2	2	2	1.04	0.93	H
ljckA	G102	0.1	14.1	0	0	0	0	-	-0.30	U
ljckA	K103	0.4	19.1	0	0	0	0	-0.26	-1.09	W
ljckA	F176	1.9	44.5	2	2	2	2	0.63	-0.16	H
ljckA	Q210	2.5	14.1	0	0	0	0	1.12	0.24	U
lvfbA	Y32	0.5	30.5	1	1	1	1	0.66	-0.09	W
lvfbA	W52	1.23	53.8	4	4	4	7	1.75	1.55	H
lvfbA	D54	1.95	32.5	2	1	1	1	-0.14	0.12	U
lvfbA	R99	0.47	49	0	0	0	0	0.73	0.51	U
lvfbA	D100	3.1	60.3	3	3	3	6	3.29	2.22	H
lvfbA	Y101	4	69.1	4	3	3	4	2.63	3.55	H
lvfbA	H1030	0.8	28.7	0	0	0	0	0.26	0.21	H
lvfbA	Y1032	1.3	55.4	7	7	7	7	1.32	1.00	H
lvfbA	Y1049	0.8	22.2	0	0	0	0	0.49	0.50	U
lvfbA	Y1050	0.4	72.1	4	4	4	10	1.52	1.28	H
lvfbA	T1053	-0.23	17.7	2	2	2	2	0.64	0.66	U
lvfbA	W1092	1.71	60.9	4	4	4	6	2.06	3.01	H
lvfbA	S1093	0.11	35.5	0	0	0	0	0.00	-0.29	U
lvfbB	D18	0.3	31.9	3	3	3	6	0.55	0.58	W
lvfbB	N19	0.3	72.3	3	3	3	6	1.04	0.84	W

Chapter 4 Hot spot interface residues

1vfbB	Y23	0.4	9.3	0	0	0	0	0	0.0	0.15	W
1vfbB	S24	0.8	53.4	2	2	2	4	4	0.78	-0.16	W
1vfbB	K116	0.7	52.6	1	1	1	1	1	0.85	-0.01	W
1vfbB	T118	0.8	33.9	2	2	1	1	1	0.1	0.05	U
1vfbB	D119	1	40.9	3	3	3	8	8	-0.2	0.35	W
1vfbB	V120	0.9	13.3	1	1	1	1	1	0.22	0.44	U
1vfbB	Q121	2.9	95.1	3	2	2	10	10	3.87	2.58	W
1vfbB	I124	1.2	23.4	1	1	1	1	1	0.46	0.40	W
1vfbB	R125	1.8	56.5	3	3	3	5	5	2.25	0.65	W
2ptcA	K15	10	165.8	3	2	2	2	2	4.16	0.33	-
3hfmA	H15	-0.44	12.9	0	0	0	0	0	0.09	0.14	U
3hfmA	Y20	4.2	66.6	2	2	2	2	2	2.72	1.46	H
3hfmA	R21	0.85	118.4	3	3	3	15	15	0.94	-1.32	H
3hfmA	W63	0.31	22.9	2	2	2	4	4	0.83	0.75	W
3hfmA	R73	-0.33	67.9	5	3	3	7	7	0.62	-0.40	W
3hfmA	L75	0.69	59	0	0	0	0	0	1.31	0.89	H
3hfmA	T89	0	15.4	0	0	0	0	0	0.17	0.08	U
3hfmA	N93	0.21	44	3	2	2	8	8	1.56	2.01	U
3hfmA	K96	6.38	51.5	4	0	0	8	8	2.13	-1.36	W
3hfmA	K97	5.5	72.2	2	2	2	4	4	1.42	0.32	W
3hfmA	I98	-0.10	2.1	0	0	0	0	0	0.10	-0.04	U
3hfmA	S100	0.26	23.2	2	2	2	4	4	-0.13	-0.11	U
3hfmA	D101	0.94	84.7	4	3	3	15	15	0.32	-2.39	H
3hfmB	Y58	1.7	45.9	6	6	6	6	6	1.77	0.55	H
3hfmB	D101	3.75	0.5	0	0	0	0	0	-0.04	0.09	-
3hhrA	P2	-0.05	75.9	2	2	2	3	3	-	0.06	H
3hhrA	T3	-0.05	0.8	0	0	0	0	0	-	-0.01	U
3hhrA	I4	0.41	68.4	0	0	0	0	0	-	-0.08	H
3hhrA	R8	0.2	88.5	3	2	2	7	7	-	-1.19	H
3hhrA	L9	-0.04	14.7	1	1	1	1	1	-	-0.01	U

Chapter 4 Hot spot interface residues

3hrA	N12	0.1	59.1	4	3	8	-	-0.26	U
3hrA	L15	0.15	70.5	3	0	1	-	-0.05	H
3hrA	R16	0.24	80.8	6	5	18	-	-0.11	W
3hrA	H18	-0.5	65.2	3	0	-1	1.51	-1.32	H
3hrA	R19	0.05	44.4	0	0	0	-	0.05	W
3hrA	H21	0.2	25.7	0	0	0	0.64	0.20	W
3hrA	Q22	-0.2	14.2	0	0	0	0.06	0.36	U
3hrA	F25	-0.4	63	0	0	0	1.68	1.08	H
3hrA	D26	-0.2	0.1	0	0	0	-	0.39	-
3hrA	Q29	-0.6	2.3	0	0	0	-	0.19	-
3hrA	Y42	0.2	92.7	2	1	0	2.09	1.34	H
3hrA	L45	1.2	57.7	2	2	2	1.13	1.36	H
3hrA	Q46	0.1	90.6	3	0	2	1.2	0.17	U
3hrA	P48	0.4	21.8	2	-1	-1	-	0.56	U
3hrA	S51	0.3	9.3	0	0	0	-0.04	0.00	-
3hrA	E56	0.4	1.2	2	2	3	-0.04	0.21	-
3hrA	S62	0.2	52	0	0	0	0.18	1.03	U
3hrA	N63	0.3	38.5	0	0	0	0.22	-0.25	W
3hrA	R64	1.6	115.2	4	3	13	1.12	1.42	W
3hrA	E65	-0.5	1.2	0	0	0	-0.05	-0.16	-
3hrA	Q68	0.6	25.2	0	0	0	0.29	-0.22	U
3hrA	Y164	0.3	25	1	1	1	0.6	-1.08	U
3hrA	R167	0.3	31.9	2	1	2	0.54	-0.64	W
3hrA	K168	-0.2	57.3	2	0	0	1.14	-2.41	U
3hrA	D171	0.8	51	4	2	2	1.4	3.94	H
3hrA	K172	2	31.8	2	2	8	0.71	0.77	U
3hrA	E174	-0.9	23.9	1	1	1	0.61	1.66	W
3hrA	T175	2	48.8	3	2	3	2.86	1.23	W
3hrA	F176	1.9	6.4	0	0	0	-	0.48	U
3hrA	R178	2.4	66.9	0	0	0	0.85	0.69	W

Chapter 4 Hot spot interface residues

3hrA	I179	0.8	31.7	0	0	0	0	0	0	1.06	0.83	W
3hrA	R183	0.5	0.3	0	0	0	0	0	0	-	-0.08	-
3hrA	C182	1.01	26.7	0	0	0	0	0	0	-0.10	0.32	U
3hrA	E186	0	5.2	0	0	0	0	0	0	-	0.06	-
3hrB	R43	2.2	28.8	2	1	2	2	2	2	3.18	2.67	-
3hrB	E44	1.8	28.5	3	3	3	3	3	3	0.05	0.02	-
3hrB	N72	0.2	9.3	0	0	0	0	0	0	-	-0.12	-
3hrB	W76	0.6	56	2	0	0	0	0	-1	2.0	2.49	-
3hrB	T77	-0.25	24.4	3	2	2	2	2	2	0.93	1.02	-
3hrB	W80	0	0.3	0	0	0	0	0	0	-	0.03	-
3hrB	S98	-0.1	1	0	0	0	0	0	0	0.14	-0.01	-
3hrB	S102	-0.2	25.2	0	0	0	0	0	0	0.33	0.96	-
3hrB	I103	1.8	6.3	0	0	0	0	0	0	0.42	0.90	-
3hrB	W104	4.5	139.2	7	6	10	6	6	10	4.98	5.17	-
3hrB	I105	2	12.6	0	0	0	0	0	0	0.15	0.40	-
3hrB	C108	0	10	0	0	0	0	0	0	-0.04	0.18	-
3hrB	E120	-0.2	22.6	0	0	0	0	0	0	0.59	-0.29	-
3hrB	K121	0.1	25.1	3	0	0	0	0	0	0.09	0.24	-
3hrB	C122	0	37.9	1	1	1	1	1	1	-0.20	0.33	-
3hrB	S124	0.2	11.2	0	0	0	0	0	0	-	-0.96	-
3hrB	D126	1	26.9	0	0	0	0	0	0	-0.04	-0.95	-
3hrB	E127	1	70	4	3	5	3	3	5	1.13	-1.54	-
3hrB	D164	1.6	16.9	4	2	2	2	2	2	0.37	0.85	-
3hrB	I165	2.2	10.2	0	0	0	0	0	0	0.06	0.15	-
3hrB	Q166	0	42	0	0	0	0	0	0	0.42	0.29	-
3hrB	K167	0	37.6	0	0	0	0	0	0	0.02	0.07	-
3hrB	W169	4.5	87.8	7	7	13	7	7	13	3.47	4.44	-
3hrB	Q216	0.9	5.9	0	0	0	0	0	0	-	0.02	-
3hrB	R217	0.2	35.9	4	0	-1	0	0	-1	0.17	-0.70	-
3hrB	N218	0.3	107.5	4	0	0	0	0	0	1.48	0.00	-

Chapter 4 Hot spot interface residues

Suffix added to the PDB entry code refers to chain name. Number added to residue symbol is the position of the residue in chain sequence. $\Delta\Delta G$ = change of binding free energy upon mutation to Alanine. ΔASA = change in water-accessible surface area of a residue upon dimer formation from monomer state. NA = number of atoms for a residue involved in inter-atomic sidechain-sidechain interactions. NA_f or NA_{unf} = number of sidechain atoms for a residue involved in favorable or unfavorable contacts. NC_f or NC_{unf} = number of favorable or unfavorable contacts. 'H' = predicted hot spot. 'W' = predicted warm residue. 'U' = predicted unimportant residue. '-' = the residue can not be identified.

4.3 Result

4.3.1 “NA”

Figure 4.3 shows percentage distributions of the three types of interface residues (hot spots, warm and unimportant residues) based on the value of NA. The percentage of hot spots increases from 15% to 50% with NA, while that of unimportant residues decreases from 60% to 23%. Interestingly, the percentage of warm residues does not significantly change with NA. This suggests that S_1S_2I interactions are prominent among hot spots. When $NA=1$, the percentage of warm residues (41%) are larger than that in the original dataset (27%) and when $NA>1$ hot spots ($>33\%$) are higher than that in the original dataset (28%). It should be noted that nearly 40% of the residues in the dataset do not participate in inter-atomic sidechain-sidechain contacts ($NA=0$). Hence, these residues can not be identified as hot spots, warm and unimportant residues using NA, NA_f-NA_{unf} and NC_f-NC_{unf} values.

