Computational analysis on protein binding and interactions-a water exclusion perspective

Liu, Qian

2012

Liu, Q. (2012). Computational analysis on protein binding and interactions-a water exclusion perspective. Doctoral thesis, Nanyang Technological University, Singapore.

<https://hdl.handle.net/10356/50692>

<https://doi.org/10.32657/10356/50692>

NANYANG TECHNOLOGICAL UNIVERSITY

**Computational Analysis on Protein
Binding and Interactions
— A Water Exclusion Perspective**

A thesis
Submitted to the School of Computer Engineering
of Nanyang Technological University

by

Qian Liu (G0702606G)

in fulfillment of the requirement
for the Degree of Doctor of Philosophy

2012

Abstract

Protein binding plays a fundamental role in biological processes and molecular functions. In this thesis, we introduce a **‘water exclusion’ perspective and a new contact definition** to conduct structural analysis on protein binding at both residue and atomic levels. In particular, our analysis is concentrated on binding hot spots in interfaces.

We first propose the hypothesis of ‘double water exclusion’ (DWE) to model binding hot spots. This hypothesis is intended to refine the seminal O-ring theory, an influential proposition stating that binding hot spots at protein interfaces are often surrounded by a ring of energetically less important residues. Beside this ring topology, DWE also describes the organizational topology of the ring-inside, energetically more important hot spot residues; this topology is not specified by the O-ring theory. Furthermore, DWE strengthens the ‘hot region’ principle and the ‘coupling’ proposition of binding hot spots.

At a residue level, the requirements for a cluster of residues to form a hot spot under the DWE hypothesis can be mathematically satisfied by a biclique subgraph of residues in a pair of interacting proteins. In our evaluation according to ASEdb (a database storing hot spots determined by wet-lab experiments), we found that these DWE bicliques are rich in true hot spot residues. Further, we suggest a Z-score method to assess the biological significance of DWE bicliques and binding hot spots. This Z-score notion is designed to capture the probability of a contact residue occurring in or contributing to protein binding interfaces. Z-score is integrated by the water exclusion hypothesis in DWE and uses crystal packing as the reference state; crystal packing is first used in pairwise potential calculation. Our evaluation shows that Z-score outperforms earlier methods for classifying binding hot spots from non-hot spots.

To study physicochemical properties of DWE hot spots and bicliques, we dissect DWE bicliques on a non-redundant dataset with 867 interactions with three different types of protein interactions. Our investigation on these DWE bicliques suggests that biological interactions have unique DWE bicliques compared to crystal packing, indicating the specificity of biological binding behaviors. Further, the unique DWE bicliques to the obligate and to non-obligate interactions also illustrate the different binding behaviors of these two types of interactions. Specifically, residues and their pairs in these DWE bicliques are found to have diverse preference for these three types of interactions.

As binding hot spots lie in the core of the stability for protein interactions, we believe that residue pairs satisfying the DWE hypothesis can sharpen the distinction of different types of protein interactions. Thus, we introduce low-ASA residue pairs and use their propensity vectors to design OringPV for distinguishing the three types of protein interactions. A low-ASA residue pair is defined as two contact residues each from one chain that have a small solvent accessible surface area (ASA). This notion of residue pairs is novel, as it first combines residue pairs with the O-ring theory and DWE. In classifying the three types of protein interactions, our OringPV method shows the superior classification ability to existing methods, such as NOXclass and DiMoVo.

These above interfacial analyses focus on residues and their contacts which are obviously based on contacts between atoms. Thus, we further dissect protein binding at an atomic level. In order to have a precise model of protein interfaces, we propose a new definition of β **contacts** for characterizing atomic organization and atomic contacts in protein 3D structures. Different from distance-cutoff contacts in previous works, β contacts have β -skeletons as their foundation, and a β contact between two atoms i and j requires that there is no other atom in a forbidden region fr between i and j . fr is similar to van der Waals spheres of atoms. In our analysis, we found that β contacts are only a small fraction of distance-cutoff atomic contacts. We then take atomic contact vector (ACV) of β contacts (a β ACV vector) to represent an interface, and use them to design β ACV for classifying homodimers from crystal packing. On four datasets, β contacts demonstrate superior classification performance, suggesting that β contacts are much more important than other distance-cutoff contacts in protein binding.

We also use β contacts to design β ACV_{ASA} for predicting binding free energy and binding hot spots. β ACV_{ASA} integrates the DWE water exclusion hypothesis into a β ACV vector to represent an alanine mutation. In our evaluation on 396 alanine mutations, β ACV_{ASA} achieves the superior prediction performance to other methods, which confirms that (i) β contacts can truly better identify critical atomic contacts than distance-based contacts, (ii) water exclusion is a necessary condition of binding hot spots, and (iii) the 20 standard amino acids have their own way of contributing to protein binding.

Finally, we apply several predictive methods, including our proposed methods, to investigate why the 2D1 antibody can cross-react with 1918 H1N1 and 2009 H1N1 and cannot with 2007 H1N1. We found that the mutations of HA1 from 1918 H1N1 to 2009 H1N1 do not harm 2D1 binding, and there are several common hot spots between the 2D1 binding to 1918 H1N1 (2D1-1918H1N1) and to 2009 H1N1 (2D1-2009H1N1). Meanwhile, β ACV also predicts these two 2D1 interactions are true biological binding. All of these are computational evidences that explain why 2D1 can cross-neutralize these two H1N1 viruses. In contrast, β ACV predicts the 2D1 binding to 2007 H1N1 (2D1-2007H1N1) as false binding. Furthermore, more than half of the predicted hot spots

in 2D1-1918H1N1 and 2D1-2009H1N1 are predicted not to greatly contribute to 2D1-2007H1N1. This explains why 2D1 cannot neutralize 2007 H1N1. This case study also suggests that both DWE and β contacts provide promising models for protein binding, which open new directions for dissection of protein folding and binding in future work.

Acknowledgments

I would like to express sincerest appreciation to my advisors, Dr. Jinyan Li and Assistant Professor Steven Chu-Hong Hoi. Dr. Li offered me this opportunity to pursue Ph.D in NTU. He gave his invaluable advice and guidance to enlighten me on deeper insight into my research topics; his seriousness in examining results and writing skills has been impressing on me. A/Prof. Hoi shares his knowledge in data mining, and gives me consistent supervision on my research works. Without the encouragement and support from Dr. Li and A/Prof. Hoi, my research could not have been here.

I am also grateful to Professor Limsoon Wong in NUS and Associate Professor Chee-Keong Kwoh who give me great help in my works.

During my study in NTU, I met many trustworthy friends such as Song Liu, Yongjin Li, Haiyan Piao, Divya and Gang Li. I thank them for their help. In particular, I wish to thank Dr. Mundra Piyushkumar A., Zhenhua Li, Shamima Banu Binte SM Rashid, Chinh Su Tran To and Liang Zhao for their useful discussion and comments in my research.

I am greatly indebted to Nanyang Technological University (NTU) and School of Computer Engineering (SCE) for financial support in pursuing my doctoral degree. I also want to express my appreciation to the staff at BIRC, especially Tan Sing Yau, for their technical support and assistance.

Finally, I want to thank my family for their continuous love, support and encouragement all along. I fail words to express my gratitude to them for all they have done for me.

Author's Publications

Journal Papers

- 1 Jinyan Li, **Qian Liu**: Double Water Exclusion: A Hypothesis Refining the O-ring Theory for the Hot Spots at Protein Interfaces. *Bioinformatics* 2009, 25(6):743—750.
- 2 **Qian Liu**, Jinyan Li: Propensity Vectors of Low-ASA Residue Pairs in the Distinction of Protein Interactions. *Proteins: Structure, Function and Bioinformatics* 2009, 78(3):589—602.
- 3 **Qian Liu**, Jinyan Li: Protein Binding Hot Spots and the Residue-Residue Pairing Preference: A Water Exclusion Perspective. *BMC Bioinformatics* 2010, 11:244.
- 4 **Qian Liu**, Steven C. H. Hoi, Su T. T. Chinh, Zhenhua Li, Chee-Keong Kwoh, Limsoon Wong and Jinyan Li: Structural Analysis of the Hot Spots in the Binding between H1N1 HA and the 2D1 Antibody: Do Mutations of H1N1 from 1918 to 2009 Affect Much on this Binding? Accepted by *Bioinformatics* 2011.
- 5 **Qian Liu**, Limsoon Wong, Jinyan Li: Z-score Biological Significance of Binding Hot Spots of Protein Interfaces by Using Crystal Packing as the Reference State. Accepted by *BBA-Proteins and Proteomics (SI:Comp. Methods for Protein Interaction)*.
- 6 **Qian Liu**, Chee Keong Kwoh, Steven C. H. Hoi: β Atomic Contacts: Identifying Critical Specific Contacts in Protein Binding Interfaces. *PLoS ONE* (Second submission).
- 7 **Qian Liu**, Chee Keong Kwoh, Jinyan Li, Limsoon Wong, Steven C. H. Hoi: Prediction of Binding Free Energy and Binding Hot Spots by Integrating Water Exclusion into β Atomic Contacts. *Bioinformatics* (to be submitted).

Conference Papers

- 8 **Qian Liu**, Yi-Ping Phoebe Chen, Jinyan Li: High Functional Coherence in k-Partite Protein Cliques of Protein Interaction Networks. *Bioinformatics and Biomedicine, 2009, Washington D.C., USA*, 111-117.
- 9 Jinyan Li, **Qian Liu**, Tao Zeng: Negative Correlations in Collaboration: Concepts and Algorithms. *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, July 25-28, 2010*, 463—472.
- 10 **Qian Liu**, Limsoon Wong, Jinyan Li: Z-score Biological Significance of Binding Hot Spots of Protein Interfaces by Using Crystal Packing as the Reference State. Accepted by *The Tenth Asia Pacific Bioinformatics Conference, Melbourne, Australia, 17-19 January 2012*.

Abbreviation and Symbol

ASA	solvent <u>A</u> ccessible <u>S</u> urface <u>A</u> rea
3D	Three(<u>3</u>)- <u>D</u> imensional
PDB	<u>P</u> rotein <u>D</u> ata <u>B</u> ank
DWE	<u>D</u> ouble <u>W</u> ater <u>E</u> xclusion
idR	<u>i</u> nteraction- <u>d</u> ominated <u>R</u> esidue
idRP	<u>i</u> nteraction- <u>d</u> ominated <u>R</u> esidue <u>P</u> air
ipR	<u>i</u> nteraction- <u>p</u> referred <u>R</u> esidue
ipRP	<u>i</u> nteraction- <u>p</u> referred <u>R</u> esidue <u>P</u> air
OringPV	learning by <u>P</u> ropensity <u>V</u> ectors of low-ASA residue pairs in <u>O</u> -ring-surrounded regions
LOOCV	<u>L</u> ease- <u>O</u> ne- <u>O</u> ut <u>C</u> ross- <u>V</u> alidation
PV	<u>P</u> ropensity <u>V</u> ector
FV	<u>F</u> requency <u>V</u> ector
BV	<u>B</u> inary <u>V</u> ector
ACV	<u>A</u> tomtic <u>C</u> ontact <u>V</u> ector
β ACV	β <u>A</u> tomtic <u>C</u> ontact <u>V</u> ector
dACV	<u>D</u> istance-based <u>A</u> tomtic <u>C</u> ontact <u>V</u> ector
β ACV _{ASA}	β <u>A</u> tomtic <u>C</u> ontact <u>V</u> ector with <u>ASA</u> integration

Contents

Abstract	i
Acknowledgments	v
Author's Publications	vii
Abbreviation and Symbol	ix
List of Tables	xvii
List of Figures	xx
1 Introduction	1
1.1 Background	1
1.1.1 Proteins: Elements, Structures and Functions	1
1.1.2 Protein Binding and Interactions	5
1.1.3 Binding Interfaces in Protein Interactions	7
1.1.4 Definition of Protein Interfaces: An Intractable But Fundamental Task in Computational Biology	11
1.2 Motivation	14
1.2.1 Computational Structural Analysis of Protein Interfaces	14
1.2.2 Limitations in Current Dissection of Protein Interfaces	16
1.2.3 Our Objectives	18
1.3 Major Contributions	18
1.4 Organization	21
1.5 A Brief Summary of the Research Experience during my Candidature	24
2 Literature Review	27
2.1 Approaches to Modeling and Prediction of Binding Hot Spots	27
2.1.1 Modeling Binding Hot Spots	27
2.1.2 Estimating $\Delta\Delta G$ of Binding Hot Spots	29
2.1.3 Identifying Binding Hot Spots Qualitatively	30
2.1.4 Databases for Binding Hot Spots	33
2.2 Approaches to Analyzing Protein Interfaces	33
2.2.1 Dissecting Various Types of Protein Interfaces Based on Physico-chemical Properties	34

2.2.2	Analyzing Residue Composition and Residue Preference in Various Types of Protein Interfaces	34
2.3	Approaches to Classifying Different Types of Protein Interactions	35
2.3.1	Identification of Biological Interactions from Crystal Packing	36
2.3.2	Classification for Different Types of Biological Interactions	36
2.3.3	Distinction of Crystal Packing and Different Types of Biological Interactions	37
3	Double Water Exclusion for Binding Hot Spots	39
3.1	Introduction	40
3.2	The Double Water Exclusion Hypothesis for Binding Hot Spots	42
3.3	Methods	43
3.3.1	ASEdb: An Evaluation Dataset for Binding Hot Spots	43
3.3.2	Mining DWE Biclique Patterns from Protein Interaction Pairs	43
3.4	Biclique Patterns are Rich in Binding Hot Spot Residues	46
3.4.1	Verifying DWE Bicliques with ASEdb	46
3.4.2	A Case Study: Validating an Example Biclique with HotSprint	48
3.5	Conclusion	50
4	Z-score Biological Significance of Binding Hot Spots and DWE Bicliques	53
4.1	Introduction	54
4.2	Z-score of Contact Residues/Hot Spots: Definitions	56
4.2.1	Preliminary Definitions	56
4.2.2	Residue Pairwise Potentials of Interface Contacts	57
4.2.3	Residue Pairwise Potentials of Nearby Contacts	58
4.2.4	Residue Pairwise Potentials of Alanine Mutations	59
4.2.5	Z-score	59
4.3	Methods	62
4.3.1	Datasets	62
4.3.2	Detailed Steps to Extract Q^c and Q^n for Residue Pairwise Potentials	63
4.3.3	The Distribution of Potential Scores of Contact Residues in Reference-State Crystal-Packing Interfaces	64
4.3.4	Evaluation Measures for Classification Performance	66
4.4	Results and Discussion	66
4.4.1	Distribution of Potential Scores and Z-scores	66
4.4.2	Classification Performance by Our Z-score Method for Binding Hot Spot Residues	70

4.4.3	Classification Performance by the Z-score Method of DWE-Biclique Binding Hot Spots	72
4.4.4	Performance Analysis According to Residue Types	75
4.4.5	Discussions on Two Confusing Features	76
4.5	Conclusions	77
5	DWE Bicliques and Their Residue and Residue Pairing Preference	79
5.1	Introduction	79
5.2	Methods	81
5.2.1	Compiling a Non-redundant Dataset	81
5.2.2	Mining Maximal DWE Bicliques for Protein Interactions	82
5.2.3	The Definitions Related to Our Composition Analysis on Residues and Residue Pairs	83
5.2.4	Measures for Comparison Analysis on the Residue Composition Differences	84
5.3	Results and Discussion	84
5.3.1	Binding Hot Spots are Small in Binding Interfaces	85
5.3.2	Common and Unique DWE Bicliques in Different Types of Interactions	85
5.3.3	Residue Composition of the DWE Bicliques for the Three Types of Protein Interfaces	94
5.3.4	Residue Pairing Preference in the DWE Bicliques for the Three Types of Protein Interfaces	98
5.4	Conclusions	101
6	Propensity Vectors of Low-ASA Residue Pairs	103
6.1	Introduction	103
6.2	Methods	106
6.2.1	Three Benchmark Datasets	106
6.2.2	Low-ASA Residue Pairs and Their Propensity Vector	107
6.2.3	OringPV: Our Classification Method	110
6.2.4	Performance Measures	111
6.3	Classification Results	111
6.3.1	LOOCV Performance by OringPV within Datasets	112
6.3.2	Propensity Vector in Comparison to Propensity Vector Without the ASA Restriction, and in Comparison to Traditional Frequency Vector	114
6.3.3	Propensity Vector in Comparison to Low-ASA Binary Vector and Low-ASA Frequency Vector	115

6.3.4	Performance Comparison of OringPV with NOXclass, DiMoVo, and Other Methods	117
6.4	Insights into Low-ASA Residue Pairs and Propensity Vectors: A Discussion Based on Misclassifications	123
6.5	Conclusion	127
7	β Atomic Contacts	129
7.1	Introduction	130
7.2	β Atomic Contacts in Protein Interfaces	132
7.2.1	Various Definitions of Binding Interfaces and Their Atomic Contacts: A Background	132
7.2.2	β Atomic Contacts: Our Definition	133
7.2.3	The Difference of β Contacts from Other Contact Definitions . . .	135
7.3	Materials and Methods	138
7.3.1	Datasets	138
7.3.2	Detecting β Atomic Contacts in Protein 3D Structures	140
7.3.3	β Atomic Contact Vectors in Protein Interfaces	141
7.3.4	Our Proposed Classifier β ACV	141
7.3.5	Evaluation Measures	142
7.4	Results and Discussion	143
7.4.1	Comparison of β Contacts with Distance-based Contacts	143
7.4.2	Analysis of the Selected Features in the Classifications	149
7.4.3	β Contacts Outperform Previous Methods in Classification of Crystal Contacts and Homodimers	155
7.4.4	β Atomic Contacts Can Identify Critical Specific Contacts: a Case Study on the HA1-2D1-Binding Dataset	156
7.5	Discussion and Future Improvement	158
7.5.1	Discussion on Infrequent and Redundant β Atomic Contact Patterns.	158
7.5.2	Future Improvement	159
7.6	Conclusion	160
8	β Contacts in Prediction of Binding Free Energy and Binding Hot Spots	161
8.1	Introduction	162
8.2	Materials and Methods	164
8.2.1	A Dataset	164
8.2.2	β Atomic Contacts in Protein Binding Interfaces	164
8.2.3	Our Prediction Method for Alanine Mutations	165
8.2.4	Separate Prediction Methods for Polar Residues and Non-Polar Residues	168

8.2.5	Ridge Regression of βACV_{ASA} and Evaluation Measures	169
8.3	Results and Discussion	170
8.3.1	β Contacts are Better than Distance-Based Contacts to Predict Binding Free Energy	170
8.3.2	Water Exclusion is a Necessary Condition of Binding Hot Spots .	171
8.3.3	Prediction Results for Polar and Non-Polar Residues Separately .	172
8.3.4	Our Method βACV_{ASA} is Superior to Several Widely-used Methods	172
8.4	Conclusion	173
9	Binding Hot Spots in the 2D1-Antibody Binding to HA1 Proteins in H1N1	177
9.1	Introduction	178
9.2	Methods	179
9.2.1	The HA-2D1 Binding Structures	179
9.2.2	Computational Methods for Predicting Hot Spots	179
9.3	Results	180
9.3.1	Energy Change Tendency of the Six Mutations	180
9.3.2	Hot Spot Residues at the Epitopes of the Two HA1s	184
9.3.3	Hot Spot Residues at the 2D1's Paratope	186
9.3.4	Analysis on the Assumed 2D1-2007HA1 Binding	190
9.3.5	Another Two Evidences	190
9.4	Conclusion	191
10	Summary and Future Work	195
10.1	Summary	195
10.2	Future Directions	198
10.2.1	Atomic Pairwise Potentials of β Contacts	198
10.2.2	A Probability Graph Model of Protein Quaternary Structures . .	199
10.2.3	Unique Atomic Contacts in Binding Interfaces	200
10.2.4	Analysis of Protein-Small Molecule Binding	200
10.2.5	Analysis of Protein Binding to DNA/RNA	201
10.2.6	Protein Docking	202
10.2.7	A More Accurate Estimation Method for $\Delta\Delta G$	202
A	A Short Tutorial on Bipartite Graphs and Their Maximal Bicliques	203
A.1	Definitions of Bipartite Graphs and of Maximal Bicliques	203
A.2	Mining Maximal Bicliques	204

B A Short Tutorial on Several Machine-Learning Algorithms Used in This Thesis **205**

B.1 SVM. 205

B.2 RFE: a Feature-Section Method. 205

B.3 Decision Tree. 206

C Supplementary Files for Chapter 4 and 8 **209**

References **244**

List of Tables

2.1	Summary of main previous works in dissection of binding hot spots . . .	28
3.1	Sensitivity of hot spots in biclique patterns	47
3.2	Conservation and ASA of residues in the example biclique	49
3.3	Conservation and ASA of ring-residues surrounding the example biclique	49
4.1	Information of three examples of mutation residues	68
4.2	Classification performance for binding hot spots ($\Delta\Delta G \geq 1$ kcal/mol) . .	73
4.3	Classification performance for binding hot spots ($\Delta\Delta G \geq 1$ kcal/mol) for each complex in ASEdb	73
4.4	Classification performance for binding hot spots ($\Delta\Delta G \geq 2$ kcal/mol) . .	74
4.5	Classification performance for DWE-biclique binding hot spots	74
4.6	Classification performance for binding hot spots for each residue type in ASEdb	75
4.7	Classification performance for binding hot spots for each residue type in BID	75
5.1	The chain-pair distribution in our non-redundant dataset	82
5.2	The fraction of DWE bicliques	85
5.3	Common and unique DWE bicliques in biological interactions and crystal packing	87
5.4	Common and unique DWE bicliques in obligate and non-obligate interactions	87
5.5	The example unique biclique to crystal packing	89
5.6	The example unique biclique to obligate interactions	89
5.7	The example unique biclique to biological interactions	91
5.8	The example unique biclique to non-obligate interactions	91
5.9	The conservation score and ASA ($\text{\AA}$) information of the unique biclique $\mathcal{G}_1 = \langle \{\text{ARG, GLY, SER}\}, \{\text{ASN, LYS}\} \rangle$	92
5.10	The conservation score and ASA ($\text{\AA}$) information of the unique biclique $\mathcal{G}_4 = \langle \{\text{GLN, GLY, SER, SER, TYP}\}, \{\text{CYS, LYS}\} \rangle$	94
5.11	Sequence similarity among chain pairs in Table 5.8	94

5.12	The difference of residue composition	97
6.1	The size of O-ring-surrounded regions	109
6.2	OringPV performance under LOOCV procedure	112
6.3	OringPV confusion matrix under LOOCV procedure for prediction of different types of protein interactions.	113
6.4	OringPV confusion matrix under LOOCV procedure for prediction of the two types of biological interactions.	113
6.5	The accuracy performance comparison among PV, BV and FV	115
6.6	The p -values of Wilcoxon signed rank test in the Bahadur dataset	117
6.7	The within-dataset performance comparison: OringPV vs NOXclass.	120
6.8	The cross-dataset performance comparison: OringPV vs DiMoVo.	121
6.9	The cross-dataset performance comparison: OringPV vs NOXclass	122
6.10	The summary of OringPV in comparison to NOXclass and DiMoVo	123
6.11	Performance trend of a classifier when training datasets change	124
7.1	The HA1-2D1-binding dataset	140
7.2	Atomic types	142
7.3	The difference of the numbers of distance-based, Voronoi-based and β atomic contacts for 114 homodimers	144
7.4	The difference of the numbers of distance-based, Voronoi-based and β atomic contacts for 314 crystal packing	144
7.5	The comparison of feature-selection classification performance	146
7.6	The comparison of classification performance on the independent datasets	147
7.7	The top 10 features of distance-based and β atomic contacts	149
7.8	Predictions on the HA1-2D1-binding dataset	157
8.1	Summary of the evaluated methods: β ACV _{ASA} and its compared methods	169
8.2	Predictions of binding hot spots by different methods	170
8.3	Predictions of binding hot spots by β ACV _{ASA} and Z-score	173
9.1	ASAs and burial levels of the six mutation residues	182
9.2	Binding hot spots in the two epitopes	182
9.3	Binding hot spots in the antibody paratopes	187
9.4	Predictions by β ACV _{ASA} for the six mutations from 1918HA1 to 2009HA1	191
9.5	Binding hot spot predictions by β ACV _{ASA} for 2D1-1918HA1 and 2D1-2009HA1 interfaces	192
9.6	HA proteins and their corresponding H1N1 viruses in 2009	192
C.1	The ASEdb dataset used in Chapter 8	209

C.2	Eight atomic types used in Chapter 8.	212
C.3	Alanine mutations in the ASEdb dataset for Chapter 8	213
C.4	Alanine mutations in ASEdb and their predictions for Chapter 4	229
C.5	Alanine mutations in BID and their predictions for Chapter 4	239

List of Figures

1.1	Twenty standard amino acids, and their structures and hydrophobicity .	3
1.2	An example of protein sequences and their 3D structures	4
1.3	An example of atomic coordinates in PDB	5
1.4	An example of solvent accessible surface area (ASA)	9
1.5	Voronoi diagram and Delaunay triangulation	9
1.6	An example of uneven distribution of binding free energy in protein interfaces	12
1.7	The relation of our interested problems and their selected applications . .	16
1.8	The organization of this thesis	22
3.1	The biclique pattern in PDB 1A8G	50
3.2	The biclique pattern in PDB 1CWQ	51
4.1	The distribution of potential scores	67
4.2	Three examples of the mutation residues in ASEdb	69
5.1	Examples of biclique location at the protein interfaces	90
5.2	The structure of the unique biclique $\mathcal{G}_4 = \langle \{\text{GLN, GLY, SER, SER, TYP}\}, \{\text{CYS, LYS}\} \rangle$	93
5.3	Residue compositions in DWE bicliques	96
5.4	Residue pair composition in DWE bicliques	99
6.1	The O-ring-surrounded region between chain F and G of 1GLA	108
6.2	The ROC curves of the classifications on the Bahadur ^{hocp} dataset	108
6.3	A visual display of the discriminative power carried by PV, BV and FV .	118
6.4	The structures of misclassified interactions	126
7.1	An example of the difference of atomic contacts according to various definitions	134
7.2	β skeletons and β contacts	136
7.3	The ASA change of atoms in β contacts and in the complement of β contacts with respect to distance-based contacts.	145
7.4	Decision tree of β contacts on the non-redundant dataset	152
7.5	Decision tree of distance-based contacts on the non-redundant dataset . .	152

7.6	Two misclassified examples in the decision tree of β contacts	154
8.1	$\Delta\Delta G$ predictions by different ACV methods	174
8.2	$\Delta\Delta G$ predictions by βACV_{ASA}^2 , FoldX and Robetta	175
8.3	An example of βACV_{ASA} predictions in 3HFM	176
9.1	The HA1 epitopes and their sequence and structural alignment	181
9.2	The hot spots in the 2D1-antibody binding interfaces	185
9.3	Three examples of (non-)hot spot predictions	189
9.4	The sequence alignment of HA proteins in different 2009 H1N1 viruses .	193

Chapter 1

Introduction

In living organisms, genes are the blueprints or library, specifying instructions for building proteins. Proteins constitute the bulk of cells, and their binding and interactions play a vital role in numerous functions and activities, such as signal transduction, enzymatic reactions, immunoreactions and inter-cellular communications. Investigation of protein binding is thus important to understand biological processes and activities in living cells. In this thesis, our research focus is on protein-protein binding.

1.1 Background

This introductory section provides basic knowledge of proteins and protein binding; it contains four subsections. First, we describe what is protein, including proteins' fundamental elements, structures and functions. In the second subsection, we present notions related to protein binding and interactions. In particular in the third subsection, we explain why protein binding interfaces are complicated but reconditely crucial in protein interactions. Finally as a work in computational biology, we give a short introduction of computational definitions on protein binding interfaces.

1.1.1 Proteins: Elements, Structures and Functions

1.1.1.1 Amino Acids

The basic biological elements of proteins are amino acids. Each amino acid contains two parts: the same backbone but different side chains. The backbone consists of three groups: an amino group (NH_2), a central carbon (the alpha carbon or CA) and a carboxyl

group (COOH). Side chains have different combinations of heavy atoms, each corresponding to one of the various types of amino acids. Heavy atoms are those atoms such as Carbon (C), Nitrogen (N), Oxygen (O) and Sulfur (S).

There are twenty standard amino acids as shown in Figure 1.1. Their 3-letter (1-letter) codes are: Ile(I), Val(V), Leu(L), Phe(F), Cys(C), Met(M), Ala(A), Gly(G), Thr(T), Ser(S), Trp(W), Tyr(Y), Pro(P), His(H), Glu(E), Gln(Q), Asp(D), Asn(N), Lys(K), and Arg(R) in the order with I as the most hydrophobic and R as the least hydrophobic according to their hydrophobicity [1]. Among them, Gly is the smallest amino acid without a side chain of any heavy atom. Ala is the second smallest amino acid and only has one heavy carbon atom CB (or the beta carbon) in its side chain; CB is the first heavy atom of side chains which has a covalent bond with CA. The side chains of the other amino acids have more number of heavy atoms. Please refer to Figure 1.1 for more details.

Because of different side chains, the twenty amino acids possess various physicochemical properties. For example, the side chains from Arg, His and Lys are positively charged, while the side chains in Asp and Glu are negatively charged; Ser, Thr, Asn and Gln have polar side chains, but Ala, Ile, Leu, Met, Phe, Trp, Tyr and Val have hydrophobic side chains and especially the side chains in Phe, Trp and Tyr contain aromatic rings. Hydrophobic amino acids generally tend not to contact with water molecules, while hydrophilic (polar and charged) amino acids prefer a high affinity for water.

1.1.1.2 What is Protein

In a real-cell environment, the amino NH_2 group of a specific amino acid can react with the carboxyl COOH group of another amino acid, forming an amino covalent bond and resulting in the release of a water molecule; this covalent bond is also called peptide bond in biology. After this condensation reaction, the involving amino acids are also referred to as (amino acid) residues. A set of residues create a linear-sequence polymer, called a peptide for a shorter and a protein for a longer as shown in the left-side sub-figure in Figure 1.2.

In vivo with a three-dimensional (3D) space, a residue of a protein can interact with some other residues according to physical, chemical and biological rules, producing a 3D structure. One example of protein 3D structures is displayed in the right-side sub-figure in Figure 1.2, whose protein sequences are in the left-side sub-figure. Generally, a protein sequence determines a unique, stable, intended and correct 3D structure (also called native conformation). It is also widely accepted that similar sequences have similar 3D structures, but similar 3D structures maybe have different protein sequences.

Proteins' structures can be often referred to from the following four distinct aspects: protein primary structures, secondary structures, tertiary structures and quaternary structures. Primary structures are protein sequences themselves, while the others

Amino Acids				
Full name	Isoleucine	Valine	Leucine	Phenylalanine
3-letter name	Ile	Val	Leu	Phe
1-letter name	I	V	L	F
Hydrophobicity	4.5	4.2	3.8	2.8
Structure
Full name	Cysteine	Methionine	Alanine	Glycine
3-letter name	Cys	Met	Ala	Gly
1-letter name	C	M	A	G
Hydrophobicity	2.5	1.9	1.8	-0.4
Structure
Full name	Threonine	Serine	Tryptophan	Tyrosine
3-letter name	Thr	Ser	Trp	Tyr
1-letter name	T	S	W	Y
Hydrophobicity	-0.7	-0.8	-0.9	-1.3
Structure
Full name	Proline	Histidine	Glutamic acid	Glutamine
3-letter name	Pro	His	Glu	Gln
1-letter name	P	H	E	Q
Hydrophobicity	-1.6	-3.2	-3.5	-3.5
Structure
Full name	Aspartic acid	Asparagine	Lysine	Arginine
3-letter name	Asp	Asn	Lys	Arg
1-letter name	D	N	K	R
Hydrophobicity	-3.5	-3.5	-3.9	-4.5
Structure

Figure 1.1: Twenty standard amino acids, and their structures and hydrophobicity. Here, the structures are from [2] and the hydrophobicity values are from [3].

Figure 1.2: An example of protein sequences and their 3D structures. The middle sub-figure represents 3D coordinates of each atom in proteins. In the right-side sub-figure, one protein is in magenta while the other is in green; the binding interface is in a sphere view for the green protein and in a stick view for the magenta.

represent different knowledge of protein 3D structures. In protein 3D structures, some segments have such favored regular shapes as alpha helix or beta sheet stabilized by hydrogen bonds, while others' shapes are irregular, specially termed as loop/turn/random coil. These local shapes are known as secondary structures of proteins. Further, a protein 3D conformation is defined as a tertiary structure if a single protein is involved, and as a quaternary structure when the conformation is about a protein complex. A protein complex is the binding or contact of two or more proteins by non-covalent bonds. It should be noted that each protein in its quaternary structures is also referred to as a protein chain or a proteomer. Please refer to the right-side sub-figure in Figure 1.2 for an example of protein secondary structures, tertiary structures and quaternary structures.

The native 3D conformations of proteins determine their functions. Failure to fold into intended shapes usually produces inactive proteins with different properties, such as toxic prions [4]. This failure may result in many diseases, including Creutzfeldt-Jakob disease, bovine spongiform encephalopathy (mad cow disease), amyloid-related illnesses such as Alzheimer's Disease and familial amyloid cardiomyopathy or polyneuropathy, as well as intracytoplasmic aggregation diseases such as Huntington's and Parkinson's disease [4]. Therefore, it is vital to determine and understand protein 3D structures.

1.1.1.3 Determination and Store of Protein 3D Structures

Many experimental methods were developed to determine protein 3D conformations, including X-ray crystallography, electron microscopy and Nuclear magnetic resonance

	1	2	3	4	5	6	7	8	Tens Place		
	1	2	3	4	5	6	7	8	Ones Place		
ATOM	171	N	ALA	A	22	-23.530	15.854	23.632	1.00	8.12	N
ATOM	172	CA	ALA	A	22	-22.101	16.127	23.631	1.00	10.49	C
ATOM	173	C	ALA	A	22	-21.714	17.207	24.627	1.00	11.94	C
ATOM	174	O	ALA	A	22	-22.515	18.047	25.045	1.00	13.86	O
ATOM	175	CB	ALA	A	22	-21.623	16.582	22.260	1.00	5.03	C
ATOM	176	N	LEU	A	23	-20.471	17.079	25.047	1.00	12.22	N
.....	Atom Name	Residue Name	Chain Name	Sequence Position	X	Y	Z	Atom Coordinates			

Figure 1.3: An example of atomic coordinates in PDB.

(NMR) spectroscopy. Among these techniques, X-ray crystallography is the most popular and prolific technique. X-ray crystallography [5] uses the following way to determine the atomic arrangement of proteins within a crystal. Firstly, according to elastic scattering, X-ray crystallography employs X-ray to produce a diffraction pattern of regularly spaced spots called a reflection. Then, several reflections are obtained through rotating the crystal. Finally, the mathematical method of Fourier transforms is used to convert the two-dimensional reflection images taken at different rotations into a three-dimensional model of the density of electrons within the crystal. According to electron density, atomic mean positions in a crystal can be determined. X-ray crystallography is relatively affordable, and can obtain high-resolution information of atoms in proteins. So far (at the time of this thesis writing in Oct 2011), it has solved 66,989 crystal structures of biological molecules according to the statistics of Protein Data Bank (PDB).

PDB is a key resource to store 3D structural data of those biological molecules, such as proteins and nucleic acids. PDB can be accessed by URL <http://www.rcsb.org/>. In each PDB entry of proteins, heavy atoms of each residue are stored with their own 3D coordinates as shown in Figure 1.3. Based on these atomic coordinates, protein 3D conformations can be visualized. For example, the right-side sub-figure in Figure 1.2 is produced according to the coordinates in the middle sub-figure of Figure 1.2. Therefore, PDB structures are primary information to investigate proteins and their structures; several important databases already derive and store the classification of PDB structural data, such as SCOP [6] and CATH [7].

1.1.2 Protein Binding and Interactions

A living organism is a dynamic integral whole, and its parts will contact with each other for performing biological functions. As a fundamental component of living organisms,

proteins definitely rarely function alone; most proteins must work in collaboration with other macromolecules, such as other proteins and DNA. Without doubt, close interactions of proteins are indispensable to fulfill molecular functions and biological processes. A good example of this is antigen-antibody interactions. When foreign objects antigens intrude, antibodies can specially identify and neutralize antigens so that immune systems can well defend the body against the attack from antigens. In this process, antigen-antibody close interactions are essential to immune systems.

Meanwhile, protein interactions are very specific: on one hand, not all segments of proteins are able to bind to other molecules, although a protein can have more than one binding segment; on the other hand, a binding segment of a protein should only contact with specific binding segments in some other proteins instead of with all binding segments or all other proteins—these specific binding segments should have a certain corresponding characteristic.

Close and specific protein interactions play a vital and fundamental role in molecular functions and biological processes, but why binding happens like this and what are binding segments are unknown yet. Thus, discovering the principle of protein interactions and the specificity of protein binding is a fundamental and challenging problem in proteomics, resulting in a lot of useful applications, such as drug design and protein engineering.

1.1.2.1 Diversity of Protein Interactions

However, protein interactions have great diversity according to various criteria. For example, some interactions are permanent and usually very stable. Thus once forming protein complexes, they only exist in their complex form in all of the lifetime of the complexes [8]. In contrast, others associate to accomplish a particular function upon a molecular stimulus and dissociate after that [8]. This kind of interactions are termed as transient interactions to be compared to permanent interactions.

Biological interactions can also be grouped depending on whether their protomers of interactions can be found or not as stable tertiary structures on their own *in vivo* [8]. On certain physiological conditions and environments, if protomers of interactions cannot be found as stable tertiary structures on their own *in vivo* [8], this kind of interactions are referred to as obligate interactions; otherwise, those interactions are non-obligate interactions. Similarly, two-state folding complexes and three-state complexes were also used to describe obligate and non-obligate interactions; this criterion depends on the different transition processes in protein folding and binding [9]: (i) in non-obligate interactions, two protomers of interactions fold separately, and then recognize each other to form the complexes. Thus, this process is three-state; (ii) in contrast, two protomers fold and bind together without the intermediate states in obligate interactions, and these interactions are two-state complexes [9] or obligomers [10].

The criterion for obligate/two-state and non-obligate/three-state complexes is similar to and confused with, but really different from, the criterion for permanent and transient interactions: (i) obligate interactions or two-state complexes are stable, and their protein chains function only in the complex form [8]. So, this kind of interactions are permanent interactions. (ii) Correspondingly, protomers in transient interactions fold separately, and they have stable tertiary structures; thus, transient interactions are non-obligate. (iii) Clearly, there exist non-obligate permanent complexes, such as antigen-antibody interactions, and the complexes of thrombin and rodniin inhibitor [8]. In these complexes, each protomer fold into a stable tertiary structure, and then the protomers come together to form protein complexes which do not dissociate again. However, the stability of complexes much depends on physiological conditions or environments: a continuum exists between obligate and non-obligate interactions, while an interaction may be mainly transient *in vivo* but become permanent under certain cellular conditions [8]. Thus, these two criteria are mostly combined in literature works, although they focus on different aspects of protein complexes.

Meanwhile, protein interactions can happen between two identical protein chains and called homodimers [8]; otherwise, the interactions are called heterodimers [8]. Homodimers are generally obligate or permanent interactions, while most of heterodimers are non-obligate or transient interactions. Furthermore in a more complicated way, protein interactions can be categorized into six subtypes [10]: intra-domain, domain-domain, homo-complexes (interfaces of transient interactions between identical protein chains), homo-obligomers (interfaces of permanent interactions between identical protein chains), hetero-obligomers (interfaces of permanent interactions between different protein chains) and hetero-complexes (interfaces of transient interactions between different protein chains). Here, a domain is a consecutive segment of protein sequences with a particular and repeatable three-dimensional structure; it may evolve, function, and exist independently of the rest of protein chains.

It is clear that these above different types of biological interactions possess their unique binding behaviors. This diversity of protein interactions necessitates individual investigation for each different type of protein interactions based on their quaternary structures.

1.1.3 Binding Interfaces in Protein Interactions

In quaternary structures of protein interactions, physicochemical properties are not uniform everywhere, such as evolutionary conservation, and solvent accessible surface area (ASA). According to the difference of solvent accessible surface area, protein quaternary structures can be divided into several segments, such as interfaces, surfaces, and intraproteins/interior. Intraproteins can further be subdivided into domain interfaces and

intradomains according to evolutionary/structural conservation or functional properties. Among these segments of protein quaternary structures, interfaces are significantly important.

In a protein complex, the region where protein chains come into contact is a binding site; or for both sides, an interface [11]. Protein binding interfaces specify how proteins involve in various activities in living cells. Thus, interfacial analysis becomes a key point to reveal the rules governing protein interactions and their cellular processes involved.

In previous works, a protein binding interface is generally considered to be composed of two relatively large, spatially close protein surfaces of atoms/residues with good geometric shape and chemical complementarity. It is also believed that the formation of protein chain interfaces is driven by various natural weak forces such as hydrogen bonds, electrostatic interactions, van der Waals forces, salt bridges and hydrophobic attractions. Thus, protein interfaces have complicated physicochemical properties.

1.1.3.1 Physicochemical Properties in Protein Binding Interfaces

In folding and binding process of proteins, a certain solvent environment is a crucial factor. Thus, it is helpful to measure how much molecules (e.g., residues and atoms) may be exposed to water/solvent environments in their protein 3D structures. To do that, solvent accessible surface area was first described by Lee & Richards in 1971. In a protein 3D structure, each heavy atom is represented by a sphere defined by its van der Waals radius; then, ASA is calculated by rolling a ball of a particular radius to probe the surface of atoms [12]. The typical radius of the rolling ball is 1.4 Å. An example of ASA for several atoms is shown in Figure 1.4. After that, the ASA sum of the certain group of atoms in a specific residue is ASA of the residue; the aggregate ASA of residues is protein ASA. Several softwares are developed to calculate ASA, and NACCESS is the well-known one [12].

It is clear that ASA is not even everywhere in protein quaternary structures: some residues/atoms are completely buried with zero ASA, while the other may be greatly exposed to the solvent; protein surfaces are heavily exposed to the solvent with larger ASA, while intraproteins/interior is little exposed to the solvent with smaller ASA; protein interfaces should be exposed to the solvent in unbound states, and little exposed in bound states. It seems that ASA is one of physical features of protein binding interfaces. It can be used to measure interface size and contact area. This measure is about the change of ASA (ΔASA) upon the formation of protein complexes, which is considered to provide a measure of binding strength [15]. Computationally for two protein chains in a protein complex, there is a difference between the sum of their ASAs each in their own tertiary structures (i.e., ASA_1 , and ASA_2) and ASA of their quaternary structures (ASA_c) upon the formation of the complex. The difference, $\Delta ASA = (ASA_1 + ASA_2 - ASA_c)/2$, is

Figure 1.4: An example of ASA for atoms. This example is from [13].

Figure 1.5: Voronoi diagram (red) and Delaunay triangulation (black). This figure is from [14].

generally referred to as interface size or contact area for the interactions. In detail, some atoms/residues may also undergo a similar change of their ASAs upon the formation of protein complexes; these atoms/residues are considered to be more likely involved in protein interfaces.

Besides ASA, other outstanding chemical, physical and geometric properties [16, 17] as follows can also be used to characterize protein 3D structures especially protein binding interfaces.

- Compositional features consisting of residue composition [18, 19, 20, 21] and propensity [22], residue pairs [23, 24, 25] and atomic pairs [26, 27, 28, 29]. These compositional features are good descriptors for different segments of proteins (e.g. interfaces, surfaces and intraproteins/interior) in protein 3D structures. Protein binding interfaces tend to have a core with hydrophobic residues.
- Geometric features including planarity, shape complementarity, circularity [15, 30] and secondary structures [16]. In a protein quaternary structure, there exists a gap between the tertiary structures of its protein chains. This kind of gaps can be measured by gap index, planarity and shape complementarity.
- Chemical features that contain hydrophobicity and polarity [15, 22, 31, 32].
- There are also other features, such as evolutionary residue conservation [17, 18]. A new protein sequence usually is not created from scratch; it is generated from existing sequences by deletion, insert and mutations of residues, and shifting of parts of old sequences. Thus, some sequences can have similar or same residues in

specific corresponding positions; these residues are more evolutionarily conserved. Interfacial residues are considered to be more conserved than surface residues.

Many of these above features have been involved in important findings. For example, interface area of biological interactions is found to be much larger than that in non-biological interactions [17, 18, 19, 22, 31, 33, 34, 35]. These above features also suggest that protein interfaces are very complicated. Here, biological interactions are those molecular binding which form in solution or in their physiological states to perform biological processes and molecular functions, while non-biological interactions are produced manually or their monomers are randomly in close vicinity without biological significance. Good examples of non-biological binding are decoys in protein docking or crystal packing in PDB—the former are artificial binding produced by computational docking algorithms, while the latter are enforced by the crystallographic packing environment and formed during the crystallization process [11], but both of them do not occur in solution or in their physiological states.

1.1.3.2 Uneven Distribution of Binding Free Energy and Binding Hot Spots in Protein Interfaces

Assume that there is free energy G with proteins in an unbound state; when proteins can contact with each other upon the formation of protein complexes, the free energy is G' . $\Delta G = G - G'$ is the change of the free energy of protein binding, or ΔG is required energy to disassemble proteins in the bound state into the unbound state. For protein interfaces, ΔG is binding free energy.

Binding free energy is not evenly distributed in protein interfaces [36]. As shown in Figure 1.6, the red part is considered to contribute most to the binding free energy, and the green part least. This uneven energetic distribution can be probed by experimental approaches. A widely-used experimental approach is alanine scanning mutagenesis [37, 38]. This approach selectively mutates an individual side chain of a residue from interfaces to Ala mostly by eliminating the side chain beyond CB. Why Ala is the substitution residue of choice is that it yet does not alter backbone conformations (as Gly or Pro can) nor does it impose extreme electrostatic or steric effects [39, 40]. Then, kinetic and thermodynamic measurements are employed to determine ΔG^{mut} of binding after mutations. Finally, the energetic contribution of individual side chains to protein binding [36] is calculated by using $\Delta\Delta G = \Delta G - \Delta G^{mut}$. $\Delta\Delta G$ is the change of binding free energy after alanine mutations. Through this way, it has been proved that there exist a small fraction of residues in protein interfaces, contributing most to binding stability and binding free energy. These residues are called ‘binding hot spots’ [38]. In literature works, residues are considered as hot spot residues if their $\Delta\Delta G$ s are not less than a threshold, such as ≥ 1.0 kcal/mol [41], or ≥ 2.0 kcal/mol [36].

Binding hot spots are considered to be very important to binding affinity and strength of protein interactions. For instance, in the complex of bovine pancreatic trypsin inhibitor (BPTI) and β -Trypsin, the mutation of Lys-15 in BPTI caused a pronounced drop in the binding energy, a 230-fold decrease in the association rate constant for the trypsin/K15A complex [42].

Binding hot spots defined by $\Delta\Delta G$ does not take the contribution from backbone atoms into account, but they provide a new opportunity to understand complicated protein binding and to pinpoint the governing principles under specific protein binding. Thus, identification of binding hot spots is particularly advantageous for many predictions in structural analysis of proteins, such as docking algorithms to construct protein quaternary structures [25], identification of protein binding sites [16, 43] and reliable predictions on interaction types for new protein complexes [18, 26, 44]. In practical applications, binding hot spots imply target candidates of drug design. For example, compared to 2007 H1N1, the mutations of 2009 H1N1 result in several new binding hot spots which increase the fatal property of 2009 H1N1 (Please refer to Chapter 9 for detail); the interfacial mutation E104D causes human triosephosphate isomerase deficiency [45]. Hence, mutations are pathogenic if they remove hot spots in expected sites or add hot spots in unexpected sites, and detecting hot spots can help design drugs against diseases.

1.1.4 Definition of Protein Interfaces: An Intractable But Fundamental Task in Computational Biology

In protein complexes, protein binding interfaces are an essential segment. Protein binding interfaces under a perfect definition should only contain those residues which have larger positive contribution to binding free energy from both backbone atoms and atoms of side chains. However, $\Delta\Delta G$ is only the contribution from atoms of side chains, while the contribution from backbone atoms is hard, if not impossible, to experimentally measure. Also, it is expensive and time-consuming to biologically identify binding free energy of interfacial residues. So to computationally analyze protein quaternary structures, definitions of protein binding interfaces and their atomic contacts should first be given based on atomic 3D coordinates of proteins.

Generally, a protein interface is defined as two parts: contact residues and neighbor residues of the contact residues as contact scaffold [46, 47, 48, 49, 50, 51, 52]. Neighbor residues are those residues which are from the same chain as contact residues and also near contact residues. The ‘near’ is mostly measured based on the spatial distance between CA atoms of contact residues and of neighbor residues [46]. Contact residues are widely defined by the following several criteria.

One definition of contact residues/atoms uses atomic distance [10, 46, 53, 54, 55, 56, 57, 58]. The measure under this definition evaluates how two residues/atoms are

Figure 1.6: An example of the uneven distribution of the binding free energy in the antibody protein interface in the PDB entry 1VFB. In this figure, the parts in black, red, skyblue and limegreen are from Chain B, and the parts in gray, blue and green are from Chain A. Those interfacial residues in yellow are not probed by the experimental alanine mutation method. $\Delta\Delta G$ of the red part is ≥ 2 kcal/mol; $\Delta\Delta G$ of the blue and skyblue part is < 2 kcal/mol and ≥ 1 kcal/mol; $\Delta\Delta G$ of the green and limegreen part is < 1 kcal/mol. The letter strings represent the specific residues in the interface: the first letter is the chain name; the letters following ‘-’ are one-letter residue types and their positions in protein sequences.

spatially close from interacting proteins. However, there is no gold standard to define ‘close’ contacts between two residues/atoms. Some works consider two residues close enough to contact if the spatial distance of their CA, or CB atoms, or the residue’s center of mass is less than a threshold [23, 59], e.g., 9 Å [46], while others take all heavy atoms of residues into account [26, 48, 49]. To define contacts between heavy atoms, two ways are commonly used in literature [23, 46]: one is to take an absolute value, e.g., 6 Å [26], as a threshold for atomic spatial distance; and the other is to take into account the van der Waals radii of atoms and use the sum of the corresponding atomic radii plus a value as a threshold—this value can be 0.5 Å [48, 49, 50, 60], 1.5 Å [61], 2.75 Å [62, 63, 64] or 3.0 Å [55].

Another definition considers the change of solvent accessible surface area (ΔASA) upon the formation of protein complexes [18, 22, 23, 30, 31, 65, 53]. Under this definition, contact residues/atoms are those residues/atoms whose ΔASA is $>0.1 \text{ \AA}^2$ [22, 31] or $>1.0 \text{ \AA}^2$ [18, 30, 66] upon the formation of complexes.

A more complicated definition of contacts is based on Voronoi diagrams of protein complexes [67, 68, 69, 70, 64, 71]. Given a set of points p , its general Voronoi diagram, $VD(p)$, is defined by Voronoi cells of $s \in p$; a Voronoi cell of an s is the region comprising all points closer to s than to any other points in p . The Delaunay triangulation of p is a graph where each node is an s and two nodes s_i and $s_j, i \neq j$, have an edge if they share a Voronoi facet. So, a Delaunay triangulation is the dual graph of a Voronoi diagram [14]. Figure 1.5 shows an example of a Voronoi diagram and its dual graph Delaunay triangulation. In 3D structures of protein complexes, each atom/residue can be considered as an s . Two atoms are considered to contact with each other if they share a facet in their Voronoi diagram or an edge in their Delaunay triangulation.

Meanwhile from an energetic perspective, van der Waals energy can also be used to define contact residues [46].

These various definitions above suggest that it is very complicated to accurately provide a computational definition for protein binding interfaces based on essential known quaternary structures of proteins. Also, the above definitions cannot capture the specific property of protein binding, because they can define large interfaces of non-biological interactions for any monomers which are in close vicinity; instead in living organisms, protein binding only happen when certain chemical, physical and biological rules are satisfied. These rules are so intractable that they are beyond the simple measures, distance and ASA in the above definitions. Hence, the above definitions are almost impossible to identify computational interfaces of protein binding which are quite similar to, if not exactly same as, what interfaces should be under the perfect definition (a definition according to the contribution of binding free energy to complex formation). This results in that the principles governing protein binding are elusive [72], although they are necessary to understand molecular functions and biological processes in living organisms.

To sum up the main points in the background, on one hand, protein interfaces are vital to biological functions in living organisms, and interfacial understanding can help in many fields and applications, such as protein engineering and drug design; on the other hand, these interfaces have a great structural and functional diversity: (i) protein interfaces have outstanding chemical, physical and geometric properties [16, 17]; (ii) various types of protein interactions possess different binding behaviors; and (iii) protein interfaces have uneven energy distribution (i.e., binding hot spots). All of these make complicated and reconditely crucial binding interfaces. Therefore, understanding of protein interfaces is a challenging and indispensable work in bioinformatics.

1.2 Motivation

One ultimate goal in bioinformatics is to unveil the governing principles buried in protein interfaces; these principles can considerably benefit structural analysis of proteins and drug design. Biological experiments are the best way of the principle discovery, since wet-lab experiments can exactly tell where binding sites and interfaces are and which residues are binding hot spots. However, protein interfaces, especially binding hot spots, intricately imply the governing principles but don't directly pinpoint. In other words, after determining binding interfaces and their binding hot spots, discovery of the governing principles is still a challenging problem. What is worse is that, due to the structural and functional diversity of protein interactions and the large dataset of protein structures, it is an arduous and exhausting task, if not impossible, for biologists to extract the rules governing protein interactions without computer aid. Therefore, many endeavors have been tried to understand specific characteristics of protein interfaces through computational methods, since the pioneer works on protein recognition in 1975 [73].

1.2.1 Computational Structural Analysis of Protein Interfaces

In the light of the structural and functional diversity of protein interfaces, computational interfacial investigation can fall into several typical problems as follows but not limited to:

- (i) **Modeling and prediction of binding hot spots** is a key problem of analyzing binding interfaces and even of structural analysis in computational biology. As discussed before, protein binding interfaces are too complicated like fog covering the governing principles. Binding hot spots occur like several lights interspersed in the fog to help us uncover what is behind the fog.

- (ii) **Statistical analysis of physicochemical properties for interfaces or for binding hot spots** aims at using statistical methods to find out what are specific properties of binding interfaces and binding hot spots, and why and how they can be interfaces and binding hot spots.
- (iii) **Prediction of protein interaction types** wants to know which binding is permanent or obligate, which is transient or non-obligate, and which is non-biological. Different types of protein interactions generally possess various binding behaviors. This classification can make interfacial analysis more effective.
- (iv) **Identification of binding sites** tries to pinpoint those atomic/residue clusters on protein surfaces which can bind to specific surfaces on other proteins for forming binding interfaces and fulfilling biological functions. The knowledge from binding hot spots and true binding interfaces can provide some hints for this identification.
- (v) **Protein docking** is a process of predicting quaternary structures based on given protein tertiary structures. Protein docking first searches potential binding interfaces based on shape complementarity with the rigid docking assumption, and then scores each potential interfaces to identify interfaces which are likeliest to be true interfaces. In this process, accurate binding site prediction and precise hot spot identification will reduce the huge search space in the first step, while both binding hot spots and the knowledge from true binding interfaces, such as interfacial pairwise potentials, can help to produce functional scores in the second step. Furthermore, the task in the second step is exactly the problem of binding type prediction: distinguishing true binding interfaces from false decoys.
- (vi) Other analyses include predictions of protein secondary and of tertiary structures, and investigation of protein-RNA/DNA binding and of protein-small molecule binding. These problems are not closely related to this thesis.

These above problems closely relate to each other. For example, identification of binding hot spots can specialize physicochemical analysis of protein interfaces, much improve predictions of protein interaction types and of binding sites, and also benefit docking algorithms. Correspondingly, interfacial physicochemical analysis can help to understand other problems; and types of protein interactions can specialize and simplify analysis of binding hot spots and interfacial physicochemical properties. As a whole, their interplay can be seen from Figure 1.7.

Figure 1.7: The relation of our interested problems in this thesis and their possible selected applications.

1.2.2 Limitations in Current Dissection of Protein Interfaces

In computational structural biology, there are a lot of interesting problems as discussed above. In this thesis, we mainly focus on problems (i), (ii) and (iii). Through literature review in Chapter 2, we found that there are many limitations in these three problems.

- (a) **The organizational topology inside O-ring:** Many works endeavored to dissect hot spots in protein interfaces. Binding hot spots were first characterized by the profound and influential O-ring theory based on the topological shape of their surrounding residues [36]. The O-ring theory confirms that the residues of the O-ring likely function as a role to occlude bulk water molecules from the hot spots inside the O-ring. However, the organizational topology of the ring-inside, energetically more important hot spot residues is not specified by the O-ring theory.
- (b) **Prediction of binding hot spots:** Considerable computational methods have been designed to predict $\Delta\Delta G$ and binding hot spots including FoldX [74, 75], Robetta [76, 77], PP_SITE [78], KFC [79], MINERVA [66], APIS [80], ISIS [81] and a web server HOTPOINT [82]. However, none of these approaches can characterize and identify binding hot spots with high prediction performance. A clear picture of binding hot spots still remains uncovered. Without doubt, binding hot spots are proving their own complexity. They cannot be accurately characterized by sole physicochemical property or even by most of existing physicochemical properties [66]. For example, although conservation is considered to be closely related

to binding hot spots [52], some works observed that conservation less help the prediction of binding hot spots [66]. Meanwhile, on the one hand, Keskin’s study [82] suggested that statistical pairwise potentials of interface residues can improve classification performance of binding hot spots; on the other hand, Huang’s study [80] reported that residue pairwise potentials do not perform well in their hot spot prediction.

(c) **Contradictory conclusions in structural analysis of binding interfaces:**

In computational structural analysis of protein interactions, accurate knowledge from binding interfaces is a starting point. Although binding interfaces are intensively investigated in literature, some conclusions about interfaces are contradictory. For example, some works reported that there was no significant change in going from interfaces to other segments of proteins, such as intraproteins [23], while another works found protein interfaces were unique and different from other segments [10, 30, 83]. Furthermore, some works found that hydrophobic clusters play a determining role in protein-protein interactions [32, 84]; in contrast, another works argued that it is the hydrophilic rather than the hydrophobic effect that is dominant in biochemical processes including protein-protein association [85].

(d) **Prediction of binding types:** Different types of protein interactions, for example permanent versus transient complexes, have various binding behaviors. When different types of protein interactions are considered on a large dataset, specific analysis of interfaces can facilitate understanding of the principles of protein binding. So, several good classification methods have been proposed for predicting different types of protein interactions, such as ACV [26], NOXclass [18] and DiMoVo [70]. But they mostly employ physicochemical properties of interface residues/atoms, e.g., residues with ASA change $>0.1 \text{ \AA}^2$ [22, 31] or $>1.0 \text{ \AA}^2$ [18, 30] upon the formation of complexes. It is well-known that binding hot spots are significantly important in the stability of proteins and hot spot residues can capture more specificity of various binding behaviors of proteins. But it is not clear how this knowledge can help the prediction of binding types.

(e) **Computational definition of binding interfaces and their contacts** is a fundamental but intractable problem in previous works and in computational structural biology. In literature, although several computational definitions for binding interfaces were proposed as discussed in Section 1.1.4, none of them exactly follows the biological meaning of protein interfaces. An evidence is crystal packing in PDB or decoys in protein docking. However, these computational definitions still will detect a larger ‘artifact interfaces’ in crystal packing, no matter the definition is based on the change of atom/residue ASA, or on atomic distance, or on Voronoi diagrams

of protein complexes. What is even worse is that it is very hard to distinguish these crystal-packing ‘binding interfaces’ from true ones. Under these computational definitions, crystal packing contacts may also make interfacial analysis full of noises. Anyway, crystal-packing interfaces clearly and intuitively expose the disadvantages of these computational definitions for protein interfaces.

1.2.3 Our Objectives

As discussed before, protein binding interfaces are complicated and reconditely crucial in living organisms. Thus, those previous works still contain many limitations, although they achieved a lot of interesting findings. Fortunately, binding hot spots are confirmed to be significantly important to protein binding interfaces. Thus in this thesis, we start our work with binding hot spots, and dissect protein binding interfaces and their binding hot spots under the following two new models.

- At a residue level, we propound the ‘Double Water Exclusion’ (DWE) hypothesis to refine the profound and influential O-ring theory and to better model binding hot spots.
- At an atomic level, we design β atomic contacts to improve those computational definitions for protein binding interfaces.

In detail, the aims of this thesis are (i) firstly to model binding hot spots by the new ‘Double Water Exclusion’ (DWE) hypothesis, incorporated with DWE bicliques; (ii) then to analyze DWE bicliques to understand specific binding behaviors of protein interfaces with different types, followed by applying the DWE hypothesis to classifying different types of protein interactions; and after that (iii) to design β contacts to analyze protein binding interfaces at an atomic level, including evaluating β contacts in estimation of $\Delta\Delta G$ of alanine mutations under DWE and in identification of biological interactions from crystal packing.

1.3 Major Contributions

The major contributions of this thesis are as follows:

1. **We propound a new hypothesis, ‘Double Water Exclusion’ (DWE), to model binding hot spots.** The DWE hypothesis refines the seminal O-ring theory by extending the water exclusion principle. In detail, DWE provides nice characterization for hot spots’ water-free environment from inside hot spots (there is no accommodation for water molecules among hot spot residues, which is not

specified by the O-ring theory) and around hot spots (exist residues forming O-ring to protect hot spot residues from the solvent attack). DWE also incorporates another two influential hot spots' propositions: the 'coupling' hypothesis and the 'hot-region' principle. In this thesis, we use maximal bicliques to model compact regions of interfacial residues under DWE. Our verification of DWE bicliques with ASEdb [86](a database storing hot spot residues that have been verified by alanine scanning mutagenesis) shows that DWE bicliques are rich in true hot spots.

2. Under DWE, we introduce Z-score to assess biological significance of interfacial residues including binding hot spots and DWE bicliques. We design Z-score for capturing the probability of a contact residue occurring in or contributing to protein binding interfaces. To do that, we take false biological interactions, i.e., crystal packing, as the reference state to calculate residue pairwise potentials under DWE. Then, we generate a null distribution of this potential score from artifact crystal packing contacts. After that, we determine the Z-score significance of a contact residue according to this null distribution given its specific potential score. Since binding hot spots contribute greatly to binding free energy, we assume that they should have large absolute values of Z-score and thus adopt Z-score to classify non-hot spots from hot spots. On the two benchmark datasets ASEdb and BID, the proposed Z-score method is mostly superior to earlier methods, demonstrating that Z-score can truly capture biological significance of binding hot spots and DWE bicliques. In particular, the Z-score method is excellent in hot spot prediction for the most hydrophobic residues, the least hydrophobic residues, as well as the aromatic residues.
3. Unique DWE bicliques are found to intuitively characterize protein interfaces. Under the DWE hypothesis, we investigate DWE bicliques for three types of protein interactions: crystal packing, obligate and non-obligate interactions. Various binding behaviors of these different types of interactions could be clearly described by biclique patterns, especially unique bicliques: (i) the unique bicliques to biological interactions can exhibit a clear picture of the specificity of protein binding behaviors; (ii) the unique bicliques to obligate and to non-obligate interactions suggest different binding behaviors in the two types of biological protein interactions. Also, the composition of residues and their pairs in DWE bicliques, such as interaction-preferred residues (ipRs) and interaction-preferred residue pairs (ipRPs), provide an intuitive picture of preference of residues and their pairs for these three types of interactions.
4. We construct a new classification method, OringPV, to classify different types of protein interactions. OringPV is based on the propensity vectors of low-ASA

residues pairs in O-ring-surrounded regions of protein binding interfaces. This novel notion of low-ASA residues pairs first combines residue pairs with the O-ring theory and DWE. The prediction ability of OringPV is evaluated on three datasets with a variety of classification tasks and various evaluation frameworks. The results demonstrate that OringPV has high prediction accuracy, outperforming the other existing classification methods, such as NOXclass [18] and DiMoVo [70].

5. **We propose a new criterion of β contacts to capture critical contacts in protein binding interfaces.** β atomic contacts have the foundation on β -skeletons, and require that there is no other atom in a forbidden region fr between i and j . fr is defined by β ; when $\beta = 1$, fr is a sphere with the midpoint of i and j as the center and the distance of i and j as the diameter. fr is similar to van der Waals spheres of atoms. This definition can detect critical “perceptually meaningful” contacts across an interface. Our investigation shows that β contacts are only a small fraction of those contacts simply based on a distance threshold. Our following two evaluations also demonstrate that β contacts are superior to distance-based contacts to be a precise model of atomic organization in protein 3D structures. Thus, β atomic contacts provide a new fundamental unit in improving other structural analyses for folding and binding of proteins and RNA/DNA.
6. Using β contacts, we design β ACV at an atomic level to distinguish crystal packing from biological interactions. In a comprehensive comparison, β ACV shows superior classification performance to several existing methods and also to the classifier using those contacts based on a distance threshold.
7. Using DWE and β contacts, we develop a simple method, β ACV_{ASA}, to predict binding free energy and binding hot spots. In particular, we design one β ACV_{ASA} for polar residues and the other for non-polar residues, since different types of residues have various physicochemical properties. On an ASEdb dataset with 396 alanine mutations, we compare β ACV_{ASA} with other methods, including ACV based on distance-cutoff contacts, ACV of β contacts but without ASA integration, Z-score, and two gold standard methods FoldX and Robetta. In the comparison, the best prediction performance of β ACV_{ASA} demonstrates that β contacts are better than distance-based contacts to capture critical atomic contacts in protein binding interfaces. It also reconfirms that DWE water exclusion is a necessary condition of protein binding hot spots.
8. We apply predictive methods, including the new methods above, to understand cross-neutralization of the 2D1-antibody binding to H1N1 influenza viruses in both 1918 and 2009. In the investigation, we find that the six mutations at the HA1’s

epitope from 1918 to 2009 should not harm their binding to 2D1. We also discover three distinguished hot spot residues, including Lys¹⁶⁶, common between the two epitopes. These common hot spots again can explain why 2D1 cross-reacted. We believe that these binding hot spot residues are mutation candidates which may help H1N1 viruses evade the immune system. Meanwhile, our β ACV also correctly classify the 2D1 binding to 1918 H1N1 and to 2009 H1N1 as true biological binding, and amazingly identify the 2D1-2007 H1N1 binding as a false interaction. The identification of these hot spot residues and the structural analysis are potentially useful to fight against H1N1 viruses.

1.4 Organization

This thesis is organized as follows. Please refer to Figure 1.8 for an intuitive detail.

In Chapter 2, we review the works in literature on protein interfaces from the following three perspectives. We give an introduction of earlier works on binding hot spots, including several influential propositions to model binding hot spots and classification/estimation methods to predict binding hot spots. After that, we present comprehensive analysis and understanding of protein interfaces in previous works. Finally, we describe several endeavors to identify the types of protein interactions.

We propound our ‘double water exclusion’ hypothesis (DWE) and DWE biclique patterns in Chapter 3. In this chapter, we introduce the DWE hypothesis, and design an algorithm to detect biclique patterns under DWE from protein interaction pairs. Then, we verify that these bicliques are rich in true hot spots according to ASEdb to highlight the significant importance of the DWE hypothesis.

To assess biological significance of binding hot spots and DWE bicliques, we design Z-score in Chapter 4. In the Z-score calculation under DWE, we use residue pairwise potentials to represent a surrounding of a contact residue with crystal packing as the reference state, and then take contact residues in crystal packing to generate a null distribution of potential scores. After that, the Z-score significance of a contact residue with a specific potential score is determined according to this null distribution. Finally, we assume that binding hot spots should have a larger absolute Z-score and employ Z-score to classify hot spot residues from non-hot spots. The classification performance is evaluated on the ASEdb and BID datasets.

In Chapter 5, we analyze DWE bicliques to understand the specificity of protein binding behaviors in three types of protein interactions: crystal packing, obligate and non-obligate interactions. In detail, we compare DWE bicliques in crystal packing and in biological binding (including obligate and non-obligate interactions) to see which DWE bicliques are signature patterns in biological binding and which are common. We also

evaluation frameworks include within-dataset, cross-dataset comparison, and leave-one-out cross-validation. OringPV’s classification performance is also compared with existing benchmark methods, such as NOXclass [18] and DiMoVo [70].

In Chapters from 3 to 6, protein binding interfaces and their binding hot spots are analyzed at a residue level. In Chapter 7, we propose β contacts to analyze protein interactions at an atomic level. In this chapter, we provide a detail description of what is β contact and discuss why β contact is better to define atomic contacts in protein binding interfaces. To evaluate β contacts, we use atomic contact vectors (ACV) of β contacts to represent interfaces of homodimers and crystal packing. Then, we take them as input of a linear SVM and of a feature-selection algorithm SVM-RFE to train our classifier β ACV. The classification performance of our β contacts is then compared with those achieved by several existing methods on four datasets. The superior classification performance of our β ACV demonstrates that β atomic contacts can truly identify critical specific contacts in protein binding interfaces.

Using β contacts, we also develop a new method, β ACV_{ASA}, to predict binding free energy and binding hot spots. In β ACV_{ASA}, we design several formula to integrate water exclusion and β contacts into an ACV for representing an alanine mutation in protein binding interfaces. In particular, we construct two β ACV_{ASA}, one for polar residues and the other for non-polar residues. After that, we compare β ACV_{ASA} with (i) ACV_{ASA} based on distance-based contacts to demonstrate which one is better to model protein binding interfaces and their hot spots, distance-based contacts or β contacts; (ii) β ACV which has no ASA integration to reconfirm whether water exclusion is a necessary condition of binding hot spots, and (iii) other existing methods, such as Robetta and FoldX, to show the prediction capacity of β ACV_{ASA}.

To summarize our works here, predictive methods, including our new methods above, are applied in Chapter 9 to understanding the 2D1-antibody binding behaviors to H1N1 influenza viruses in 1918, 2007 and 2009. In detail, binding hot spots and non-hot spots are identified for the three binding interfaces of 2D1 to 1918 H1N1, to 2007 H1N1 and to 2009 H1N1. It is interesting to see whether the interfacial mutations from 1918 H1N1 to 2009 H1N1 destruct the 2D1 binding, and which common interfacial residues greatly contribute to the 2D1 cross-neutralization of 1918 H1N1 and 2009 H1N1. In contrast, why 2D1 cannot bind to 2007 H1N1 is also computationally uncovered.

Finally in Chapter 10, we conclude this thesis, and give the interesting problems in the future.

At the end part of this thesis, the Appendix A is provided for a simple tutorial on maximal bicliques. The formal definitions of bipartite graphs and their bicliques are described, followed by a brief introduction of the LCM-MBC method for mining maximal bicliques. The Appendix B introduces some machine-learning algorithms used in this thesis, while the Appendix C provides some supplementary information for Chapter 4 and 8.

1.5 A Brief Summary of the Research Experience during my Candidature

I started my PhD study in January 2008. In the first year of my candidature, I learned a lot Bioinformatics from the courses such as ‘Methods & Tools of Proteomics’ and ‘Introductory Bioinformatics’; I am strongly attracted by the miraculous 3D structures of proteins in living organisms: the beautiful α -helix and β -sheet, the Ramachandran plot and the unique folding of similar protein sequences. All of them extremely fascinate me and I decided to focus on 3D structural analysis of proteins during my candidature.

Since the second half year of 2008, I worked under Dr. Jinyan Li on the hypothesis of ‘double water exclusion’ (DWE) to refine the influential O-ring theory. Under the supervision of Dr. Li, I implemented the algorithms to compile a non-redundant dataset of obligate and non-obligate interactions from PDB and to mine biclique patterns under DWE, and then analyzed these bicliques and evaluated them on the ASEdb dataset. This work finally was submitted to Bioinformatics and published in 2009. Chapter 3 is the main part of this work.

During this DWE work, I realized that the distinction of different types of protein interactions (obligate vs non-obligate, transient vs permanent, and/or true biological binding vs false biological binding) will remove noise protein binding from PDB and then help specific analysis of protein binding. However, previous works on this task had a lot of false positives and false negatives. According to the O-ring and DWE theories, I felt that those residues on O-ring could harm the classification performance in previous works. Hence, I proposed to use propensity vector of those ring-inside residues for this classification task and finally improved the performance. This work in Chapter 6 was published in *Proteins: Structure, Function and Bioinformatics* in 2009.