4.3.2 “ NA_f-NA_{unf} ”

Figure 4.4A shows percentage distributions of hot spots, warm and unimportant residues based on NA_f-NA_{unf} . The percentage of unimportant residues decreases with the increase in NA_f-NA_{unf} , and that of hot spots increases. The percentage of warm residues is not significantly affected by NA_f-NA_{unf} . We also show that when $NA_f-NA_{unf}>1$, the percentage of hot spots is larger than the fraction (28%) of hot spots in the original dataset.

Figure 4.3. Percentage distribution of hot spots, warm and unimportant residues in 296 interface residues obtained from the Alanine Scanning Energetics database (ASEdb) (Thorn and Bogan 2001), based on the value NA (the number of atoms for a residue involved in sidechain-sidechain interactions across protein-protein interface). The first column shows the percentage of the three types in the 296 residues. The number of residues is 114 for NA=0, 34 for NA=1, 52 for NA=2, 46 for NA=3 and 50 for NA>3. White: hot spots; gray: warm residues; black: unimportant residues.

Figure 4.4. Percentage distribution of the three residue types in 182 interface residues with $NA > 0$ (Unimportant residues: 64; warm residues: 52; hot spots: 66) based on the value of (A) NA_f-NA_{unf} (the difference between numbers of sidechain atoms for a residue involved in favorable and unfavorable contacts). The number of residues is for 24 for $NA_f-NA_{unf} < 1$, 45 for $NA_f-NA_{unf} = 1$, 54 for $NA_f-NA_{unf} = 2$, 35 for $NA_f-NA_{unf} = 3$ and 24 for $NA_f-NA_{unf} > 4$ (B) NC_f-NC_{unf} (the difference between numbers of favorable and unfavorable contacts). The number of residues is 17 for $NC_f-NC_{unf} < 1$, 32 for $NC_f-NC_{unf} = 1$, 30 for $NC_f-NC_{unf} = 2$, 21 for $NC_f-NC_{unf} = 3$, 82 for $NC_f-NC_{unf} > 3$. The first column in each graph shows the percentages of the three types in 296 interface residues. White: hot spots; gray: warm residues; black: unimportant residues.

4.3.3 “ NC_f - NC_{unf} ”

Figure 4.4B shows percentage distribution of the three types of interface residues types based on NC_f - NC_{unf} . The percentage of unimportant residues decreases with the increase in NC_f - NC_{unf} , and hot spots increases. The percentage of warm residues does not significantly change with NC_f - NC_{unf} . It was also found that the percentage of hot spots is high when NC_f - $NC_{unf} \geq 2$, in comparison to the fraction (28%) of hot spots in the original dataset.

4.4 Discussion

4.4.1 A ‘hot spot’ prediction approach

Our results show that the fraction of hot spots increases and unimportant residues decreases with increase in NA , NA_f - NA_{unf} and NC_f - NC_{unf} . However, the fraction of warm residues is not significantly affected by the three parameters. Thus, hot spots are preferentially involved in strong and energetically favorable sidechain-sidechain interactions, and unimportant residues tend to participate in weak and energetically unfavorable sidechain-sidechain interactions. Here, we used NA , NA_f - NA_{unf} and NC_f - NC_{unf} to develop a method to identify hot spots using interface residues in structural complexes. We classified the residues in our dataset using a combination of the three parameters. This is based on the observation that hot spots are prevailing in the residues with $NA > 1$, NA_f - $NA_{unf} > 1$ or NC_f - $NC_{unf} > 1$, and unimportant residues are predominant in those with $NA = 0$, NA_f - $NA_{unf} \leq 1$ or NC_f - $NC_{unf} \leq 1$ (Figure 4.3 and 4.4).

Chapter 4 Hot spot interface residues

Table 4.4 shows that the percentages of unimportant residues when (i) $NA=0$, (ii) $NA=1 \& NA_f - NA_{unf} \leq 1 \& NC_f - NC_{unf} \leq 1$, and (vi) $NA > 0 \& NA_f - NA_{unf} \leq 1 \& NC_f - NC_{unf} \leq 1$ are larger than that in original dataset; and hot spots in (iii) $NA=1 \& NA_f - NA_{unf} \leq 1 \& NC_f - NC_{unf} \geq 2$, (vii) $NA > 1 \& NA_f - NA_{unf} \leq 1 \& NC_f - NC_{unf} \geq 2$, and (ix) $NA > 1 \& NA_f - NA_{unf} \geq 2 \& NC_f - NC_{unf} \geq 2$ are higher than original dataset. Thus, the residues with $NC_f - NC_{unf} \geq 2$ could be predicted as hot spots and those with $NC_f - NC_{unf} \leq 1$ as unimportant residues. Therefore, these observations find application in the development of an expert system, like prediction model for the identification of hot spots from structural complexes.

4.4.2 Comparison with other ‘hot spot’ predication methods

We evaluated our ‘hot spot’ prediction method by comparison with other three ‘hot spot’ prediction methods: (1) PP_SITE (Gao et al. 2004), (2) Alanine Scanning method developed by Kortemme and coworkers (Kortemme and Baker 2002) and (3) FOLDEF (Guerois et al. 2002). The PP_SITE method is structure-based, while the other two are energy-based. We assessed the performance of the four methods on the differentiation between hot spots and unimportant residues. Since the PP_SITE classifies interface residues as three types (hot spots, warm and unimportant residues) and its predicted warm residues include 43% of experimental hot spots (Gao et al. 2004), we considered its predicted warm residues as predicted hot spots here. For the Alanine Scanning method and the FOLDEF method, we considered interface residues with calculated

Chapter 4 Hot spot interface residues

$\Delta\Delta G \geq 1 \text{ Kcal}\cdot\text{mol}^{-1}$ as predicted hot spots and other residues as predicted unimportant residues.

Chapter 4 Hot spot interface residues

Table 4.4. Classification of the residues in the dataset using NA, NA_f-NA_{unf} and NC_f-NC_{unf}

Groups	Interface residues				Total (296)
	Hot-spot 83 (28%)	Warm 80 (27%)	Unimportant 133 (45%)		
(i) NA=0	17 (15%)	28 (25%)	69 (60%)		114
NA=1					
(ii) NA _f -NA _{unf} ≤ 1 & NC _f -NC _{unf} ≤ 1	2 (8%)	10 (40%)	13 (53%)		25
(iii) NA _f -NA _{unf} ≤ 1 & NC _f -NC _{unf} ≥ 2	4 (44%)	4 (44%)	1 (11%)		9
(iv) NA _f -NA _{unf} ≥ 2 & NC _f -NC _{unf} ≤ 1	0	0	0		0
(v) NA _f -NA _{unf} ≥ 2 & NC _f -NC _{unf} ≥ 2	0	0	0		0
NA>1					
(vi) NA _f -NA _{unf} ≤ 1 & NC _f -NC _{unf} ≤ 1	4 (17%)	6 (25%)	14 (58%)		24
(vii) NA _f -NA _{unf} ≤ 1 & NC _f -NC _{unf} ≥ 2	6 (55%)	2 (18%)	3 (27%)		11
(viii) NA _f -NA _{unf} ≥ 2 & NC _f -NC _{unf} ≤ 1	0	0	0		0
(ix) NA _f -NA _{unf} ≥ 2 & NC _f -NC _{unf} ≥ 2	50 (44%)	30 (27%)	33 (29%)		113

Chapter 4 Hot spot interface residues

We employed our dataset to evaluate the four methods, as it contains most of alanine-mutated interface residues available. The FOLDEF and our method identified all the 296 residues, while the PP_SITE method identified 226 residues and Alanine scanning method identified 261 residues (See Table 4.3). Then, we retained the identified residues which belong to experimental hot spots and unimportant residues (FOLDEF and our method: 215 residues; PP_SITE: 160 residues; Alanine scanning: 187 residues). Finally, for each method, we calculated sensitivity (SN), specificity (SP), positive predictive value (PPV), negative predictive value (NPV) and average successful rate (formula: $\frac{\text{number of true positive} + \text{number of true negative}}{\text{number of positive} + \text{number of negative}}$) for hot spot prediction (Table 4.5). The FOLDEF and our method had large average successful rate (71%-72%), compared to the other two methods (PP_SITE: 66%; Alanine Scanning: 68%). Thus, the FOLDEF and our method can effectively distinguish between hot spots and unimportant residues. Moreover, our method was compared with the FOLDEF. Our method efficiently identified hot spots (SN=72%; SP=72%), while the FOLDEF efficiently identified unimportant residues (SN=45%; SP=88%). In addition, the PP_SITE correctly identified the most host spots (SN=90%) in these methods, however it could not effectively differentiate unimportant residues from hot spots (SP=37%). It agreed with the conclusions drawn by Gao *et al.* that the PP_SITE over-estimated unimportant residues. From these analyses, we can see that our method has remarkable hot spot prediction accuracy relative to the prevailing prediction approaches.

Chapter 4 Hot spot interface residues

Table 4.5. Evaluation of ‘hot spot’ prediction approaches

	Our approach	PP_SITE	Alanine Scanning	FOLDEF
SN	72%	90%	60%	45%
SP	72%	47%	74%	88%
PPV	62%	57%	64%	70%
NPV	81%	86%	71%	72%
Averaged successful rate	72%	66%	68%	71%
% of warm residues predicted as hot spots	45%	60%	42%	21%

The four prediction methods were assessed by comparing their performance on the differentiation between hot spots and unimportant residues. Warm residues were disregarded. SN=sensitivity; SP=specificity; PPV= positive predictive value; NPV= native predictive value; average successful rate

$$= \frac{\text{number of true positive} + \text{number of true negative}}{\text{number of positive} + \text{number of negative}}$$

hot spots by the method PP_SITE were regarded as predicted hot spots here and the evaluation is based on the PP_SITE prediction result with surface punishment (Gao et al. 2004). For the Alanine Scanning method and the FOLDEF method, we considered interface residues with calculated $\Delta\Delta G \geq 1 \text{ Kcal}\cdot\text{mol}^{-1}$ as predicted hot spots and other residues as predicted unimportant residues.