The good results in these two works inspired me to investigate the DWE biclique patterns and their residue composition in different types of protein binding, since different types of protein interactions have various binding behaviors. In this work, some findings included interesting unique bicliques to different types of protein binding and various interpretable residue compositions. This work in Chapter 5 was published in *BMC bioinformatics* in 2010.

In the above analysis, Dr. Li and I found that it is more meaningful to output DWE biclique patterns with high statistical and biological significance. I noticed that crystal packing is a nice reference state to produce biologically significant patterns. With this idea in mind, I proposed Z-score in Chapter 4, which was accepted by The Tenth Asia Pacific Bioinformatics Conference 2012 and will be published in the supplementary issue in *BBA-Proteins and Proteomics* in 2012.

The above publications made me feel that previous works of both mine and others are still far away from solving the puzzles of protein 3D structures. I decided to dive into an

atomic level instead of staying in a residue level. The reason is simple: atomic contacts are fundamental in residue contacts and further in protein 3D structures. If atomic contacts are not defined precisely, there are a lot of unexpected noises in the high-level (including the residue level) structural analysis. As discussed in Section 1.1.4, atomic 3D organization is not well defined. In previous works, some defined atomic contacts have direct contact areas while others are almost completely interrupted by other atoms. But both received equal attention. I finally conceived of β contact to precisely describe atomic organization in protein 3D structures, and applied this concept to the classification of interaction types and binding free energy prediction of alanine mutations. These works in Chapters 7 and 8 were submitted to Bioinformatics and PLoS ONE.

Meanwhile, I am always trying to apply my methods to infectious diseases. By chance from BIRC student symposium on Bioinformatics in October 2011, I learned that 1918 H1N1 and 2009 H1N1 are very interesting and should be a nice case study of my works. Hence, I used my previous methods to investigate why 2D1 can identify both of the H1N1 viruses, although there are several mutations. This work in Chapter 9 was submitted to Bioinformatics and published in 2011.

In addition, I am also interested in protein interaction networks, protein structure predictions, and time-course gene expression. One work on protein-protein interaction networks was published in the conference of Bioinformatics and Biomedicine (BIBM) 2009, while another work on time-course gene expression was published in the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD) 2010. These two works are not closely related to the topic in this thesis and not included.

Chapter 2

Literature Review

Protein interfaces are very complicated to analyze; they can be dissected from several perspectives, as shown in Figure 1.7 in Chapter 1. In this section, we review previous works that are closely related to our interested problems in this thesis. We first present the models and prediction methods for binding hot spots. Then, we describe statistical analysis of interfacial properties. Finally, we provide an introduction of the methods in literature to classify different types of protein interactions.

2.1 Approaches to Modeling and Prediction of Binding Hot Spots

Previous works dissecting binding hot spots can fall into three groups with different purposes: (i) some studies tried to characterize topological properties of binding hot spots, e.g., the O-ring theory; (ii) many other works endeavored to estimate binding free energies or to predict binding hot spots qualitatively, for example, Robetta [76], FoldX [75] and KFC [79]; and (iii) several other authors built databases/servers based on wet-lab methods or computational methods to assist the analysis of binding hot spots, such as ASEdb and Hotsprint. Since 2000, several surveys [11, 87, 88] summarized computational analysis of binding hot spots, aiming at facilitating development of more effective prediction methods for binding free energy and binding hot spots.

2.1.1 Modeling Binding Hot Spots

Hot spots of binding free energy were initially conceptualized by Clackson and Wells (1995), supported by an outstanding finding that a “functional epitope” (hot spot) be-

Table 2.1: Summary of main previous works in dissection of binding hot spots.

objective	previous works	findings and properties
	[38]	the pioneering works of binding hot spots
Characteristic of hot spots	[36]	the O-ring theory
	[89, 90]	high correlation of structurally conserved residues and the experimental enrichment
	[48]	hot spots tend to couple binding interfaces
	[52]	hot regions of binding hot spots
Estimating $\Delta\Delta G$	Robetta [76]	decomposing $\Delta\Delta G$ into several terms
	FoldX [74, 75]	such as hydrogen bonds,
	CC/PBSA [91]	Coulomb electrostatics, Lennard Jones potential
Identify hot spots	PP_SITE [78]	
	MINERVA [66]	quantifying different levels of protein information, including structure, sequence and molecular interactions
	KFC [79, 92]	
	Hotsprint [60]	using simple rules
	HotPoint [82]	
	HSPred [93, 94]	
	APIS [80]	machine-learning approaches
	PCRPI [95]	
	GCR [64]	
	[96]	graph-based approaches
	[97]	
	[41]	based on tertiary structures
	ISIS [81]	based on primary structures
Database of hot spots	ASEdb [86]	each mutation has a quantitative $\Delta\Delta G$
	BID [98]	each mutation has a qualitative measure: ‘Strong’, ‘Intermediate’, ‘Weak’ or ‘Insignificant’

tween human growth hormone and its receptor accounts for more than three-quarters of the binding free energy [38]. This hot spot was found to be a central hydrophobic region dominated by two Trps and also geometrically surrounded by energetically less important contact residues that are generally hydrophilic and partially hydrated [38].

On the basis of the pioneering observations and studies of Clackson and Wells (1995), Bogan and Thorn (1998) investigated 402 alanine mutations and was the first to formalize a hypothesis named the O-ring theory to more intuitively characterize the topological shape of the surrounding residues of binding hot spots [36]. The O-ring theory confirmed that occlusion of solvent is a necessary condition for highly energetic interactions (hot spots) and the residues on O-ring likely function to occlude bulk water molecules from binding hot spots.

Following this step, after observing high correlation between structurally conserved interface residues and the experimental enrichment energies [89, 90], Nussinov and her colleges suggested two specific organizations of hot spots in binding interfaces [48, 52]: on one hand, both experimental hot spots and conserved residues tend to couple a two-chain interface with higher local packing density [48]; on the other hand, binding sites (one side of the interfaces) might have several ‘hot regions’, which are locally tightly packed regions containing the clustered, networked, structurally conserved residues [52].

Although the above works provided some nice topology organizations of binding hot spots, none of them demonstrate what is the topology between hot spot residues and how two hot spots ‘couple’ across binding interfaces. This problem was discussed in (a) of Section 1.2.2.

2.1.2 Estimating $\Delta\Delta G$ of Binding Hot Spots

With the above topological organization and understanding of binding hot spots, prediction methods were designed to calculate $\Delta\Delta G$ of residue mutations or qualitatively identify binding hot spots. The knowledge used in these prediction methods can be categorized into three kinds [74]: (i) one is from physical effective energy functions; (ii) another is statistical potentials of residue or atom contacts based on protein databases; and (iii) the other is empirical information obtained from protein engineering experiments. Based on these kinds of knowledge, atomistic simulations were used to calculate $\Delta\Delta G$ of residue mutations. Atomistic simulations apply such strategies as the rigorous free energy perturbation [99], thermodynamic integration [100], and the approximate approaches (e.g., MM-PBSA) [100, 101].

However, atomistic simulations are time-consuming. The most rapid approaches to the estimation of $\Delta\Delta G$ of residue mutations are the empirical or knowledge-based (statistical) approaches in conjunction with simple physical models [76]. The well-known examples include FoldX and Robetta. Robetta is a simple physical model for estimating binding energy of hot spots [76, 77]. This method uses all heavy atoms and polar hydrogens to represent proteins. Then, it computationally mutated a residue into alanine (a residue type Ala), and then repacked a local sphere of 5 Å radius of the site of the mutation. Binding free energy changes upon mutations were calculated by an energy function which linearly combines such terms as Lennard-Jones potentials, an orientation-dependent hydrogen bond potential, Coulomb electrostatics, and an implicit solvation model.

Similarly, FoldX [74, 75] uses a linear combination of empirical terms to calculate binding free energy. It uses several empirical terms similar to Robetta, such as hydrogen bonds (water-intermediate hydrogen bonds included) and Coulomb electrostatics; other empirical terms used by FoldX are hydrophobic and polar solvation, the Van der Waals

terms and so on. The weights of these empirical terms were optimized by empirical mutant data of protein experiments.

Recently, Benedix *et al.* proposed a structure-based method, CC/PBSA for fast estimation of the effect of mutations for protein folding and binding [91]. Given a protein structure, no matter it has a wild type or a mutant, CC/PBSA uses the program Concoord to generate alternative protein conformations. Concoord [102] first finds pairwise interatomic distance bounds, and then starts from random coordinates and iteratively corrects the coordinates until satisfying all distance constraints. After that, CC/PBSA considers the following terms to design its energy function: Coulomb electrostatics, polar solvation free energy, a Lennard Jones potential for solute-solute interactions and protein-protein interaction surface. Several of these terms are similar to those used in Robetta and FoldX.

All of these computational methods achieved good prediction performance based on experimental mutations. For example, the overall correlation between the observed and Robetta-calculated changes in binding free energy had an average unsigned error of 1.06 kcal/mol for interface mutations [76]; and on 367 mutant of 9 protein complexes, CC/PBSA obtained the correlation with $R = 0.79$ and $\delta = 1.19$ kcal mol⁻¹.

2.1.3 Identifying Binding Hot Spots Qualitatively

However, predicted energies by the above methods still have a large discrepancy from experimentally measured energy changes [66]. Thus, computational methods were designed to qualitatively identify binding hot spots using those generated knowledge from protein structures, such as atomic contacts or evolutionary/structural conservation; the used protein structures can be quaternary structures, tertiary structures, or even primary structures with orderly increasing difficulty to accurately predict binding hot spots.

2.1.3.1 Hot spot Identification based on Quaternary Structures

Most of previous prediction methods of binding hot spots were designed based on known protein quaternary structures. Gao *et al.* (2004) [78] developed PP_SITE using hydrogen bonds, hydrophobic and van der Waals interactions to qualitatively estimate individual contribution of each interfacial residue to protein binding. This method correctly predicted 75 hot spot residues with 88% success rate (recall). Recently, Cho *et al* [66] combined 54 multifaceted features to develop their predictive model, MINERVA, for interaction hot spots; these features are composed of different levels of protein information, including structure, sequence and molecular interactions. They then used a decision tree to select the best subset of features, and treated them as input of SVM to build their classifier MINERVA. They claimed that MINERVA is better than Robetta, FoldEF and KFC. In their analysis, they found that weighted atomic packing density, relative surface

area burial and weighted hydrophobicity are the top 3 features for identifying hot spots; π -related interactions, hydrogen bonds, and salt bridges are observed to be closely related to binding hot spots as expected.

Hot Spot Identification based on Simple rules KFC [79, 92] uses a decision-tree model to produce some rules for classifying binding hot spots. KFC comprises K-FADE (based on shape specificity features calculated by the Fast Atomic Density Evaluator) and K-CON (based on biochemical contact features). K-FADE and K-CON are two decision tree classifiers to improve the ability of hot spot prediction. These two decision trees employ the following features to represent a residue: physical and chemical features, shape specificity, and biochemical contacts such as atomic contacts, hydrogen bonds and salt bridges. Similarly based on simple rules, Hotsprint [60] detects hot spots using some thresholds of conservation, ASA, and residue propensity, while HotPoint [82] combines conservation, ASA and residue pairwise potentials to produce hot spot rules.

Machine-learning Algorithms to Identify Hot Spots Since 2009, machine-learning algorithms, such as SVM and Bayesian networks, have been actively applied to improve prediction performance of binding hot spots. Lise *et al.* [93] considered energetic terms such as van der Waals potentials, solvation energy, hydrogen bonds and Coulomb electrostatics, and took them as input features of SVM to classify binding hot spots. They found that transductive SVM can achieve the best performance with precision 56% and recall 65% for binding hot spots with $\Delta\Delta G \geq 2$ kcal mol⁻¹. Further, they treated predictions involving either an arginine or a glutamic residue separately, and improved the classification performance with precision 61% and recall 69% [94]. It seems that different types of residues may have their unique way of contributing to protein binding. It should be better to construct individual hot spot classifiers for each type of residues when enough mutations are available.

Xia *et al.* [80] introduced an ensemble classifier of nine SVM classifiers to predict binding hot spots. This method investigated a wide variety of 62 features from a combination of protein sequences and structure information, and used F-score to select non-redundant and relevant features for SVM classifiers. In particular, this method combined a new feature, protrusion index with solvent accessibility, to significantly improve the prediction performance.

Assi *et al.* used Bayesian Networks to design a novel probabilistic method [95] for binding hot spots. This method combines three main sources of information. One source is energetic term of computational alanine scanning by FoldX, another source is evolutionary determinants of mutated residues and of their contact residues, while the other source is structural environment such as proportion of side chain atoms involving contacts and the ratio of the number of structurally neighbor residues over the average number of neighbor residues of the type of mutated residues. This method achieved 0.71 F1 score on a BID dataset.

Network-based Approaches Different from these feature-based methods above, graphs and networks of residues are also used to investigate the topological organization of binding hot spots. Similar to O-ring, Li and Li proposed a novel descriptor of atoms and residues, called burial level, to enhance hot spot prediction performance [64]. By this method, they built an atomic contact graph for a protein complex, where vertices are atoms and edges are atom contacts. They defined the burial level of an atom in this graph as the length of the shortest path from this atom to its nearest exposed atom to the bulk solvent. The burial level of an atom or a residue indicates the extent it is buried inside the complex. As hot spot residues are protected by O-rings [36, 62], hot spot residues always have low ASA and high burial levels. But Li and Li claimed that a high burial level seems to be more sufficient than ASA: there are very few highly buried interfacial residues that are not hot spot residues. Based on this concept, a GCR model is built for binding hot spot prediction [64] and has achieved good performance.

Similarly, network-based approach [96] is also used to identify the properties of key interfacial residues. In this approach, protein complexes are represented by graphs with residues as vertexes and residue contacts as edges. On a dataset with 48 dimer complexes, these protein complex graphs were proved to exhibit characteristics that resemble a small-world network; in these networks, 83% of predicted highly central residues were found to correspond to or directly contact an experimentally determined hot spot. Also, Tuncbag *et al.* used graphs to visualize residue contact networks of protein interfaces. Edges in this graph are weighted according to an energy function derived from knowledge-based potentials. Then, they constructed min-cut tree to simplify and summarize contact graphs [97]. They observed that binding hot spots are the highest degree node and also in a few paths in the min-cut tree.

2.1.3.2 Hot spot Identification based on Tertiary or Primary Structures

All of the above analyses require protein quaternary structures. Grosdidier and Recio [41] tried to identify binding hot spots from protein tertiary structures. They used computational docking to produce docking solutions for protein tertiary structures; then they estimated normalized interfacial propensity (NIP) for each residue according to averaged ASA of 100 lowest-energy docking structures. Those residues with higher NIP were considered to be predicted hot spots. This method was reported to have comparative prediction performance to those methods based on protein quaternary structures.

Further, another work ISIS [81] endeavored to predict binding hot spots from amino acid sequences, without determined 3D structures of proteins. ISIS employed features are those kinds of information generated from amino acid sequences, such as sequence environment of residues, evolutionary profile, predicted secondary structures and solvent accessibility, and conservation. In this method, almost all binding residues predicted by ISIS were found to experimentally have significant effect on protein binding, while more than 90% of the negative predictions contribute little to protein binding.

2.1.4 Databases for Binding Hot Spots

To facilitate the analysis of binding hot spots, several databases of binding hot spots were built. ASEdb (Alanine Scanning Energetics Database) [86] and BID (Binding Interface Database) [98] store binding hot spots determined by wet-lab experiments. These two databases were widely used to verify the effectiveness of hot spot prediction and estimation. The difference of these two databases is that hot spot residues in ASEdb have quantitative $\Delta\Delta G$, while binding strength in BID is qualitative measures such as ‘Strong’, ‘Intermediate’, ‘Weak’ or ‘Insignificant’.

However, the size of binding hot spots in these two databases is very limited, since wet-lab experiments are expensive and time-consuming. Thus, several servers or databases of computational hot spots were also constructed, for example, KFC server [79] and Hotsprint database [60]. HotSprint [60] stores those computational hot spots for 35776 protein interfaces among 49512 protein interfaces extracted from the multi-chain structures in PDB (as of February 2006); those computational hot spots were derived based on residues conservation score, propensity, and ASA, and they are highly correlated with the experimental hot spots with a sensitivity of 76%.

As a summary, these previous works above are grouped in Table 2.1 for a quick review. Table 2.1 also suggests that although there have been a lot of works on the analysis of binding hot spots, the principles underlying binding hot spots are not yet clearly elaborated. In particular, the prediction performance is still low, even when almost of all available sequence, structural and molecular-contact information and features are used, as reviewed in Section 2.1.2 and 2.1.3. Meanwhile, there are conflicting observations for some features in terms of whether a feature can help uncover the puzzles behind binding hot spots. These problem were also discussed in (b) of Section 1.2.2.

2.2 Approaches to Analyzing Protein Interfaces

Since the earlier works on analyzing protein interfaces originated from 1970s [73], the analysis of protein interfaces, especially comparison of interfaces with other segments in protein 3D structures including protein surfaces, is widely investigated to uncover the specific knowledge in protein binding. Glaser *et al.* [23] investigated a non-redundant set of 621 protein interfaces and found that residue frequencies and pairing preference had no difference at protein interfaces. Saha *et al.* [20] produced the environment for each residue, unsymmetrical propensity values of 20 residues to be in contact with a given residue in three types of protein interactions; then, they similarly concluded that there was no appreciable change of residue environment in going from tertiary structures to interfaces.

However, there are contradictory conclusions about the comparative analysis for binding interfaces. Ofra and Rost [10] estimated the preferences of residue contacts based on six types of protein interfaces, and found that the preferences differed remarkably between the six types of interfaces, including intradomain, domain interfaces and different types of protein interfaces. This problem was also discussed in (c) of Section 1.2.2.

This inconsistent conclusion is partly attributed to the great structural and functional diversity of protein interfaces, such as obligate interactions and non-obligate interactions. That is, these different types of interactions possess various binding behaviors. Therefore, it is necessary to analyze protein interfaces with different interaction types separately—including crystal packing—for obtaining specific knowledge of protein binding.

2.2.1 Dissecting Various Types of Protein Interfaces Based on Physicochemical Properties

Jones and Thornton did the impressive work to dissect interfacial difference in various types of protein interfaces [15]. In their work and many other following works, physicochemical properties in different types of interactions were analyzed and highlighted to reflect their biological roles. Generally, the interfacial physicochemical properties can be divided into several kinds [16] as follows. One kind are physical features, i.e., interface size and contact area. Biological interfaces were reported to have larger interface area than non-biological interactions [17, 19, 22, 31, 33, 34, 35]. Permanent complexes seemed to have averagely larger interface areas than transient complexes [15, 18, 31], and “standard size” of contact area in heterocomplexes was in the range of 1200–2000 Å² [65].

The second kind are chemical features, such as hydrophobicity and polarity. The interfaces in obligate interactions were predominantly non-polar residues [19] and characterized to be more hydrophobic [15, 22, 31, 32].

The third kind are geometric features, including planarity, shape complementarity, circularity [15, 30] and secondary structures [16]. Most of homodimer interfaces showed more shape complementarity and less planarity than heterocomplex interfaces [15, 18].

2.2.2 Analyzing Residue Composition and Residue Preference in Various Types of Protein Interfaces

Another kind of physicochemical properties in interfaces are interfacial residue conservation, residue composition and preference, propensity of residue pairs, and atomic pairs [22, 26, 27, 28, 83]. Different types of interfaces were observed to vary greatly in composition of residues and their pairs [10, 15, 18, 21]. Biological interactions were detected to be more conserved than crystal packing [17, 18] and abundant in aromatic

residues [18]. Besides, biological interactions had different residue composition from the rest of protein surfaces [22, 30, 83], while crystal packing possessed similar composition to the rest of protein surfaces as a whole [35].

For two different types of biological interfaces, the hydrophobic residues preferred the interfaces of homodimers rather than those of heterocomplexes, while polar residues had higher propensity in the heterocomplex interfaces than homodimer interfaces [15]. Identical amino acids were also observed to frequently occur in obligate interactions [10, 20, 21]. In particular, Bogan and Thorn (1998) also found that binding hot spots were enriched in tryptophan, tyrosine and arginine [36].

Although intensive studies have focused on protein interfaces, the review indicated that none of single property can make a clear distinction among different types of interfaces. It seems that no strict and simple rules were found underlying protein interactions based on these existing physicochemical properties.

2.3 Approaches to Classifying Different Types of Protein Interactions

From Jones and Thornton's work [15], it is clearly seen that there exists close relation between various binding behaviors and different types of interactions. Knowing the types of protein interactions can help to specify and simplify understanding of protein interfaces. Meanwhile, false interactions, i.e., crystal packing, exist in protein structure databases such as PDB. These false interactions should be removed when analyzing the structures of protein complexes. However, for those structures determined by the techniques such as X-ray crystallography, it is not easy to infer the types of protein interactions (including crystal packing) without ambiguity. Computational methods can help this and use machine-learning algorithms to automatically construct powerful classifiers after these interactions are represented by helpful features discussed in Section 2.2.

Generally, the process of computational classification methods can be divided into three essential steps. First, interfaces are defined according to contact requirement [26, 27], or ASA changes [18], or Voronoi tessellation [70]. Second, interfacial physicochemical properties are extracted, such as contact size, conservation and amino acid composition. Third, statistical analysis is performed or machine-learning approaches are trained for classifying interaction types.

These classification methods can perform various tasks, such as identification of biological interactions from crystal packing, or classification of different types of biological interactions, or both.

2.3.1 Identification of Biological Interactions from Crystal Packing

The first task here is to review the methods of identifying biological interactions from crystal packing. Thornton and their colleges [27] made statistical analysis based on atom-pair frequencies across interfaces to differentiate monomeric and homodimeric proteins. Later, they improved the performance to an accuracy 98.3% with the help of contact conservation [17]. Similarly, Ponstingl *et al.* [28] inferred protein quaternary structures from crystal structures in PDB by scoring crystal interactions by size and atom pair frequencies. Liu *et al.* [103] distinguished homodimers from monomers based on atomic potential of mean force, packing density, contact size, and geometric complementarity.

With a different definition of interfaces, Bernauer *et al.* developed structure-based scoring functions [44, 104] and later the DiMoVo method [70] for discriminating crystallographic and biological protein-protein interactions. Their key idea is to use a Voronoi tessellation to nicely describe the geometric and physicochemical complementarities of protein interfaces. On residue-based Voronoi diagrams of protein structures, DiMoVo [70] extracts 106 features from protein interfaces. Then an RBF-kernel SVM is used to select the best feature set on the Bahadur dataset. Its feature-selection process is a backward-removal method [70] as follows. Given a data matrix D with features $F = F_0$, this method removes $f \in F$ one by one and calculates classification performance. f whose removal results in a highest performance is eliminated from F , that is, $F_i = F_{i-1} - f$, $F = F_i$ and i starts from 1 to $\|F\|$. This process continues until F has only two features. Finally, a feature set F_i is selected if it has the highest performance. So, this process checks $\frac{\|F\|(\|F\|-1)}{2}$ potential subsets of features.

2.3.2 Classification for Different Types of Biological Interactions

Computational methods are also designed to specifically distinguish types of biological interactions, including different types of transient interactions.

Cho *et al.* [105] designed 18 molecular interaction types using canonical and non-canonical interactions. Then, they constructed kNN with frequency vectors of each interaction type as input to differentiate the types of transient interactions, such as protease-inhibitors, antibody-antigens and signaling proteins.

Recently, Association Rule Based Classification (ARBC) was also proposed to predict four types of interactions [106]. ARBC used association rule generation and posterior classification to address pattern discovery of the interaction sites. ARBC's advantage was that the discovered association rules could be easily understood and interpreted.

2.3.3 Distinction of Crystal Packing and Different Types of Biological Interactions

More powerful methods are proposed both to classify crystal packing from biological interactions and to differentiate two types of biological interactions. Mintseris and Weng [26] described protein interactions with atomic contact vectors (ACV) after grouping atoms into 18 atomic types. They reported that ACV achieved exciting performance to differentiate both biological interactions from crystal packing and permanent complexes from transient complexes. Later, Block *et al.* [107] also used an ACV method [26] to design their classifiers. They employed three feature selection methods to extract discriminating features from permanent and transient complexes and from homodimers from crystal packing. The feature selection methods include Relief F filter approach, Genetic algorithms and a backward-removal feature-selection method similar to DiMoVo. Then, they applied four machine-learning approaches including decision tree, SVM, Naive Bayes and K-nearest neighbors on the Ponstingl dataset and on the dataset of two types of biological protein interactions used in [26]. Their classifiers classified those interactions in Ponstingl with up to 94.8% accuracy and distinguished the permanent/transient complexes with up to 93.6% accuracy.

Using other interfacial descriptors instead of ACV, Zhu *et al.* [18] built NOXclass to distinguish obligate, non-obligate interactions and crystal packing. NOXclass incorporated descriptors of interfaces, such as interface area, amino acid composition and gap volume, into SVM. It was reported that NOXclass achieved an accuracy of 91.8% to distinguish these three types of interactions.

The classification performance of these above methods is high. The problem is that these methods often focus on all residues/atoms in interfaces. It is widely accepted that only a small fraction of interfaces are actually essential for recognition or binding [81]. This fraction of interface residues or atoms should sharpen the distinction of different types of interactions, but fewer methods utilize this information in the prediction. This problem was also discussed in (d) of Section 1.2.2.

Chapter 3

Double Water Exclusion: a Hypothesis Refining the O-ring Theory for Binding Hot Spots at Protein Interfaces

A binding hot spot is defined as a cluster of residues at protein interfaces that are energetically important in protein complex formation [38, 36]. Although binding hot spots typically constitute only a small subset of interfacial residues, they contribute most to binding affinity and strength of protein interactions. Mutations of hot spot residues will significantly decrease the stability of protein complexes and can greatly reduce the binding affinity and strength of protein interactions. Thus, understanding hot spot residues is a fundamental and imperative problem in molecular biology (see two surveys [87, 108]).

Previous works have already suggested several topological properties of binding hot spots, such as O-ring and hot regions. But the ring-inside organization of binding hot spots is not yet specified. In this chapter, we conceptualize the ‘Double Water Exclusion’ (DWE) hypothesis to model a binding hot spot in the protein interface. Theoretically, DWE refines the O-ring theory and further requires that there is no enough space to accommodate free water molecules between binding hot spot residues at the protein interface. Given a protein interface, a DWE biclique is used to detect a cluster of the interfacial residues satisfying the DWE requirement. The verification shows that DWE bicliques are enriched in true hot spots.

3.1 Introduction

Anatomy, characterization, and statistical analysis of protein-protein interfaces have been broadly and extensively studied under the principles [15, 73] of protein-protein recognition and interactions. A protein binding interface composes of two relatively large, spatially close protein surfaces with good geometric shape and chemical complementarity. The formation of protein chain interfaces is driven by various natural forces such as van der Waals contacts and electrostatic interactions, resulting in the removal of water molecules from the binding sites [109, 110]. This water-removal-then-binding has been well characterized by the influential ‘O’-ring theory [65, 36, 86, 108, 87], which is also called ‘water exclusion’ hypothesis. It highlights that the stability of a complex is mainly determined by only a small number of energetically outstanding residues; it also reveals that these ‘hot spot’ residues are usually located at the center of the interface and surrounded by energetically less important residues that shape like an O-ring to occlude bulk water molecules from the hot spot, abstractly called the O-ring theory. A lab-verified example of the O-ring theory was reported in the pioneering work [38] in which alanine-scanning mutagenesis was used to determine the binding hot spots between human-growth hormones and human-growth hormone-binding proteins. The O-ring theory is profound. However, the organizational topology of the ring-inside, energetically more important hot spot residues is uncertain and not specified by the O-ring theory.

To refine this long-standing theory, we suggest a ‘double water exclusion’ hypothesis to characterize the topological organization of residues in a hot spot and their neighboring residues. On one side, we agree with the O-ring theory that there should exist a ring of residues surrounding the hot spot for avoiding the invasion of water molecules after the complex formation; On the other hand, we suppose that the hot spot itself is water-free. Given a protein interface, we propose to use contact residues that are *densely* interacted in a compact region to model a hot spot. (Here, a pair of residues contact to each other if there exists a pair of atoms whose distance is below the sum of their corresponding van der Waals radii plus the diameter (2.75 Å) of a water molecule.) We take the classical graph term ‘biclique’ [111] to denote this molecular topology, and call it ‘biclique pattern’. Formally, a biclique pattern between two chains in a protein complex consists of two *maximal* groups of residues (one from each chain) such that every residue contacts with all residues in the other group. Thus, if a biclique pattern is observed from a protein complex, then all possible pairs of the residues from the two sides are spatially so close that there is no sufficient room between them to accommodate any free water molecule. That is, this is an interface structure exhibiting a zero-water tolerance property and forming a collective force of multiple, dense atom-atom pairs. Such a structure is therefore useful for lowering the local dielectric constant and enhancing specific electrostatic and hydrogen bond interactions to strengthen the stability of the binding. On the other

hand, by the maximality requirement of biclique pattern, any neighbor residue at one side of the pattern disassociates with at least one residue at the other side. This makes the neighbor residues flexible to form a fence preventing the biclique pattern residues from solvent partially or entirely. Thus, if every residue in a biclique pattern has a small ASA in the complex, then this biclique pattern reflects best the spirit of ‘double water exclusion’—all the inner hot spot residues are organized as a biclique without holding any free water molecule, while there exist a ring of residues blocking the solvent accessibility to the hot spot.

Our biclique pattern concept integrates ideas not only from the O-ring theory but also from the ‘coupling’ hypothesis [48] which is another insightful proposition about hot spot residues. On one hand, the ‘coupling’ hypothesis bridges the relation between hot spot residues from the two sides of an interface, stating that hot spot residues tend to couple across a two-chain interface. On the other hand, as said in the O-ring theory, energetically important hot spot residues are often clustered, and usually located at the center of less important residues. Combining these two observations, it can be seen that the central hot spot residues of an O-ring structure are ‘coupled and interacting’ across the interface. Our biclique pattern concept exactly refines this hybrid idea by specifying that the coupled and interacting hot spot residues are a maximal cluster of residues that have a full ‘water exclusion’ adjacency.

The biclique pattern concept can also be supported by the influential work by [52] where the organization of structurally conserved hot spot residues has been studied. [52] used the term ‘hot region’ to describe assemblies of hot spot residues that are located within densely packed regions with a network of interactions; their analysis over 44 interface clusters with 568 hot spot residues shows that 79% of the hot spot residues were found to be in the hot regions. By the definition of bicliques, it can be seen that the all-versus-all full adjacency required by a biclique pattern is the most optimal network of interactions for a region having two groups of residues. Therefore, our biclique pattern concept theoretically strengthens and signifies the hot region proposition. The maximality requirement also guarantees that the size of a biclique pattern is as large as possible, increasing the binding stability as much as possible. These enhancements and consistency suggest that our theoretical top-down ‘biclique’ approach meets the empirical bottom-up ‘region’ approach at a point where both demonstrate a spatially well organized topology of hot spot residues.

This work is also related to previous studies on contact residues and binding interfaces. Wolfson, Nussinov, and their coworkers [46, 47, 48, 49, 50, 51, 52] defined a protein interface as two parts: contact residues and the neighbor residues of the contact residues. A pair of residues (one from each chain) is defined as ‘contacting’ across the interface if the distance between any two atoms of the two residues is less than the sum of their corresponding van der Waals radii plus 0.5 Å; A residue from the same chain of a contact

residue is defined to be a ‘nearby’ residue of this contact residue if the distance between their two C^α atoms is under 6 Å. This category of definitions also include several variations as proposed in [10, 53, 54, 55, 56, 57, 58]. The key difference from these existing works is that we investigate clusters of contact residues that are densely interacted like a biclique in a compact region where the compactness is measured by the water exclusion principle. Besides, we use the small complex ASA constraint to narrow down the scope of biclique patterns, so that the biclique patterns have a good correspondence to hot spots, and have the best spirit of the ‘double water exclusion’ hypothesis.

Here, we address the relation of biclique patterns and verified hot spots in literature. We want to see whether our ‘double water exclusion’ truly refines the O-ring theory. In our method, we detect biclique patterns by the LCM-MBC algorithm [112] from the complexes in ASEdb. For those interesting patterns with smaller complex ASA, we believe that they are likely to be a hot spot.

3.2 The Double Water Exclusion Hypothesis for Binding Hot Spots

In the ideal DWE hypothesis, a binding hot spot should be completely water-free in protein interfaces. This water-free assumption actually is too strict, and in most protein complexes, this hypothesis produces few interfacial residues for analysis. So, the water-free assumption should be relaxed in real protein binding interfaces. According to the extent of the relaxation of the water-free assumption, there are other two hypotheses as follows. One of the relaxed water exclusion hypotheses is that a binding hot spot contact no free water. That is, there may be some immobilized water molecules in protein 3D structures; these water molecules are so heavily involved in protein folding and binding that they are an integral part of proteins. This is an immobilized-water free hypothesis. The other relaxed water exclusion hypothesis is that the amount of water molecules inside a hot spot is proportional to its phylogenetic evolution progression towards the perfect water-free binding.

In this chapter, we focus on the investigation of the ‘double water exclusion’ hypothesis only; this is roughly equivalent to the immobilized-water exclusion hypothesis, but water molecules are considered implicitly. This investigation is applicable in protein complexes both with and without water information, since many of real 3D structures of protein complexes have no water information in PDB. For explicitly taking water information into account, please refer to the work done by our another group member [64]. The other (and bigger) relaxed hypotheses are left as an open question for interested researchers and ourselves in future research.

3.3 Methods

3.3.1 ASEdb: An Evaluation Dataset for Binding Hot Spots

The evaluation dataset is from the ASEdb database. ASEdb [86] is a searchable database containing hot spot residues that have been verified by the experiments of alanine scanning mutagenesis. Although this is a small database due to experimental difficulties (very slow and labor-intensive), the data quality is high.

This ASEdb database contains 28 protein complexes. However, only 13 of them are matched with PDB entries: 1A4Y, 1AHW, 1BRS, 1BXI, 1CBW, 1DAN, 1DVF, 1GC1, 1JCK, 1VFB, 2PTC, 2HFM and 3HHR. The other 15 protein complexes' structure information is not available from PDB, or not well matched (2 of them). This pre-processing result is consistent with a literature work [78]. In these 13 protein complexes, 439 alanine-mutated residues have been tested, including 268 residues whose $\Delta\Delta G < 0.5$ kcal/mol, 89 residues whose $\Delta\Delta G$ is between 0.5 and 1.5 kcal/mol (called *warm* residues), 30 residues whose $\Delta\Delta G$ is between 1.5 and 2.0 kcal/mol (called *very warm* residues), and 52 residues whose $\Delta\Delta G \geq 2$ kcal/mol (called *hot* residues).

3.3.2 Mining DWE Biclique Patterns from Protein Interaction Pairs

Given a set of protein interaction pairs, it is computationally challenging to efficiently identify biclique patterns due to the maximality constraint and the all-versus-all adjacency. In this work, we take an efficient approach to the localization of biclique patterns at every pair of interacting protein chains. Our algorithm consists of two main computational steps: (i) the construction of DWE bipartites from protein interaction pairs based on the residues' 3D-coordinate information within protein complexes, and (ii) the discovery of maximal bicliques from constructed DWE bipartites. To help to understand the basic concepts of bipartite graphs and their bicliques here, please refer to Appendix A for a short tutorial.

3.3.2.1 Constructing DWE Bipartites for Protein Interactions

Given two interacting polypeptide chains C_1 and C_2 , according to the DWE hypothesis, we define its DWE bipartite as a bipartite graph $G = \langle V_1, V_2, E \rangle$, where (i) the vertices in V_1 and in V_2 represent the amino acids from C_1 and C_2 respectively; (ii) the relative accessibility of all residues in $V_i, i = 1, 2$, is less than a certain threshold T_{asa} ; and (iii) E represents all residue contacts between V_1 and V_2 , and every residue in V_i must contact at least one residue in $V_j, i, j = 1, 2$ and $i \neq j$.

In detail, to construct the DWE bipartite for a given pair C_1 and C_2 , we can firstly construct a bipartite graph from this protein interaction; and then filter out those residues in the bipartite graph by using the constraint of residue accessible surface area.

Constructing Bipartite Graphs Each pair of chains can be transformed into a bipartite graph according to the contact requirement of the DWE bipartites below. In this work, two amino acids from V_1 and V_2 are considered as contact if the minimum of the distances of atoms from these two amino acids is less than the sum of van der Waals radii of the corresponding atoms plus a certain threshold. To ascertain that there is no free water between interacting residue pairs, this threshold, denoted as T_d , is set to van der Waals diameter of water molecules (2.75 Å).

In other words, residue r_{ik} of C_i and residue r_{jl} of C_j , $i, j=1, 2$ and $i \neq j$, contact if and only if the minimal distance among those distances between the atoms of r_{ik} and the atoms of r_{jl} is less than T_d . Here, all heavy atoms in backbone and side chains of amino acids are used. The distance between a pair of atoms $a_{i'}$ from r_{ik} and $a_{j'}$ from r_{jl} is calculated by: $d=d(a_{i'}, a_{j'})-r(a_{i'})-r(a_{j'})$ where $d(a_{i'}, a_{j'})$ is the spatial distance of $a_{i'}$ and $a_{j'}$, and $r(a_{k'})$ is van der Waals radius of $a_{k'}$, $k'=i'$ or j' . Suppose we are given m_h number of protein interactions, the bipartite graph database can be denoted by $P = \{G^i = \langle V_1^i, V_2^i, E^i \rangle | i = 1, 2, \dots, m_h\}$ where G^i represent the protein interactions.

Accessibility Filtering The constructed bipartite graphs P are further processed by using the constraint of water accessible surface area of residues. We take NACCESS [12] to produce the relative ASA for each residue in a protein interaction. Here, the relative ASA of a residue i is the fraction of its ASA over its maximum ASA (The maximum ASA of a residue i is the ASA of its residue type r_i in an extended Ala- r_i -Ala tripeptide.). The remaining ones are only those residues whose relative accessible surface area is less than a certain threshold T_{asa} . In this work, T_{asa} is set as 36%. Theoretically, this threshold should be close to zero, but in real case, it is too strict, leaving too small number of residues for statistical analysis. On the other hand, some hot spot residues have relatively large ASA [86]. How to determine the optimal threshold is still a difficult problem.

The resulting bipartite graphs are called DWE bipartites of protein interactions, denoted by $H = \{G_{dwe}^i = \langle V_1^i, V_2^i, E^i \rangle | i = 1, 2, \dots, m_h\}$ where all bipartites satisfy the DWE requirement due to that (i) the distance of atoms of interacting residues is less than T_d , and there is no room to accommodate free water molecules within interactions; (ii) all residues in bipartite graphs have smaller areas to contact with water molecules.

3.3.2.2 Mining Maximal DWE Bicliques From DWE Bipartites

A biclique is a special bipartite graph where each residue in one partite contacts with every residue in the other partite. A DWE biclique is a biclique from a DWE bipartite. Maximal DWE bicliques $\mathcal{G} = \langle V_1, V_2, E \rangle$ are DWE bicliques where there is no other DWE biclique $\mathcal{G}'$ containing $\mathcal{G}$. In this work, maximal DWE bicliques $\mathcal{G}$ are abbreviated to

$\mathcal{G}=\langle V_1, V_2 \rangle$ without E due to the constraint of all-versus-all interactions between the residues in V_1 and V_2 . In protein interactions, a maximal DWE biclique $\mathcal{G}$ represents densely interacted residue pairs in a compact region.

Given a DWE bipartite $G_{dwe}^i=\langle V_1^i, V_2^i, E^i \rangle$, we take the LCM-MBC algorithm [112] to mine maximal DWE bicliques. The LCM-MBC algorithm needs two parameters: p and q , $p \leq q$. Suppose that V_i is with less vertices than V_j in a maximal DWE biclique $\mathcal{G}$, $i, j=1$ or 2 and $i \neq j$, p is the minimum size of V_i and q is the minimum size of V_j . That is, the LCM-MBC algorithm filters out those $\mathcal{G}$ in which the minimum size of V_i is less than p or the minimum size of V_j is less than q . Assume that the LCM-MBC algorithm detects n maximal bicliques from H , denoted as $M = \{\mathcal{G}_j = \langle V_1, V_2 \rangle | j = 1, 2, \dots, n\}$ where items $\mathcal{G}_j$ are maximal bicliques with amino acids as their vertices.

Not every $\mathcal{G}_j$ in M is useful for our analysis due to that some bicliques are infrequent and random. Therefore, for each DWE biclique $\mathcal{G}_j$, we enumerate H to get its occurrence in protein interactions. Here, a biclique occurs in an interaction if all the residues in this biclique are in the DWE bipartite of this interaction and these residues also maintain the same biclique structure of full contacts. If the occurrence of $\mathcal{G}_j$ is not less than a threshold sup , $\mathcal{G}_j$ is considered to be interesting and frequent; these maximal bicliques are referred to as DWE bicliques. However, the space of possible bicliques is too large. Take bicliques of one partite with 2 residues and of the other partite with 3 residues for example, there are $323,400((C_{20}^2 + 20) \times (C_{20}^3 + 20^2))$ possible bicliques and 20^5 biclique instances if each residue is considered to be independent. In a dataset with only 868 interactions, if each residue is with equal probability, the maximum of the expected support levels for these bicliques is $0.027 (10 \times 868 / 20^5)$, much less than 1. That is, those bicliques whose support is equal to 1 also have a higher support than what they are expected. They have a lower support likely due to that there are limited sample interaction pairs and larger biclique space. The larger biclique space requires a huge number of biological interaction pairs to produce meaningful frequent bicliques. Thus, although producing a larger dataset is a solution, it is quite a time-consuming task if it is a manual work. There is also no gold standard to determine sup . The setting of sup should be decided by the application of bicliques.

Finally, the whole procedure is shown in Algorithm 1 to mine DWE bicliques from a given set of protein interactions.

We note that maximal biclique and its generalized form quasi-biclique subgraph have recently emerged in bioinformatics studies for characterizing protein-protein interactions and their networks [21, 113, 114]. Here, we have explored a novel use of bicliques in structural bioinformatics.

Algorithm 1 Algorithm to mine DWE bicliques from a given set of protein interactions

Require: A set of pairs of interacting polypeptide chains $I_{ppi} = \{\langle C_1^i, C_2^i \rangle, i = 1, 2, \dots, m_{ppi}\}$;
 p is the size threshold for one side of a biclique pattern,
 q the size threshold for the other side;
 T_{asa} is the threshold of residue relative accessible surface area;
 sup is the minimum occurrence of a biclique pattern.

- 1: convert $I_{ppi} = \{\langle C_1^i, C_2^i \rangle, i = 1, 2, \dots, m_{ppi}\}$ into a set of bipartite graphs $P = \{G^i = \langle V_1^i, V_2^i, E^i \rangle | i = 1, 2, \dots, m_h\}$
- 2: produce residue's ASA using NACCESS.
- 3: obtain DWE bipartites $H = \{G_{dwe}^i = \langle V_1^i, V_2^i, E^i \rangle | i = 1, 2, \dots, m_h\}$ by filtering P with the threshold T_{asa}
- 4: use LCM-MBC to mine maximal bicliques from H
- 5: Let $M = \{\mathcal{G}_j = \langle V_1, V_2 \rangle | j = 1, 2, \dots, n\}$ be the set of all maximal bicliques discovered
- 6: **for all** $\mathcal{G}_j \in M$ **do**
- 7: $count = 0$
- 8: **for all** G_{dwe}^i **do**
- 9: **if** $\mathcal{G}_j$ is a subgraph of G_{dwe}^i **then**
- 10: $count++$
- 11: **end if**
- 12: **end for**
- 13: **if** $count \geq sup$ **then**
- 14: $\mathcal{G}_j$ is a DWE biclique
- 15: **end if**
- 16: **end for**
- 17: output all DWE bicliques

3.4 Biclique Patterns are Rich in Binding Hot Spot Residues

To evaluate our hypothesis, we detect biclique patterns from 13 protein complexes in ASEdb by using the algorithm in Section 3.3.2 with the setting $p = q = 2$, $sup = 1$. (We note that under this setting the biclique residues constitute around 10% of the total residues in a chain pair on average over the 13 protein complexes.) Our biclique patterns contain a total of 576 residues. After that, we make queries first to the ASEdb database to test whether our biclique pattern residues are hot spot residues, and then to the HotSprint [60] database.

3.4.1 Verifying DWE Bicliques with ASEdb

Of the 52 hot residues of ASEdb, 94.2% of them (49) are also our biclique residues; of the 30 very warm residues of ASEdb, 80% of them (24) are also our biclique residues; of the 89 warm residues of ASEdb, 64% of them (57) are covered by our biclique residues; and

Table 3.1: Numbers of *very warm* and *hot spot* residues of the 13 protein complexes stored in ASEdb in comparison to those contained in our biclique patterns. The larger $\Delta\Delta G$ the residues are, more likely those are in our biclique patterns.

PDB	$\Delta\Delta G \geq 1.5$ kcal/mol		$\Delta\Delta G \geq 2.0$ kcal/mol	
	In ASEdb	In biclique	In ASEdb	In biclique
1A4Y	4	4	3	3
1AHW	1	1	1	1
1BRS	10	9	9	8
1BXI	7	7	6	6
1CBW	1	1	1	1
1DAN	6	6	3	3
1DVF	19	18	8	8
1GC1	1	0	0	0
1JCK	6	5	4	4
1VFB	6	5	3	3
2PTC	1	1	1	1
3HFM	6	5	5	4
3HHR	14	11	8	7
Total	82	73	52	49
sensitivity	$73/82 = 89\%$		$49/52 = 94\%$	

of the 268 residues whose $\Delta\Delta G < 0.5$, 31.7% of them (85) are contained in our biclique residues. Therefore, we can note that lab-verified *hot spot* residues are almost all (49 out of 52) contained in our biclique patterns; thus we can conjecture that almost all hot spot residues of ASEdb satisfy the property of double water exclusion. When residues are energetically becoming less important from the range *very warm* to *warm* and to the range $\Delta\Delta G < 0.5$, their likelihood to be contained in our biclique patterns becomes smaller and smaller from 80% to 64%, and to 31.7%. Therefore, we can also conjecture that our biclique patterns are less likely to contain energetically less important residues. Table 3.1 details the breakdown of this result for the 13 protein complexes.

We are unable to determine the accuracy of our biclique residues that are also true hot spot residues, as there are 361 out of our 576 biclique residues that have not been tested by the alanine scanning mutagenesis — the 361 biclique residues may be or may not be true hot residues. However, we have at least verified based on ASEdb that if a residue is a hot spot residue, then it can be covered by our biclique pattern. That means, double water exclusion is a necessary condition for a true hot spot residue. However, our biclique patterns also contain some energetically less important residues. This leaves us room to refine the hypothesis of double water exclusion again to further narrow down biclique patterns so that those energetically less important biclique residues can be filtered.

If the total 268 residues ($\Delta\Delta G < 0.5$) are taken as a negative set, the specificity of our method is $1 - 85/268 = 68.3\%$; If the total 133 residues ($\Delta\Delta G \leq 0$) are taken as a negative set, then our biclique residues contain 41 of them, thus the specificity of our method is $1 - 41/133 = 69.1\%$. These results indicate that our biclique residues did not entirely exclude energetically less important residues. However, we would like to note that overall, the double water exclusion hypothesis and biclique patterns can provide helpful insight into protein binding and thus are still biologically acceptable. Alanine scanning mutagenesis, the main biotechnology adopted in ASEdb, examines the binding free energy change of a single residue upon its mutation to alanine. This experimental method does not solve the cooperativity and additivity problems of energetically important residues. Recently, alanine shaving technologies [87] have been proposed. Alanine shaving is a process of making multiple simultaneous alanine mutations. Thus, cooperativity and additivity can be measured by comparing the simultaneous free energy change to the sum of the free energy changes attributed to single mutations. Therefore, some single mutations that are less important under alanine scanning mutagenesis can be energetically important under alanine shaving. That is, some energetically less important single mutants can be considered as an integral part of a hot spot. We would also like to note that as the ‘double water exclusion’ mechanism underlines our biclique patterns, residues in a biclique pattern (especially the one having high occurrence and with multiple residues in each side) are therefore excellent candidate residues for the simultaneous shaving.

3.4.2 A Case Study: Validating an Example Biclique with HotSprint

HotSprint is a new database which does not provide automatic query service yet over large number of PDB entries. Therefore, we report our manual query search results and present examples to illustrate the ‘double water exclusion’ mechanism that refines the ‘O’-ring theory.

Our used simple example of a biclique pattern is detected from PDB 1A8G, consisting of 5 residues (Leu5-Pro9-Leu24-Thr26-Leu97) from chain A (total 99 residues), and two residues (Thr26-Leu97) from chain B (total 99 residues). According to the HotSprint database (by the default setting), only Leu5 was not in the hot spot, the other 6 residues are all hot residues. The conservation score [50], monomer ASA, and complex ASA for these 7 residues are listed in Table 3.2.

We can see from this table that these 7 residues are all buried in the complex with a very small ASA ranging from 0.0-2.63 Å². Therefore, this compact cluster of residues is almost entirely not accessible to solvent, namely, there exists a ring of neighbor residues constructing a shelter to prevent the hot spot residues from free water molecules. This is the necessary condition for a residue to become a hot spot residue [36, 108] by the

Table 3.2: The conservation score and ASA information of 7 residues in a biclique pattern in the interface between chain A and B of PDB entry 1A8G.

Pos.	Name	Cons. Score	ASA in chain	ASA in complex
Biclique residues in 1A8G chain A				
5	LEU	6	126.88	1.27
9	PRO	7	22.39	2.63
24	LEU	7	23.91	0
26	THR	7	69.56	0.39
97	LEU	7	127.3	0
Biclique residues in 1A8G chain B				
26	THR	7	63.71	0.13
97	LEU	7	121.82	0.03

Table 3.3: The conservation score and ASA information of the ring residues surrounding a biclique pattern.

neighbouring residues in 1A8G chain A				
Pos.	Name	Cons. Score	ASA in chain	ASA in complex
1	PRO	7	149.94	94.77
2	GLN	7	165.3	122.92
3	ILE	6	77	31.69
4	THR	6	85.36	67.53
25	ASP	7	36.03	13.31
27	GLY	7	74.11	32.99
95	CYS	7	59.46	2.97
96	THR	7	101.82	28.45
98	ASN	7	126.99	26.36

O-ring theory. The neighbor residues from 1A8G chain A are Pro1, Gln2, Ile3, Thr4, Asp25, Gly27, Cys95, Thr96, and Asn98. Their conservation score, monomer ASA, and complex ASA are shown at Table 3.3. These neighbor residues (except Cys95) all have a large complex ASA, implying that, together with the neighbor residues from chain B, they indeed form a residue wall to prevent the central biclique pattern from any free water molecule. This is a nice picture showing an O-ring structure with a ‘double water exclusion’ mechanism: the central biclique residues form a compact water-free hot spot, while the neighboring residues form an O-ring to occlude any solvent accessibility.

This nice picture is depicted in Figure 3.1 which was plotted by the PyMOL software [115]. A lock-and-key architecture can be clearly observed in the left panel: The five hot spot residues from chain A form the lock (or groove), while the two hot spot residues from chain B act as a key. This is in agreement with the mechanism of ‘anchoring residues’ in protein-protein interactions [116], which explains the kinetically low

3.1.a Biclique ‘water exclusion’

3.1.b O-ring ‘water exclusion’

Figure 3.1: The biclique pattern in PDB 1A8G with the five hot spot residues—Leu5-Pro9-Leu24-Thr26-Leu97—in chain A, and two hot spot residues—Thr26-Leu97—in chain B (best viewed in color). (a) The biclique shaped in 3D space like a groove-anchor, exhibiting an inner ‘water exclusion’. (b) The biclique as a hot spot embedded in the binding interface between chain A and chain B, surrounded by neighbor residues of large ASA.

structural rearrangement of the residues during the formation of complex.

As mentioned, this biclique pattern also occurs between chain A of PDB 1CWQ at LEU92-LEU93-LEU94-THR89-PRO91 and chain B at THR90-LEU93. (1CWQ has 248 residues in each of its two chains.) Though this protein complex is quite different from the 1A8G complex in both structure and sequence, the shape of this biclique pattern is very similar in these two protein complexes as expected. See Figure 3.2 and compare it to Figure 3.1.

We manually searched HotSprint and found that all or almost all residues of this biclique pattern are hot spot residues in 1CWQ, 1E0P, and 6UPJ; and also in many redundant obligate PDB entries such as 1BDQ, 1KZK, 1LZQ, 1RL8, 1SGU, 1SH9, 1SP5, 1U8G 1W5V, 1W5W, 1W5X, 1W5Y, 1WBK, 1WBM, 1XL2, 1XL5, 1YTG, 1YTH, 1ZTZ, 2BQV, etc.

3.5 Conclusion

For the study of hot spots at protein binding interfaces, we have proposed a ‘double water exclusion’ hypothesis to refine the influential O-ring theory by further assuming that binding hot spots are water-free inside their residues (See Section 3.2). We take

3.2.a Biclique ‘water exclusion’

3.2.b O-ring ‘water exclusion’

Figure 3.2: The biclique pattern in PDB 1CWQ with the five hot spot residues—LEU92-LEU93-LEU94-THR89-PRO91—in chain A, and two hot spot residues—THR90-LEU93—in chain B (best viewed in color).

the classical graph term ‘maximal biclique’ to model the molecular topology of contact residues tightly packed in close vicinity between a pair of interacting protein chains. We term this molecular topology as ‘biclique pattern’ to emphasize the collective force of the multiple, dense atom-atom pairs (See Section 3.3.2). Based on PDB structural data, we have addressed the relation with computational and experimental hot spot residues. We have verified that many biclique residues are hot spot residues through queries to the HotSprint and ASEdb databases when the complex ASA of the residues is small (See Section 3.4). This result supports our ‘double water exclusion’ mechanism; and this result also strongly suggests that the influential ‘O’-ring theory, the ‘coupling’ tendency, together with our biclique pattern concept can provide a road map to deeply understand the binding affinity of protein interactions.

Chapter 4

Z-score Biological Significance of Binding Hot Spots in Protein Interfaces by Using Crystal Packing as the Reference State

In molecular biology, it is a fundamental problem to characterize binding hot spots in protein interfaces. In the previous chapter, we propose DWE and DWE biclique to model a binding hot spot. However, there is no statistical assessment of biological significance for DWE bicliques. So in Chapter 3, we use thresholds of biclique size and of biclique occurrence to obtain interesting DWE bicliques. But it is difficult to provide a gold standard to filter biclique noise (i.e., those bicliques with small size or low occurrence). This may be due to the following two facts. One is the larger space of biclique patterns and the quite limited available protein quaternary structures; thus, biclique occurrence is conditionally based on the size of analyzed structures. The other is that the number of the contacts of a residue are confined by its surrounding 3D space. Due to this spatial constraint, it is hard to definitely claim that the bicliques of larger size are always better than those of smaller size. To overcome these problems, the statistical and biological assessment of DWE bicliques may be an option.

Meanwhile, there are few studies to assess the biological significance of binding hot spots, although many experimental and computational methods have been proposed to distinguish between binding hot spots and non-hot spots.

Thus in this chapter, we introduce the notion of biological significance of a contact residue or of a residue cluster for capturing the probability of the residue occurring in

or contributing to protein binding interfaces. We take a statistical Z-score approach to the assessment of the biological significance. The method has three main steps. First, the potential score of a residue is defined by using a knowledge-based potential function with relative ASA calculations under DWE. Second, a null distribution of this potential score is then generated from artifact crystal packing contacts. After that, the Z-score significance of a contact residue with a specific potential score is determined according to this null distribution. We hypothesize that residues at binding hot spots have big absolute values of Z-score, as they contribute greatly to binding free energy. In accordance to this hypothesis, we propose to use Z-score to predict whether a residue is a hot spot residue in a binding interface or not. Comparison with previously reported methods on two benchmark datasets, ASEdb and BID, shows that this Z-score method is mostly superior to earlier methods. In particular, the Z-score method is excellent in hot spot prediction for the most hydrophobic residues, the least hydrophobic residues, as well as the aromatic residues. Another advantage of Z-score over existing prediction methods is its simplicity; so incorporating our Z-score method into applications such as protein docking should be easy.

4.1 Introduction

Characterization of binding hot spots in protein interfaces is a long-standing and fundamental research in structural biology. Accurate identification of binding hot spots can deepen our understanding of the mechanisms behind protein binding, which is also helpful in practical applications such as protein engineering and drug design.

Identification of binding hot spots can be achieved by wet-lab methods such as alanine-scanning mutagenesis [38] and alanine-shaving [117]. The experimental data have been compiled and stored at ASEdb [86] or at BID [98]. However, these methods are labour intensive, expensive, and time consuming. So, computational characterization and identification of binding hot spots have been widely undertaken to complement wet-lab experiments.

As discussed in Chapter 2, several past studies have examined binding hot spots based on the topological, evolutionary and compositional properties of hot spots [36, 52, 48, 62]. The ‘O-ring’ theory [36, 62] suggests that binding hot spots are surrounded by energetically less important residues that form an O-ring shape. The ‘coupling’ and ‘hot region’ hypotheses suggest that binding hot spots are networked, highly packed, and structurally conserved residue clusters [52], and that they tend to couple across a two-chain interface [48].

Atomistic simulations have been used for estimating binding free energy of binding hot spots. FoldX [74, 75] and Robetta [76, 77] are methods that use these ideas in predicting

binding hot spots. However, atomistic simulations are slow. Another shortcoming is that the predicted energies have a wide discrepancy from experimentally measured energy changes [66].

Other computational methods have also been proposed for identifying hot spots by using sequence, structural, and/or molecule-contact knowledge generated from protein primary structures [81], tertiary structures [41] or quaternary structures [78, 118, 92, 79, 60, 66, 93, 82, 80]. Sophisticated machine learning classifiers are also widely used such as PP_SITE [78], KFC [79], MINERVA [66], APIS [80], ISIS [81] and a web server HOTPOINT [82].

Despite efforts in the past, classification performance is still poor even when a vast amount of sequence, structural and molecular-contact information and features are used [66, 80]. There are also some contradictory observations. For example, residue conservation is found to have a close correspondence to binding hot spots [52]. Yet the conservation does not help significantly in distinguishing hot spot residues from non-hot spot residues [66, 80]. Another example is whether knowledge-based statistical residue pairwise potentials are useful for predicting binding hot spots. On the one hand, Keskin’s study [82] suggests that statistical pairwise potentials of interface residues can improve classification performance. On the other hand, Huang’s study [80] reports that residue pairwise potentials do not perform well.

This chapter presents the notion of Z-score biological significance of contact residues. This notion of Z-score biological significance is intended to capture—based on a statistical Z-score approach—the probability of a contact residue occurring in or contributing to protein binding interfaces. Our method is as follows. First, the *potential score* of a contact residue is defined by using a knowledge-based potential function with relative ASA calculations. A null potential score distribution is then estimated from crystal-packing artifact interfaces. Finally, the Z-score biological significance of a contact residue with a specific potential score is determined according to this null distribution. We hypothesize that true binding hot spots should have high absolute values of Z-score, and thus propose to use Z-score to identify contact residues located at binding hot spots.

Our ideas and methods are tested and evaluated on hot spot residues and non-hot spot residues compiled in two widely used datasets ASEdb and BID [66]. It is verified that most of these binding hot spots have a high absolute value of Z-score biological significance. Moreover, the probability of a contact residue occurring in or contributing to a protein-binding interface is generally reflected by the Z-score biological significance of the residue. For hot spot residues with $\Delta\Delta G \geq 1$ kcal/mol, our method achieves F1 values (*Please refer to Equation 4.10 for its definition*) of 0.674 and 0.693 respectively on the two datasets using a threshold of the absolute value of Z-score > 1 (i.e., > 1 standard deviation apart from the mean potential score of contact residues in the crystal-packing reference states). This F1 performance is mostly superior to previous methods such as Robetta,

FoldX, KFC and MINERVA [66]. In particular, the Z-score method is excellent in hot spot prediction for the most hydrophobic residues, the least hydrophobic residues, as well as the aromatic residues. Another advantage of Z-score over existing prediction methods is its simplicity; so incorporating our Z-score method into applications such as protein docking should be easy. A server for the Z-score calculation of interfacial residues in a given protein complex is available at <http://155.69.2.25/liuqian/indexz.py> (visit from outside NTU), or at <http://172.21.100.239/liuqian/indexz.py> (visit from inside NTU), or <http://sunim1.sce.ntu.edu.sg/liuqian/indexz.py>.

4.2 Z-score of Contact Residues/Hot Spots: Definitions

As mentioned earlier, the Z-score biological significance measures how far away a contact residue of a hot spot is from crystal-packing contacts in terms of a potential score. Thus to define Z-score biological significance, we need to establish a reference set of contacts and specify the potential score. For the reference set, we use contacts in crystal packing, which are residues from different chains that are spatially close to each other but do not result in any biological interactions. To determine a potential score, it is important to note that the interface of a protein complex consists of interface-contact residues and their nearby residues [46]. These nearby residues provide scaffolds for the binding interface. Thus to obtain a comprehensive potential score for contact residues, we explore two sources of residue pairwise potentials—one from the interface contacts and the other from the nearby contacts. We further consider accessible surface area in the potential score. Details are presented below.

4.2.1 Preliminary Definitions

A residue pair, one from each chain, is defined as a contact pair if the two residues have at least one atom from each residue contacting each other. It is also referred to as an interface contact or a contact for short. The nearby residues are referred to as those residues that are from the same chain as a contact residue, and their CA atoms also contact with the CA atom of the contact residue. We call them *nearby contacts*. A residue is considered to be buried if its relative ASA is less than a threshold T_{asa} ; otherwise, it is an exposed residue. Similarly, a contact of a pair of residues—regardless of whether it is an interface contact or a nearby contact—is considered to be exposed if one of the contact residues is exposed; otherwise, this contact is buried.

4.2.2 Residue Pairwise Potentials of Interface Contacts

Knowledge-based statistical pairwise potentials have been widely used in a variety of problems in computational biology such as protein folding, protein interactions and protein docking [119, 120]. The most common form of pairwise potentials of (r_i, r_j) is a log-likelihood ratio as shown in Equation 4.1.

$$q_{r_i r_j}^c = \log \frac{p^c(r_i, r_j|O)}{p^c(r_i, r_j|R)} \quad (4.1)$$

where r_* is a residue or an atom type of $*$, $p^c(r_i, r_j|O)$ is the observed contact probability of pair (r_i, r_j) , and $p^c(r_i, r_j|R)$ is the corresponding expected probability in the reference states. The value of $q_{r_i r_j}^c$ can be considered as a statistical assessment of the likelihood for the contact pair (r_i, r_j) in an observed environment—for example, an environment of significant biological contacts occurring as reported in the literature.

A good definition of reference states is difficult but critical for pairwise potentials. An idea is to assume that the occurrence of residue types r_i and r_j is independent. Under this assumption, $p^c(r_i, r_j|R) = p(r_i|R)p(r_j|R)$, where $p(*|R)$ is the marginal probability of $*$ [25, 120]. However, this independence assumption does not hold in most biological situations. For example, past works have shown that protein interfaces prefer hydrophobic residues while protein surfaces prefer hydrophilic residues; so, the occurrence of a residue in protein interfaces or surfaces is surrounding-dependent, regardless of whether it contributes to protein binding or not.

To overcome this limitation, we take crystal packing as the reference state. Crystal-packing contacts are false interactions in crystallographic packing environments [11]. Past works have also revealed a significant difference between protein surfaces and interfaces, as well as a high similarity between protein surfaces and crystal packing [30, 83, 22, 121, 35]. That is, binding sites in crystal packing are similar to protein surfaces, and thus their residue composition is surrounding-dependent and not random. However, crystal-packing binding sites are artifact. The interface-contacts at crystal-packing binding sites are random according to crystal-packing definitions. Therefore, it is reasonable to take crystal packing as the reference state, because surrounding-dependent artifact interfaces in crystal packing can overcome the independence assumption in past works.

Our idea is similar to Vajda’s idea [119], where decoys are used as a reference state to develop pairwise potentials for improving the performance of protein-docking methods. Crystal-packing contacts are more helpful than decoys, as this reference state and the contrasting partner—the biological interactions—are both taken from the natural crystallization process. With crystal packing as the reference state, we can verify that biological interactions and their interface-contact residues are not formed by chance as crystal packing do.

We consider all of the 20 standard residue types in calculating residue pairwise potentials. We use matrix Q^c to represent statistical knowledge between residue-type pairs:

$$Q^c = \left\{ q_{r_i r_j}^c = \log \frac{p^c(r_i, r_j|O)}{p^c(r_i, r_j|R)} \mid r_i, r_j = 1, 2, \dots, 20 \right\} \quad (4.2)$$

A contact potential for residue i , denoted by $pp_{c,i}^{ASA}$, is defined by summing all pairwise potentials $q_{r_i r_j}^c$, namely, $pp_{c,i}^{ASA} = \sum_{j \in S_c} q_{r_i r_j}^c$, where S_c is the set of residues from other chains which have interface-contacts with the residue i in the protein complex.

The idea above for determining residue contact potentials assumes that all interface-contact residues are equally important. However, the influential O-ring theory [36, 62] and most of the computational methods for hot spot prediction [60, 66, 93, 82, 80] have observed that energetically important hot spot residues almost all have a small ASA. ASA is also considered to be a critical feature in other protein applications, such as identification of biological interactions from non-biological interactions [22, 27].

To integrate ASA information, we propose Equation 4.3 to calculate the potential of a contact residue i :

$$pp_{c,i}^{ASA} = (1 - A_i) \sum_{j \in S_c} q_{r_i r_j}^c (1 - A_j) \quad (4.3)$$

where A_* is the relative ASA of $*$ in the interface of a protein complex. It can be seen from Equation 4.3 that the larger a residue's relative ASA is, the less contribution it makes to the pairwise potential. This is based on the influential O-ring theory stating that residues on an O-ring, in order to avoid water attack on ring-inside residues, usually have larger ASAs and are energetically less important.

4.2.3 Residue Pairwise Potentials of Nearby Contacts

Residue contacts in biological interfaces are not in isolation without support from other residues. Instead, nearby residues that surround the interface-contact residues form the supportive scaffold for protein interfaces [46]. Similar to residue pairwise potentials for interface-contact residues, we develop an idea for determining residue pairwise potentials for the nearby contacts in protein binding sites as shown in Equation 4.4.

$$q_{r_i r_j}^n = \log \frac{p^n(r_i, r_j|O)}{p^n(r_i, r_j|R)} \quad (4.4)$$

where r_i and r_j , $p^n(r_i, r_j|O)$, and $p^n(r_i, r_j|R)$ have similar meaning to those in Equation 4.1. But residues i with type r_i and j with type r_j have a nearby contact instead of an interface contact.

We use matrix Q^n to denote the 20×20 residue-type pairwise potentials for nearby contacts:

$$Q^n = \left\{ q_{r_i r_j}^n = \log \frac{p^n(r_i, r_j | O)}{p^n(r_i, r_j | R)} \mid r_i, r_j = 1, 2, \dots, 20 \right\} \quad (4.5)$$

We also integrate ASA into pairwise potentials of nearby contacts. Thus, the nearby-contact pairwise potential of a residue i is not just the sum of $q_{r_i r_j}^n$ in Q^n . Instead, it is calculated by Equation 4.6.

$$pp_{n,i}^{ASA} = (1 - A_i) \sum_{j \in S_n} q_{r_i r_j}^n (1 - A_j) \quad (4.6)$$

where A_* is the relative ASA of $*$ in the interface of a protein complex; j ranges over the set S_n of residues from the same chain as i that have nearby contacts with i . Note that the residues i and j must be involved in the interface contacts.

4.2.4 Residue Pairwise Potentials of Alanine Mutations

A hot spot residue is a residue whose alanine mutation can lead to a big change in binding free energy. So, it is useful to calculate the change in pairwise potential after the alanine mutations for hot spot residues. Based on Equations 4.3 and 4.6, the total potential for a residue i is defined as $pp_i^{ASA} = pp_{c,i}^{ASA} + pp_{n,i}^{ASA}$. The potential change after its alanine mutation is defined by Equation 4.7.

$$pp_i^{mut} = pp_i^{ASA} - pp_{ala}^{ASA} \quad (4.7)$$

where pp_{ala}^{ASA} is the pairwise potential when the residue i is replaced by an alanine residue. pp_i^{mut} is called the *mutation residue pairwise potential* for i . It is also referred to in a short form as a (pairwise) potential (score) when used in estimating a potential score distribution and in calculating Z-scores.

4.2.5 Z-score

In this section, we first describe an estimation for the distribution of pairwise potentials for contact residues in crystal packing. Then we move to define a Z-score biological significance for contact residues, especially for hot spot residues. As introduced, this Z-score is a probability assessment method against the contacts observed from the non-biological interactions of crystal packing.

Our null hypothesis H_0 is that a potential score is from the potential distribution of contact residues in crystal packing. The alternative hypothesis H_1 is that it is not. Under H_0 , we need the distribution of the potential scores for the contact residues in

crystal packing, which can be estimated from a sample set of residue potentials from crystal-packing interfaces.

For H_0 , we calculate the potential scores for contact residues in a dataset of crystal packing using Equation 4.7. These scores are denoted by $R = \{pp_i^{mut} | i = 1, 2, \dots, N_r\}$, where N_r is the number of contact residues in a reference-state dataset. As can be observed from Figure 4.1(a), these scores can be assumed to follow a normal-like distribution. So, we obtain R 's distribution with the mean $\bar{m} = \frac{\sum_i^{N_r} pp_i^{mut}}{N_r}$ and the standard deviation $\delta^{std} = \sqrt{\frac{\sum_i^{N_r} (pp_i^{mut} - \bar{m})^2}{N_r - 1}}$. Here, we note why the distribution under the null hypothesis is only estimated based on crystal packing: on one hand, we want to sharpen the difference of potentials from false binding interfaces and true binding interfaces; on the other hand, we only would like to test whether a residue is significantly not from crystal-packing interfaces according to its potential score.

4.2.5.1 Z-score of a Contact/Hot Spot Residue

For each mutation of a contact/hot spot residue i , its mutation potential pp_i^{mut} is determined by Equation 4.7. Its Z-score biological significance is defined as

$$Z\text{-score}(i) = \frac{pp_i^{mut} - \bar{m}}{\delta^{std}} \quad (4.8)$$

This Z-score tells how far away a residue's potential score is from that of crystal packing. In other words, if a residue has an absolute value of $Z\text{-score} > \zeta$, it can be predicted as a hot spot residue. The reasons for using the absolute values of $Z\text{-score}$ (i.e., $||Z\text{-score}||$) are as follows. First, in both the case $Z\text{-score} > \zeta$ and the case $Z\text{-score} < -\zeta$, the surrounding of the residue changes a lot compared to the reference state of crystal packing. Second, as can be observed from the residue pairwise matrices, hydrophobic residues are more frequent in biological binding and generally have positive potentials, while hydrophilic residues are infrequent and generally have negative potentials correspondingly. Third, hydrophobic residues usually have small numbers of neighbors with similar hydrophobicity in crystal packing, while the surroundings of hydrophobic residues in biological binding interfaces can neighbor with large numbers of hydrophobic residues to form a hydrophobic cluster. This environmental change for hydrophobic residues often leads to $> \zeta$ scores. In contrast, the surroundings of a hydrophilic core in biological binding interfaces usually have a hydrophobic ring, but it is absent in crystal packing contacts. This leads to $< -\zeta$ scores. We note that a hydrophilic residue with $Z\text{-score} < -\zeta$ only means biological contacts occurring. Hydrophilic residues should also involve such significant contacts as hydrogen bonds to contribute to protein binding greatly. These significant contacts can be considered in the future.

The absolute value of pairwise potentials was also used previously by [82, 97] in the same application domain of hot spot prediction. Actually, Tuncbag *et al.* [82] used several residue features including statistical pairwise potentials of interface residues to improve classification performance. Our method is different in two aspects. One is that we use crystal packing as the reference state to extract pairwise potentials. Crystal packing is more reasonable as a reference state than the assumption that residue contacts occur independently. Second, we integrate ASA into pairwise potentials and propose to use Z-score values for the prediction.