4.4.3 Investigation of misidentified hot spots by our method

Out of the 83 hot spots in our dataset, 23 were not predicted, 17 of which do not have sidechain-sidechain interactions ($NA=0$) and the remaining five have not significantly energetic contribution to sidechain-sidechain interactions ($NC_f-NC_{unf}=1$). It seems that energetic contribution of these hot spots to protein-protein interaction could not be reflected by their participation in inter-monomeric sidechain-sidechain interaction. In order to understand how the 23 misidentified hot spots contribute to protein-protein interaction, they were investigated and several reasons were found. (1) Some of them interact with interfacial water molecules to enhance the stability of protein-protein interaction. For instance, the residue D51 in the protein Im9 (PDB:1BXI) hydrogen bonds two interfacial water molecules buried in cavities (Kuhlmann et al. 2000). (2) Some neighbor on hot spots with $NC_f-NC_{unf} \geq 2$. It is possible that their mutations to alanine influence their neighboring hot spot's conformation which reduces protein-protein interaction. In human growth hormone-hGH receptor complex (PDB: 3hhr), misidentified hot spots I103 ($\Delta\Delta G = 1.8 \text{ Kcal}\cdot\text{mol}^{-1}$) and I105 ($\Delta\Delta G = 2.0 \text{ Kcal}\cdot\text{mol}^{-1}$) neighbor hot spot W104 ($\Delta\Delta G = 4.5 \text{ Kcal}\cdot\text{mol}^{-1}$). (3) Some play roles in stabilizing monomer structure so that their mutations to alanine disrupt monomer conformation which weakens protein-protein interactions, such as the residue D58 in Tissue Factor (PDB: 1DAN) (Kelley et al. 1995). (4) Some contribute to protein-protein interaction by participating in backbone-backbone or backbone-sidechain interactions. The residue K15 in protein Basic Pancreatic Trypsin Inhibitor (PDB: 1CBW) are involved in three backbone-backbone hydrogen bonds (Leder et al. 1998). Similarly, the

Chapter 4 Hot spot interface residues

residues N23 and Q120 in Staphylococcal enterotoxin C3 (PDB: 1JCK) form hydrogen bonds with backbone atoms of interacting monomer T cell antigen receptor V_{β} (Scheidig et al. 1997).

4.5 Conclusion

It is known that binding free energy of protein-protein interaction is mainly contributed by hot spot interface residues. Here, we investigate the characteristics of hot spots by examining inter-atomic sidechain-sidechain interactions across the interface, using a dataset of 296 alanine-mutated interface residues. Results show that hot spots prefer to participate in strong and energetically favorable sidechain-sidechain interactions. Based on the observation, we describe a simple ‘hot spot’ prediction model, which has better accuracy to the ‘hot spot’ prediction approaches available. Moreover, we investigated the hot spots misidentified by our methods and found several reasons. We will revise our method to enhance its prediction accuracy.

Chapter 5

Distinguishing between specific and non-specific protein-protein interfaces

5.1 Introduction

Cells contain proteins that collide and interact at every instant. Most of protein-protein interactions are short-lived and insignificant, while a few interactions form functionally relevant assemblies, due to specific non-covalent interactions encoded in 3D structure. The differentiation between specific and non-specific interfaces on a structural basis is an important task in current studies of protein-protein interactions. Various *in silico* methods have been developed to predict protein-protein interactions based mainly on shape complementarity and energy function; however, they are far from satisfactory (Janin et al. 2003). On the other hand, several parameters have been proposed to explain the features of specific interfaces; however, most break down when applied to the distinction between specific and non-specific interfaces. Therefore, it is necessary to find effective parameters. To that purpose, sets of protein complexes illustrating the specific interactions and crystal-packing monomer pairs (illustrating non-specific interactions) have been assembled. Their obvious difference is that complexes have much larger interfaces than monomer pairs (Janin and Rodier 1995; Carugo and Argos 1997; Janin 1997; Ponstingl et al. 2000). Nevertheless, there are exceptions to the rule,

Chapter 5 Specific and non-specific interfaces

and parameters other than interface size are worth considering. Recently, Bahadur *et al.* have collected 188 pairs of monomers with interface $>800\text{\AA}^2$, where 103 pairs have 2-fold symmetry and form “crystal dimers” that cannot be distinguished from real dimers based on interface size (Bahadur et al. 2004). They compared the monomers with 122 homodimers to analyze the difference between non-specific and specific interactions. They found that based on two parameters (fraction of buried atoms and non-polar interface area), 88% of the homodimers and 77% of the monomers are identified. The success rate increase to 93-95% when the parameter residue propensity (RP) score of the interfaces is taken into consideration. The stirring observations bring about our interest to explore other effective parameters on the structural basis.

5.2 Methods and results

5.2.1 Datasets of specific and non-specific interactions

We used that datasets of Bahadur et al. in this study (Bahadur et al. 2004) (Table 5.1). Non-specific interactions are represented by the dataset of large crystal-packing interfaces comprising 188 unique monomer pairs with interface area (IA) $>800\text{\AA}^2$, where 103 have 2-fold symmetry and called crystal dimers. Specific interactions are represented by homodimer dataset (122 non-redundant homodimers) and heterodimer dataset (70 heterodimers).

Chapter 5 Specific and non-specific interfaces

Table 5.1 Datasets representing specific interfaces and non-specific interfaces

PDB entries for homodimers (122)									
12as	1a3c	1a4i	1a4u	1aa7	1ad3	1ade	1af5	1afw	1ajs
1alo	1amk	1aor	1aq6	1auo	1b3a	1b5e	1b67	1b8a	1b8j
1bam	1bbh	1bd0	1bif	1biq	1bis	1bjw	1bkp	1bmd	1brw
1bsl	1bsr	1buo	1bxg	1bxk	1cdc	1cg2	1chm	1cmb	1cnz
1coz	1csh	1ctt	1cvu	1czj	1daa	1dor	1dpg	1dqs	1dxg
1e98	1ebh	1fl3	1fip	1fro	1gvp	1hhp	1hjr	1hss	1hxp
1icw	1imb	1isa	1ivy	1jhg	1jsg	1kba	1kpf	1lyn	1m6p
1mkb	1mor	1nox	1nse	1nsy	1oac	1opy	1pgt	1pre	1qfh
1qhi	1qr2	1r2f	1reg	1rfb	1rpo	1ses	1slt	1smn	1smt
1sox	1tc1	1tox	1trk	1uby	1utg	1vfr	1vok	1wtl	1xso
2arc	2ccy	2hdh	2ilk	2lig	2mcg	2nac	2ohx	2spc	2sqc
2tct	2tgi	3dap	3grs	3sdh	3ssi	4cha	4kbp	5csm	5rub
8prk	9wga								
PDB entries for heterodimers (70)									
1a0o	1a2k	1acb	1agr	1aip	1ak4	1ao7	1atn	1avw	1brs
1bth	1cbw	1cho	1cse	1dan	1dfj	1dhk	1dkg	1dvf	1efn
1efu	1fbi	1fc2	1fin	1fle	1fss	1gg2	1gla	1got	1gua
1hia	1hwg	1iai	1igc	1jhl	1kb5	1mct	1mel	1mlc	1nca
1nfd	1nmb	1nsn	1osp	1ppf	1qfu	1seb	1stf	1tbq	1tgs
1toc	1tx4	1udi	1vfb	1ycs	1ydr	1yqv	2btf	2jel	2kai
2pcc	2ptc	2sic	2sni	2trc	3hfm	3sgb	3tpi	4cpa	4htc
PDB entries for crystal-packing monomers: 2-fold symmetry (103)									
13pk	1a7v_1	1a7v_2	1ad5	1afk	1ag9	1ah7_1	1ako_1	1ako_2	1amu_1
1atl	1aw7	1ayl	1blj_1	1blj_2	1b3j	1bc2	1bea	1bin	1bkz
1bs2_1	1byo	1c02	1caq	1ck7_1	1ck7_2	1cki_1	1clu	1cqx	1dsu
1dys	1e0s_1	1e0s_2	1ehy	1epa	1ewf_1	1feh	1fgk	1fjm	1fkd
1fmt_1	1fmt_2	1g2a	1gar	1gjm	1hf8	1hfv_1	1ilr_1	1kpt	1kwa_1
1mpg_1	1mpg_2	1mss	1naw_1	1naw_2	1np4	1pbg_1	1pbg_2	1pda_1	1ppo_1
1ppo_2	1ppo_3	1qaz	1qci	1qdm	1qha	1qjp	1qme_1	1qpa	1qtq_1
1qtq_2	1rb3	1rhs_1	1rne	1shk	1the_1	1the_2	1tht	1toa	1ton
1urp_1	1urp_2	1vbt	1xgs	256b	256l	2acy	2atj_1	2bc2	2bls_1
2erc	2g3p	2ihl	2mbr	2scp	2shp	2tps	2ugi	3mht	3pmg
5tss	830c_1	8pti							
PDB entries for crystal-packing monomers: No 2-fold symmetry (85)									
1a12_1	1a12_2	1a39	1a3y_1	1a3y_2	1a6q	1a7t	1ae9	1af7	1ah7_2
1ajk	1akz	1amu_2	1aq0	1aqz	1ayi	1b24	1b80	1b8x	1be0
1bf6	1bg0	1bgc	1bs2_2	1cfy	1cki_2	1dxm	1dz4_1	1dz4_2	1elp
1ewf_2	1ewf_3	1fmt_3	1fvk_1	1fvk_2	1hfv_2	1hfv_3	1hrn_1	1hrn_2	1hsl

Chapter 5 Specific and non-specific interfaces

li4g_1	li4g_2	lihb	lilr_2	linp_1	linp_2	ljfr	lkfs_1	lkfs_2	lkwa_2
lmbi	lml1	lmwc	lnmt	lnuc	lome	lovw	lpda_2	lpmi	lpva
lqme_2	lqnt_1	lqnt_2	lqs8	lqsn	lrge	lrhs_2	lsw6_1	lsw6_2	ltrn
lvjw	vlz_1	vlz_2	lxca	lygh	lzin	2atj_2	2bls_2	2end	2nap
2rn2	3cms	3ngl_1	3ngl_2	830c_2					

5.2.2 Monomer Interface area and monomer surface area

We estimated protein-protein binding by measuring the fraction of monomer interface area (MIA) to monomer surface area (MSA). For homodimers and crystal-packing monomers, the MIA (MSA) is the interface (surface) area of any monomers. Interface area and surface area of the small monomer in heterodimers (e.g. antigen in antigen-antibody pair and inhibitor in enzyme-inhibitor pair) are considered as MIA and MSA for heterodimers. Figure 5.1 shows the relationship between MIA and MSA for homodimers, heterodimers and crystal-packing monomers. It is found that for the homodimers, heterodimers and crystal-packing monomers with the same MSA, MIA of homodimers and heterodimers is generally larger than MIA of crystal-packing monomers. Table 5.2 lists the fraction of MIA to MSA. The mean for homodimers and heterodimers is significantly larger than that for crystal-packing monomers. It indicates that specific interactions require some extent of surface burying for stability.

Figure 5.1 Relationship between monomer interface area (MIA) and surface area (MSA).

Chapter 5 Specific and non-specific interfaces

Table 5.2 Average properties of protein-protein interfaces

Protein	Crystal packing			Homodimers	Heterodimers
	Crystal dimers	Others	All		
Sample size	103	85	188	122	70
MIA/MSA	0.08	0.06	0.07	0.20	0.19
s.d.	0.04	0.03	0.03	0.11	0.10
Hot-spot residues	13	11	12	41	19
s.d.	6	5	6	24	9

5.2.2 Hot-spot interface residues

Protein-protein binding free energy is not distributed evenly over interface residues but mostly contributed by hot spot residues, as showed by alanine scanning mutagenesis. In Chapter 4, we analyzed interface residues using inter-atomic sidechain-sidechain contacts and proposed that the residues with $NC_i-NC_{il} \geq 2$ were predicted as hot-spot residues. Here, we applied the proposal for calculating hot-spot residues at specific and non-specific interfaces. Table 5.2 lists that the means hot-spots for specific interfaces are larger than that for non-specific interfaces. Figure 5.2 shows that 75% of specific interfaces have more than 20 hot-spot residues, while 92% of non-specific interface has less than 20 hot-spot residues. It was found that the number of hot-spot residues is linearly correlated with MIA for homodimers ($R^2=0.89$) and for heterodimers ($R^2=0.80$). Averagely, there is one hot-spot residue per $93 \text{ \AA}^2$ of MIA for homodimers and per $101 \text{ \AA}^2$ of MIA for heterodimers. However, the number is little correlated for monomers ($R^2=0.48$ for crystal dimers and 0.42 for other monomers) and there is one hot-spot residue per $126 \text{ \AA}^2$ of MIA.