4.2.5.2 Z-score of a DWE-Biclique Hot Spot

A protein-binding hot spot is a small cluster of residues [36], involving complicated cooperativity and additivity among the contact residues, especially among binding residues that are energetically important. Biological experiments such as alanine shaving [117, 87] can identify a true hot spot by mutating more than one residue at one time. It is commonly understood that some single mutations that are less important under alanine scanning mutagenesis can be an integral part of a hot spot.

However, it is costly to perform such shaving due to the combinatorial complexity. So, computational studies have been made to model the topological organization of hot spots, including the O-ring theory [36] and the ‘hot region’ proposition [48]. In Chapter 3, we use a DWE biclique to model a hot spot [62]. The DWE-biclique model uses a bipartite graph to represent the interaction between a pair of protein chains; and then it narrows down to subgraphs each with a full and dense connection between the two sets of residues in a compact region, called a ‘biclique’. On the one hand, the DWE-biclique model agrees with the ‘O-ring’ theory that there exist residues on the O-ring to protect DWE-biclique residues from solvent attack. On the other hand, this model assumes that this full and dense connection is so close spatially that there is no space inside to accommodate any free water molecule. So, the DWE biclique describes a reasonable topological organization of hot spot residues.

To examine the biological significance of DWE-biclique hot spots, we assume that the potential scores pp_*^{mut} for biclique residues $*$ follow an independent and identical normal distribution with mean $\bar{m}$ and standard deviation δ^{std} . Given a DWE-biclique hot spot hs [62] with n_1 biclique residues from one chain and n_2 biclique residues from the other, its potential score $pp_{hs}^{mut} = \sum_i^{n_1} pp_i^{mut} + \sum_j^{n_2} pp_j^{mut}$ follows a normal distribution with mean $\bar{m}^{hs} = (n_1 + n_2)\bar{m}$ and standard deviation $\delta_{hs}^{std} = \delta^{std}\sqrt{n_1^2 + n_2^2}$. The Z-score biological significance of a DWE-biclique hot spot (biclique Z-score for short) is then defined as

$$Z-score_{hs} = \frac{pp_{hs}^{mut} - \bar{m}_{hs}}{\delta_{hs}^{std}} \quad (4.9)$$

Then those DWE bicliques with large absolute values of $Z-score_{hs}$ can be considered as candidates of alanine shaving.

4.3 Methods

Our computational steps for determining the Z-scores of residues are as follows: (i) Count the residue pairwise interface-contacts (r_i, r_j) from a dataset of biological interfaces as well as pairwise contacts from a dataset of crystal packing; and then calculate the pairwise potentials of interface contacts (r_i, r_j) for the matrix Q^c by using the two sources of statistical knowledge. (ii) Count residue pairwise nearby-contacts (r_i, r_j) from the same dataset of biological interfaces as well as pairwise nearby-contacts from the same dataset of crystal packing; and after that, calculate pairwise potentials of nearby contacts (r_i, r_j) for the matrix Q^n by using the two sources of statistical knowledge. (iii) Estimate the distribution of the mutation potential scores for the contact residues from another dataset of crystal-packing interfaces whose proteins have low sequence similarity to those in the datasets used in (i). The contact residues here are in artificial interfaces with no significant contribution to protein binding, and thus their mutation potential scores are a good reference state of the mutations of binding hot spots. (iv) Finalize the Z-scores for contact/hot spot residues based on the estimated distribution.

4.3.1 Datasets

As required above, we need two types of datasets—viz., protein interaction datasets for extracting residue pairwise potentials and datasets containing binding hot spots.

To extract residue pairwise potentials, it is necessary to obtain the number of residue pairwise contacts in the datasets of biological interfaces as well as in the artifact interfaces of the reference-state crystal packing.

We used the biological interfaces from the dataset of biological interactions reported in [26, 122]. This dataset consists of 212 transient and 115 obligate protein complexes without redundancy. We refer to this dataset as BIO. We used BIO to calculate the frequency of interface-contact residue pairs and the frequency of nearby residue pairs for residue pairwise potentials.

The crystal packing contacts are from three published datasets: the BNCPCS dataset [18], the Bahadur dataset [22], and the Ponstingl dataset [27]. Poupon’s team compiled these three datasets together and removed those crystal packing contacts with more than 30% sequence identities [70]. There are in total 49, 94, and 178 crystal packing contacts in the Ponstingl, BNCPCS, and Bahadur dataset respectively. After removing some special items [70], there are 315 crystal packing contacts used in this work.

4.3.1.1 Datasets Containing True Binding Hot Spots

We use two datasets ASEdb and BID, which are previously used in [66], to assess the performance of our Z-score method for identifying binding hot spots. ASEdb is mainly

from the ASEdb database [86] containing 17 protein-protein complexes [66]. According to interface contacts here, ASEdb contains 265 mutations as shown in Table C.4; among them, 119 are hot spot residues with $\Delta\Delta G \geq 1$ kcal/mol and 146 are energetically unimportant residues. If those residues with $\Delta\Delta G \geq 2$ kcal/mol are considered as hot spot residues, then there are 65 hot spot residues and 200 non-hot spot residues in ASEdb.

BID contains 127 alanine-mutations in 18 protein complexes [98] according to interface contacts here (please refer to Table C.5). BID contains 64 hot spot residues labeled as ‘strong’ or as ‘intermediate’, and 63 non-hot spot residues. If those residues labeled as ‘strong’ are considered as hot spot residues, then BID contains 39 hot spot residues and 88 residues that are energetically less important.

4.3.2 Detailed Steps to Extract Q^c and Q^n for Residue Pairwise Potentials

Recall that residue pairwise potentials are composed of two parts: the interface-contact pairwise potentials and the nearby-contact pairwise potentials. An interface-contact of two residues, one from each chain, must have at least two atoms, one from each residue, whose spatial distance is smaller than the sum of their van der Waals radii plus 2.75 Å as used in Chapter 3. The two CA atoms from nearby-contact residues are required to have a distance less than 8.5 Å [123, 124]. As recommended by [82], we set T_{asa} to 20% to determine whether a residue or a residue pair is buried or not. The ASA information for each protein complex and its interacting chains are determined by NACCESS [12].

4.3.2.1 Compute Q^c of the Residue Pairwise Potentials for Interface Contacts

To obtain residue pairwise potentials for interface contacts, we calculate the number of residue contacts $f^c(r_i, r_j|O) = F_{r_i r_j}^{bio}$ in the interfaces from a dataset of biological interactions, and similarly produce $f^c(r_i, r_j|R) = F_{r_i r_j}^{cp}$ from a set of crystal packing contacts. The computational steps to produce the residue pairwise potentials for interface contacts are as follows. (i) For each protein interaction of chains c_1 and c_2 in BIO, we obtain its contact residues and contact residue pairs according to the contact distance requirement above. (ii) NACCESS is then used to produce the ASA for each contact residue in the protein complexes. After that, those exposed contact residue pairs are filtered out. (iii) The residue pairwise frequency matrix for BIO is calculated based on the resultant contact residue pairs as follows:

$$P_{bio}^c = \left\{ p^c(r_i, r_j|O) = \frac{F_{r_i r_j}^{bio}}{\sum_{r_m}^{20} \sum_{r_n}^{20} F_{r_m r_n}^{bio}} \mid r_i, r_j = 1, 2, \dots, 20 \right\}$$

Using a similar process to steps (i) and (iii) above, we obtain the residue pairwise frequency matrix for crystal packing from the BNCPCS and Bahadur datasets, as formulated by

$$P_{cp}^c = \left\{ p^c(r_i, r_j | R) = \frac{F_{r_i r_j}^{cp}}{\sum_{r_m}^{20} \sum_{r_n}^{20} F_{r_m r_n}^{cp}} \mid r_i, r_j = 1, 2, \dots, 20 \right\}.$$

So, the residue pairwise potentials $q_{r_i r_j}^c$ in Equation 4.1 and Q^c in Equation 4.2 for interface contacts can be calculated by $q_{r_i r_j}^c = \log \frac{p^c(r_i, r_j | O)}{p^c(r_i, r_j | R)}$, $r_i, r_j = 1, 2, \dots, 20$.

4.3.2.2 Determine Q^n of the Residue Pairwise Potentials for Nearby Contacts

The calculation steps for the residue pairwise potentials of nearby contacts in binding sites are similar to those for interface contacts. At the first step, we obtain the contact residues and their nearby contacts according to the nearby distance requirement above for each protein binding interface in BIO. At the second step, NACCESS is used to produce ASA for each involved residue in protein complexes. At the third step, we filter out those exposed nearby-residue pairs in protein complexes. At the final step, the residue pairwise frequency matrix for BIO is calculated based on the resultant buried nearby residue pairs, denoted as

$$P_{bio}^n = \left\{ p^n(r_i, r_j | O) = \frac{F_{r_i r_j}^{bio}}{\sum_{r_m}^{20} \sum_{r_n}^{20} F_{r_m r_n}^{bio}} \mid r_i, r_j = 1, 2, \dots, 20 \right\}$$

where $F_{r_i r_j}^{bio}$ is the occurrence frequency of residue type pair (r_i, r_j) for nearby contacts in all BIO interactions.

Also, by using the first and final step, we obtain the residue pairwise frequency matrix based on all the nearby contacts of contact residues in the crystal-packing surfaces from the BNCPCS and Bahadur datasets, as formulated by

$$P_{cp}^n = \left\{ p^n(r_i, r_j | R) = \frac{F_{r_i r_j}^{cp}}{\sum_{r_m}^{20} \sum_{r_n}^{20} F_{r_m r_n}^{cp}} \mid r_i, r_j = 1, 2, \dots, 20 \right\}.$$

Finally, the residue pairwise potentials $q_{r_i r_j}^n$ in Equation 4.4 and Q^n in Equation 4.5 for nearby contacts can be calculated by $q_{r_i r_j}^n = \log \frac{p^n(r_i, r_j | O)}{p^n(r_i, r_j | R)}$, $r_i, r_j = 1, 2, \dots, 20$.

4.3.3 The Distribution of Potential Scores of Contact Residues in Reference-State Crystal-Packing Interfaces

To get the Z-score biological significance for a contact residue, the distribution of potential scores of contact residues in the reference-state crystal-packing interfaces has to be

estimated first. In this work, we calculate the potential scores of all contact residues in the Ponstingl dataset as the sample scores, because protein interactions in Ponstingl have lower sequence identities compared to those in the BNCPCS or the Bahadur datasets. The latter two datasets are used to extract residue pairwise potential matrices Q^c and Q^n .

To compute the Z-score for a contact residue i in an interface (e.g., the crystal-packing interfaces from Ponstingl), we first obtain its interface-contact residue set S_c and its nearby-contact residue set S_n . Then its ASA is produced using NACCESS. After that, we calculate $pp_{c,i}^{ASA}$ by Equation 4.3 with S_c and $pp_{n,i}^{ASA}$ by Equation 4.6 with S_n , and obtain its total potential pp_i^{mut} by using Equation 4.7. The potential scores of the contact residues in Ponstingl form a sample set $R = \{pp_i^{mut} | i = 1, 2, \dots, N_r\}$, where N_r is the number of the contact residues in Ponstingl. We assume that R follows a normal distribution. Then, we compute R 's mean $\bar{m}$ and standard deviation δ_{std} after discarding, from the two tails of R , 5% of the samples (i.e., 2.5% for each of the tails). These sample scores are not used to calculate $\bar{m}$ and δ_{std} because they might be outliers.

4.3.3.1 Calculate the Biological Significance of Binding Hot Spot Residues

Similar to the calculation process above for the contact residues in Ponstingl, we determine the residue potential pp_i^{mut} for each mutation i in ASEdb and BID. Then, the Z-score biological significance of mutation residues are produced by using Equation 4.8. It is expected that a hot spot residue has a large deviation from R 's mean $\bar{m}$.

4.3.3.2 Calculate the Biological Significance of DWE-Biclique Binding Hot Spots

DWE bicliques [62, 63] from the protein interactions stored in the ASEdb and BID datasets are identified by using the Algorithm in Section 3.3.2. Finally, $Z\text{-score}_{hs}$ of these DWE bicliques is calculated using Equation 4.9. Those bicliques with large absolute values of $Z\text{-score}_{hs}$ are considered as DWE-biclique hot spots.

4.3.4 Evaluation Measures for Classification Performance

Generally, several measures as shown in Equation 4.10 can be used in evaluation of classification.

$$\begin{aligned}
 precision(p.) &= \frac{TP}{TP + FP} \\
 recall(r.) &= \frac{TP}{TP + FN} \\
 specificity(sp.) &= \frac{TN}{TN + FP} \\
 accuracy(acc.) &= \frac{TP + TN}{TP + TN + FP + FN} \\
 F1 &= \frac{2 \times precision \times recall}{precision + recall}
 \end{aligned} \tag{4.10}$$

where TP, FP, TN and FN are true positives, false positives, true negatives and false negatives respectively. So, *precision* is the number of correct positive predictions divided by the number of positive predictions, *recall* or *sensitivity* is the fraction of correct positive predictions over all positives, *specificity* is the fraction of correctly predicted negatives over all negative cases, while *accuracy* is the number of correctly predicted positive and negative cases divided by the number of all cases. Meanwhile, according to [125], The AUC value of a binary test is equal to half of the sum of *r.* and *sp.*.

In this chapter, hot spot residues are considered as the positive case (T) and non-hot spot residues as the negative case (F); *precision*, *recall* and *F1* are used to measure Z-score performance in the classification of binding hot spots. These three measures are also used in [66, 82, 80] with the same definitions.

4.4 Results and Discussion

4.4.1 Distribution of Potential Scores and Z-scores

We computed the potential scores as well as their Z-scores for all of the alanine-mutations in ASEdb and BID. Figure 4.1 shows a histogram distribution of the potential scores in the Ponstingl crystal packing interfaces, as well as a relationship between $\Delta\Delta G$ (*the change of the binding free energy*) of the alanine-mutations in ASEdb and their Z-scores.

We can see from Figure 4.1(a) that almost all 942 contact residues in crystal packing interfaces have a potential score around 0 ranging from -1 to 1 . These small absolute values of Z-score indicate that contact residues in crystal packing are similar to protein surfaces without large biological significance of binding. Note that these assessed residues are all from Ponstingl, whose protein sequences have a low sequence similarity to those

Figure 4.1: The distribution of potential scores. (a) The histogram distribution of the pairwise potentials for contact residues in the crystal-packing interfaces. The continuous curve is the shape of the histogram. (b) The scatter of the potential scores and $\Delta\Delta G$ of the mutations in ASEdb. In (b), the potential scores less than -5 are reset to -5, and those more than 5 reset to 5 for a better visual clarity; the Z-scores of the dot, dash and solid lines are ± 1 , ± 1.5 and ± 2 , respectively, meaning a ± 1 , ± 1.5 and ± 2 standard deviation from the mean; 'x' are those mutations with $\Delta\Delta G < 1$ kcal/mol, while 'o' are those mutations with $\Delta\Delta G \geq 1$ kcal/mol; each label on the X axis is a triplet with a potential score followed by the corresponding Z-score and p-value.

Table 4.1: Information of three examples of mutation residues. ASA has a unit $\text{\AA}^2$.

PDB	Chain	Residue	Position	$\Delta\Delta G$ (kcal/mol)	Potential	Z-score	p-value	ASA in monomer	ASA in complex
2PTC	I	Lys	15	10	-12.35	-16.57	<1E-324	201.7	0.6
1A22	A	Ser	62	0.1	-0.47	-0.43	0.664	96.9	22.8
1DAN	T	Phe	50	0.4	2.02	2.95	0.003	128	27.9

in the datasets for extracting the residue pairwise potential matrices. It means that Figure 4.1(a) contains few overfitting problems. However, there are still some residues with large absolute potential scores as shown in Figure 4.1(a). This may be due to some crystal-packing binding sites in Ponstingl having some overlap with biological binding sites where the residue scores have larger deviation from the surface contact residues in crystal packing.

Figure 4.1(b) exhibits a bowl shape with a small range around the mean at the bowl bottom, although the upper edges of the bowl are a little divergent from the perfect convergent bowl edges. At the bottom, with Z-scores ranging from -2 to 2 , most mutations have a $\Delta\Delta G$ less than 2 kcal/mol with a negative precision value (NPV) 0.85 . Here, NPV is the percentage of correct predictions of non-hot spots (i.e., the number of correct predictions of non-hot spot residues divided by the number of non-hot spot residue predictions). Further, 96.3% of those mutations with Z-scores from -1 to 1 are non-hot spot residues with $\Delta\Delta G < 2$ kcal/mol. Generally, the larger the deviation from the mean is, the more likely a residue mutation is to have a high $\Delta\Delta G$.

4.4.1.1 Z-scores of Mutation Residues in ASEdb: Three Examples

The first example is Lys at position 15 in chain I of the PDB 2PTC complex. This residue is an interface residue between chain E and chain I in the complex. Its mutation results in a large change of 10 kcal/mol binding free energy ($\Delta\Delta G = 10$ kcal/mol). As shown in Figure 4.2, this residue is deeply buried in the complex with a sharp ASA decrease from $201.7 \text{ \AA}^2$ to $0.6 \text{ \AA}^2$ after the complex formation (Table 4.1). The potential score of this residue is -12.35 , which is the largest absolute value of potential scores among all the mutations in ASEdb. Its Z-score is -16.57 , corresponding to a p-value that is less than $1\text{E-}324$. These scores suggest that this residue has a very different surroundings from non-binding surfaces and it is most likely to contribute significantly to binding free energy.

On the other hand, the residue Ser at position 62 in chain A of the PDB 1A22 complex has a small potential score of only 0.47 , corresponding to a Z-score of -0.43 and a p-value of 0.664 (Table 4.1), although this residue also has a large ASA change from $96.9 \text{ \AA}^2$ to $22.8 \text{ \AA}^2$ in the formation of the complex. This low Z-score and large p-value indicate that

Figure 4.2: Three examples of the mutation residues in ASEdb. In the first row, the mutation residues are in red and in ‘sticks’ view, and at the second row they are in ‘spheres’ view. The colors except red indicate atoms in different chains. (a) and (d) are for Lys at position 15 in Chain I of the PDB 2PTC complex; (b) and (e) are for Ser at position 62 in Chain A of the PDB 1A22 complex; (c) and (f) are for Phe at position 50 in Chain T of the PDB 1DAN complex.

the surroundings of this residue differs only slightly from non-binding surfaces and that it is unlikely to make contribution to protein binding. In fact, the $\Delta\Delta G$ of this residue is pretty low (0.1 kcal/mol) according to ASEdb.

The residue Phe at position 50 in chain T of the PDB 1DAN complex also has a large ASA change from 128 Å² to 26.9 Å² for the complex 1DAN formation (Table 4.1). Its potential score is 2.02, corresponding to a Z-score of 2.95 and a p-value of 0.003. That is, the Z-score suggests that it is a potential hot spot residue. In fact, Robetta, FOLDX, MINERVA and HOTPOINT [82] also predict this residue as a hot spot residue. Obviously, this is a false prediction, because the $\Delta\Delta G$ of this residue is not high, 0.4 kcal/mol according to ASEdb. However, this false prediction is interpretable from Figure 4.2. The side chain (i.e an aromatic ring π) of the residue Phe involved is in contact with the partner protein. This results in its large ASA change. It seems that atomic contacts (e.g. a π -involving contact here) have a specific role with varying importance to protein binding. They should be considered separately in atom-pairwise or residue-pairwise potentials in our future work.

4.4.2 Classification Performance by Our Z-score Method for Binding Hot Spot Residues

Z-scores can be used to identify hot spot residues from non-hot spot residues as briefly illustrated in the three examples above. If the Z-score of a binding residue is bigger than a threshold ζ or smaller than $-\zeta$, it is predicted as a hot spot residue; otherwise, it is classified as a non-hot spot residue. We call this Z-score classification method. In this work, the classification performance by our Z-score method is compared to those achieved by earlier methods such as Robetta [76, 77], FOLDX [74, 75], KFC [79], MINERVA [66] and HotPoint [82]. The comparison study is conducted on ASEdb and BID [66] under two different thresholds for defining hot spot residues (i.e., $\Delta\Delta G \geq 1$ kcal/mol or $\Delta\Delta G \geq 2$ kcal/mol in ASEdb). As these two datasets are the same as those used by MINERVA [66], the performance by the earlier methods is copied from [66]. Note that the threshold ζ used in our Z-score method is not tuned or optimized for hot spot classification.

The comparison results are presented in Tables 4.2 and 4.4.

The hot spot residues in Table 4.2 are residues defined as those mutations with $\Delta\Delta G \geq 1$ kcal/mol in ASEdb, or those labeled as ‘intermediate’ or ‘strong’ in BID. By our Z-score method, a mutation residue is predicted to be energetically important if its absolute Z-score is bigger than 1. It can be seen from Table 4.2 that the classification performance by our Z-score method is superior to that of FOLDX or KFC according to the overall F1 performance measure. In Table 4.2, FOLDX’s F1 values are 0.585 on ASEdb and 0.532 on BID; F1 values of our Z-score method are 0.674 on ASEdb and 0.693 on BID, much higher than the FOLDX’s F1 values. F1 value of our Z-score method is

also 0.21 higher than those of KFC on the ASEdb dataset, and 0.266 higher on the BID dataset.

Our Z-score method also outperforms Robetta and MINERVA in terms of F1 values. Robetta is a gold standard in the evaluation of hot spot prediction. This method was originally designed to predict hot spot residues with $\Delta\Delta G \geq 1$ kcal/mol. In Table 4.2 on ASEdb, Robetta’s F1 value is 0.643, lower than the F1 value (0.674) of our Z-score method; on BID, Robetta has a F1 value of 0.613, 0.08 less than the F1 value of our Z-score method. MINERVA has a slightly lower F1 value than our Z-score method on ASEdb, but its F1 value on BID is 0.121 less than the F1 value achieved by our Z-score method. Overall, our Z-score method delivers the best classification performance for identifying hot spots with $\Delta\Delta G \geq 1$ kcal/mol.

The classification performance of Z-score, Robetta and FOLDX for each complex in ASEdb are presented in Table 4.3. According to the F1 values in this table, Z-score wins on 9 complexes, while Robetta wins on 4 complexes and FOLDX wins on 3 complexes. Robetta and FOLDX have the same high F1 on 1NMB, and the three methods are in a tie on 1AHW and 1F47. Again, Z-score has better performance than Robetta and Foldx.

Hot spot residues presented in Table 4.4 are those residues defined with $\Delta\Delta G \geq 2$ kcal/mol in ASEdb, or those labeled as ‘strong’ in BID. The performance by our Z-score method (where ζ is set as 1 or 1.5) is reported in Table 4.4. On the ASEdb dataset, the Z-score method has F1 values that are higher than FOLDX and HotPoint, comparable to Robetta and KFC, but less than MINERVA. However, on the independent test dataset BID, MINERVA’s F1 value decreases to a similar level as our Z-score method. On the BID dataset, Z-score’s F1 value is much higher than those F1 values achieved by Robetta, FOLDX, KFC and HotPoint.

Note that the performance of the Z-score method shown in Table 4.2 is much better than that in Table 4.4. Actually, these two results are not directly comparable because different thresholds are used to define hot spot residues— $\Delta\Delta G \geq 1$ kcal/mol is used in Table 4.2, while $\Delta\Delta G \geq 2$ kcal/mol is used in Table 4.4 on ASEdb. Nevertheless, this performance discrepancy can deepen our understanding on the relationship between hot spots and their pairwise potentials and Z-scores. One reason for this performance difference may be attributed to that the Z-scores just provide a way to assess the extent of a residue surroundings deviating from a non-biological environment. ‘Non-biological environments’ here indicates that there is much less possibility of biological binding occurring. So, a reasonable deviation of potential scores from a non-biological environment can lead to mutations with $\Delta\Delta G$ not less than 1 kcal/mol. But to reach a higher $\Delta\Delta G$ of mutations, some other factors—such as hydrogen bonding and π -involving interactions—may also be necessary besides a biological surrounding. Thus, to achieve a better performance on predicting hot spots with $\Delta\Delta G \geq 2$ kcal/mol, more factors including molecular contacts should be considered.

Our Z-score method has a higher recall but a lower precision, while other methods have higher precisions with lower recalls. It is good to combine our method with the other four methods Robetta, FOLDX, KFC and/or MINERVA following the idea of meta-learning [126]. HotPoint is not considered here as both HotPoint and Z-score take advantage of pairwise potentials and their combination may contain redundant information. We propose the following classification process: Given a mutation, it is predicted as a hot spot residue if and only if both the Z-score method and one of the other four methods predict it as a hot spot residue. It can be seen from Tables 4.2 and 4.4 (marked with a superscript ^c) that the classification performance is improved on ASEdb for both $\Delta\Delta G$ thresholds. On BID, although the precision is still good, the poor recalls of the four methods led to a lower recall value of the meta-learning approach than using our Z-score method alone. As a result, it is hard to obtain an impressive F1-measure improvement.

Meanwhile, we also tried another numerical combination. Under this approach, each mutation residue is represented by a vector with five features which are numerical prediction outputs of Robetta and FOLDX, binary predictions of KFC and MINERVA, and the absolute value of Z-score. Then, SVM with an RBF kernel function in LibSVM [127], the freely available SVM library, is used to perform twenty-fold cross-validation on the ASEdb dataset. The RBF best parameters are determined using the default grid search in LibSVM. This approach could not achieve a better performance for $\Delta\Delta G \geq 1$ kcal/mol hot spot prediction. But for $\Delta\Delta G \geq 2$ kcal/mol hot spot prediction, this combination has recall 0.554, higher precision 0.857 and F1 value 0.673.

In summary, our Z-score method alone is already very powerful for the identification of protein binding hot spots, especially for hot spot residues with $\Delta\Delta G \geq 1$ kcal/mol.

4.4.3 Classification Performance by the Z-score Method of DWE-Biclique Binding Hot Spots

It has been reported that there exist ‘hot regions’ or residue clusters (hot spots) in protein interfaces [52]. Residues within a region of this kind contribute cooperatively to the stability of protein complexes, and additively between regions. These hot spots can be detected by biological experiments such as alanine shaving technologies [117, 87]. DWE bicliques [62, 63] are modeled as computational hot spots, and they are likely to be ‘hot regions’ and candidates of alanine shaving. So, we would like to verify Z-score biological significance of DWE bicliques.

The performance by the biclique Z-scores on predicting binding hot spots is shown in Table 4.5. This Z-score’s classification performance is comparable to those in Tables 4.2 and 4.4. However, we would like to point out that hot spot residues from the ASEdb and BID dataset are determined by alanine *scanning* which mutates each residue separately, while DWE bicliques are candidates of alanine *shaving* for mutating a cluster

Table 4.2: Classification performance for binding hot spots with $\Delta\Delta G \geq 1$ kcal/mol from the ASEdb dataset, and for residues labeled with ‘intermediate’ or ‘strong’ in the BID dataset.

Methods	ASEdb			BID		
	precision	recall	F1	precision	recall	F1
KFC	0.68	0.35	0.46	0.76	0.30	0.43
FOLDX	0.75	0.48	0.59	0.64	0.45	0.53
Robetta	0.67	0.62	0.64	0.72	0.53	0.61
MINERVA ¹	0.75	0.6	0.66	0.73	0.47	0.57
$\ Z\text{-score}\ > 1$	<i>0.59</i>	<i>0.78</i>	0.67	<i>0.61</i>	<i>0.81</i>	0.69
$\ Z\text{-score}\ ^c > 1$	<i>0.67</i>	<i>0.69</i>	<i>0.68</i>	<i>0.69</i>	<i>0.63</i>	<i>0.56</i>

¹ MINERVA is trained based on hot spots with $\Delta\Delta G \geq 1$ kcal/mol on ASEdb [66].

^c combines our Z-score method with all the other four literature methods. Please refer to the last third paragraph of Section 4.4.2 for the detail of the combination.

Table 4.3: Classification performance for binding hot spots with $\Delta\Delta G \geq 1$ kcal/mol for each complex from the ASEdb dataset.

PDB	Z-score			Robetta			FOLDX		
	precision	recall	F1	precision	recall	F1	precision	recall	F1
1A22	0.86	0.35	0.50	0.50	0.39	0.44	0.43	0.46	0.44
1A4Y	0.83	0.33	0.48	0.83	0.36	0.50	1.00	0.86	0.92
1AHW	0.25	1.00	0.40	0.25	1.00	0.40	0.25	1.00	0.40
1BRS	0.80	0.89	0.84	0.90	0.90	0.90	0.50	1.00	0.67
1BXI	0.60	0.86	0.71	0.40	0.80	0.53	0.40	0.80	0.53
1CBW	1.00	0.33	0.50	0.00	0.00	0.00	0.00	0.00	0.00
1DAN	1.00	0.25	0.40	1.00	0.67	0.80	0.50	0.33	0.40
1DVF	0.60	0.82	0.69	0.20	0.75	0.32	0.20	1.00	0.33
1F47	0.75	0.60	0.67	0.75	0.60	0.67	0.75	0.60	0.67
1FC2	1.00	1.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
1FCC	1.00	1.00	1.00	0.80	1.00	0.89	0.80	1.00	0.89
1GC1	1.00	0.50	0.67	0.67	0.33	0.44	0.33	0.33	0.33
1JRH	0.69	0.69	0.69	0.88	0.93	0.90	0.56	1.00	0.72
1NMB	0.00	0.00	0.00	1.00	1.00	1.00	1.00	1.00	1.00
1VFB	1.00	0.53	0.69	0.67	0.67	0.67	0.67	0.86	0.75
2PTC	1.00	1.00	1.00	1.00	1.00	1.00	0.00	0.00	0.00
3HFM	0.94	0.94	0.94	0.75	0.92	0.83	0.44	0.88	0.58

Table 4.4: Classification performance for binding hot spots with $\Delta\Delta G \geq 2$ kcal/mol from the ASEdb dataset, and for residues labeled with ‘strong’ in the BID dataset.

Methods	ASEdb			BID		
	precision	recall	F1	precision	recall	F1
KFC	0.58	0.55	0.57	0.48	0.31	0.38
FOLDX	0.59	0.31	0.40	0.48	0.26	0.33
Robetta	0.62	0.51	0.56	0.52	0.33	0.41
MINERVA ²	0.73	0.58	0.65	0.65	0.44	0.52
HotPoint	0.49	0.52	0.51	0.5	0.59	0.54
$\ Z\text{-score}\ > 1$	<i>0.39</i>	<i>0.94</i>	<i>0.55</i>	<i>0.41</i>	<i>0.90</i>	<i>0.57</i>
$\ Z\text{-score}\ > 1.5$	<i>0.47</i>	<i>0.82</i>	<i>0.59</i>	<i>0.4</i>	<i>0.69</i>	<i>0.51</i>
$\ Z\text{-score}\ ^c > 1$	<i>0.54</i>	<i>0.77</i>	<i>0.64</i>	<i>0.51</i>	<i>0.64</i>	0.57

² MINERVA is trained based on hot spots with $\Delta\Delta G \geq 2$ kcal/mol on ASEdb [66].

^c combines our Z-score method with all the other four literature methods.

of residues at one time. Furthermore, the Z-scores for DWE bicliques are induced from the estimation of the potential distribution for single residues—induced errors may exist in the Z-score calculation of DWE bicliques. So, we do not expect an improvement of prediction performance when Z-scores go from single residues to DWE bicliques.

Table 4.5: Classification performance for DWE-biclique binding hot spots.

Methods	ASEdb			BID		
	precision	recall	F1	precision	recall	F1
$\ Z\text{-score}\ ^1 > 1$	0.59	0.78	0.67	0.61	0.81	0.69
$\ Z\text{-score}_{hs,20}\ ^1 > 0.5$	0.60	0.68	0.64	0.63	0.73	0.68
$\ Z\text{-score}_{hs,36}\ ^1 > 0.5$	0.53	0.91	0.67	0.57	0.81	0.67
$\ Z\text{-score}\ ^2 > 1$	0.39	0.94	0.55	0.41	0.90	0.57
$\ Z\text{-score}\ ^2 > 1.5$	0.47	0.82	0.59	0.40	0.69	0.51
$\ Z\text{-score}_{hs,20}\ ^2 > 0.5$	0.41	0.86	0.56	0.40	0.77	0.53
$\ Z\text{-score}_{hs,36}\ ^2 > 1(2)$	0.36	0.86	0.51	0.39(0.55)	0.74(0.59)	0.51(0.57)

¹ binding hot spots with $\Delta\Delta G \geq 1$ kcal/mol from the ASEdb dataset, or labeled with ‘intermediate’ or ‘strong’ in the BID dataset.

² for binding hot spots with $\Delta\Delta G \geq 2$ kcal/mol from the ASEdb dataset, or labeled with ‘strong’ in the BID dataset.

₂₀ means $T_{asa} = 20\%$, and ₃₆ means $T_{asa} = 36\%$.

Table 4.6: Classification performance for binding hot spots with $\Delta\Delta G \geq 1$ kcal/mol from the ASEdb dataset.

residue	N_m	precision	recall	F1
ILE	10	0.67	1	0.8
VAL	7	0.5	0.67	0.57
LEU	8	0.83	0.83	0.83
PHE	4	0.5	1	0.67
CYS	1	0	0	-
MET	0	0	0	-
ALA	0	0	0	-
GLY	2	0	0	-
THR	19	0	0	-
SER	23	0	0	-
TRP	20	0.68	1	0.81
TYR	33	0.68	0.83	0.75
PRO	1	0	0	-
HIS	12	1	0.33	0.5
GLU	22	0.44	0.88	0.58
GLN	14	0.14	0.5	0.22
ASP	26	0.8	0.706	0.75
ASN	15	0.5	0.8	0.62
LYS	25	0.52	1	0.69
ARG	23	0.56	0.9	0.69

Table 4.7: Classification performance for binding hot spots labeled with ‘intermediate’ or ‘strong’ in the BID dataset.

residue	N_m	precision	recall	F1
ILE	6	0.6	1	0.75
VAL	5	0.67	1	0.8
LEU	10	0.5	1	0.67
PHE	13	0.69	1	0.82
CYS	1	0	0	-
MET	2	1	1	1
ALA	1	0	0	-
GLY	5	1	0.2	0.33
THR	6	1	1	1
SER	5	0	0	-
TRP	5	1	1	1
TYR	7	0.71	1	0.83
PRO	3	1	0.5	0.67
HIS	6	1	0.33	0.5
GLU	11	0.29	0.67	0.4
GLN	3	0	0	-
ASP	7	0.83	1	0.91
ASN	2	0	0	-
LYS	14	0.39	1	0.56
ARG	15	0.7	0.7	0.7

4.4.4 Performance Analysis According to Residue Types

To further understand the advantages and disadvantages of the Z-score method, we break down the performance result of Table 4.2 according to residue types. This breakdown result is presented in Tables 4.6 and 4.7, where the residues are ordered by their hydrophobicity with ILE as the most hydrophobic and ARG the least hydrophobic, and N_m stands for the number of mutations for each residue type.

It can be seen from Tables 4.6 and 4.7 that different residue types have various classification performance. After grouping the twenty residue types according to their hydrophobicity, a finding is that the Z-score classification method has a good and stable performance on three residue groups in both ASEdb and BID. One group covers the most hydrophobic residues: ILE, VAL, LEU, and PHE; another group contains ASP, ASN, LYS, and ARG which are the least hydrophobic. These two residue groups have a great surrounding change when they are converted from the environment of protein surfaces to protein binding interfaces. Recall that the residue potential score uses crystal-packing

contacts as background and is derived by using this surrounding change. This is why Z-scores have a better distinguishing power for predicting whether these residues are hot spot residues or not.

The third group containing TRP and TYR can also be clearly classified by the Z-score method with the highest F1 values. On the one hand, these two residues are believed to be enriched in hot spots; on the other hand, TRP and TYR prefer protein binding interfaces to protein surfaces. As their pairwise potentials exactly capture this preference, these hot spot residues can be identified with an excellent accuracy.

However, the Z-score method does not perform well on residues such as CYS, ALA, MET, GLY, THR, SER, PRO, GLN and GLU. There are two possible reasons: one is that there are too few mutations of these residue types for our analysis; the other is that these residues do not show significant difference between protein surfaces and protein interfaces.

In summary, the Z-score classification method is specifically excellent in hot spot prediction for the most hydrophobic residues, for the least hydrophobic residues, as well as for the aromatic residues.

4.4.5 Discussions on Two Confusing Features

High complexity of protein interfaces has given rise to contradictory conclusions about protein interactions and their binding hot spots in the literature. Here, we provide a brief discussion on two features from these conflicting observations.