Figure 5.2 Distribution of predicted hot-spot interface residues.

5.3 Discussion

5.3.1 Distinguishing between specific and non-specific interfaces

It was found that mean MIA/MSA for homodimers and heterodimers is larger than that for monomers. It indicates that the stability of protein-protein complexes require a certain extent of surface area burying. The value MIA/MSA was considered as a parameter to distinguish between specific and non-specific interfaces.

We calculated the number of hot-spot residue at interface using the prediction method proposed in Chapter 4. The number for specific interfaces is greater than that for non-specific interfaces (Table 5.2). The number of hot-spot interface residues ($N_{\text{hot-spot}}$) is correlated with MIA for specific interfaces, while it has little correlation for non-specific interfaces. Thus, it could discriminate specific and non-specific interfaces which are undistinguishable by MIA. It is found that MIA/MSA and $N_{\text{hot-spot}}$ are not correlated ($R^2=0.25$). They are also uncorrelated ($R^2<0.17$) to the three parameters applied in the study of Bahadur *et al.* (Bahadur et al. 2004). Thus, we selected them as the parameters for the discrimination between specific and non-specific interfaces.

5.3.2 Specific interfaces vs non-specific interfaces

Table 5.3 and Figure 5.3 summarize the results of a selection achieved by applying MIA/MSA and $N_{\text{hot-spot}}$ criteria to the 188 monomers and 122 homodimers. The diagram of Figure 5.3 is split into three regions (i.e. M, monomer; D, dimer; U, undecided). The

Chapter 5 Specific and non-specific interfaces

limits of the regions were selected manually and, with larger datasets, they could be refined to optimize the distinction between homodimers and crystal-packing monomers. Region M includes 81% of the monomers, identified correctly as monomers; region D includes 80% of the homodimers, identified correctly as dimers. Only 5% of the monomers and the homodimers are misclassified at that stage. The success rate is rather similar to the two parameters (Fraction of buried atoms and Non-polar interface area) developed by Bahadur *et al.* Region U contains the remaining 15% of the sample. We employed the parameter (RP with the cut-off 1.5) developed by Bahadur *et al.* to sort out the interfaces. 83% of the 46 proteins in region U are correctly assigned a quaternary structure. In addition, the RP criterion was employed to detect some misclassified proteins in M and D regions. The procedure has an overall rate of 96% for the monomers and 93% for the homodimers. Based on the three parameters 15 proteins were misclassified (Table 5.3), while 17 proteins were misclassified based on the parameters reported by Bahadur *et al.* Thus, the two parameters (MIA/MSA and $N_{\text{hot-spot}}$) are effective to distinguish between homodimers and crystal-packing interfaces.

Figure 5.3 Identification of homodimers and crystal-packing monomers. In a scatter plot of the fraction of buried surface area MIA/MSA versus $N_{\text{hot-spot}}$ (the number of predicted hot-spot interface residues). Region M stands for ‘monomer’, D for ‘dimer’ and U for ‘undecided’ (their definitions given in Table 5.3).

Chapter 5 Specific and non-specific interfaces

Table 5.3 Discriminating between homodimers and crystal-packing monomers

Protein	Classification ^a				Total
	Monomer ^b	Dimer ^c	Undecided ^d	Misclassified	
Crystal dimers ^e	78	8 (3)	17 (2)	5	103
Monomers with no 2-fold ^e	74	1 (1)	10 (1)	2	85
Total monomers ^e	152	9 (4)	27 (3)	7	188
Homodimers ^f	6 (3)	97	19 (5)	8	122
Total				15	310

^a Based on MIA/MSA (the fraction of monomer interface area to monomer surface area), $N_{\text{hot-spot}}$ (the number of hot-spot interface residues), and RP (residue propensity score) derived from the study of Bahadur et al. (Bahadur et al. 2004).

^b Region M of Figure 3: either ($\text{MIA/MSA} < 0.07$ and $N_{\text{hot-spot}} < 23$) or ($\text{MIA/MSA} < 0.17$ and $N_{\text{hot-spot}} < 13$).

^c Region D of Figure 3: either ($\text{MIA/MSA} > 0.17$) or ($N_{\text{hot-spot}} < 23$).

^d Region U of Figure 3: either ($0.07 < \text{MIA/MSA} < 0.17$ and $13 < N_{\text{hot-spot}} < 23$).

^e In parentheses, number of misclassified monomers with $\text{RP} > 1.5$. Crystal dimers misclassified in D region are 1b3j, 1qjp and 1xgs. Crystal dimers misclassified in U region are 1rne and 2erc. Monomers with no 2-fold misclassified in D region is 1qs8 and in U region is 1fvk_1.

^f In parentheses, number of misclassified monomers with $\text{RP} < 1.5$. Homodimers misclassified in M region are 1reg, 2ccy and 2sqc. Homodimers misclassified in U region are 1a3c, 1aa7, 1af5, 1bbh and 3sdh.

Chapter 5 Specific and non-specific interfaces

If the five parameters are combined and applied to the distinction, the overall success rate is 97% for specific interfaces and 99% for non-specific interfaces. Based on these parameters, five proteins are misclassified (i.e. homodimer: 1a3c, 1reg, 2sqc and 3sdh; crystal dimer: 1b3j). Their unusual properties suggest that the oligomeric state in solution of these proteins should be reinvestigated.

Figure 5.4 summarizes the results by applying the two parameters (MIA/MSA and Nhot-spot) for distinguishing between the 188 monomers and 70 heterodimers. 60% of heterodimers are correctly identified in D region, while 16% of heterodimers are misclassified as monomers in M region. The misclassified rate is less than that (25%) reported by Bahadur and coworkers (Bahadur et al. 2004). In addition, 89% of antigen-antibody complexes are located in M and U regions. It suggests that antigen-antibody complex interfaces have few hot-spot interface residues and the buried surface area is relatively small. It is consistent with the fact that antigen-antibody complex binding site is the most difficult for prediction in current protein docking algorithms (the methods to predict protein-protein binding sites).

Figure 5.4 Identification of heterodimers and crystal-packing monomers. In a scatter plot of the fraction of buried surface area MIA/MSA versus $N_{\text{hot-spot}}$. Regions M, D and U are defined as in Table 5.3.

5.4 Conclusion

We compare the geometric and physical-chemical properties of interfaces involved in specific and non-specific interfaces based on 3D structure. We found that the number of hot-spot residues for specific interfaces are more than that for non-specific interfaces. Specific interactions require burying more surface area for binding and stability. We develop two parameters (1. the number of hot-spot interface residues and 2. the fraction of monomer interface area to monomer surface area) to identify specific interfaces. A simple combination of the two parameters assigns the quaternary structure of 80% of the homodimers and 81% of the monomers in our dataset correctly. The two parameters also can efficiently identify heterodimers, although they are less powerful for antigen-antibody complex identification. Thus, both parameters could be used to distinguish specific interactions from non-specific interactions.

Chapter 6

Structural difference between 2S (2 state) and 3S (3 state) homodimers

6.1 Introduction

Equilibrium denaturation experiments (using temperature and chemical agents) are employed to analyze the unfolding of proteins. These studies are useful in understanding monomeric protein folding. Recently, such techniques have been used to study the mechanism of homodimer formation (Neet and Timm 1994). Dimer folding involves both intra-molecular and inter-molecular interaction, unlike monomer folding that involves only intra-molecular interaction. It is known that some dimers denature from native dimer to unfolded monomers with no thermodynamically stable intermediates, whereas others have folded intermediates during the process (Neet and Timm 1994; Mazzini et al. 2002; Mei et al. 2005). Based on the unfolding patterns, homodimers are known to exist in three different states. They are (1) two-state (2S), (2) three-state with dimeric intermediate (3SDI) and (3) three-state with monomeric intermediate (3SMI). 2S refers to $N_2 \leftrightarrow 2U$ mechanism, 3SDI refers to $N_2 \leftrightarrow I_2 \leftrightarrow 2U$ and 3SMI refers to $N_2 \leftrightarrow 2I \leftrightarrow 2U$, where N_2 is the native dimer state, I is the intermediate monomeric species, I_2 is the intermediate dimeric species, and U is unfolded monomeric state. 3SDI and 3SMI are commonly considered as three-state (3S). It is found that 2S interfaces are similar to protein cores and 3SMI interfaces resemble the monomer surfaces (Tsai et al.

Chapter 6 2S and 3S homodimers

1997). 2S and 3SMI dimerization were also studied by following the evolution of two identical 20-letter residue chains within the framework of a lattice model, using Monte Carlo simulations (Tiana and Broglia 2002). It is found that folding of 2S sequences depend on a significantly larger number of conserved amino acids than 3SMI sequences. The effects of the monomer and interface geometry on 2S and 3S association mechanism were also studied by the energetically minimally frustrated Gō model (Levy et al. 2004). It is found that the native protein 3D structure is the major factor that governs the choice of binding mechanism.

Mei and colleagues investigated the importance of 2S and 3S dimers using structural data and folding data (Mei et al. 2005). Apiyo and colleagues proposed (based on 13 oligomeric proteins) that small obligomers (molecular mass < 20 kDa) unfold through two-state (Apiyo et al. 2001). On the other hand, large obligomers (molecular mass > 35 kDa) unfold through oligomeric intermediate and those with intermediate size unfold through monomeric intermediate. Moreover, Levy and colleagues proposed (using 21 homodimers) that 2S and 3SMI dimers can be effectively classified based on the ratio of intra-molecular/inter-molecular contacts and interface hydrophobicity (Levy et al. 2004). Here, we created an extended dataset of 41 homodimers (2S: 25; 3SDI: 6; and 3SMI: 10) to design a methodology for the discrimination of 2S, 3SDI and 3SMI dimers using 3D structural properties.

6.2 Methods

6.2.1 Dataset creation

We created a dataset consisting of 41 homodimer structures (2S: 25; 3SDI: 6; and 3SMI: 10) from Protein Databank (PDB) (Berman et al. 2000). The unfolding pathways for these dimers observed using thermodynamic experiments were obtained from literature (Table 6.1). The selected dimers have at least 40 residues per monomer.

6.2.2 Analyses of 2S and 3S homodimers

Interface area

The solvent accessible surface area (ASA) was computed using the program NACCESS (Hubbard 1993). Dimeric interface area (IA) (Jones and Thornton 1996) was calculated based on the difference (Δ ASA) between ASA of both protomers, on their own and in complex. We then calculated subunit interface area (IA/2), due to the twofold symmetry of homodimer complexation.

Interior, interface and exterior residues

Homodimer residues were classified into three categories (interior, interface and exterior) based on relative ASA. The percentage relative ASA was obtained by dividing the accessible surface area by the total surface area of a side-chain in an extended conformation in the tripeptide GXG. Exterior residues were defined as having a relative ASA > 5%, interior residues were defined as having a relative ASA < 5% and interface

Chapter 6 2S and 3S homodimers

residues were defined satisfying the conditions $\Delta\text{ASA} > 1\text{\AA}^2$ & relative ASA < 5%. The 5% cut-off was optimized elsewhere by Miller *et al.* (Miller et al. 1987).

Fraction of interface to interior Hydrophobicity (F_{hp})

F_{hp} (Fraction of interface to interior hydrophobicity) was defined by the equation $(H_{inf}-H_{ext})/(H_{int}-H_{ext})$, where H_{int} is interior hydrophobicity, H_{inf} is interface hydrophobicity and H_{ext} is exterior hydrophobicity. The individual hydrophobicity values were calculated using the equation $\Sigma n_i h_i / \Sigma n_i$, where n_i is the number of residue type i and h_i is hydrophobicity value (based on SES (solvent excluded surface) & SAS (solvent accessible surface)) of type I, as described elsewhere (Pacios 2001).

Small and large homodimers

By definition, small homodimers were defined as those with ML (monomer length) less than the dataset mean length (185 residues). By definition, large homodimers were defined as those with ML larger than the dataset mean length (185 residues).

Homodimers with small and large IA/2

By definition, homodimers with small IA/2 were defined as those whose IA/2 is less than the dataset mean IA/2 (1424 $\text{\AA}^2$). By definition, homodimers with large IA/2 were defined as those whose IA/2 is larger than the dataset mean IA/2 (1424 $\text{\AA}^2$).

Table 6.1 Dataset of homodimeric proteins divided into three groups according to their unfolding pathways

PDB ID	Chain	Protein name	Cofactors	Source	ML (aa)	IA/2 (Å ²)	Hydrophobicity			Reference
							H _{int}	H _{inf}	H _{ext}	
2S (25)										
2cpg 1arr 1rop 5cro	A&B	transcriptional repressor CpG	-	Streptococcus agalactiae	45	1632	0.37	0.68	0.23	(Wales et al. 2004)
	A&B	arc repressor	-	Bacteriophage P22	53	1962	0.47	0.58	0.23	(Milla and Sauer 1994)
	(Sym)	repressor of protein Rop	-	E. coli	63	1345	0.41	0.51	0.24	(Steif et al. 1993)
	A&C	Cro repressor	-	Bacteriophage lambda	66	648	0.49	0.73	0.29	(Jana et al. 1997)
1bfm	A&B	Histone B	-	Methanothermus fervidus	69	1593	0.50	0.72	0.30	(Topping and Gloss 2004)
1a7g	(Sym)	E2 DNA-binding domain	-	HPV strain 16E2	82	918	0.6	0.52	0.29	(Mok et al. 1996)
1vqb	(Sym)	gene V protein	-	Bacteriophage f1	87	850	0.58	0.66	0.31	(Liang and Terwilliger 1991)
1b8z	A&B	histone-like protein HU	-	Thermotoga maritima	90	1894	0.26	0.67	0.23	(Ruiz-Sanz et al. 2004)
1ety	A&B	FIS protein	-	E. coli	98	2079	0.49	0.5	0.25	(Topping et al. 2004)
1y7q	A&B	SCAN domain of ZNF 174	-	Homo sapiens	98	1508	0.67	0.54	0.26	(Stone et al. 2002)
1a8g	A&B	HIV-1 protease	-	HIV type 1	99	1785	0.63	0.49	0.33	(Grant et al. 1992)
1siv	A&B	SIV protease	-	SIV	99	1684	0.59	0.53	0.33	(Grant et al. 1992)
1vub	A&B	CcdB	-	E. coli	101	1074	0.51	0.36	0.33	(Bajaj et al. 2004)
1cmb	A&B	Met repressor	-	E. coli	104	1813	0.35	0.54	0.27	(Johnson et al. 1992)
3ssi	(Sym)	subtilisin inhibitor	-	Streptomyces	108	866	0.51	0.46	0.32	(Tamura et al. 1995)
1wrp	(Sym)	trp repressor	-	albogriseolus	108	2243	0.69	0.54	0.29	(Gloss et al. 2001)
1bet	(Sym)	β-nerve growth factor	-	Mus musculus	107	1366	0.47	0.47	0.31	(Timm et al. 1994)
1buo	(Sym)	Btb domain from PLZF	-							(Li et al. 1997)
1oh0	A&B	protein ketosteroid isomerase	-	Homo sapiens	121	1972	0.56	0.41	0.28	
		class π glutathione	-	Pseudomonas putida	131	1036	0.49	0.41	0.31	(Kim et al. 2001)
2gsr	A&B	s-transferase	-	Sus scrofa	207	1331	0.5	0.3	0.27	(Dirr and Reinemer 1991)
1gsd	A&B	glutathione transferase A1-1	-	Homo sapiens	208	1477	0.57	0.4	0.27	(Wallace et al. 1998)

Chapter 6 2S and 3S homodimers

1ga	(Sym)	glutathione transferase	-	Mn & Ca ion	Schistosoma japonica	218	1505	0.5	0.42	0.26	(Kaplan et al. 1997)
2bqp	A&B	pea lectin triosephosphate isomerase	-		Garden pea	234	955	0.49	0.37	0.27	(Ahmad et al. 1998)
1hti	A&B	Nh(3)-dependent	-		Homo sapiens	248	1685	0.54	0.35	0.27	(Mainfroid et al. 1996)
1eel	A&B	Nad(+) synthetase	-		Bacillus subtilis	271	2507	0.51	0.43	0.24	(Yang et al. 2004)
3SDI (6)											
1mul	(Sym)	histone-like protein hu- α	-		E. coli	90	1739	0.47	0.61	0.3	(Ramstein et al. 2003)
1hqo	A&B	Ure2 Protein	-		Saccharomyces cerevisiae	258	1656	0.54	0.44	0.3	(Zhu et al. 2003)
1psc	A&B	parathion hydrolase 3-isopropylmalate dehydrogenase	-	Cd ion	Brevundimonas diminuta	329	1353	0.5	0.3	0.3	(Grimsley et al. 1997)
1cm7	A&B	ascorbate oxidase	-		E. coli	363	2317	0.5	0.47	0.28	(Motono et al. 1999)
1aoz	A&C	SecA	-	Cu ion	Green zucchini Mycobacterium tuberculosis	552	1817	0.43	0.47	0.29	(Mei et al. 1997)
1nl3	A&B		-			835	1351	0.46	0.64	0.28	(Doyle et al. 2000)
3SMI (10)											
1a43	(Sym)	C-terminal domain of HIV-1 capsid protein	-		HIV type 1	72	921	0.47	0.42	0.26	(Mateu 2002)
1ql	A&B	lysine-49 phospholipase A2	-		Bothrops jararacussu Desulfovibrio desulfuricans	121	432	0.38	0.17	0.27	(Ruller et al. 2003)
1dfx	(Sym)	desulfoferrodoxin cu, zn superoxide dismutase	-	Fe & Ca ion	Photobacterium leiognathi	125	1472	0.44	0.44	0.29	(Apiyo et al. 2001)
1yai	B&C	cu, zn superoxide dismutase	-	Cu & Zn ion	Homo sapiens	151	309	0.48	0.41	0.28	(Malvezzi-Campeggi et al. 1999)
1spd	A&B	cAMP receptor protein glutathione-s-transferase	-	Cu & Zn ion		154	658	0.49	0.4	0.28	(Stroppolo et al. 2000)
1run	A&B	tyrosyl-tRNA synthetase	-		E. coli	197	1542	0.66	0.47	0.28	(Malecki and Wasylewski 1997)
1lgs	A&B	prostatic acid	-		Homo sapiens	209	1197	0.5	0.28	0.3	(Aceto et al. 1992)
1tya	(Sym)		-		Bacillus stearothermophilus	319	1513	0.48	0.55	0.26	(Park and Bedouelle 1998)
1nd5	A&B		-		Homo sapiens	354	1512	0.43	0.44	0.27	(Wojciak et al. 2003)

Chapter 6 2S and 3S homodimers

2crk	(Sym)	phosphatase creatine kinase	-	Oryctolagus cuniculus	381	1119	0.46	0.18	0.25	(Liang et al. 2003)
Average for 2S										
SD							0.51	0.50	0.28	
							0.10	0.12	0.03	
Average for 3SDI										
SD							0.48	0.49	0.29	
							0.04	0.14	0.01	
Average for 3SMI										
SD							0.48	0.38	0.27	
							0.07	0.13	0.02	

ML=monomer length; IA/2 = subunit interface area. 2S=two-state; 3SDI=three-state with dimeric intermediate; 3SMI=three-state with monomeric intermediate. SIV = Simian immunodeficiency virus; HIV=Human immunodeficiency virus; HPV=Human papillomavirus; Ccldb = controller of cell division or death B protein; PLZF=promyelocytic leukemia zinc finger protein; FIS=factor for inversion stimulation. (sym) indicates that the dimer is generated from a single chain in the PDB by Protein Quaternary Structure Server (PQS)(Henrick and Thornton 1998). Interior hydrophobicity (H_{int}), interface hydrophobicity (H_{int}) and exterior hydrophobicity (H_{ext}) for each dimer were calculated, separately, by the single equation $\sum n_i h_i / \sum n_i$, where n_i is the number of residue type i and h_i is ASA hydrophobicity factor of residue type i from Pacios (Pacios 2001).

6.3 Results

6.3.1 Distribution of 2S and 3S on monomer length and subunit

interface area

Figure 6.1 shows the distribution of 2S and 3S in the Cartesian plane consisting of ML (monomer length) and IA/2 (subunit interface area). It shows that 76% of small proteins form 2S and 60% of large proteins form 3S homodimers. Figure 6.1 also shows that 68% of 2S have large interface area and 45% of 3S have small interface area. 2S have ML in the range of 45-270 residues and 3S have ML in the range of 70-850 residues. However, 3SMI lie within 90-380 residues and 3SDI lie within 70-850 residues. 2S and 3S dimers have significantly different ML range ($p = 0.05$ in F test). Nonetheless, 2S and 3SMI have similar ML range ($p = 0.05$ in F test). The dataset mean ML is 185 residues. This lies between 2S mean (125 residues) and 3S mean (282 residues). Data also show that 2S and 3S ML means are different ($p < 0.05$). The mean ML for 3SDI is 405 and this is much greater than the mean ML for 2S (125) and 3SMI (208).

The IA/2 range for 2S (650 - 2500 Å²) and 3S (300 - 2317 Å²) are overlapping and are not significantly different ($p = 0.21$). However, 3SMI and 3SDI are distinguished by the IA/2 range ($p < 0.05$). 3SMI having small IA/2 range (300-1550 Å²) and 3SDI having large IA/2 range (1350-2317 Å²) are distinguished from each other. The dataset mean for IA/2 is 1424 Å², which lies between 2S mean (1509 Å²) and 3S mean (1239 Å²).