One is on the effect of the feature *residue pairwise potentials* in the identification of hot spots at protein interfaces. Keskin’s research team proposed an empirical approach by combining conservation, ASA and statistical pairwise residue potentials for predicting binding hot spots [82]. Their approach is based on simple rules of these three features. They found this approach outperformed other prediction methods including learning-based methods. They also claimed that residue solvent and pairwise potentials are the most discriminating features in hot spot predictions. However, Huang’s team had a different conclusion for residue potentials [80]. Huang’s team conducted an analysis on a wide variety of sixty-two sequence and structural features, and identified nine features to develop an SVM-based ensemble classifier to predict hot spots. They found that residue pairwise potentials did not perform well in their study.

In this work, the calculation of the Z-scores of hot spots is derived from ASA-integrated pairwise potential scores. The excellent classification result demonstrated that this Z-score is a distinguishing feature for hot spot prediction. So, we believe that both ASA and residue pairwise potentials are excellent features for hot spot prediction, although ASA seems to contribute slightly more than pairwise potentials—for example in the prediction of mutations with $\Delta\Delta G \geq 2$ kcal/mol hot spots, ASA’s F1 value is 0.53

while the F1 value of pairwise potentials is 0.52. Here, the F1 value for ASA is calculated for those predicted hot spots which are buried, and the hot spot prediction process of pairwise potentials is just similar to that of Z-score.

The second is: which one is dominant in protein-protein association, the hydrophilic or the hydrophobic effect? Many studies have focused on this problem. In [32, 84], they examined sets of protein complexes, and found that hydrophobic clusters played a determining role in protein-protein interactions. In contrast, Ben-Naim did a comprehensive analysis on contributions to the standard free energy by protein-protein association, and argued that it is the hydrophilic rather than the hydrophobic effect that was dominant in biochemical processes including protein-protein association [85].

From Figure 4.1(b), we can see that: (i) for those residues with a Z-score less than -2.0, 3 are Arg, 2 are Asn, 7 are Asp, 2 are Gln, 6 are Glu and 21 are Lys. All of them are hydrophilic residues and none is hydrophobic. (ii) Among those residues with a Z-score bigger than 2.0, 5 are Ile, 2 are Leu, 3 are Phe, 18 are Trp, 9 are Tyr and there is only 1 Cys. Most of them are hydrophobic and aromatic residues. In both the former residues and the latter residues, the majority (59% and 71% respectively) are hot spot residues whose $\Delta\Delta G \geq 1$ kcal/mol. This indicates that many big Z-score hydrophilic residues and hydrophobic residues have great surrounding change upon complex formation, and both types can make remarkable contribution to binding free energy. Therefore, whether the effect is significantly important in protein-protein association depends not only on whether it is a hydrophilic or hydrophobic effect but also on the contact surroundings. In this work, protein surfaces and crystal packing together provide a good background of whether a contact surrounding is biologically significant or not. So, the hydrophilic and/or the hydrophobic effect that deviates from the non-biologically significant surroundings may play an important role in protein-protein association.

4.5 Conclusions

There are few works addressing the biological significance of binding hot spots. In this chapter, we have proposed a statistical Z-score method to assess the biological significance of binding hot spots by combining residue accessible surface area and residue pairwise potentials. With crystal-packing contacts as the background reference state, we have assessed the Z-score biological significance of contact residues (See Section 4.2). To show the utility of this score, we have used Z-scores to classify hot spot residues from non-hot spots stored at two benchmark datasets ASEdb and BID. We have shown that our Z-score method can identify hot spot residues from non-hot spots with a better F1 performance than the previous methods (See Section 4.4.2). Specifically, our Z-score method is excellent in hot spot prediction for the most hydrophobic residues, the least

hydrophobic residues, as well as the aromatic residues (See Section 4.4.4). This good performance demonstrates that most of binding hot spots truly have high biological significance, and that the Z-score biological significance can predict whether a contact residue has a higher probability or not of occurring in or contributing to protein binding interfaces.

Chapter 5

DWE Bicliques and Their Residue and Residue Pairing Preference in Different Types of Protein Interactions

5.1 Introduction

As discussed in Chapter 3, a protein binding hot spot is generally a small cluster of residues [36] tightly packed at the center of the interface between two interacting proteins; though a hot spot constitutes a small fraction of the interface, it contributes most to the binding stability and free energy. A hot spot of binding energy was initially conceptualized by Clackson and Wells (1995), supported by an important finding that a ‘functional epitope’ (a hot spot) between human growth hormone and its receptor accounts for more than three-quarters of the binding free energy. This hot spot was found to be geometrically surrounded by less important contact residues that are generally hydrophilic and partially hydrated [38]. On the basis of these pioneering observations and studies, Bogan and Thorn (1998) formalized more intuitively a hypothesis named O-ring theory to characterize the topological shape of the surrounding residues [36]. The O-ring theory points out that the residues of the O-ring likely play a role to occlude bulk water molecules from the hot spots. Thus, the O-ring theory is also known as ‘water exclusion’ hypothesis [65, 36, 86, 109, 110, 108, 87]. This theory is profound and influential. However, the organizational topology of the ring-inside, energetically more important hot spot residues is not specified by the O-ring theory.

In Chapter 3, we investigate the spatial adjacency and vicinity of these hot spot residues and propose a hypothesis called ‘double water exclusion’ (DWE) to refine the O-ring theory. At one hand, the DWE hypothesis agrees with the O-ring principle that there should exist a ring of residues surrounding the hot spot for avoiding the invasion of water molecules after the complex formation; on the other hand, the DWE hypothesis affirms that the hot spot itself is water-free, having a zero-tolerance to free water molecules. In fact, the DWE hypothesis shares a light with the ‘coupling’ proposition [48] which is another insightful theory about hot spot residues, and it also theoretically strengthens the influential ‘hot region’ principle [52].

Given a protein complex, the requirements for a cluster of interfacial residues to form a binding hot spot under the DWE hypothesis can be mathematically satisfied by a biclique subgraph in Chapter 3 if a vertex is used to represent a residue, an edge to indicate a close distance between two residues, and a bipartite graph to represent a pair of interacting proteins. We term these hot spots as DWE bicliques, and note that in a DWE biclique, residues from one chain all have full connection with the residues from the other chain. In Chapter 3 by applying to the ASEdb repository [86] and the Hotsprint database [60], we found that these DWE bicliques are rich in true hot spot residues.

With the perspective change from the O-ring theory to the DWE hypothesis, it is interesting to study the physicochemical properties of DWE biclique hot spots, and to compare with those [36, 31, 65] under the O-ring theory. Specifically in this chapter, we examine those DWE bicliques that are unique to crystal packing contacts, or unique to biological interactions including obligate and non-obligate interactions. As discussed in Chapter 1, obligate and non-obligate interactions depend on various factors for promoting complex formation although some factors are common, while the non-specific crystal packing should have different properties from these two kinds of biological interactions. So at the residue level, it is of our primary interests to see which DWE bicliques are signature binding hot spots of the different types of protein interactions, and which are common. For demonstrating their such difference, given two types of interactions (e.g., biological interactions and crystal packing contacts), a unique DWE biclique is defined as a DWE biclique that frequently occurs in one type of interactions (e.g., biological interactions) but is absent in the other type (e.g., crystal packing). We also examine the residue composition of DWE bicliques, and the distribution of residue pairs and their pairing preference within DWE bicliques.

Residue composition and residue pairing preference are two fundamental physicochemical properties for protein folding [25, 24] and protein interactions [23, 20]. Both of them have been broadly used in comparative analyses based on protein structure data. One aspect of those comparative analyses is on different segments within protein complex structures, such as comparison between intraproteins and protein interfaces [20],

or interdomain comparison [10]; another angle is on different types of protein interactions [15, 31, 22]. De *et al.* [19] and Lukman *et al.* [21] analyzed residue contacts of obligate complexes and non-obligatory/transient protein complexes, and reported that the interaction patterns of these two types of complexes were different. Ofra and Rost [10] studied the preferences of residue contacts in a more complicated way for six types of protein interfaces. One finding is that the preferences differed remarkably between the six types of interfaces, such as intra-domain, domain-domain, homo-obligomers, homo-complexes, hetero-obligomers and hetero-complexes. However, in almost all previous works, the individual residues and the residue pairs were dissected based on the whole interfaces. In this work, we analyze the composition of residues and residue pairs in DWE bicliques, the energetically important part of each interface.

We propose to use interaction-preferred residues (ipRs) and interaction-preferred residue pairs (ipRPs) to describe the binding specificity for different types of interactions (obligate interactions, non-obligate interactions and crystal packing contacts). Our important findings include: (i) in the obligate interactions, hydrophobic ipRs and ipRP contacts involving only hydrophobic residues are widely conserved, no anti-ipRs are hydrophobic, and the contacts involving only hydrophilic residues are depleted in ipRPs; (ii) aromatic ipRs and their ipRPs much prefer to the biological interactions, especially to the non-obligate interactions; (iii) hydrophilic ipRs and ipRPs involving only hydrophilic residues seem to be rich in crystal packing contacts; the anti-ipRs of crystal packing, such as Met, Trp and Cys, are just the ipRs of the obligate or non-obligate interactions. So, these ipRs and ipRPs in DWE bicliques provide a clear distinction for the specific binding behaviors of biological interactions as well as crystal packing contacts.

5.2 Methods

To mine DWE bicliques from the three types of protein interactions, we firstly produce a non-redundant dataset from the compiled sets of protein interactions in previous works. Then, DWE bicliques are detected from those interacted chain pairs as the descriptors of protein interactions.

5.2.1 Compiling a Non-redundant Dataset

The data used in this chapter contain three types of protein interactions: obligate interactions, non-obligate interactions and crystal packing contacts. All of them are obtained from previously published literature works. (i) The obligate interactions are from the obligate interactions used in [18, 122], as well as the homodimers used in [22, 27]; (ii) the non-obligate interactions comprise the non-obligate interactions used by [18, 122],

and the protein complexes used in [22]; and (iii) the crystal packing contacts are those from [18, 22] and the monomers used in [27].

To get rid of the redundancy within each type of protein interactions, we remove those redundant interaction pairs with high similarity. Let C_1^i and C_2^i , and C_1^j and C_2^j , $i \neq j$, be two pairs of protein interaction chain pairs in one type of interactions, these two pairs are redundant if $score(C_1^i, C_1^j) \geq s$ and $score(C_2^i, C_2^j) \geq s$, or $score(C_1^i, C_2^j) \geq s$ and $score(C_2^i, C_1^j) \geq s$, where the *score* function is a sequence similarity score of two protein sequences and it can be produced by the BLAST software (downloadable from NCBI <http://www.ncbi.nlm.nih.gov/BLAST/download.shtml>) without filtering of low compositional complexity, and $s = 90\%$ here. This redundancy removing process resulted in a non-redundant data set comprising 291 crystal packing contacts, 289 non-obligate interactions and 287 obligate interactions. The distribution of these interactions under different s value ranges is shown in Table 5.1. At one hand, it is clear in Table 5.1 that most of them have a low similarity of $s = 40\%$ or below. On the other hand, the detected bicliques from chain pairs with high similarity are actually different. For example, in Table 5.8, although chain E of 2PTC and chain E of 1TAB are the identical, the occurring bicliques are involved with residues of different positions in the interaction partner chains.

Table 5.1: The chain-pair distribution in our non-redundant dataset according to the similarity.

	Similarity region						Total
	(80,90]	(70,80]	(60,70]	(50,60]	(40,50]	[0,40]	
Obligate	0	1	1	3	11	271	287
Non-obligate	3	4	6	5	15	256	289
Crystal Packing	0	3	1	4	9	274	291

5.2.2 Mining Maximal DWE Bicliques for Protein Interactions

We take the Algorithm in Section 3.3.2 to mine maximal DWE bicliques from the above dataset with the setting $p = 2, q = 3, sup = 2$. Why $sup = 2$ is that bicliques occurring once cannot show the specificity of binding behaviors in different types of protein interactions. We would like to note that the Z-score biological significance for DWE bicliques in Chapter 4 should be a better solution to filtering out uninteresting and random DWE bicliques here. However, as discussed in Chapter 4, Z-score for DWE bicliques is induced from the estimation of the potential distribution for single residues—induced errors may exist in the Z-score calculation of DWE bicliques; thus, Z-score of DWE bicliques needs further improving. Meanwhile, the datasets of crystal packing and biological interactions

used in the Z-score calculation of Chapter 4 are almost same as the datasets used in this chapter for mining DWE bicliques. To use Z-score for producing interesting bicliques in this chapter, another considerable dataset should be compiled for crystal packing and biological interactions. Thus, we still use the Algorithm in Section 3.3.2 to produce interesting DWE bicliques instead of using Z-score.

5.2.3 The Definitions Related to Our Composition Analysis on Residues and Residue Pairs

We calculate the frequencies of residues and residue pairs in DWE bicliques for each type of protein interactions. We define interaction-dominated patterns, idR and idPR, as follows. Given an amino acid in one type of interactions, if it is at the top of residue rank in the frequency descending order, this amino acid maybe contributes more to this type of protein interactions and is defined as an interaction-dominated residue (idR). In this work, idRs are defined as the top seven residues (a little more than one third of all the twenty standard residues). If a pair of contact residues has top frequencies in one type of protein interfaces, this pair is termed as an interaction-dominated amino acid pair (idRP).

Interaction-dominated patterns alone might not clearly exhibit the preferred ways for different types of interactions, as the percentages of the twenty standard residues in nature are not equal. Thus, we define interaction-preferred patterns to help the dissection on protein interfaces. Assume that the percentage of amino acids for the complete database from release 55.0 of the Swiss-Prot database (<http://cn.expasy.org/sprot/relnotes>) is our background residue composition, then

- (i) we define a residue as an **ipR** in one type of interactions, if the frequency of the residue is higher than its frequency in the background residue composition. As this amino acid prefers to this type of protein interactions, we name it an interaction-preferred residue (ipR). In this work, the extent of such preference is measured by the ratio of the residue frequencies in each type of protein interfaces over the corresponding residue percentage in the background composition. This ratio can indicate how much the residues prefer a certain type of protein interfaces. Thus, in a given type of protein interactions, if the ratio of a residue is greater than a ratio threshold, this residue is an ipR. To reduce the influence of random errors, this ratio threshold here is set to 1.05 rather than 1. In contrast, the residues whose ratios are less than another ratio threshold are called anti-ipR residues. In this work, the second ratio threshold is set to 0.8. Please note that Ofra and Rost [10] also took the percentage of amino acids from the complete database of the Swiss-Prot database as a background residue composition to measure the preference of residues.

- (ii) we define a residue pair as an interaction-preferred residue pair (**ipRP**) if it has much higher frequencies than random cases. In this work, the frequencies of residue pairs (r_i, r_j) in the random case are calculated as $\begin{cases} f_{r_i} \times f_{r_j} & \text{if } i = j \\ 2 \times f_{r_i} \times f_{r_j} & \text{if } i \neq j \end{cases}$ where f_{r_i} and f_{r_j} are the background percentage of residues r_i and r_j . Here, there is an assumption in the random case: each amino acid occurs in protein interfaces independently. Similar to the definition of ipRs, frequency ratios of residue pairs are calculated by the frequencies of residue pairs in each type of interfaces over those in the random case. These ratios can suggest the specificity of ipRPs in the different types of interactions. However, there are too many ipRPs. In this work, we focus on the residue pairs with the highest frequency ratios, and ipRPs are only referred to as those residue pairs whose natural logarithm of $r+1$ is larger than 1.05 where r is their frequency ratio.

We would like to point out that only patterns in biological interactions are interesting, and patterns in crystal packing are used as reference for comparison.

5.2.4 Measures for Comparison Analysis on the Residue Composition Differences

In this work, we employ Euclidean distance Δf as defined in [31, 65] to measure the difference of residue percent compositions in different types of protein interactions:

$$\Delta f = \frac{\sum_{i=1}^{20} (f_i - f'_i)^2}{19} \quad (5.1)$$

where f and f' are percent composition of the twenty standard amino acids. Meanwhile, correlation coefficient [10], CC , is also used to compare residue ratio compositions:

$$CC = \frac{\sum_{i=1}^{20} (r_i^x - \bar{r}^x) \times (r_i^y - \bar{r}^y)}{\sqrt{\sum_{i=1}^{20} (r_i^x - \bar{r}^x)^2 \times \sum_{i=1}^{20} (r_i^y - \bar{r}^y)^2}} \quad (5.2)$$

where r^x and r^y are frequent ratio vector for the twenty standard amino acids, and $\bar{r}$ is the mean of the corresponding r_i s.

5.3 Results and Discussion

On the non-redundant data set consisting of 287 obligate interactions, 289 non-obligate interactions and 291 crystal packing contacts in Section 5.2, we identified 1580, 725 and 208 DWE bicliques, respectively. All of these DWE bicliques have at least two residues in the smaller side and at least three residues in the bigger side, and occur in at least

two interactions. There are 26 obligate interactions, 75 non-obligate interactions and 159 crystal packing that do not contain any DWE bicliques. Please note that by ‘containing no DWE bicliques’, we mean that these interactions do not contain any DWE bicliques of high frequency or of big size; they may contain some bicliques with low support (only one occurrence) or with a small size (for example, those bicliques with only one residue from each protein chain).

The remaining results are organized into four parts in this section. Firstly, we show that a binding hot spot is usually a very small area in comparison to its binding interface. Then, we present the distribution of DWE bicliques among the three types of protein interactions to provide an intuitive way for understanding the protein interfaces. After that, we conduct statistical analysis on the composition of the residues and on the pairing preference of residue pairs in our DWE bicliques.

5.3.1 Binding Hot Spots are Small in Binding Interfaces

We calculated the fractions of residues in DWE bicliques over interface residues. As expected, we found that binding hot spots are small in binding interfaces as shown in Table 5.2. In this table, all the DWE bicliques are from the biological interactions; in the first column, the numbers split by ‘-’ mean that the residue number of the smaller partite in DWE bicliques must be not less than the left number, and that the residue number of the larger partite must be not less than the right number; in the following two columns, $_C$ means that the fraction is over the whole contact residues, and $_I$ means that the fraction is over the whole interface residues. Here, interface residues comprise *contact residues* and *their nearby residues* [46]. The definitions for contact residues and their nearby residues are similar to those in Chapter 4.2.1 and 4.3.2. The only difference is that the threshold for nearby residues is 6 Å [46], smaller than 8.5 Å [123, 124] used in Chapter 4.3.2. Although there are fewer nearby residues when this threshold is 6 Å than is 8.5 Å, the conclusion still hold true: binding hot spots are small in binding interfaces.

Table 5.2: The fraction of DWE bicliques over the interfaces.

Minimum Biclique Size	Fraction $_C$ (%)	Fraction $_I$ (%)
1-2	71.99	39.33
2-2	50.23	27.44
2-3	44.69	24.42

5.3.2 Common and Unique DWE Bicliques in Different Types of Interactions

In this work, unique DWE bicliques are defined based on bicliques’ *support* information. Given a set of interactions containing one or more types of interactions, a DWE biclique

may have multiple occurrence in these interactions. We call the number of occurrence the *support* level of this biclique. We also use *support ratio* r to help the comparisons of DWE bicliques in different types of interactions. Suppose that a biclique has a higher support S_h in one type c_h and a lower support S_l in the contrasting type c_l , $r = S_h/S_l$ indicates the preference of this biclique between these two types. In the extreme cases when the support ratio r is infinite (i.e., $S_l=0$), this biclique is more interesting and it is defined as a unique biclique. In other words, this kind of unique bicliques have zero occurrence in one entire type of protein interactions but have multiple, perhaps very high, occurrence in the other type. To avoid some possible noise patterns, we require that $S_h + S_l$ should be not less than 2 which is consistent with the biclique filtering constraint presented in Section 5.2. Thus for unique bicliques, S_h is always equal to or greater than 2 since $S_l = 0$.

Obligate and non-obligate interactions are two types of distinct biological interactions, each possessing their specific binding behaviors. However, crystal packing contacts are randomly formed by chance. We present DWE bicliques that are unique to biological interactions (c_{p+}) when compared to crystal packing contacts (c_{n-}), and those unique to crystal packing when compared to biological interactions. Meanwhile, the common bicliques, those with support ratio around 1.0, between biological interactions and crystal packing are also presented. Then we report common and unique DWE bicliques between obligate (c_{p+}) and non-obligate interactions (c_{n-}) to demonstrate their distinct binding specificity. After that, a case study is followed to examine deep structural details of unique DWE bicliques. Here, the symbol c_{p+} stands for a ‘positive’ set, while the symbol c_{n-} stands for a ‘negative’ set. If the higher support S_h of a biclique is in the positive set, then we denote the support ratio r as S_h/S_l ; otherwise if the higher support S_h is in the negative set, we denote r as $-S_h/S_l$.

5.3.2.1 Common and Unique DWE Bicliques: Biological Interactions vs Crystal Packing Contacts

The distribution of DWE bicliques on the range of their support ratios from -INF to +INF is shown in Table 5.3. It can be noted that the type of crystal packing contacts contains much fewer favorite DWE bicliques than the type of biological interactions (columns 2 and 3 vs columns 5-8). In particular, there are only 13 DWE bicliques unique to crystal packing, counting to only 0.79% of the total DWE bicliques occurring in the biological and crystal packing interactions. In contrast, the type of biological interactions is rich of DWE bicliques, including unique bicliques. Table 5.3 shows that about 86.2% bicliques are unique bicliques to biological interactions. Few DWE bicliques contained by crystal packing but abundant DWE bicliques in biological interactions are interpretable with the following reasons: (i) crystal packing are constructed randomly based on protein

surfaces during the crystallization process, and their artifact interfaces cannot form much repeatable DWE bicliques; (ii) residues in biological interfaces are found to be more conserved than crystal packing and than the rest of protein surfaces [18, 17], and thus biological interactions can easily form stable and repeatable biclique structures.

Table 5.3: Common and unique DWE bicliques in biological interactions and crystal packing.

(-)crystal packing				biological interactions(+)			
support level	range of support ratios						
	-INF	-2	1	2	3	>=4	INF
	-INF	(-2.5,-1.5]	(-1.5,1.5)	[1.5,2.5)	[2.5,3.5)	[3.5,7.5)	INF
2	13	0	152	0	0	0	1144
3	0	3	0	38	0	0	207
4	0	0	0	0	13	0	40
>=5	0	0	0	1	7	6	23
Total	13	3	152	39	20	6	1414

5.3.2.2 Common and Unique DWE Bicliques: Obligate vs Non-obligate Interactions

Table 5.4 shows the number of common and unique bicliques for the obligate and non-obligate interactions. About 43.5% (641) of them are unique bicliques for the type of obligate interactions, and 16.6% (244) of them are for the type of non-obligate interactions. These unique bicliques can demonstrate the intuitive difference of binding behaviors in the two types of biological interactions, although there exist overlapping bicliques occurring in both obligate and non-obligate interactions. Detailed information of important unique DWE bicliques will be provided in the next subsection.

Table 5.4: Common and unique DWE bicliques in obligate and non-obligate interactions.

	(-)non-obligate interactions				obligate interactions(+)		
support level	range of support ratios						
	-INF	-3	-2	1	2	>=3	INF
	-INF	(-3.5,-2.5]	(-2.5,-1.5]	(-1.5,1.5)	[1.5,2.5)	[2.5,5.5)	INF
2	223	0	0	436	0	0	523
3	17	0	38	0	76	0	90
4	2	5	0	6	0	12	19
>=5	2	0	1	0	3	11	9
Total	244	5	39	442	79	23	641

5.3.2.3 Related Evidence for Unique Biclques

The existence of unique bicliques in the three types of interactions is in agreement with the observations in Chapter 6 for classifying three types of interactions (i.e., crystal packing, obligate and non-obligate interactions). In Chapter 6, we compare classification performance between our propensity vectors, binary vectors and frequency vectors on three literature datasets under three evaluation frameworks; all of the propensity vectors, binary vectors and frequency vectors contain 3-dimensional summary information of protein interactions and another 210 dimensions for residue pairs. The difference is that the values of 210-dimensional residue pairs in binary vectors indicate whether a certain residue pair occurs or not in a certain corresponding protein interaction, and those in frequency vectors are the frequency of corresponding residue pairs in protein interactions, while those in propensity vectors are the propensity value of corresponding residue pairs in protein interactions.

Our comparison results of these three vectors in Chapter 6 suggest the following two evidences to support our concept of unique bicliques: (i) In almost all cases, the performances under the DWE hypothesis are better than those on all interface residues. That is, the whole interfaces may cover more noise features among different types of interactions, while DWE bicliques can remove some noise patterns and pinpoint distinct features. (ii) Binary vectors have a similar high classification performance to, sometimes a bit higher than, the frequency vectors. That similar performance maybe implies that certain combination patterns of residue pairs rather than their frequency are signature features for different types of protein interactions. When DWE bicliques are detected from DWE bipartites of protein interactions, these DWE bicliques are likely to be this kind of the combination patterns. In fact, the union of bicliques, possibly not maximal bicliques, might also be one kind of the combination patterns.

5.3.2.4 Unique Biclques: A Case Study

We present four unique DWE bicliques and study their structural properties. They are: (i) a unique biclique $\mathcal{G}_1 = \langle \{\text{GLY, GLY, LYS}\}, \{\text{GLN, THR}\} \rangle$, which occurs only in two crystal packing contacts (Table 5.5); (ii) a unique biclique $\mathcal{G}_2 = \langle \{\text{LEU, THR, VAL}\}, \{\text{LEU, VAL}\} \rangle$, which occurs only in five obligate interactions as shown in Table 5.6; a DWE biclique $\mathcal{G}_3 = \langle \{\text{ALA, LEU, VAL}\}, \{\text{LEU, TYR}\} \rangle$ which is a unique biclique to the type of biological interactions occurring in two obligate and three non-obligate interactions (Table 5.7), and (iv) a unique biclique $\mathcal{G}_4 = \langle \{\text{GLN, GLY, SER, SER, TYP}\}, \{\text{CYS, LYS}\} \rangle$ contained in only 6 non-obligate interactions (Table 5.8). At the first column of these four tables, the first four letters represent PDB entry identifiers, if necessary, followed by ‘-’ and two interaction protein chains which are separated by ‘:’. At the columns 2 and 3 of these four tables, the strings split by ‘-’ are residue

Table 5.5: Interactions involving the unique biclique $\mathcal{G}_1 = \langle \{\text{GLY}, \text{GLY}, \text{LYS}\}, \{\text{GLN}, \text{THR}\} \rangle$ of crystal packing.

PDB entry	First Chains	Second Chains
1BG0	GLY117-LYS235-GLY236	GLN196-THR197
2ACY	GLY15-LYS16-GLY45	GLN44-THR46

Table 5.6: Interactions involving the unique biclique $\mathcal{G}_2 = \langle \{\text{LEU}, \text{THR}, \text{VAL}\}, \{\text{LEU}, \text{VAL}\} \rangle$ of obligate interactions.

PDB entry	First Chains	Second Chains
1AD3-A:B	VAL396-THR398-LEU399	VAL86-LEU97
1B7B-A:C	THR109-VAL110-LEU111	VAL110-LEU111
1B8J-A:B	LEU33-THR81-VAL430	LEU80-VAL430
1LDJ-A:B	VAL539-LEU540-THR580	VAL30-LEU32
1QOE-A:B	LEU250-VAL261-THR262	LEU250-VAL261

types followed by their corresponding positions in the amino acid sequences, representing specific residues from the two interacting protein chains.

We take these examples to highlight that the uniqueness of DWE bicliques matches to different interfacial properties of the three types of protein interactions in terms of polarity, hydrophobicity, the composition of residues and residue pairs in protein interfaces. As mentioned, the unique biclique $\mathcal{G}_1$ (Table 5.5) has a support of only 2. In fact, the maximum support of the unique bicliques that occur in the type of crystal packing contacts is 2. While the maximum support of the unique bicliques in the biological interactions is larger, 7 for the obligate bicliques and 6 for the non-obligate bicliques. Thus, the biclique structures in the biological interactions are more stable and repeatable than those in the crystal packing contacts. The four examples of unique bicliques also give a glance at residue composition in the three types of interactions. As shown in Table 5.5, the unique biclique $\mathcal{G}_1$ of crystal packing consists of more polar and hydrophilic residues, such as GLY/LYS and GLN/THR, while the unique biclique $\mathcal{G}_2$ to the biological interactions comprises more hydrophobic residues, such as LEU and VAL as shown in Table 5.6. Table 5.6 also indicates that the contacts of identical residues easily occur in obligate interactions. However, this is found less in non-obligate interactions as shown in Table 5.7 and Table 5.8.

We would also like to present the conservation scores and the residues' ASA of unique bicliques in specific PDB entries. We take $\mathcal{G}_1$ in the crystal packing 2ACY in Table 5.9 and $\mathcal{G}_4$ in the transient interaction 2PTC in Table 5.10 as example. In these two tables, the conservation score is taken from the website of rate4site [128], while ASA is calculated by NACCESS [12]. The location of these two bicliques at the protein interfaces are displayed in Figure 5.1.

5.1.a Biclique location at interface of Chain I in 2PTC

5.1.b Biclique location at interface of Chain E in 2PTC

5.1.c Biclique location at interface of Chain B in 2ACY

5.1.d Biclique location at interface of Chain A in 2ACY

Figure 5.1: Examples of biclique location at the protein interfaces. All residues in this figure are shown in ‘spheres’ and ‘surface’ view with the magenta color; the biclique residues are in green; the parts without ‘surface’ view are the interfaces; the complexes of each row can be constructed by rotating the left figure counterclockwise and the right clockwise round the axis of the middle vertical line.

Table 5.7: Interactions involving the DWE biclique $\mathcal{G}_3 = \langle \{\text{ALA}, \text{LEU}, \text{VAL}\}, \{\text{LEU}, \text{TYR}\} \rangle$ occurring in both obligate and non-obligate interactions.

PDB entry	First Chains	Second Chains	Interaction Type
1AT3-A:B	LEU214-ALA217-VAL218	TYR124-LEU213	Obligate
1GO3-E:F	VAL12-ALA17-LEU41	TYR2-LEU75	Obligate
1DOA-A:B	VAL36-ALA59-LEU67	LEU48-TYR51	Non-obligate
1EVT-A:C	LEU165-ALA167-VAL168	TYR15-LEU135	Non-obligate
1JSU-A:C	VAL18-ALA31-LEU83	LEU84-TYR88	Non-obligate

 Table 5.8: Interactions involving the unique biclique $\mathcal{G}_4 = \langle \{\text{GLN}, \text{GLY}, \text{SER}, \text{SER}, \text{TYP}\}, \{\text{CYS}, \text{LYS}\} \rangle$ of non-obligate interactions.

PDB entry	First Chains	Second Chains
1EJA-A:B	GLN192-SER195-SER214-TRP215-GLY216	CYS33-LYS34
1TAB-E:I	GLN192-SER214-TRP215-GLY216-SER217	CYS24-LYS26
1TGS-Z:I	GLN192-SER214-TRP215-GLY216-SER217	CYS16-LYS18
2BTC-E:I	GLN192-SER214-TRP215-GLY216-SER217	CYS503-LYS505
2PTC-E:I	GLN192-SER195-SER214-TRP215-GLY216	CYS14-LYS15
1BTH-H:P	GLN192-SER195-SER214-TRP215-GLY216	CYS14-LYS15

It can be seen from Table 5.10 that the biclique residues in $\mathcal{G}_4$ have relatively small ASA in the 2PTC complex and larger ASA change upon complex formation. This can be easily understood from Figure 5.1: Figure 5.1(a) and Figure 5.1(b) clearly show that the biclique residues in the biological interface 2PTC are buried. The ASA of LYS15 in chain I decreases from 201.71 Å to 0.6 Å, indicating that this biclique is closer to the interface center than to the rim of interface in Figure 5.1(a). However in the crystal packing 2ACY, both partites of $\mathcal{G}_1$ have residues with relatively larger ASA, more than 50 Å. So, the biclique residues in chain B in Figure 5.1(c) are a little away from the interface center.

As shown in Table 5.9 and Table 5.10, both bicliques contain residues with high conservation scores. However, three of the seven residues in $\mathcal{G}_4$ have a conservation score less than 5. For example, the conservation score of LYS15 in chain I of 2PTC is 3, but this residue contributes greatly to the formation of the complex 2PTC — its mutation results in a big binding free energy change (10 kcal/mol) according to ASEdb [86]. This observation might give a hint that although residue conservation is one of major factors contributing to frequent bicliques, frequent unique bicliques to biological interactions can capture more specific evidence for understanding complex formation than the conservation alone, such as ASA, residue physicochemical properties, and tightly packing residue contact.

Table 5.9: The conservation score and ASA ($\text{\AA}$) information of the unique biclique $\mathcal{G}_1 = \langle \{\text{ARG, GLY, SER}\}, \{\text{ASN, LYS}\} \rangle$ in PDB entry 2ACY

Pos.	Residue	Cons. Score	ASA in chain	ASA in complex
15	GLY	9	35.33	0.0
16	LYS	6	131.14	53.37
45	GLY	9	36.32	0.2
44	GLN	6	147.86	54.17
46	THR	7	10.97	0.0

Next, we present our sequence and structural analysis results on the unique biclique $\mathcal{G}_4$. As shown in Table 5.8, $\mathcal{G}_4$ occurs only in six non-obligate interactions in six different PDB protein complexes. These six interactions are all about trypsins/trypsinogen interacting with different types of inhibitors in different organisms. For example, 1TGS is about ‘three-dimensional structure of the complex between pancreatic secretory inhibitor (kazal type) and trypsinogen’, and 2BTC is about ‘bovine trypsin in complex with squash seed inhibitor (cucurbita pepo trypsin inhibitor II)’. The sequence similarities of the six interacting chain pairs are as follows. Chain E of 1TAB, chain Z of 1TGS, chain E of 2BTC and chain E of 2PTC are identical chains. In comparison to this identical chain, chain A of 1EJA has only 83% sequence similarity and chain H of 1BTH possesses only about 36% sequence similarity. The sequence similarity among the other chains of these interactions is very low except two identical chains P in 1BTH and I in 2PTC. Overall, there are no two pairs of interactions whose sequence similarity is larger than 40%. That is, there is no pair sequence redundancy in these six non-obligate interactions. We also note that although the two chain Es in 2BTC and in 2PTC are identical chains, the specific residues involved in $\mathcal{G}_4$ are not the same due to the low similarity between their partner chains. Residue SER in 2BTC is in the position 217, while it is in the position 195 in 2PTC. So, both bicliques are interesting to show. The details of these sequence similarities are provided in Table 5.11. The computational steps for determining the sequence similarity between two sequence pairs can be found in Section 5.2.

The 3D structures of this DWE biclique in the six different PDB protein complexes are displayed and compared in Figure 5.2. The 3D shape of these structures looks highly similar to each other with a common lock-and-key topology [114]. Since this stable topology is repetitive in six non-obligate interactions, it is worthy of further investigation to see whether this group of residues in this biclique is closely related to or involved in the above mentioned protein functions.

Another interesting question is: which residues in this DWE biclique are energetically outstanding. As mentioned, Lys in the sequence position 15 of chain I in 2PTC is a wet-lab confirmed hot spot residue with an extremely high energy (10 kcal/mol) according to ASEdb [86]. This may suggest that the Lys residue is also a hot spot residue in the other 5 interacting chain pairs.

5.2.a Biclique Structure at Chain H and P in 1BTH

5.2.b Biclique Structure at Chain A and B in 1EJA

5.2.c Biclique Structure at Chain E and I in 2PTC

5.2.d Biclique Structure at Chain Z and I in 1TGS

5.2.e Biclique Structure at Chain E and I in 2BTC

5.2.f Biclique Structure at Chain E and I in 1TAB

Figure 5.2: The structure of the unique biclique $\mathcal{G}_4 = \langle \{\text{GLN}, \text{GLY}, \text{SER}, \text{SER}, \text{TYP}\}, \{\text{CYS}, \text{LYS}\} \rangle$ in six interaction chain pairs. In these figures, residues from one chain in blue color are shown in the ‘mesh’ view, and residues from the other chain in green color are shown in the ‘spheres’ view.

Table 5.10: The conservation score and ASA (Å) information of the unique biclique $\mathcal{G}_4 = \{\text{GLN, GLY, SER, SER, TYP}\}, \{\text{CYS, LYS}\}$ in chain E and I of PDB entry 2PTC.