Chapter 6 2S and 3S homodimers

Interestingly, the 3SMI mean ($1068 \text{ \AA}^2$) is close to 3S mean IA/2 ($p = 0.25$) and 3SDI mean ($1705 \text{ \AA}^2$) is close to 2S mean IA/2 ($p = 0.35$).

In Figure 6.1, the distribution of 2S and 3S were divided into four regions (G1 to G4) based on the dataset mean of ML and IA/2. Entries in G1 are small proteins with large IA/2 and entries in G4 are large proteins with small IA/2 (refer to methodology section for definition of small and large proteins). However, entries in G2 are small proteins with small IA/2 and those in G3 are large proteins with large IA/2. This grouping shows 84% of homodimers in G1 are 2S and 67% of homodimers in G4 are 3S. Nevertheless, homodimers in G3 there are 40% 2S and 60% 3S. Homodimers in G2 have 67% 2S and 33% 3S. It should be observed that 3S in G2 are solely 3SMI. The results show that 2S and 3S are distinctly and prevalently distinguished in G1 and G4 but not as much in G2 and G3. The distribution of 2S and 3S in regions G1 to G4 provide insight to their structural preference in terms of ML and IA/2.

Figure 6.1. Correlation between monomer length (ML) and subunit interface area (IA/2) for three groups of homodimers. 2S: two-state; 3SDI: three-state with dimeric intermediate; 3SMI: three-state with monomeric intermediate. The two dash lines through 185 aa and 1424Å² represent mean monomer length and mean IA/2 for all homodimers, respectively. They classify the dimers into four regions (G1, G2, G3 and G4). The distributions of 2S, 3SDI and 3SMI dimers are given for each region. The value within parentheses is hydrophobicity factor (F_{hp}), calculated by the equation (H_{int} - H_{ext})/(H_{int} - H_{ext}), where H_{int} is interface hydrophobicity, H_{int} is interior hydrophobicity and H_{ext} is exterior hydrophobicity.

6.3.2 Exterior, interior and interface hydrophobicity

Table 6.1 gives the hydrophobicity of interior, interface and exterior residues for 2S, 3SDI and 3SMI. It also gives the mean hydrophobicity of interior, interface and exterior residues for 2S, 3SDI and 3SMI in the dataset. Very small 2S (≤ 90 residues) have greater interface hydrophobicity compared to interior hydrophobicity. However, this is not true with very large 2S (> 90 residues). It is also interesting to observe that majority of 3SMI have less interface hydrophobicity compared to interior hydrophobicity. Nonetheless, this is not true with a majority of 3SDI. Table 6.1 shows that the mean interface hydrophobicity values satisfy a condition ($2S > 3SDI > 3SMI$). However, the mean interior hydrophobicity satisfy a different condition ($2S > (3SDI = 3SMI)$). The ratio of interface to interior hydrophobicity is ~ 1 for 2S and 3SDI, while it is < 1 for 3SMI.

6.3.3 Factor of interface to interior hydrophobicity (F_{hp}) value

Figure 6.1 shows that 92% of entries in G1 have high F_{hp} value (> 0.5) and 83% of entries in G4 have low F_{hp} value (< 0.5). It also shows that 3S in G1 have high F_{hp} value and 2S in G4 have low F_{hp} value. Interestingly, 75% of entries in G2 have high F_{hp} value and 78% of entries in G3 have high F_{hp} value. Moreover, Figure 6.1 show that 91% 2S in G1 have high F_{hp} value and 75% 3S in G4 have low F_{hp} value. However, 100% 3S (2 entries) in G1 have high F_{hp} value and 100% 2S (2 entries) in G4 have low F_{hp} value. In G2, 75% of 2S have high F_{hp} value and 67% of 3S have high F_{hp} value. Nonetheless, 100% 3S have high F_{hp} value and 50% of 2S have high F_{hp} value in G3. The mean F_{hp}

Chapter 6 2S and 3S homodimers

value for 2S and 3SDI is 1, while it is 0.5 for 3SMI. Thus, the distribution of 2S and 3S in the G1 to G4 regions is described.

6.3.4 Correlation of the three parameters

The three parameters (ML, IA/2 and F_{hp}) were employed to differentiate between three different groups of homodimers (2S, 3SMI and 3SDI). The correlation between every two parameters for each group was calculated (Table 6.2). The small correlation coefficients indicate that the three parameters were not correlated.

Table 6.2 Correlation coefficient (R^2) between parameters

	ML vs. F_{hp}	ML vs. IA/2	IA/2 vs. F_{hp}
2S homodimers	0.251	0.023	0.014
3SDI homodimers	0.130	0.089	0.001
3SMI homodimers	0.008	0.250	0.232

6.4 Discussion

The mechanism of homodimer folding and binding has been investigated using denaturation experiments. Three-dimensional structures are also available for many of these homodimers with known folding and binding mechanisms (Table 6.1). The folding and binding homodimer data collected from the literature is classified into three 2S, 3SMI and 3SDI. The study of homodimer folding and binding using energy models is computational intensive and time consuming. Alternatively, study on their folding and binding using structural data is found useful (Mei et al. 2005). Recently, Mei and colleagues documented the differences between 2S, 3SMI and 3SDI homodimers using 3S structure data (Mei et al. 2005). The study provided structural insight to the mechanism of 2S and 3S folding. However, the analysis did not document parameters to differentiate 2S, 3SMI and 3SDI homodimers using structural data. In this study, we study an extended dataset of homodimer complexes to distinguish 2S and 3S homodimers using structural features. Results show that 76% of small proteins are 2S homodimers and 60% of large proteins are 3S homodimers. Thus, protein size plays an important role in determining the pathways of homodimer folding and binding. The result also shows that 68% of 2S have large subunit interface area and 45% of 3S have small subunit interface area. These observations suggest the importance of protein size and subunit interface area in determining the mechanism of homodimer formation.

The distribution of 2S and 3S in the G1 and G4 regions of Figure 6.1 show difference between them based on protein size, subunit interface area and F_{hp} . In G1, 84% dimers

Chapter 6 2S and 3S homodimers

are 2S and 92% of dimers have high F_{hp} (> 0.5). Thus, entries with high F_{hp} are grouped in G1 and this region represents small proteins with large subunit interface area. Moreover, 91% of 2S in G1 have high F_{hp} . This implies that a majority of small proteins with large subunit interface area and high F_{hp} are 2S. 3S in G1 have high F_{hp} and this explains the presence of exceptional 3S entries in G1. Similarly, 66% of dimers are 3S and 83% of dimers have low F_{hp} (< 0.5) in G4. Thus, entries with low F_{hp} are grouped into G4 and this region represents large proteins with small subunit interface area. Furthermore, 75% 3S in G4 have low F_{hp} . 2S in G4 have low F_{hp} and this explains the presence of unusual 2S entries in G4. Entries in G2 and G3 have a mixture of 2S and 3S with low and high F_{hp} values. This is different to the distribution in G1 and G4. 100% 3S and 50% 2S in G3 have high F_{hp} and thus dimers in G3 are not distinguished by their folding mechanisms using structural parameters. The mean F_{hp} for 2S and 3SDI is 1, while it is 0.5 for 3SMI. The similarity between 2S and 3SDI in F_{hp} is interesting. It implies that binding after folding displayed by 3SMI resembles the association of protein-protein complexes (Lo Conte et al. 1999). However, the cooperative folding-binding displayed by 2S and 3SDI resembles a single-chain folding.

Thus, we show that small homodimers with large interface area and high F_{hp} are prevalently 2S. Similarly, large homodimers with small interface area and low F_{hp} are prevalently 3S. Hence, it is possible to distinguish 2S and 3S dimers using 3D structural data. However, small homodimers with small interface area and large homodimers with large interface area are not significantly distinguished into 2S and 3S using structural

Chapter 6 2S and 3S homodimers

parameters ML, IA/2 and F_{hp} . It should be noted that the conclusion made in the report are based on a limited set of homodimers given in Table 6.1.

6.5 Conclusion

The mechanisms of homodimer complexations have implications in drug discovery. However, elucidation of homodimer mechanism using unfolding experiments is difficult. Prediction of homodimer folding and binding using structural data has application in target validation. Here, we show that small proteins with large interface area and high F_{hp} form 2S. We also show that large proteins with small interface area and low F_{hp} form 3S. Therefore, it is feasible to differentiate 2S and 3S homodimers using structural data.

Chapter 7

Conclusions

The significance of interface ligands, sidechain-sidechain prevalence at the interfaces, “hot spot” interface residues and binding are discussed in the thesis using structural complexes. It is observed that dimer interfaces having ligands are less hydrophobic with small interface area compared to those without ligands and the effect of ligands occupying $\leq 7\%$ interface area is negligible on dimer interactions. In addition, protein-protein complexes are found to have more inter-atomic contacts across the interface, more ‘hot spot’ interface residues and larger fraction of interface area to surface area compared to non-specific complexes. Moreover, a simple ‘hot spot’ prediction approach is developed which has better accuracy than the prevailing ‘hot spot’ prediction methods. Also, the analysis of dimer binding (2S and 3S) shows that a combination of structural properties monomer length, subunit interface area and ratio of interface to interior hydrophobicity can predominately distinguish 2S and 3S dimers. These observations provide insight to protein-protein interactions and have utility in the development of models for interaction site/partner prediction.

List of Publications

1. 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.
2. Li, L., Gan, GK., Sakharkar, MK., Kanguane, P. 2005. Is the effect of ligands at homodimer interface on interface properties negligible? Asia-Pacific Bio-Informatics Conference. Singapore, 42-49.
3. 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.
4. Cui, ZH., Gan, GK., Li, L., Sakharkar, MK., Kanguane, P. 2005. Protein subunit interfaces: heterodimers versus homodimers. *Bioinformation* **1(2)**: 28-39.
5. 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.
6. Yu. YT, Li, L., Sakharkar, MK., Kanguane, P. 2006. Insight into gene fusion from molecular dynamics simulation of fused and un-fused IGPS (Imidazole Glycerol Phosphate Synthetase). *Bioinformation* **1(3)**: 99-104.
7. Li, L., Zhao, B., Gan, GK., Cui, ZH., Sakharkar, MK., Kanguane, P. Identification of hot-spot residues at protein-protein interface. 2006. *Bioinformation* **1(4)**:121-126.

References

- Aceto, A., Caccuri, A.M., Sacchetta, P., Bucciarelli, T., Dragani, B., Rosato, N., Federici, G., and Di Ilio, C. 1992. Dissociation and unfolding of Pi-class glutathione transferase. Evidence for a monomeric inactive intermediate. *Biochem J* **285** (Pt 1): 241-245.
- Ahmad, N., Srinivas, V.R., Reddy, G.B., and Surolia, A. 1998. Thermodynamic characterization of the conformational stability of the homodimeric protein, pea lectin. *Biochemistry* **37**: 16765-16772.
- Apiyo, D., Jones, K., Guidry, J., and Wittung-Stafshede, P. 2001. Equilibrium unfolding of dimeric desulfoferrodoxin involves a monomeric intermediate: iron cofactors dissociate after polypeptide unfolding. *Biochemistry* **40**: 4940-4948.
- Bahadur, R.P., Chakrabarti, P., Rodier, F., and Janin, J. 2003. Dissecting subunit interfaces in homodimeric proteins. *Proteins* **53**: 708-719.
- Bahadur, R.P., Chakrabarti, P., Rodier, F., and Janin, J. 2004. A dissection of specific and non-specific protein-protein interfaces. *J Mol Biol* **336**: 943-955.
- Bajaj, K., Chakshusmathi, G., Bachhawat-Sikder, K., Surolia, A., and Varadarajan, R. 2004. Thermodynamic characterization of monomeric and dimeric forms of CcdB (controller of cell division or death B protein). *Biochem J* **380**: 409-417.
- Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. 2000. The Protein Data Bank. *Nucleic Acids Res* **28**: 235-242.

References

-
- Bogan, A.A., and Thorn, K.S. 1998. Anatomy of hot spots in protein interfaces. *J Mol Biol* **280**: 1-9.
- Brinda, K.V., Kannan, N., and Vishveshwara, S. 2002. Analysis of homodimeric protein interfaces by graph-spectral methods. *Protein Eng* **15**: 265-277.
- Brinda, K.V., and Vishveshwara, S. 2005. Oligomeric protein structure networks: insights into protein-protein interactions. *BMC Bioinformatics* **6**: 296.
- Carugo, O., and Argos, P. 1997. Protein-protein crystal-packing contacts. *Protein Sci* **6**: 2261-2263.
- Castro, M.J., and Anderson, S. 1996. Alanine point-mutations in the reactive region of bovine pancreatic trypsin inhibitor: effects on the kinetics and thermodynamics of binding to beta-trypsin and alpha-chymotrypsin. *Biochemistry* **35**: 11435-11446.
- Chakrabarti, P., and Janin, J. 2002. Dissecting protein-protein recognition sites. *Proteins* **47**: 334-343.
- Chothia, C. 1976. The nature of the accessible and buried surfaces in proteins. *J Mol Biol* **105**: 1-12.
- Clackson, T., and Wells, J.A. 1995. A hot spot of binding energy in a hormone-receptor interface. *Science* **267**: 383-386.
- del Sol, A., Fujihashi, H., and O'Meara, P. 2005. Topology of small-world networks of protein-protein complex structures. *Bioinformatics* **21**: 1311-1315.
- del Sol, A., and O'Meara, P. 2005. Small-world network approach to identify key residues in protein-protein interaction. *Proteins* **58**: 672-682.

References

- Dirr, H.W., and Reinemer, P. 1991. Equilibrium unfolding of class pi glutathione S-transferase. *Biochem Biophys Res Commun* **180**: 294-300.
- Doyle, S.M., Braswell, E.H., and Teschke, C.M. 2000. SecA folds via a dimeric intermediate. *Biochemistry* **39**: 11667-11676.
- Fields, S., and Sternglanz, R. 1994. The two-hybrid system: an assay for protein-protein interactions. *Trends Genet* **10**: 286-292.
- Gao, Y., Wang, R., and Lai, L. 2004. Structure-based method for analyzing protein-protein interfaces. *J Mol Model (Online)* **10**: 44-54.
- Glaser, F., Steinberg, D.M., Vakser, I.A., and Ben-Tal, N. 2001. Residue frequencies and pairing preferences at protein-protein interfaces. *Proteins* **43**: 89-102.
- Gloss, L.M., Simler, B.R., and Matthews, C.R. 2001. Rough energy landscapes in protein folding: dimeric E. coli Trp repressor folds through three parallel channels. *J Mol Biol* **312**: 1121-1134.
- Grant, S.K., Deckman, I.C., Culp, J.S., Minnich, M.D., Brooks, I.S., Hensley, P., Debouck, C., and Meek, T.D. 1992. Use of protein unfolding studies to determine the conformational and dimeric stabilities of HIV-1 and SIV proteases. *Biochemistry* **31**: 9491-9501.
- Grimsley, J.K., Scholtz, J.M., Pace, C.N., and Wild, J.R. 1997. Organophosphorus hydrolase is a remarkably stable enzyme that unfolds through a homodimeric intermediate. *Biochemistry* **36**: 14366-14374.
- Guerois, R., Nielsen, J.E., and Serrano, L. 2002. Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. *J Mol Biol* **320**: 369-387.

References

-
- Haliloglu, T., Keskin, O., Ma, B., and Nussinov, R. 2005. How similar are protein folding and protein binding nuclei? Examination of vibrational motions of energy hot spots and conserved residues. *Biophys J* **88**: 1552-1559.
- Halperin, I., Wolfson, H., and Nussinov, R. 2004. Protein-protein interactions; coupling of structurally conserved residues and of hot spots across interfaces. Implications for docking. *Structure (Camb)* **12**: 1027-1038.
- Henrick, K., and Thornton, J.M. 1998. PQS: a protein quaternary structure file server. *Trends Biochem Sci* **23**: 358-361.
- Hubbard. 1993. 'NACCESS', computer program, Department of biochemistry and Molecular Biology, University College London.
- Jackson, R.M. 1999. Comparison of protein-protein interactions in serine protease-inhibitor and antibody-antigen complexes: implications for the protein docking problem. *Protein Sci* **8**: 603-613.
- Jana, R., Hazbun, T.R., Mollah, A.K., and Mossing, M.C. 1997. A folded monomeric intermediate in the formation of lambda Cro dimer-DNA complexes. *J Mol Biol* **273**: 402-416.
- Janin, J. 1997. Specific versus non-specific contacts in protein crystals. *Nat Struct Biol* **4**: 973-974.
- Janin, J., Henrick, K., Moult, J., Eyck, L.T., Sternberg, M.J., Vajda, S., Vakser, I., and Wodak, S.J. 2003. CAPRI: a Critical Assessment of PRedicted Interactions. *Proteins* **52**: 2-9.
- Janin, J., and Rodier, F. 1995. Protein-protein interaction at crystal contacts. *Proteins* **23**: 580-587.

References

- Jeong, H., Mason, S.P., Barabasi, A.L., and Oltvai, Z.N. 2001. Lethality and centrality in protein networks. *Nature* **411**: 41-42.
- Johnson, C.M., Cooper, A., and Stockley, P.G. 1992. Differential scanning calorimetry of thermal unfolding of the methionine repressor protein (MetJ) from *Escherichia coli*. *Biochemistry* **31**: 9717-9724.
- Jones, S., and Thornton, J.M. 1995. Protein-protein interactions: a review of protein dimer structures. *Prog Biophys Mol Biol* **63**: 31-65.
- Jones, S., and Thornton, J.M. 1996. Principles of protein-protein interactions. *Proc Natl Acad Sci U S A* **93**: 13-20.
- Jones, S., and Thornton, J.M. 1997. Analysis of protein-protein interaction sites using surface patches. *J Mol Biol* **272**: 121-132.
- Kaplan, W., Husler, P., Klump, H., Erhardt, J., Sluis-Cremer, N., and Dirr, H. 1997. Conformational stability of pGEX-expressed *Schistosoma japonicum* glutathione S-transferase: a detoxification enzyme and fusion-protein affinity tag. *Protein Sci* **6**: 399-406.
- Kelley, R.F., Costas, K.E., O'Connell, M.P., and Lazarus, R.A. 1995. Analysis of the factor VIIa binding site on human tissue factor: effects of tissue factor mutations on the kinetics and thermodynamics of binding. *Biochemistry* **34**: 10383-10392.
- Kim, D.H., Nam, G.H., Jang, D.S., Yun, S., Choi, G., Lee, H.C., and Choi, K.Y. 2001. Roles of dimerization in folding and stability of ketosteroid isomerase from *Pseudomonas putida* biotype B. *Protein Sci* **10**: 741-752.
- Kortemme, T., and Baker, D. 2002. A simple physical model for binding energy hot spots in protein-protein complexes. *Proc Natl Acad Sci U S A* **99**: 14116-14121.

References

-
- Kortemme, T., Kim, D.E., and Baker, D. 2004. Computational alanine scanning of protein-protein interfaces. *Sci STKE* **2004**: pl2.
- Kuhlmann, U.C., Pommer, A.J., Moore, G.R., James, R., and Kleanthous, C. 2000. Specificity in protein-protein interactions: the structural basis for dual recognition in endonuclease colicin-immunity protein complexes. *J Mol Biol* **301**: 1163-1178.
- Leder, L., Llera, A., Lavoie, P.M., Lebedeva, M.I., Li, H., Sekaly, R.P., Bohach, G.A., Gahr, P.J., Schlievert, P.M., Karjalainen, K., et al. 1998. A mutational analysis of the binding of staphylococcal enterotoxins B and C3 to the T cell receptor beta chain and major histocompatibility complex class II. *J Exp Med* **187**: 823-833.
- Levy, Y., Wolynes, P.G., and Onuchic, J.N. 2004. Protein topology determines binding mechanism. *Proc Natl Acad Sci U S A* **101**: 511-516.
- Li, X., Keskin, O., Ma, B., Nussinov, R., and Liang, J. 2004. Protein-Protein Interactions: Hot Spots and Structurally Conserved Residues often Locate in Complemented Pockets that Pre-organized in the Unbound States: Implications for Docking. *J Mol Biol* **344**: 781-795.
- Li, X., Lopez-Guisa, J.M., Ninan, N., Weiner, E.J., Rauscher, F.J., 3rd, and Marmorstein, R. 1997. Overexpression, purification, characterization, and crystallization of the BTB/POZ domain from the PLZF oncoprotein. *J Biol Chem* **272**: 27324-27329.
- Liang, H., and Terwilliger, T.C. 1991. Reversible denaturation of the gene V protein of bacteriophage f1. *Biochemistry* **30**: 2772-2782.

References

- Liang, Y., Du, F., Sanglier, S., Zhou, B.R., Xia, Y., Van Dorselaer, A., Maechling, C., Kilhoffer, M.C., and Haiech, J. 2003. Unfolding of rabbit muscle creatine kinase induced by acid. A study using electrospray ionization mass spectrometry, isothermal titration calorimetry, and fluorescence spectroscopy. *J Biol Chem* **278**: 30098-30105.
- Lo Conte, L., Chothia, C., and Janin, J. 1999. The atomic structure of protein-protein recognition sites. *J Mol Biol* **285**: 2177-2198.
- Ma, B., Elkayam, T., Wolfson, H., and Nussinov, R. 2003. Protein-protein interactions: structurally conserved residues distinguish between binding sites and exposed protein surfaces. *Proc Natl Acad Sci U S A* **100**: 5772-5777.
- Mainfroid, V., Terpstra, P., Beauregard, M., Frere, J.M., Mande, S.C., Hol, W.G., Martial, J.A., and Goraj, K. 1996. Three hTIM mutants that provide new insights on why TIM is a dimer. *J Mol Biol* **257**: 441-456.
- Malecki, J., and Wasylewski, Z. 1997. Stability and kinetics of unfolding and refolding of cAMP receptor protein from Escherichia coli. *Eur J Biochem* **243**: 660-669.
- Malvezzi-Campeggi, F., Stroppolo, M.E., Mei, G., Rosato, N., and Desideri, A. 1999. Evidence of stable monomeric species in the unfolding of Cu,Zn superoxide dismutase from Photobacterium leiognathi. *Arch Biochem Biophys* **370**: 201-207.
- Marcotte, E.M., Pellegrini, M., Thompson, M.J., Yeates, T.O., and Eisenberg, D. 1999. A combined algorithm for genome-wide prediction of protein function. *Nature* **402**: 83-86.