Pos.	Residue	Cons.	Score	ASA in chain	ASA in complex
residues in 2PTC chain E					
192	GLN	1		116.42	20.18
195	SER	9		19.16	0.0
214	SER	9		10.96	2.06
215	TRP	4		51.96	5.44
216	GLY	9		29.15	1.49
residues in 2PTC chain I					
14	CYS	9		55.63	0.0
15	LYS	3		201.71	0.6

Table 5.11: Sequence similarity (%) among corresponding chains of the protein interaction pairs in Table 5.8.

<i>PDB entries</i>	<i>1EJA:B</i>		<i>1TAB:I</i>		<i>1TGS:I</i>		<i>2BTC:I</i>		<i>2PTC:I</i>		<i>1BTH:P</i>	
1EJA:A	A	<i>B</i>	-		-		-		-		-	
1TAB:E	83		E	<i>I</i>	-		-		-		-	
1TGS:Z	83		100		Z	<i>I</i>	-		-		-	
2BTC:E	83		100		100		E	<i>I</i>	-		-	
2PTC:E	83		100		100		100		E	<i>I</i>	<i>100</i>	
1BTH:H	37		36		36		36		36		H	<i>P</i>

The *italic* half is for the sequence similarity among these inhibitors, while the **bold-face** half is for the sequence similarity among trypsins/trypsinogen; ‘-’ means no significant sequence alignment.

5.3.3 Residue Composition of the DWE Bicliques for the Three Types of Protein Interfaces

The residue composition of protein binding interfaces or binding sites has been intensively studied previously [15, 31, 22, 19, 21, 129]. The composition of residues and residue pairs in DWE bicliques are studied by the current work in order to understand whether protein binding hot spots change their residue composition under the constraint of ‘double water exclusion’ hypothesis. We focus on the preference and tendency of residues to the specific types of interactions, as well as the preference and tendency of residue pairs. We would like to note that the composition of residues and their pairs in unique bicliques may be more interesting than those in DWE bicliques to show the residue preference in different types of interactions, because unique bicliques only occur in one type of interactions. But our investigation shows that there is no significant change for the

composition of residues and their pairs in going from DWE bicliques to unique bicliques. This situation may be due to (i) that unique bicliques dominate DWE bicliques, and/or (ii) that common bicliques among the different types of interactions, especially those with larger support ratios, may also cover useful patterns for understanding protein binding behaviors. Therefore in the following two subsections, our analysis on the residues and their pairs in DWE bicliques is not on unique bicliques alone.

We begin our analysis on the interaction-dominated residues (short for idRs) and interaction-preferred residues (short for ipRs). A residue is an idR in a type of interactions if its percent frequency in a set of DWE bicliques for this type of interactions is high; while a residue is an ipR in a type of interactions if its frequency ratio over the background is high. See Section 5.2 for the detailed definitions of idRs, ipRs and anti-ipRs. Figure 5.3(a) shows the frequency information of the twenty amino acids in our DWE bicliques for the three types of interactions, and Figure 5.3(b) displays the frequency ratio information of the twenty amino acids with reference to their background frequencies.

From Figure 5.3(a), we can see that the idRs for the obligate interactions are Leu (14.4%), Ala (8.20%), Val (7.63%), Gly (6.78%), Ile (6.68%), Arg (6.68%) and Phe (6.38%) ordered by their frequencies. Five of these residues are hydrophobic except Gly and Arg. However, Arg is broadly considered to be the richest in hot spots [36]. Gly's frequency ratio over its background percentage is 0.96, near to 1. That is, the abundance of Gly in nature makes Gly become an idR. Thus, we can make a conjecture that the binding hot spots of obligate interactions are dominated by hydrophobic residues. This point agrees to the frequency ratio trend of the ipRs of the obligate interactions as shown in Figure 5.3(b). The ipRs in these obligate interactions are Tyr (1.87), Phe(1.64), Leu(1.49), Met(1.38), Arg(1.22), Trp(1.17), Ile(1.13), Val(1.12) and Thr(1.07) according to the ranking of their frequency ratios. Six of these ipRs are hydrophobic except Tyr, Arg and Thr. However, Tyr is an aromatic residue which can form π - π /cation- π interactions to stabilize protein binding. This is also why all aromatic residues, Tyr, Phe and Trp, are ipRs of obligate interactions. Another interesting observation from Figure 5.3(b) is that the anti-ipRs are Cys, Gln, Ser, Glu, Asp and Lys—none of them is hydrophobic. Therefore, all these observations are consistent, and indicate that the binding hot spots of obligate interactions are hydrophobic and stable.

Different from the obligate interactions, the binding hot spots of the non-obligate interactions contain only three hydrophobic idRs (Leu, Val and Ile), in addition to three hydrophilic idRs (Gly, Tyr and Ser) and one basic idR (Arg). It seems that non-obligate interactions are generally less hydrophobic than obligate interactions. The ipRs of these non-obligate interactions have a similar composition to their idRs, including three hydrophilic residues (Trp, Phe and Leu), three hydrophobic residues (Tyr, Cys and Gly), and two basic residues (His and Arg). Three aromatic residues, especially Trp and Tyr, seem to have a higher propensity to non-obligate interactions.

5.3.a Frequencies of the residues in the DWE bicliques of the three types of protein interactions.

5.3.b Ratios of residue frequencies of the DWE bicliques of the three types of protein interactions over the background frequencies.

Figure 5.3: Residue compositions in DWE bicliques are different in the three types of protein interactions: obligate, non-obligate interactions and crystal packing. Here, the residues are ordered according to their hydrophobicity with Ile as the most hydrophobic and Arg as the least hydrophobic. The background residue frequencies are the percentage of the amino acids for the complete database from the release 55.0 of Swiss-Prot.

The idRs of crystal packing are Leu (10.65%), Gly (8.46%), Arg (7.68%), Glu (7.16%), Ala (6.98%), Ile (5.85%) and Tyr (5.76). But the anti-ipRs of crystal packing, such as Met, Trp and Cys, become ipR residues of obligate or non-obligate interactions. As expected, biological interactions have different residue preference from crystal packing in their DWE bicliques.

5.3.3.1 Quantifying the Difference of Residue Composition for the Three Types of Protein Interactions

We take two ways to quantify the residue composition difference between different interaction types. One is a Euclidean distance Δf [31, 65] as described by Equation 5.1 in Section 5.2 to measure the difference of residue percent composition in the three types of protein interactions; the other is a correlation coefficient CC [10] as described by Equation 5.2 in Section 5.2 mainly to compare different residue ratio composition. The comparison result is presented in Table 5.12.

Table 5.12: The difference of residue composition in the three types of protein interactions.

Interaction Types	Obligate	Non-obligate	Crystal Packing
Obligate		<i>0.5219</i>	<i>0.4640</i>
Non-obligate	2.56%		<i>0.0561</i>
Crystal Packing	2.87%	2.06%	
Background	3.70% (<i>0.78</i>)	2.922% (<i>0.70</i>)	1.52% (<i>0.88</i>)

The *italic* numbers are for correlation coefficient, and others for Euclidean distance.

It is not surprising to see that the residue composition of all the three types is highly correlated to the background residue composition in the Swiss-Prot database with $CC > 0.7$ [10]; however, biological interactions have larger Euclidean distance from the background residue composition with $\Delta f=3.7\%$ for the obligate interactions and $\Delta f=2.922\%$ for the non-obligate interactions. The Euclidean distance of residue percent composition in the three types of interactions is also large. For example, this Euclidean distance between the obligate and non-obligate interactions is 2.56%, while that between the cores of protein-protein complexes and homodimers is 2.0% [31]. More information about Euclidean distance is shown in Table 5.12.

We can understand from Table 5.12 that frequency ratio of residue composition in the three types of protein interactions has very low correlation coefficient, especially between non-obligate interactions and crystal packing with $CC = 0.0561$. The exceptionally larger correlation coefficient but larger Euclidean distance ($\Delta f=2.87\%$) between crystal packing and obligate interactions is partly, if not mainly, due to that most of crystal packing in

the analyzed dataset are based on identical chains while most of obligate interactions are homodimers (homodimers are biological binding between identical chains also).

5.3.3.2 Comparison on Residue Composition between the O-ring and DWE Hypothesis

Our analysis result on the residue composition of our DWE bicliques is in agreement with the influential study by Bogan and Thorn [36] who investigated the binding hot spots of protein interfaces under the O-ring hypothesis. Bogan and Thorn [36] found that hot spots are abundant with Trp, Tyr and Arg. We also found that these three residues are actually ipRs for both obligate and non-obligate interactions.

Similar to Bogan and Thorn’s method, Janin and her colleagues had a study for identifying the core and rim from a protein interface. They found that aromatic residues have high propensity values in the core of protein-protein recognition sites [65]. For homodimeric proteins, their another work [31] pointed out that aliphatic and aromatic residues are very rich in the binding hot spots. All these results are consistent with ours as shown in Figure 5.3 and discussed in Section 5.3.3.

Therefore, we can see that when protein binding hot spots are refined from the O-ring theory to the double water exclusion hypothesis, the composition properties are inherited and some properties are more enlightened and sharpened.

5.3.4 Residue Pairing Preference in the DWE Bicliques for the Three Types of Protein Interfaces

A DWE biclique can contain many residue pairs. We are interested in those residue pairs that dominate, with high frequency, the binding hot spots of a type of protein interactions. We term this kind of residue pairs as interaction-dominated residue pairs (or idRPs for short). Meanwhile, we also examine interaction-preferred residue pairs (ipRPs). (See exact definitions for idRPs and ipRPs in Section 5.2.)

The composition information of all possible 210 residue pairs in our DWE bicliques is displayed in Figure 5.4. It can be seen that the obligate interactions are dominated by the contacts of hydrophobic residues. Taking the idRP group of I-V-L-F as example, the total frequencies of ten idRP contacts within this group are 18.22%, 10.74% and 11.74% for the obligate interactions, non-obligate interactions and crystal packing, respectively. The contact frequency of this most hydrophobic group in the obligate interactions is much higher than those in the other two types of interactions.

The ipRPs sharpen the difference of residue pairs in the three types of interactions. In the obligate interactions, most ipRPs are from four groups: (i) the contacts of identical residues, especially the residues with hydrophobicity not less than Tyr (Y)—this

Figure 5.4: Residue pair composition in DWE bicliques for crystal packing, non-obligate interactions and obligate interactions. (a), (b) and (c) show the frequency matrixes of residue pairs in these three types of interactions; (d), (e) and (f) are the matrixes of natural logarithm of frequency ratios for residue pairs in these three types of interactions. These figures are symmetric matrixes of residue pairs where rows and columns represent different amino acids, and the residues are ordered according to their hydrophobicity with I as the most hydrophobic and R as the least hydrophobic. In these figures, the colors from blue to red mean the values from smallest to largest, and the similar colors mean the similar value in the second row.

observation agrees with the discussion in [21] where identical residues more likely contact themselves in obligate interactions; (ii) the interacting pairs between aliphatic residues, Ile (I), Val (V) and Leu (L)—all these residues are most hydrophobic; (iii) the contacts between aromatic residues, Tyr (Y), Trp (W) and Phe (F); and (iv) the contact pairs between aromatic residues and Arg(A)/aliphatic residues. Aromatic residues are much involved in ipRPs due to that they easily form π - π /cation- π contacts which are vital to the stability of biological interactions. Besides the above ipRPs, another three ipRPs of obligate interactions are residue pairs Met(M)-Leu(L), Ala(A)-Leu and Met-Phe(F).

All these ipRPs shape an interesting distribution as outlined in Figure 5.4(f). Most of the ipRPs are located at the top-left corner of Figure 5.4(f), an area on the top of and at the left of Tyr (Y) included. We call this area ipRP area. There are fewer ipRPs outside this area, while those ipRP exceptions outside the ipRP area are the identical contacts of Asn (N) and Arg (R), and the interactions between Arg (R) and aromatic residues. That is, when taking the aromatic residue Tyr (Y) as the dividing line for the columns and also for the rows in Figure 5.4(f), ipRPs are depleted in the top-right corner (and also the bottom-left corner due to the symmetry of Figure 5.4(f)). The bottom-right corner also has rare ipRPs where the least hydrophilic residues are solely involved in the contacts. Such an ipRP distribution suggests that it is the very hydrophobic contacts that much prefer to the obligate interactions.

Similarly, in the non-obligate interactions, there are also very fewer ipRPs in the bottom-right corner of Figure 5.4(e) except the contact between Arg (R) and Asp (D). In the non-obligate interactions, the ipRPs are mainly from the contacts involving Trp (W), Tyr (Y), Phe (F), Cys (C), His (H) and Leu (L), specially the contacts involving W and Y. Of these contact residues, three are aromatic residues (Y, W and F), and two are nonpolar (hydrophilic C and hydrophobic L). Cys-involved ipRPs are expected due to that Cys contains a sulfate atom and can form disulfate bridges to stabilize the protein interactions. The reason why H is also involved in the ipRPs may be that H is sometimes categorized into aromatic residues [83] and likely possesses some properties of aromatic residues in certain environments. In conclusion, residue pairs involved by hydrophobic and aromatic residues are abundant in the two types of biological interactions, indicating the importance of these ipRPs in specific binding behaviors.

In contrast to biological interactions, the ipRP distribution for the crystal packing contacts is completely opposite. Crystal packing contacts have more ipRPs of hydrophilic contacts (at the bottom-right corner of Figure 5.4(d)) and fewer ipRPs at the top-left corner. The top-right corner (and also the bottom-left corner due to the symmetry) of Figure 5.4(d) has more ipRPs than the top-left corner does.

5.4 Conclusions

With the integration of the influential O-ring theory and the insightful ‘coupling proposition’, DWE (double water exclusion) is a more comprehensive hypothesis for modeling protein binding hot spots. In this work, we constructed DWE bipartites from interacting protein chains under the constraints of both residue contacts and residue accessibility (See Section 5.2.2). Biclique patterns were then detected for each type of protein interactions. Our comparative analysis on DWE bicliques suggested that there do exist unique bicliques in the three types of interactions. Compared to crystal packing, those unique bicliques only occurring to biological interactions made it much clear that the biological binding behaviors have strong specificity (See Section 5.3.2.1). The unique bicliques in the obligate and non-obligate interactions also confirmed the different binding behaviors in these two types of biological interactions (See Section 5.3.2.2). Therefore, the idea of DWE bicliques provides a new way to the study on protein interfaces.

The composition of residues and the composition of residue pairs, in particular ipRs and ipRPs, did reveal the deep characteristics of these types of interactions (See Section 5.3.3 and 5.3.4). The protomers of obligate interactions fold and bind at the same time. Obligate interfaces need hydrophobic residues to form their interior cores, similar to the cores of protein tertiary structures in the same folding-binding process. Therefore in the obligate interactions, hydrophobic residues were greatly involved in ipRs and ipRPs, while none of the anti-ipRs of the obligate interactions was hydrophobic. Also in this process of protein folding and binding in a solvent environment, hydrophilic and polar residues prefer protein solvent surface than hydrophobic core, and the contacts involved by hydrophilic residues of obligate interactions were thus depleted in ipRPs.

On the other hand, two protomers in non-obligate interactions fold separately. They then come together to bind upon a specific molecular stimulus, and may dissociate after that. In an unbound form of non-obligate protomers, their interface surfaces have to contact with the solvent, and less hydrophobic residues are necessary [19] to keep the stability of unbound non-obligate protomers. So, the hydrophobic ipRPs in non-obligate interactions are much less than in obligate interactions. To compensate for the decrease of hydrophobic ipRPs in non-obligate interactions, aromatic residues are rich in non-obligate interfaces. Aromatic residues, such as Trp, Tyr and His sometimes, can contribute protein binding through the hydrophobic effect. Meanwhile, aromatic residues do not result in a large entropic penalty because they have few rotatable bonds [36]. This is why aromatic ipRs and their ipRPs were observed to be abundant in biological interactions, especially in non-obligate interfaces.

In contrast, hydrophilic ipRs and ipRPs were affluent in crystal packing to stabilize monomers in crystal packing without interactions. The anti-ipRs of crystal packing, such as Met, Trp and Cys, were just the ipRs of the biological interactions, indicating

they can form significant atom contacts to greatly increase the probability of biological interactions, such as π involving contacts and disulphide bridges.

In summary, these ipRs and ipRPs in DWE bicliques are excellent indicators for the specificity analysis of biological binding behaviors. They can be used to identify biological interactions from crystal packing and classify different types of biological interactions, such as obligate and non-obligate interactions [61]. The identification of obligate or non-obligate interactions can help docking algorithm to remove the noise of produced decoys.

Chapter 6

Propensity Vectors of Low-ASA Residue Pairs in the Distinction of Protein Interactions

In Chapter 3 and 5, the DWE hypothesis has proved to be helpful in interfacial analysis based on the following two observations: (i) DWE bicliques are verified to be enriched in true hot spots, and (ii) DWE bicliques capture the incorporated physicochemical properties in different types of interfaces. Meanwhile, it can be seen from literature works that different types of protein interactions possess their specific binding behaviors, and the physicochemical properties of binding interfaces are complicated. Thus, one of the fundamental problems in computational structural biology is to characterize the types of protein interactions by using their structure information at a residue or atom level.

With the above knowledge in mind, an immediate idea is whether the DWE hypothesis can help the distinction of different types of interactions. In this chapter, we employ the O-ring and DWE hypotheses to design the OringPV classification method for distinguishing crystal packing and the two different types of biological interactions.

6.1 Introduction

Close interactions in protein complexes are necessary and indispensable for proteins to fulfill molecular functions and biological processes. Protein complexes have different types with various binding behaviors. These protein complexes that are determined by the popular and prolific technique X-ray crystallography can be broadly classified into

crystal packing or biological interaction according to the biological reality of the contacts. Crystal packing interactions/contacts are false interactions in crystallographic packing environments [11]. As discussed in Chapter 1, biological interactions have been carefully studied and categorized into sub-groups according to various criteria, such as permanent versus transient complexes on the basis of the lifetime of protein complexes [8], and homo-oligomers versus hetero-oligomers according to whether protein interactions occur between identical chains or not [8]. Depending on whether the protomers of interactions can be found or not as stable structures on their own *in vivo* [8], biological interactions can also be grouped into obligate or non-obligate interactions [8]. Focusing on different transition processes in protein folding and binding [9], two-state folding complexes and three-state complexes were also used to describe obligate and non-obligate interactions. Further, Ofra and Rost’s work [10] categorized biological interactions in a more complicated way with six subtypes. The different types of biological interactions possess their unique binding behaviors. For example, obligate interactions are stable, and their protein chains function only in the complex form [8]; however, the protomers in transient interactions associate to accomplish a particular function upon a molecular stimulus and dissociate after that [8].

Therefore, deep understanding of these binding behaviors can be particularly useful for reliable predictions on the types of interactions in new protein complexes [18, 26, 44], and it is also helpful to identify protein binding sites [16, 43] and for docking algorithms to construct protein quaternary structures [25].

Outstanding chemical, physical and geometric properties [16, 17] have been extensively explored in literature to describe and characterize protein interactions and their binding interfaces, such as physical features interface area, and residue composition. Many of them have been involved in important findings. For example, biological interactions are found to be significantly different in residue composition from the rest of protein surfaces [22, 30, 83, 121], while crystal packing possess similar composition to the rest of protein surfaces [35]. The physical feature, interface area of biological interactions, is found to be much larger than that in non-biological interactions [17, 18, 19, 22, 31, 33, 34, 35].

These properties have also been integrated by many classification algorithms to discriminate biological interactions from crystal packing [22, 103], and to distinguish different types of biological interactions [18, 26, 107]. The good examples are DiMoVo and NOXclass. Recently, atom/residue pairs were conceptualized and used in the distinction of protein interactions [25, 26, 27, 28, 107], e.g., by the ACV (atom contact vector) method [26].

The common approach adopted by these classification methods is that the classification properties are all taken from interface residues whose surface accessibility change is $>0.1 \text{ \AA}^2$ [22, 31] or $>1.0 \text{ \AA}^2$ [18, 30] upon the formation of complexes. In this work,

we further narrow down the scope of interfacial residues to concentrate on those of low solvent exposure. Our new notion is called *low-ASA residue pairs*. A low-ASA residue pair is defined as two contact residues whose ASA is very small. The notion of low-ASA residue pairs is in agreement with the influential O-ring hypothesis [36, 86, 108, 87], and one of its successors, the insightful “coupling” proposition [48]. The O-ring theory states that the binding ‘hot spot’ residues are usually clustered and located at the center of the interfaces, and they are often surrounded by energetically less important residues shaped like an O-ring for occluding water molecules. The “coupling” theory [48] highlights that the hot spot residues are always coupled to each other with a short distance between the two sides of the interface. Therefore, given a low-ASA residue pair, the two residues are both buried by O-ring residues (i.e., the residues on the O-ring), and the spatial compactness between them is very tight. Thus, low-ASA residue pairs probably form a special area that is richer of hot spot residues than the other areas of the interface. As binding hot spots lie in the core of the stability for protein interactions [36], we believe that low-ASA residue pairs can sharpen the difference between different types of protein interactions. With low-ASA residue pairs, an immediate ease is to accurately identify crystal packing, because the interfaces of crystal packing contain few fully buried atoms [22].

A residue in a low-ASA residue pair may have multiple partners. Assume (A_1, B_1) is a low-ASA residue pair in an interacting chain pair, it is often the case that (A_1, B_2) , (A_1, B_3) , or even (A_1, B_4) is also a low-ASA residue pair. With this regard, we introduce O-ring-surrounded regions. Given an interacting chain pair, its O-ring-surrounded region is the union of all low-ASA residue pairs of this interacting chain pair.

We propose to construct a propensity vector to characterize the interaction behavior of an interacting chain pair by using all of its low-ASA residue pairs as features. The propensity vector of the low-ASA residue pairs consists of 213 feature elements: three elements are used for describing the summary information of its O-ring-surrounded region, and the remaining 210 elements are reserved for the propensity values of all possible low-ASA residue pairs contained in the O-ring-surrounded region against the rest of protein surfaces. Assume we are given n number of protein interactions in a classification problem, then n propensity vectors will be constructed accordingly. Each of these propensity vectors will be labeled with the types of biological interactions, or crystal packing in the training data.

Propensity vectors of our low-ASA residue pairs are related to but different from the residue-pair method proposed in [24, 25]. Firstly, our residue pairs are low-ASA residue pairs satisfying the O-ring theory and the coupling proposition, while residue pairs of [23, 24, 25] are just interface residue pairs which include residues outside the O-ring surrounded regions. Secondly, we calculate propensity values of the low-ASA residue pairs which are totally different from frequency values of residue pairs as used in [24, 25].

Propensity values of residues are more competitive to frequency values to improve classification performance as early observed by Bahadur *et al.* [22]. Finally, propensity vectors do not require the assumptions of additivity as [24, 25] required. Our propensity vector is also different from the frequency vector of atom pairs proposed in ACV [26, 107]. Compared with ACV, our vectors take into account more biologically useful properties, such as residue composition and propensity of residue pairs, for signifying the physicochemical properties of interfaces.

The discriminating power of propensity vectors of protein interactions are tested on three benchmark datasets [18, 22, 27]. In the experiments, we consider a variety of classification tasks, such as distinction between biological interactions and crystal packing, distinction within biological interactions (e.g., obligate vs non-obligate interactions) and the 3-class classification problem. The performance evaluation is also measured under a variety of frameworks, including within-dataset comparison, cross-dataset generalization capability test, and LOOCV (leave-one-out cross-validation). Our comprehensive comparison results have shown that low-ASA residue pairs and the propensity vector description of protein interactions are truly strong in the prediction of protein interaction types. In particular, many cross-dataset generalization capability tests have achieved excellent recalls and overall accuracies, much outperforming existing benchmark methods.

6.2 Methods

In this section, we give an overview to the three test datasets. Then we present a formal definition for low-ASA residue pairs and O-ring-surrounded regions, followed by a description of how to compute a propensity vector of the low-ASA residue pairs for a protein interaction. We also introduce our classification method, OringPV, and describe performance evaluation measurements.

6.2.1 Three Benchmark Datasets

The first benchmark dataset is the BNCP-CS dataset [18] which comprises 75 obligate interactions, 62 non-obligate interactions and 106 crystal packing. Zhu *et al.* [18] tested their NOXClass method on this dataset to predict the three types of protein interactions. Here, the obligate and non-obligate interactions in the BNCP-CS dataset are specially denoted by BNCP-CS^{bio}, and the set of obligate interactions and crystal packing are denoted by BNCP-CS^{hocp}.

The second dataset is the Ponstingl dataset [27]. It consists of 95 monomers and 76 homodimers. On this dataset, Ponstingl *et al.* tested their score schemes [27] for the distinction between homodimeric proteins and monomeric proteins, and Mintseris and Weng tested their ACV method [26].

The third dataset is the Bahadur dataset [22] which contains 70 heterodimeric complexes (also termed as protein-protein complexes in [83]), 122 homodimeric proteins and 188 crystal packing. The homodimeric proteins and crystal packing of this dataset were used previously in [22, 70]. In this work, the homodimeric proteins and crystal packing are specially denoted by Bahadur^{hosp}, and the heterodimeric complexes and homodimeric proteins in the Bahadur dataset are denoted by Bahadur^{bio} dataset.

6.2.2 Low-ASA Residue Pairs and Their Propensity Vector

Let C_1 and C_2 be a pair of interacting polypeptide chains, two residues $r_1 \in C_1$ and $r_2 \in C_2$ are defined as a **low-ASA residue pair**, if (i) the ASA of r_1 and of r_2 are both small, ideally with a value close to zero, namely no much contact with water solvent; (ii) the minimum of the atom distances between r_1 and r_2 is less than a threshold plus the van der Waals radii of the corresponding atoms. Here, we again set this threshold value, denoted by T_d , as a real number less than the van der Waals diameter (2.75 Å) of water molecules. The first criteria of this definition captures the idea of the influential O-ring theory, indicating that the two residues r_1 and r_2 should satisfy a proposed condition [36, 87] for them to be in a hot spot. The second criteria best follows the spirit of the coupling proposition, emphasizing the importance of a water-free distance between two contact residues. The notion of low-ASA residue pairs also more-or-less shares a light with our “double water exclusion” hypothesis in Chapter 3.

The **O-ring-surrounded region** of an interacting chain pair is the union of all low-ASA residue pairs of this interacting chain pair. Usually, such a region is covered by one O-ring, i.e., a mono-island region. However, sometimes, two subsets of low-ASA residue pairs in the O-ring-surrounded region of an interacting chain pair may not share any common residues with each other. Such a region may be covered by two or more O-rings, i.e., a multi-island region. Nevertheless, we treat the union of all low-ASA residue pairs as the O-ring-surrounded region of the chain pair. Figure 6.1 shows an example of O-ring-surrounded region which is located at the binding site between chain F and G of PDB entry 1GLA. This O-ring-surrounded region consists of 24 low-ASA residue pairs involving 14 residues at chain F and 8 residues at chain G. Of the 24 low-ASA residue pairs, some are duplicates. For example, residue pair (THR, VAL) occurs four times, residue pair (PHE, THR) occurs three times, and both residue pair (ILE, PHE) and (ILE, SER) occur twice. All the rest occur only once in this O-ring-surrounded region. Actually, the size of O-ring-surrounded regions varies greatly among protein interfaces, especially among different types of protein interfaces. This can be seen from Table 6.1 which shows the size information of O-ring-surrounded regions for the interactions in the BNCP-CS dataset. We note that a size of zero means the O-ring-surrounded region is an empty set of residue.

Figure 6.1: The O-ring-surrounded region between chain F and G of 1GLA in the ‘spheres’ view in PyMOL with the color red and magenta. The residues in green and blue are interfacial residues of large ASA which are filtered and are not used to form low-ASA residue pairs.

Figure 6.2: The ROC curves of the classifications between homodimeric proteins and monomers on the Bahadur^{hocp} dataset for frequency vectors, binary vectors and propensity vectors.

Table 6.1: The size of O-ring-surrounded regions for the BNCP-CS dataset.

Interaction Type	number of residues (and residue pairs)		
	minimum	average	maximum
Crystal packing	0 (0)	13 ± 7.4	40 (46)
Transient	12 (13)	45 ± 20.6	146 (214)
Obligate	23 (28)	83 ± 53.2	288 (393)

Given an interacting polypeptide chain pair C_1 and C_2 , computationally, the process to locate its O-ring-surrounded region is similar to the Algorithm 1 in Section 3.3.2. To help understand this process, we also present two steps to locate low-ASA residue pairs as follows:

- (1) we take the NACCESS software [12] to remove those residues that have a relative accessible surface area in the complexed form bigger than a threshold. In this work, this threshold is set as 36% following the one recommended by [62].
- (2) Let C'_1 and C'_2 be the residue sets of C_1 and C_2 , respectively, after the residue removal by step 1. Let $r_i \in C'_1$ and $r_j \in C'_2$, we calculate the distance between all possible atom pairs of r_i and r_j . If the minimal distance is less than their van der Waals radii plus T_d , then r_i and r_j is a low-ASA residue pair.

We would like to also point out that there is no gold standard about how to determine an optimal T_d . We have tried to set T_d as every value from 0.5 Å to 2.75 Å with step 0.25 Å. We found that the classification performance had only small variation across most of these situations when T_d changed. In this work, we only report the results when T_d was set as 1.5 Å. Note that under this setting ($T_d=1.5$ Å), the performance was not always the best on all of the datasets.

Construction of our Propensity Vectors: Given an interacting protein chain pair, C_1 and C_2 , we construct a propensity vector based on the low-ASA residue pairs of this interaction. This propensity vector consists of two parts. (i) At the *summary part*, there are three feature elements for describing the summary information of the O-ring-surrounded region—two numbers (each for one chain) of the contact residues from the O-ring-surrounded region, and the total number of low-ASA residue pairs; (ii) at the *propensity part*, there are 210 elements ($C_{20}^2 + 20=210$) each for describing the propensity value of one of the all possible residue pairs (r_i, r_j) of 20 standard residues, $i \leq j, i, j = 1, \dots, 20$, between the O-ring-surrounded region and the surface area of this protein chain pair. The propensity value of a residue pair (r_i, r_j) is calculated by

$$p_{(r_i, r_j)} = \log\left(\frac{f_{(r_i, r_j)}}{(f_{r_i}^1 f_{r_j}^2 + f_{r_i}^2 f_{r_j}^1)/2} + 1\right) \quad (6.1)$$

where $f_{(r_i, r_j)}$ is the frequency fraction of the residue pair (r_i, r_j) in the O-ring-surrounded region; $f_{r_k}^c$, $k=i, j$, is the frequency fraction of the residue r_k in the surface residues of the protein chain C_c ($c=1, 2$); the number 2 is used to compensate the double expected count of the pairs for surface residues. A residue is considered as a surface residue if its relative accessibility is greater than 25% [130]. Note that our $f_{r_k}^c$ is based on the protein surface residues instead of interface residues [23]. Reasons why we calculate propensity values of residue pairs by using $f_{r_k}^c$ in the protein surface residues include: (i) The interface between two proteins in crystal packing occurs by chance, and the residue composition in the interface is similar to that in the rest of protein surfaces [35]; (ii) However, at the interfaces of biological interactions, the residue composition is statistically different from protein surfaces [30, 83, 22, 121]. Thus, the interfaces of biological interactions and crystal packing have different propensities to be compared with protein surfaces.

We take an example to show how the propensity value of a feature is obtained. Residue pair (THR, VAL) is a low-ASA residue pair in the O-ring-surrounded region shown in Figure 6.1. This residue pair occurs four times in this region, so $f_{(THR, VAL)}$ is 0.167 (4/24). And then $f_{(THR)}^F=0.076$, $f_{(THR)}^G=0.0286$, $f_{(VAL)}^F=0.038$ and $f_{(VAL)}^G=0.0514$. Therefore, the propensity value of this residue pair, $p_{(THR, VAL)} = \log\left(\frac{0.167}{(0.076*0.0514+0.038*0.0286)/2} + 1\right) = 4.22$.

Construction of binary vectors and frequency vectors: In parallel, a binary vector of low-ASA residue pairs and a frequency vector of low-ASA residue pairs are also proposed in comparison to the above propensity vector. The binary vector and the frequency vector both have the same 213 feature elements as the propensity vector does. The only difference is on the feature values. The values of the 210 feature elements (i.e., the second part of the vector) of a binary vector indicate whether the residue pairs occur in the O-ring-surrounded region or not; similarly, the 210 feature values of the frequency vector are the frequency values of the residue pairs in the O-ring-surrounded region. These binary vectors and frequency vectors are called low-ASA binary vectors and low-ASA frequency vectors respectively in this work. They are different from those defined in [23, 24, 25] which determine the feature values by using the entire interface residues (without considering the water accessibility restriction). However, we use residues in O-ring-surrounded region which are the energetically most important subset of residues in the binding. In this work, the binary vectors and the frequency vectors based on the entire interfaces (without considering the water accessibility restriction) are termed *traditional* binary vectors and *traditional* frequency vectors respectively.

6.2.3 OringPV: Our Classification Method

Given an interaction classification task, we first construct a propensity vector for every interaction in the training data and also in the test data. Then, we take these training propensity vectors as input and feed to Support Vector Machine (SVM) [131] to

build a classifier. As low-ASA residue pairs are heavily involved in the construction of the propensity vectors, we name this learning process OringPV (short for learning by Propensity Vectors of low-ASA residue pairs in O-ring-surrounded regions).

In this work, all classification tasks are performed by running the *LibSVM* software package [127] which contains an implementation of the SVM learning method. A Radial Basis (RBF) kernel function was chosen in the training. To determine optimized C and γ for the RBF kernel functions, a grid search heuristics [132] was imposed on training data with 10-fold cross-validation.

For the test datasets above that have three types of protein interactions, such as the BNCP-CS dataset and the Bahadur dataset, our OringPV method employs a two-stage SVM: the first-stage SVM is used to discriminate crystal packing and biological interactions, and the second-stage SVM is subsequently to differentiate different types of biological interactions.

6.2.4 Performance Measures

To quantify sensitivity performance of classification methods, we denote biological interactions as positive set and crystal packing as negative set when identifying biological interactions from non-biological interactions, and one type of biological interactions (e.g., obligate interactions in the BNCP-CS dataset) as positive set and the other type of biological interactions (e.g., non-obligate interactions in the BNCP-CS dataset) as negative set in the process of distinguishing two types of biological interactions. Then, as shown in Equation 4.10, classification performance is evaluated also based on *sensitivity* (the fraction of correctly predicted positive interactions over all positive interactions), *precision* (the percentage of correctly predicted positive interactions over all predicted interactions), *specificity* (the fraction of correctly predicted negative interactions over all negative interactions), *accuracy* (the number of correctly predicted positive and negative interactions divided by the number of all interactions), as well as Receiver Operating Characteristics (ROC) curves and their Area Under the ROC curves (AUC).

6.3 Classification Results

Our experiments are conducted under the following four aspects of considerations:

- Leave-one-out cross-validation (LOOCV) within the datasets for showing the outstanding capability of our OringPV method to distinguish different types of protein interactions.
- The comparison of the propensity vectors with low-ASA binary vectors, low-ASA frequency vectors, and the vectors without considering the water accessibility restriction to show the subtle and deep discriminating power of the propensity vectors.

Table 6.2: The overall performance of our OringPV method under LOOCV procedure. The numbers in parentheses are for classification performance by the traditional frequency vectors.