References

-
- Massova, I., and Kollman, P.A. 1999. Computational alanine scanning to probe protein-protein interactions: A novel approach to evaluate binding free energies. *J. Am. Chem. Soc.* **121**: 8133-8143.
- Mateu, M.G. 2002. Conformational stability of dimeric and monomeric forms of the C-terminal domain of human immunodeficiency virus-1 capsid protein. *J Mol Biol* **318**: 519-531.
- Mazzini, A., Maia, A., Parisi, M., Sorbi, R.T., Ramoni, R., Grolli, S., and Favilla, R. 2002. Reversible unfolding of bovine odorant binding protein induced by guanidinium hydrochloride at neutral pH. *Biochim Biophys Acta* **1599**: 90-101.
- Mei, G., Di Venere, A., Buganza, M., Vecchini, P., Rosato, N., and Finazzi-Agro, A. 1997. Role of quaternary structure in the stability of dimeric proteins: the case of ascorbate oxidase. *Biochemistry* **36**: 10917-10922.
- Mei, G., Di Venere, A., Rosato, N., and Finazzi-Agro, A. 2005. The importance of being dimeric. *Febs J* **272**: 16-27.
- Mika, S., and Rost, B. 2003. UniqueProt: Creating representative protein sequence sets. *Nucleic Acids Res* **31**: 3789-3791.
- Milla, M.E., and Sauer, R.T. 1994. P22 Arc repressor: folding kinetics of a single-domain, dimeric protein. *Biochemistry* **33**: 1125-1133.
- Miller, B.G., Hassell, A.M., Wolfenden, R., Milburn, M.V., and Short, S.A. 2000. Anatomy of a proficient enzyme: the structure of orotidine 5'-monophosphate decarboxylase in the presence and absence of a potential transition state analog. *Proc Natl Acad Sci U S A* **97**: 2011-2016.

References

- Miller, S., Lesk, A.M., Janin, J., and Chothia, C. 1987. The accessible surface area and stability of oligomeric proteins. *Nature* **328**: 834-836.
- Mintseris, J., and Weng, Z. 2003. Atomic contact vectors in protein-protein recognition. *Proteins* **53**: 629-639.
- Mok, Y.K., de Prat Gay, G., Butler, P.J., and Bycroft, M. 1996. Equilibrium dissociation and unfolding of the dimeric human papillomavirus strain-16 E2 DNA-binding domain. *Protein Sci* **5**: 310-319.
- Moont, G., Gabb, H.A., and Sternberg, M.J. 1999. Use of pair potentials across protein interfaces in screening predicted docked complexes. *Proteins* **35**: 364-373.
- Motono, C., Yamagishi, A., and Oshima, T. 1999. Urea-induced unfolding and conformational stability of 3-isopropylmalate dehydrogenase from the Thermophile thermus thermophilus and its mesophilic counterpart from Escherichia coli. *Biochemistry* **38**: 1332-1337.
- Neet, K.E., and Timm, D.E. 1994. Conformational stability of dimeric proteins: quantitative studies by equilibrium denaturation. *Protein Sci* **3**: 2167-2174.
- Nooren, I.M., and Thornton, J.M. 2003. Structural characterisation and functional significance of transient protein-protein interactions. *J Mol Biol* **325**: 991-1018.
- Ofran, Y., and Rost, B. 2003. Analysing six types of protein-protein interfaces. *J Mol Biol* **325**: 377-387.
- Pacios, L.F. 2001. Distinct molecular surfaces and hydrophobicity of amino acid residues in proteins. *J Chem Inf Comput Sci* **41**: 1427-1435.

References

-
- Park, Y.C., and Bedouelle, H. 1998. Dimeric tyrosyl-tRNA synthetase from *Bacillus stearothermophilus* unfolds through a monomeric intermediate. A quantitative analysis under equilibrium conditions. *J Biol Chem* **273**: 18052-18059.
- Phizicky, E.M., and Fields, S. 1995. Protein-protein interactions: methods for detection and analysis. *Microbiol Rev* **59**: 94-123.
- Ponstingl, H., Henrick, K., and Thornton, J.M. 2000. Discriminating between homodimeric and monomeric proteins in the crystalline state. *Proteins* **41**: 47-57.
- Ramstein, J., Hervouet, N., Coste, F., Zelwer, C., Oberto, J., and Castaing, B. 2003. Evidence of a thermal unfolding dimeric intermediate for the *Escherichia coli* histone-like HU proteins: thermodynamics and structure. *J Mol Biol* **331**: 101-121.
- Reichmann, D., Rahat, O., Albeck, S., Meged, R., Dym, O., and Schreiber, G. 2005. The modular architecture of protein-protein binding interfaces. *Proc Natl Acad Sci U S A* **102**: 57-62.
- Ruiz-Sanz, J., Filimonov, V.V., Christodoulou, E., Vorgias, C.E., and Mateo, P.L. 2004. Thermodynamic analysis of the unfolding and stability of the dimeric DNA-binding protein HU from the hyperthermophilic eubacterium *Thermotoga maritima* and its E34D mutant. *Eur J Biochem* **271**: 1497-1507.
- Ruller, R., Ferreira, T.L., de Oliveira, A.H., and Ward, R.J. 2003. Chemical denaturation of a homodimeric lysine-49 phospholipase A2: a stable dimer interface and a native monomeric intermediate. *Arch Biochem Biophys* **411**: 112-120.

References

-
- Scheidig, A.J., Hynes, T.R., Pelletier, L.A., Wells, J.A., and Kossiakoff, A.A. 1997. Crystal structures of bovine chymotrypsin and trypsin complexed to the inhibitor domain of Alzheimer's amyloid beta-protein precursor (APPI) and basic pancreatic trypsin inhibitor (BPTI): engineering of inhibitors with altered specificities. *Protein Sci* **6**: 1806-1824.
- Sobolev, V., Wade, R.C., Vriend, G., and Edelman, M. 1996. Molecular docking using surface complementarity. *Proteins* **25**: 120-129.
- Steif, C., Weber, P., Hinz, H.J., Flossdorf, J., Cesareni, G., and Kokkinidis, M. 1993. Subunit interactions provide a significant contribution to the stability of the dimeric four-alpha-helical-bundle protein ROP. *Biochemistry* **32**: 3867-3876.
- Stone, J.R., Maki, J.L., Blacklow, S.C., and Collins, T. 2002. The SCAN domain of ZNF174 is a dimer. *J Biol Chem* **277**: 5448-5452.
- Stroppolo, M.E., Malvezzi-Campeggi, F., Mei, G., Rosato, N., and Desideri, A. 2000. Role of the tertiary and quaternary structures in the stability of dimeric copper, zinc superoxide dismutases. *Arch Biochem Biophys* **377**: 215-218.
- Tamura, A., Kojima, S., Miura, K., and Sturtevant, J.M. 1995. A thermodynamic study of mutant forms of Streptomyces subtilisin inhibitor. II. Replacements at the interface of dimer formation, Val13. *J Mol Biol* **249**: 636-645.
- Thorn, K.S., and Bogan, A.A. 2001. ASEdb: a database of alanine mutations and their effects on the free energy of binding in protein interactions. *Bioinformatics* **17**: 284-285.
- Tiana, G., and Broglia, R.A. 2002. Design and folding of dimeric proteins. *Proteins* **49**: 82-94.

References

- Timm, D.E., de Haseth, P.L., and Neet, K.E. 1994. Comparative equilibrium denaturation studies of the neurotrophins: nerve growth factor, brain-derived neurotrophic factor, neurotrophin 3, and neurotrophin 4/5. *Biochemistry* **33**: 4667-4676.
- Tong, A.H., Drees, B., Nardelli, G., Bader, G.D., Brannetti, B., Castagnoli, L., Evangelista, M., Ferracuti, S., Nelson, B., Paoluzi, S., et al. 2002. A combined experimental and computational strategy to define protein interaction networks for peptide recognition modules. *Science* **295**: 321-324.
- Topping, T.B., and Gloss, L.M. 2004. Stability and folding mechanism of mesophilic, thermophilic and hyperthermophilic archaeal histones: the importance of folding intermediates. *J Mol Biol* **342**: 247-260.
- Topping, T.B., Hoch, D.A., and Gloss, L.M. 2004. Folding mechanism of FIS, the intertwined, dimeric factor for inversion stimulation. *J Mol Biol* **335**: 1065-1081.
- Tsai, C.J., Xu, D., and Nussinov, R. 1997. Structural motifs at protein-protein interfaces: protein cores versus two-state and three-state model complexes. *Protein Sci* **6**: 1793-1805.
- Tsai, J., Taylor, R., Chothia, C., and Gerstein, M. 1999. The packing density in proteins: standard radii and volumes. *J Mol Biol* **290**: 253-266.
- Verkhivker, G.M., Bouzida, D., Gehlhaar, D.K., Rejto, P.A., Freer, S.T., and Rose, P.W. 2002. Monte Carlo simulations of the peptide recognition at the consensus binding site of the constant fragment of human immunoglobulin G: the energy

References

- hr/>
- landscape analysis of a hot spot at the intermolecular interface. *Proteins* **48**: 539-557.
- Wales, T.E., Richardson, J.S., and Fitzgerald, M.C. 2004. Facile chemical synthesis and equilibrium unfolding properties of CopG. *Protein Sci* **13**: 1918-1926.
- Wallace, L.A., Sluis-Cremer, N., and Dirr, H.W. 1998. Equilibrium and kinetic unfolding properties of dimeric human glutathione transferase A1-1. *Biochemistry* **37**: 5320-5328.
- Wells, J.A. 1991. Systematic mutational analyses of protein-protein interfaces. *Methods Enzymol* **202**: 390-411.
- Wojciak, P., Mazurkiewicz, A., Bakalova, A., and Kuciel, R. 2003. Equilibrium unfolding of dimeric human prostatic acid phosphatase involves an inactive monomeric intermediate. *Int J Biol Macromol* **32**: 43-54.
- Xu, D., Tsai, C.J., and Nussinov, R. 1997. Hydrogen bonds and salt bridges across protein-protein interfaces. *Protein Eng* **10**: 999-1012.
- Yang, Z.W., Tendian, S.W., Carson, W.M., Brouillette, W.J., Delucas, L.J., and Brouillette, C.G. 2004. Dimethyl sulfoxide at 2.5% (v/v) alters the structural cooperativity and unfolding mechanism of dimeric bacterial NAD⁺ synthetase. *Protein Sci* **13**: 830-841.
- Zhu, L., Zhang, X.J., Wang, L.Y., Zhou, J.M., and Perrett, S. 2003. Relationship between stability of folding intermediates and amyloid formation for the yeast prion Ure2p: a quantitative analysis of the effects of pH and buffer system. *J Mol Biol* **328**: 235-254.