Dataset and Interaction Types		Sensitivity (%)	Specificity (%)	Precision (%)	Accuracy(%) / (AUC)		
					SVM Stage 1	SVM Stage 2	SVM Overall
BNCP-CS	OB	90.7(88.2)	97.0(92.9)	93.2(84.8)	96.7(95.1)	87.9(82.9)	92.2(88.1)
	NO	90.3(79.0)	93.4(93.4)	82.4(80.3)			
	CP	94.3(93.4)	98.5(96.4)	98.0(95.2)			
Bahadur	HO	83.6(81.1)	96.9(88.0)	92.7(76.2)	87.9(78.7)	87.4(73.8)	85.8(75.5)
	NO	71.4(51.4)	95.5(94.5)	78.1(67.9)			
	CP	92.6(80.9)	83.3(76.6)	84.5(77.2)			
Ponstingl	HO	88.2(81.6)	-	93.1(88.6)	-	0.915(0.866)	91.8(87.1)
	CP	94.7(91.6)	-	90.9(86.1)			
Bahadur ^{hosp}	HO	85.2(80.3)	-	96.3(84.5)	-	0.916(0.854)	92.9(86.5)
	CP	97.9(90.4)	-	91.1(87.6)			
BNCP-CS ^{bio}	OB	89.3(86.7)	-	94.4(86.7)	-	0.914(0.853)	91.2(85.4)
	NO	93.5(83.9)	-	87.9(83.9)			
Bahadur ^{bio}	HO	95.9(90.2)	-	96.7(92.4)	-	0.951(0.887)	95.3(89.1)
	NO	94.3(87.1)	-	93.0(83.6)			

‘OB’, ‘NO’, ‘CP’ and ‘HO’ represent obligate interactions, non-obligate interactions (heterodimeric complexes), crystal packing and homodimeric interactions, respectively; ‘-’ means the values are not applicable. The **bold** numbers are AUC values (See Section 4.3.4 for the calculation formula.).

- Within-dataset comparison with benchmark classification methods such as the NOXclass method [18], the DiMoVo method [70], and other methods [22, 27, 26]. The evaluation frameworks for each dataset strictly follow those set by these literature methods.
- Cross-dataset test for comparison between our OringPV method and the literature methods. We use one dataset for training and other datasets for performance testing. It is similar to the so called independent or blind data testing scheme. This is a more reliable approach to testing a classifier’s generalization capability.

6.3.1 LOOCV Performance by OringPV within Datasets

The OringPV’s sensitivity, precision, specificity and accuracy are presented in Table 6.2, while the corresponding confusion matrix results are reported in Table 6.3 and Table 6.4.

The results in Table 6.2 together with some results from Table 6.3 demonstrate a strong capability of distinguishing biological interactions and crystal packing by our OringPV method. For example, it can effectively identify homodimers from monomers with an accuracy of 91.8% on the Ponstingl dataset and 92.9% on the Bahadur^{hosp} dataset. For differentiating biological interactions from crystal packing, the accuracy of OringPV

Table 6.3: OringPV confusion matrix under LOOCV procedure for prediction of different types of protein interactions.

Dataset	Prediction				Total
		OB	NO	CP	
BNCP-CS	OB	68(67)	7(8)	0(1)	75
	NO	4(9)	56(49)	2(4)	62
	CP	1(3)	5(4)	100(99)	106
	Total	73(79)	68(61)	102(104)	243
Bahadur	OB	102(99)	5(4)	15(19)	122
	NO	3(8)	50(36)	17(26)	70
	CP	5(23)	9(13)	174(152)	188
	Total	110(130)	64(53)	206(197)	380
Ponstingl		HO		CP	Total
	HO	67(62)		9(14)	76
	CP	5(8)		90(87)	95
	Total	72(70)		99(101)	171
Bahadur ^{hosp}	HO	104(98)		18(24)	122
	CP	4(18)		184(170)	188
	Total	108(116)		202(194)	310

‘OB’, ‘NO’, ‘CP’ and ‘HO’ represent obligate interactions, non-obligate interactions, crystal packing and homodimeric proteins, respectively. The numbers in parentheses are for classification performance of the traditional frequency vectors.

Table 6.4: OringPV confusion matrix under LOOCV procedure for prediction of the two types of biological interactions.

Dataset	Prediction			Total
		OB	NO	
BNCP-CS ^{bio}	OB	67(65)	8(10)	75
	NO	4(10)	58(52)	62
	Total	71(75)	66(62)	137
Bahadur ^{bio}	OB	117(110)	5(12)	122
	NO	4(9)	66(61)	70
	Total	121(119)	71(73)	192

‘OB’, ‘NO’ and the numbers in parentheses have the same meanings as those in Table 6.3.

can reach a level as high as 96.7% on the BNCP-CS dataset and 87.9% on the Bahadur dataset. Therefore, we can conjecture that low-ASA residue pairs and the propensity

vectors can capture the signature patterns of biological interactions, and our OringPV method can identify them from crystal packing with high accuracy.

Outstanding classification performance by OringPV to distinguish between two types of biological interactions is also shown in Table 6.2, and the corresponding confusion matrix results are reported in Table 6.4. The point we want to make here is: when an interaction is confirmed as biological interaction, OringPV can exactly tell which type of biological interaction it belongs to with high accuracy. For example, only 12 interactions were misclassified out of 137 interactions in the BNCP-CS^{bio} dataset, and only 9 misclassified in the 192 interactions of the Bahadur^{bio} dataset, which means 91.2% and 95.3% accuracy respectively.

We calculated the ROC curves and their AUC values of OringPV in every above individual classification experiment with the optimal C and γ . Again, it is confirmed that OringPV can well characterize the distinction among different protein interactions. For example, in Figure 6.2, the AUC values of classification performance for low-ASA frequency vectors, low-ASA binary vectors and our propensity vectors on the Bahadur^{hosp} dataset are as high as 0.9571, 0.9573 and 0.9789 respectively.

6.3.2 Propensity Vector in Comparison to Propensity Vector Without the ASA Restriction, and in Comparison to Traditional Frequency Vector

Recall that a low-ASA residue pair is a contact residue pair that satisfies the ASA restriction, i.e. its ASA is required to be small. This ASA restriction is sometimes referred to as residue filtering. In most cases, the classification performance of OringPV can be improved a lot if we take low-ASA residue pairs instead of using the residue pairs that do not apply the ASA filtering, as shown in column 2 and 5 in Table 6.5. In particular, on the Bahadur^{hosp} dataset, the accuracy of OringPV is improved from 86.8% to 92.9% when the ASA filtering is applied to concentrate on low-ASA residue pairs. It seems that the ASA filtering is very sensitive to the performance improvement for identifying biological interactions from crystal packing. However, for distinguishing the two types of biological interactions, the filtering cannot improve the performance as it does when classifying biological and non-biological interactions. This difference is probably attributed to the degree of the ASA filtering. The ASA filtering removes about 17.9% residue pairs from the contact residue pairs of the biological interactions, and 31.8% residue pairs from crystal packing. In other words, the binding interfaces of biological interactions much more satisfy the requirement of O-ring-surrounded regions than non-biological interactions. Thus, the ASA filtering can achieve more improvement for identifying biological interactions from crystal packing and has lesser impact on classification of two types of biological interactions.

Table 6.5: The accuracy performance comparison among propensity vector (PV), binary vector (BV) and frequency vector (FV) with/without ASA restriction.

Dataset/SVM-stage		PV ^{OR}	BV ^{OR}	FV ^{OR}	PV ^{NonOR}	BV ^{NonOR}	FV ^{NonOR}
Ponstingl		91.8	86.5	83.6	89.5	88.3	87.1
Bahadur	SVM 1	87.9	82.6	82.6	85.2	81.6	78.7
	SVM 2	87.4	81.7	79.0	84.5	82.0	73.7
	overall SVM	85.8	81.1	79.7	83.2	81.0	75.5
Bahadur ^{hosp}		92.9	87.4	88.4	86.8	85.8	86.5
Bahadur ^{bio}		95.3	94.8	89.5	96.8	95.3	89.0
BNCP-CS		92.2	89.7	90.9	93.8	89.3	88.1

‘SVM 1’, ‘SVM 2’ and ‘overall SVM’ represent the first stage, the second stage of a SVM classifier and the whole classifier. ^{OR/NonOR} means that the vectors are with/without ASA restriction.

The numbers in the parentheses of Table 6.2, Table 6.3 and Table 6.4 are the classification performance achieved by the traditional frequency vectors on the three benchmark datasets. (See the definition of the traditional frequency vectors at Section 6.2.2.) This performance comparison suggests that the propensity vectors can much outperform the traditional frequency vectors and improve the classification accuracy by 3% to 10% in most cases. For example, on the Bahadur dataset, the accuracy is improved from 75.5% to 85.8% by the propensity vectors.

6.3.3 Propensity Vector in Comparison to Low-ASA Binary Vector and Low-ASA Frequency Vector

The performance of the propensity vectors is also compared to those achieved by low-ASA binary vectors and low-ASA frequency vectors, as shown in column 3 vs column 6, and in column 4 vs column 7 in Table 6.5. First of all, the performance between the binary and frequency vector is very similar to each other, though in most cases the binary vector can achieve a bit better performance than the frequency vector. In fact, the accuracy difference between these two kinds of vectors is in the range from -2% to 6%. Take the performance on the Bahadur^{bio} dataset as an example, the binary vector method has an accuracy of 94.8% and outperforms the frequency vectors (89.5%). But on the Bahadur^{hosp} dataset, the frequency vectors have an 88.4% accuracy, slightly better than the accuracy 87.4% of the binary vector method.

OringPV outperforms both the binary vector and the frequency vector for almost every classification task involved in this work. For example, on the Bahadur dataset, our OringPV method achieved an accuracy of 85.8%, much higher than 81.1% by the binary vectors and 79.7% by the frequency vectors. The superior performance by OringPV is possibly attributed to two biological observations: (i) the residue composition

in biological interaction interfaces is different from that in the rest of protein surfaces, but those in non-biological interactions are similar to that in the rest of protein surfaces [22, 30, 83, 121]; (ii) the residue-residue pairing preference in obligate and transient interactions was also found to be different [21]. Thus, it is only propensity vectors rather than binary vectors or frequency vectors that can translate the ideas behind these facts into sharp discrimination power of a classifier, while binary vectors and frequency vectors are concentrated on the binding hot spots (O-ring-surrounded regions) only and unable to capture the relative difference between the binding interfaces and the rest of protein surfaces.

For a visual display of the subtle and deep discriminating power provided by OringPV, we draw a picture according to the average feature value of every residue pair within different interaction types. Let (r_i, r_j) be a residue pair and N be the total number of interactions in a class, for example, in the homodimeric class, then this average is calculated by $V_{(r_i, r_j)} = \frac{\sum_{k=1}^N V_{(r_i, r_j)}^k}{N}$, where $V_{(r_i, r_j)}^k$ is the propensity value, or binary value, or frequency value of the residue pair (r_i, r_j) of the k th interaction, depending on what kind of vectors is used in the classification. Similarly, we calculate such averages for the heterodimeric class, and for the crystal packing class. The visualization for the Bahadur dataset is shown in Figures 6.3 where in each figure, each row and column represents a different residue, and the residues are ordered according to their hydrophobicity with I as the most hydrophobic and R as the least hydrophobic; the colors from blue to red indicate the magnitude of values from the smallest to the largest. We would like to note that the figures of propensity vectors and of frequency vectors are similar to those in Figure 5.4 in Chapter 5; but the figures in Figure 5.4 are statistical analysis based on filtered DWE bicliques, while these figures here are based on the whole regions inside O-rings.

We also employ Wilcoxon signed rank test [133] to calculate the statistical significance of the difference for each pairs of vectors with the same kind. To show the process of the signed rank test, let's take for example propensity vectors of homodimeric interactions and crystal packing. Assume their propensity vector representations are $P^1 = \{p_{(r_i, r_j)}^1\}$ and $P^2 = \{p_{(r_i, r_j)}^2\}$ respectively where (r_i, r_j) , $i \leq j$, $i, j = 1, \dots, 20$, are residue pairs of residue r_i and residue r_j , and $p_{(r_i, r_j)}$ is propensity value. Then the significance level of their difference, p -value, is produced by (i) calculating the value difference of each feature $p_{(r_i, r_j)}^1$ and $p_{(r_i, r_j)}^2$, and ranking every feature based on the absolute value of the difference; (ii) restoring the signs of the differences to the ranks for obtaining the signed ranks; (iii) summing these ranks with positive signs to W_+ ; (iv) calculating the normalized test statistics by $T_{value} = \frac{W_+ - E(W_+)}{\sqrt{Var(W_+)}}$, where $E(W_+) = \frac{n(n+1)}{4}$, $Var(W_+) = \frac{n(n+1)(2n+1)}{24}$, and n is the number of features with nonzero difference values; (v) obtaining p -value according to t-distribution with degree of freedom n , and value T_{value} ; theoretically, the smaller

Table 6.6: The p -values of Wilcoxon signed rank test for three pairs of the three types of interactions based on propensity vectors, binary vectors or frequency vectors.

Vectors	p -values		
	CP-PX	CP-HO	HO-PX
Propensity Vectors	1.11E-16	<1E-324	<1E-324
Binary Vectors	<1E-324	<1E-324	4.61E-5
Frequency vectors	2.0E-15	<1E-324	<1E-324

‘CP’, ‘PX’ and ‘HO’ represent crystal packing, heterodimeric complexes and homodimeric proteins respectively in the Bahadur dataset. Each column is one pair of the three different types of interactions.

a p -value is, the more likely a pair of vectors have significant difference. The similar way above can also be used to calculate p -value for the other two pairs of propensity vectors, and for pairs of binary vectors or of frequency vectors. Finally, the p -values for three pairs of three types of interactions based on propensity vectors, binary vectors and frequency vectors are shown in Table 6.6 where 3 p -values in each row are for the pair of crystal packing and heterodimeric complexes, the pair of crystal packing and homodimeric proteins, and the pair of heterodimeric complexes and homodimeric proteins, respectively.

It can be noted from Figure 6.3 and Table 6.6 that: (i) the propensity vectors make clear distinction between the three types of protein interactions, and the smallest difference among three pairs of propensity vectors has p -value 1.11E-16 (see the three subfigures at the first row of Figure 6.3); (ii) the binary vectors show the difference between biological and non-biological interactions, but they sometimes confuse homodimeric proteins with heterodimeric complexes with p -value 4.61E-5 (see the middle three subfigures); (iii) the frequency vectors can clearly differentiate homodimeric proteins from others, but they maybe mislead the discrimination between crystal packing and heterodimeric complexes with p -value 2.0E-15 (see the three subfigures at the third row).

Overall in summary, the vectors with the ASA filtering, especially our OringPV method, outperform the vectors without the ASA restriction in most cases. In fact, both the pair propensity idea and the ASA filtering can sharpen the difference for the distinction of different types of interactions.

6.3.4 Performance Comparison of OringPV with NOXclass, DiMoVo, and Other Methods

We have taken two comparison approaches: one is within-dataset comparison, and the second is cross-dataset comparison. In these experiments, the thresholds for relative ASA

6.3.a Propensity vectors in the crystal packing class

6.3.b Propensity vectors in heterodimeric class

6.3.c Propensity vectors in homodimeric class

6.3.d Binary vectors in crystal packing class

6.3.e Binary vectors in heterodimeric class

6.3.f Binary vectors in homodimeric class

6.3.g Frequency vectors in crystal packing class

6.3.h Frequency vectors in heterodimeric class

6.3.i Frequency vectors in homodimeric class

Figure 6.3: A visual display of the discriminative power carried by the three types of vectors (propensity vector, low-ASA binary vector, and low-ASA frequency vector) for the distinction of protein interactions in the Bahadur dataset. Each cell indicates the average feature value of different types of vectors in different classes.

and for atomic distance (T_d) are always fixed at 36% and 1.5 Å respectively. Parameters C and γ in the RBF kernel function of SVM are optimized in the training process of OringPV by cross-validation as what exactly done by the existing methods.

6.3.4.1 Within-Dataset Comparison with NOXclass, DiMoVo, and Other Methods

NOXclass [18] is a highly accurate algorithm trained on the BNCP-CS dataset to differentiate obligate from non-obligate interactions, and it also identifies crystal packing interactions. For a fair comparison between the performance of OringPV and NOXclass on the BNCP-CS dataset, we take the two evaluation frameworks (EF) originally set by the NOXclass method in [18].

Under the first evaluation framework (EF1), the whole BNCP-CS dataset was used in choosing optimal parameters of SVM; then the performance of the classifier is measured also on the same data. (This approach is called maximized training dataset method [18].) The main objective of this kind of learning is to see whether a classifier can have an optimistic learning on all existing data despite a possible overfitting problem. The result under this evaluation framework is shown in Table 6.7. Overall in the distinction of the three types of protein interactions, OringPV achieved an accuracy of 94.7%, higher than NOXclass' 91.8%. This accuracy improvement is attributed to the outstanding performance of OringPV for distinguishing the two types of biological interactions (91.5% versus NOXclass' 86.4%).

The second evaluation framework (EF2) takes into consideration of testing performance. It was set as a 5-time 3-fold cross-validation integrated by a 10-fold cross-validation for training parameter selection [18]. Under this framework, the BNCP-CS dataset is randomly divided into three parts: iteratively each of the three parts for testing, and the other two parts for training in which the selection of parameter values is optimized by 10-fold cross-validation. Then this procedure is repeated five times to get an average performance of the classifier. The result is shown in Table 6.7. It can be seen that our OringPV method has improved NOXclass' performance significantly from 83.1% to 91.4%. Again, this significant improvement comes from the sharper distinction capability of OringPV for distinguishing the two types of biological interactions.

On the Ponstingl dataset, a score cut-off method [27] misclassified 12 interactions, achieving an accuracy of 93.0%. The ACV method [26] (without symmetric consideration) also achieved the same level 93.0% accuracy. Our OringPV method achieved an accuracy of 94.7% (under the same evaluation framework). This is a slightly better performance than the two existing benchmark methods. Furthermore, the generalized error rate of 200 bootstrap samples for the atom-pair scoring schemes is 12.5% [27], which is worse than our OringPV method's 8.2% error rate under the LOOCV procedure.

Table 6.7: The within-dataset performance comparison: OringPV vs NOXclass.

Evaluation Framework	Method	Accuracy(%)		
		SVM 1	SVM 2	overall SVM
EF1	NOXclass	97.9	86.4	91.8
	OringPV	97.5	91.5	94.7
EF2	NOXclass	94.5	75.2	83.1
	OringPV	97.0	86.0	91.4

‘SVM 1’ and ‘SVM 2’ represent the first stage and the second stage of a SVM classifier. The performance of NOXclass is taken from [18].

On the Bahadur^{hosp} dataset, Bahadur *et al.* tested the performance of a new score cut-off method [22], which is a method based on interface area, shape and atomic packing density, residue propensity, etc. This score cut-off method [22] misclassified 17 of the total 310 interactions, while our OringPV method misclassified 18 interactions. More recently, Bernauer *et al.* [70] proposed a new method called DiMoVo. DiMoVo is a binary classifier developed specifically to discriminate crystal packing and biological interactions. It was evaluated on the Bahadur^{hosp} dataset and achieved a 95% accuracy under the LOOCV procedure with 0.5 as the cut-off score. Our OringPV method achieved a comparable performance (an accuracy of 94.2%) with optimal parameters.

With all these within-dataset comparison results, we can note that OringPV is much more accurate than NOXclass, and it is comparable to the score cut-off methods and DiMoVo. Our superior performance over these score cut-off methods and DiMoVo is presented in the following subsection.

6.3.4.2 Cross-Dataset Test for Performance Comparison

As introduced, cross-dataset test refers to an evaluation framework for a classifier where two datasets (usually from different authors) are given: one is used for training the classifier, and the other is for testing. This is a less-biased assessment to demonstrate the high reliability and generalization capability of a classifier. In this subsection, we compare the performance of OringPV under this evaluation framework with NOXclass and DiMoVo. The reason for choosing only NOXclass and DiMoVo is because their executable codes are available from the authors. However, both NOXclass and DiMoVo are final, user-end software programs built-in with the whole BNCP-CS and Bahadur^{hosp} dataset respectively in the training. To our best knowledge, we are unable to train them again on a different dataset, but they can be used to get a test accuracy for new datasets.

For a fair comparison to DiMoVo, we trained OringPV on the whole Bahadur^{hosp} dataset as well. Table 6.8 shows the test performance of both OringPV and DiMoVo on

Table 6.8: The cross-dataset performance comparison: OringPV vs DiMoVo.

Tested Dataset	Method	Recall CP	Recall HO	Accuracy(%)	AUC
BNCP-CS ^{hosp}	DiMoVo	98.1	76.0	89.0	0.871
	OringPV	99.1	94.7	97.2	0.969
Ponstingl	DiMoVo	97.9	92.1	95.3	0.95
	OringPV	96.8	96.1	96.5	0.965

‘Recall CP’ and ‘Recall HO’ represent the recall of crystal packing and of homodimeric interactions respectively. The performance of DiMoVo is obtained from the website “<http://cgal.inria.fr/DiMoVo/>”.

the Ponstingl dataset and on the BNCP-CS^{hosp} dataset for distinguishing between obligate (homodimeric) interactions and crystal packing (monomers). We can see that on the Ponstingl dataset, our OringPV method and the DiMoVo method achieved a comparable accuracy and recall. However, on the BNCP-CS^{hosp} dataset, our OringPV method had a much better overall performance than the DiMoVo method (97.2% versus 89.0%), and especially on the homodimeric interactions, a significant recall improvement is from DiMoVo’s 76% to 94.7%. This poor performance by DiMoVo is no surprise, and it is in agreement with the limitations of DiMoVo as discussed by the authors previously [70].

NOXclass is capable of conducting two kinds of classification tasks: (i) binary distinction between obligate and non-obligate interactions, and (ii) binary distinction between crystal packing (monomers) and biological interactions. For task (i), we used the whole BNCP-CS^{bio} dataset as OringPV’s training data, and the Bahadur^{bio} dataset was used as the independent testing dataset to assess the prediction performance of both NOXclass and OringPV. Based on the result shown in Table 6.9, we can see that the performance of OringPV is tremendously better than NOXclass. In particular, the recall rate on the non-obligate interactions is 20 points higher, and the accuracy is 13 points higher. Such an excellent performance is almost maintained by OringPV when it was trained on the Bahadur^{bio} dataset and tested on the BNCP-CS^{bio} dataset (shown in third row in Table 6.9). We were unable to report NOXclass’ performance for this case, as its trained model is fixed on the BNCP-CS dataset only.

For task (ii), we trained OringPV on the BNCP-CS dataset and tested on the Ponstingl dataset. The accuracies by NOXclass and by OringPV are 86% and 87.7% respectively. However, when these two classifiers were used to predict whether the interactions in the Bahadur dataset are or not crystal packing, both performances were not good [18]. It may be due to that the non-biological interactions in the Bahadur dataset have large interfaces [18] similar to the size of biological interactions. This causes difficulties for the classifiers to learn necessary information to clearly classify biological and non-biological

Table 6.9: The cross-dataset performance comparison between OringPV and NOXclass to distinguish the two types of biological interactions.

Tested Dataset	Method	Recall OB	Recall NO	Accuracy(%)	AUC
Bahadur ^{bio}	NOXclass	84.4	78.6	82.3	0.815
	OringPV	93.4	98.6	95.3	0.96
BNCP-CS ^{bio}	OringPV	97.3	79.0	89.1	0.882

‘Recall OB’ and ‘Recall NO’ represent the recall of obligate (homodimeric) interactions and non-obligate interactions (heterodimeric complexes) respectively. The performance of NOXclass is taken from [18].

interactions when they are trained on the BNCP-CS dataset whose crystal packing generally have smaller interfaces than biological interactions.

We have further examined the capability of distinguishing homodimers and monomers by OringPV. We merged the two datasets of Bahadur^{hocp} and BNCP-CS^{hocp} as training data for OringPV. Then OringPV was tested on the Ponstingl dataset. The accuracy reached to 98.8%, an almost perfect accuracy (namely, only 1 homodimer and 1 monomer misclassified).

We note that the interpretation of the cross-dataset test results should be taken with some caution. The concern is the inter-dataset redundancy. Actually in this work, the redundancy among these three datasets does not play an important impact in the performance evaluation. This point can be verified as follows. Among the datasets used here, there is no any redundancy within a dataset and there is some redundancy between two datasets; if the inter-dataset redundancy can improve classification performance, OringPV should always have lower LOOCV performance in a dataset than the cross-dataset performance when it is trained on this dataset. In fact, (i) OringPV’s LOOCV accuracy (91.8%) on the Ponstingl dataset is less than the cross-dataset accuracy (97.7%) on the Ponstingl dataset when OringPV is trained on the Bahadur^{hocp} dataset. This may speculate some redundancy concern over the Ponstingl dataset and the Bahadur^{hocp} dataset. (ii) However, OringPV’s LOOCV accuracy (92.9%) on the Bahadur^{hocp} dataset is higher than the cross-dataset accuracy on the Bahadur^{hocp} dataset when OringPV is trained on the Ponstingl dataset or on the BNCP-CS^{hocp} dataset. (The corresponding accuracy is 90% or 85.8%.) OringPV has the highest LOOCV accuracy.

To conclude Section 6.3.4, we use Table 6.10 to summarize the various and critical comparison results.

Table 6.10: The summary of OringPV in comparison to NOXclass and DiMoVo.

Tested Dataset (Training on BNCP-CS)	OringPV vs NOXclass (%)		OringPV vs DiMoVo (%)		Tested Dataset (Training on Bahadur ^{hosp})
	OringPV	NOXclass	OringPV	DiMoVo	
BNCP-CS	<i>94.7^a (91.4^b)</i>	<i>91.8^a (83.1^b)</i>	97.2	89.0	BNCP-CS ^{hosp}
Bahadur ^{bio}	95.3	82.3	<i>94.2</i>	<i>95.0</i>	Bahadur ^{hosp}
Ponstingl	87.7	86	96.5	95.3	Ponstingl

The *italic numbers* are for within-dataset comparison and the others for cross-dataset comparison. ^a and ^b stand for EF1 and EF2 respectively.

6.4 Insights into Low-ASA Residue Pairs and Propensity Vectors: A Discussion Based on Misclassifications

The comprehensive comparison results have already shown that OringPV is a highly accurate classifier to distinguish different types of protein interactions no matter the performance evaluation is by within-dataset, cross-dataset, or LOOCV. However, the focus on this section is different: we discuss why some interactions are still misclassified and give insights into the reasons.

We start with an interesting observation on the performance change when OringPV turned back to make predictions on the BNCP-CS^{bio} dataset. Recall from Table 6.9 that OringPV achieved a testing accuracy of 95.3% on the Bahadur^{bio} dataset when it was trained on the BNCP-CS^{bio} dataset. However, its testing accuracy reduced to 89.1% on the BNCP-CS^{bio} dataset when it was trained on the Bahadur^{bio} dataset. To find out the reason behind this discrepancy, we examined the size of the O-ring-surrounded regions of the interactions in these two datasets. We obtained that on average, both the O-ring-surrounded region size of the interactions and the size variance in the BNCP-CS^{bio} dataset are much bigger than those in the Bahadur^{bio} dataset. This indicates that the diversity of the O-ring-surrounded regions in the BNCP-CS^{bio} dataset seems to overwhelm the cases in the Bahadur^{bio} dataset. Therefore, the testing performance on the BNCP-CS^{bio} dataset can be sacrificed as OringPV may not learn enough from the less-diversified Bahadur^{bio} dataset.

The second observation is about the big change of the testing performance on the same dataset when OringPV's training data is shift from one dataset to another. This can be seen from Table 6.11 that: (i) the testing performance on Bahadur^{hosp} changes from 85.8% to 90% to 92.9% if OringPV's training data is switched from BNCP-CS^{hosp} to the Ponstingl dataset and to Bahadur^{hosp} (LOOCV is used if training data is the same as test data); (ii) the testing performance on the Ponstingl dataset changes from 88.3% to 91.8% to 97.7% if OringPV's training data is switched from BNCP-CS^{hosp} to

Table 6.11: Performance trend of a classifier when training datasets change.

Tested Dataset	Training Dataset (Accuracy %)		
	BNCP-CS ^{hosp}	Ponstingl	Bahadur ^{hosp}
Bahadur ^{hosp}	85.8	90	<i>92.9</i>
Ponstingl	88.3	<i>91.8</i>	<i>97.7</i>
BNCP-CS ^{hosp}	<i>98.3</i>	98.3	98.3

The *italic numbers* are for within-dataset comparison and others for cross-dataset comparison. In this table, the accuracies are a little different from Table 6.10 due to that the training datasets and training frameworks in Table 6.10 are under the same ones as what exactly done by the existing methods, such as leave-one-out learning with a 5-fold cross-validation procedure when OringPV is compared with DiMoVo in Table 6.10. But OringPV is trained with leave-one-out cross-validation here.

the Ponstingl dataset to Bahadur^{hosp}; (iii) however, the OringPV’s testing performance on BNCP-CS^{hosp} has no significant change and maintains at the high level of a 98.3% accuracy when the training dataset is switched from BNCP-CS^{hosp} to the Ponstingl dataset and to Bahadur^{hosp}. It seems that Bahadur^{hosp} is the most accountable training dataset, and the Ponstingl dataset is the second most accountable, while BNCP-CS^{hosp} is less reliable, in this special classification task of distinguishing between obligate (homodimeric) interactions and crystal packing.

To understand the deep reasons, we examine two factors: the average size of O-ring-surrounded regions in each dataset, and the propensity vector values of O-ring-surrounded regions with various sizes. Our examination shows that the average size for O-ring-surrounded regions of the crystal packing in the Bahadur^{hosp} dataset and the size variance are both bigger than those in the Ponstingl dataset, which are both bigger than those in BNCP-CS^{hosp}. On the contrary, the average size for O-ring-surrounded regions of the homodimeric interactions in the Bahadur^{hosp} dataset and the size variance are both similar to those in the Ponstingl dataset, which are both smaller than those in BNCP-CS^{hosp}. This means that crystal packing and homodimeric interactions within the Bahadur^{hosp} dataset are more complicated though separable than those within BNCP-CS^{hosp} and the Ponstingl dataset are. Therefore, OringPV, especially propensity part of propensity vectors, can learn better if its training data is switched to the Bahadur^{hosp} dataset. On the other hand, smaller O-ring-surrounded regions are easier to produce random sharp propensity values in a vector due to that the propensity vectors have 210 dimensions of residue pairs together describing the whole interfaces. Further, the dataset with smaller O-ring-surrounded regions and also with smaller size variance might make the propensity part of propensity vectors less helpful in the classification process. It is why the testing performance is better when OringPV is trained on the datasets whose

crystal packing have larger O-ring-surrounded regions and larger size variance, such as the Bahadur^{hocp} dataset, than when OringPV is trained on the datasets whose crystal packing have smaller O-ring-surrounded regions and smaller size variance, such as the BNCP-CS^{hocp} dataset.

Our third observation is as follows. A biological interaction is easy to be wrongly predicted as crystal packing if its O-ring-surrounded region is small, and a crystal packing is also easy to be wrongly predicted as biological interaction if its O-ring-surrounded region is large. We show three examples taken from the BNCP-CS dataset in the process of identifying biological interactions from crystal packing. Under LOOCV, OringPV misclassifies 8 interactions (2 non-obligate interactions are wrongly predicted as non-biological interactions, and 1 crystal packing as obligate, and 5 crystal packing as non-obligate interactions). If OringPV is trained on the Bahadur dataset, it wrongly classifies 12 interactions (2 obligate interactions and 7 non-obligate interactions are grouped into non-biological interactions, and 1 and 2 crystal packing into obligate and non-obligate interactions respectively). Under these two prediction approaches, there are two common biological interactions misclassified as non-biological interactions, and one common crystal packing that is wrongly predicted as a biological interaction. Interestingly, DiMoVo also wrongly classifies these three interactions.

Of the two wrongly predicted biological interactions, one is in the PDB entry 1ARO between chains L (T7 lysozyme) and P (T7 RNA polymerase), and the other is in 2PCB (A complex between electron transfer partners) between chains A (Cytochrome C peroxidase) and B (Cytochrome C). Figure 6.4(a)(b) show their structures. Recall that the propensity vectors are based on the number of O-ring-surrounded residues and the number of surface residues. We examine the number of surface residues, N_S , of 1ARO and of 2PCB, in comparison to the number of residues, N_O , in their O-ring-surrounded regions. The investigation indicates that N_O/N_S of 1ARO and of 2PCB are the minimum in all biological interactions. A smaller N_O/N_S means that residue pairs in the O-ring-surrounded region are too fewer to make propensity values outstanding. We also observed that when N_O/N_S is less than 0.1, 3 out of 4 biological interactions are wrongly classified. The misclassified prediction is likely due to that these O-ring-surrounded regions of misclassified interactions are so small that the propensity part of propensity vectors has many random sharp propensity values, thus misleading the true propensity properties of those residue pairs.

The PDB entry of the one commonly misclassified crystal packing is 1M7G (APS kinase from *Penicillium Chrysogenum*: ternary structure with ADP and APS) whose structure is shown in Figure 6.4.c. Similar to the above data analysis, N_O/N_S of 1M7G is found with the maximum N_O/N_S in all crystal packing. Furthermore, there are 7 misclassified non-biological interactions (*by the LOOCV procedure or by OringPV trained on the Bahadur dataset*) among the top 11 maximum of N_O/N_S descending ranking of

6.4.a A misclassified biological interaction at chain L and P in 1ARO

6.4.b A misclassified biological interaction at chain A and B in 2PCB

6.4.c The misclassified crystal packing in 1M7G

Figure 6.4: The structures of misclassified interactions. The interactions in (a) and (b) are biological interactions but wrongly predicted to crystal packing; the interaction in (c) is crystal packing but wrongly predicted to biological interaction. The O-ring-surrounded regions are in the ‘spheres’ view in PyMOL with red and blue colors.

crystal packing. The N_O/N_S values of these 7 misclassified crystal packing are all larger than 0.13. A possible reason is that crystal packing can occasionally form ‘abnormal’ large O-ring-surrounded regions, which computationally results in propensity values similar to those of biological interactions.

6.5 Conclusion

In this work, we have introduced low-ASA residue pairs and O-ring-surrounded regions (See Section 6.2.2). The biological principle of this notion is based on the long standing O-ring theory and the coupling proposition. The water accessibility restriction and the contact distance are both considered in the definition of low-ASA residue pairs. Thus, with this definition, the properties of binding hot spot residues are fully integrated into such residue pairs. We also introduced propensity vectors of low-ASA residue pairs and have suggested to use these propensity vectors to characterize the different types of protein interactions.

The OringPV method, our newly proposed learning scheme with propensity vectors as the input of SVM, has shown excellent performance in the prediction of the three types of protein interactions. The experiments are conducted on three benchmark datasets: the BNCP-CS dataset [18], the Ponstingl dataset [27], and the Bahadur dataset [22]. The performance is evaluated under the LOOCV procedure (See Section 6.3.1), and also under the comparison frameworks such as within/cross-dataset tests in comparison to widely accepted literature methods, including NOXClass [18] and DiMoVo [70] (See Section 6.3.4). The evaluation results demonstrate that the propensity vectors can signify important characteristics of protein interactions, and OringPV is highly accurate to identify biological interactions from non-biological interactions and to distinguish different types of biological interactions.

As a future work, OringPV perhaps can be used to determine and rank the fitness scores of all possible binding structures constructed by docking algorithms [134]. We also consider to apply low-ASA residue pairs and the propensity idea to deal with hot spot or interface prediction problems. In fact, these problems are similar to the current one, though they are beyond the scope of the current work. One consideration is that we construct propensity vectors for interacting chain pairs that contain a hot spot as the current work does. Second, we construct propensity vectors for interacting chain pairs where non-hot spots are identified. These two classes of propensity vectors, labeled with hot spot or non-hot spot, can be then used to train a classifier to predict whether a cluster of contact residue pairs is or not a hot spot. However, one difficulty of this future work is that we are lack of experimental data of non-hot spots. Perhaps, one-class learning algorithms [135] are useful. We leave all these details for readers who are interested in these problems.

Chapter 7

β Atomic Contacts: Identifying Critical Specific Contacts in Protein Binding Interfaces

From the chapters 3 to 6, the DWE hypothesis is proposed to analyze protein interfaces and their binding hot spots at a residue level. Without doubt, residue contacts in binding interfaces are based on atomic contacts, and atomic contacts can capture more detail of protein binding interfaces. Thus, in the following two chapters 7 and 8, protein binding is investigated at an atomic level under DWE.

Meanwhile the previous chapters and also previous works typically defined a protein binding interface as a cluster of spatially close atoms and their contacts but neglecting the atomic local organization of contacts. Hence, a lot of atomic contacts defined are actually accommodate an atom almost between them. Under these definitions, efficient structural analysis of protein binding is very difficult, if not impossible and misleading, since atomic contacts are a fundamental unit in 3D structures. For example, these non-specific definitions may detect larger ‘binding interfaces’ in crystal packing in Chapter 6, and what is even worse is that it is very hard to exactly distinguish these crystal-packing ‘binding interfaces’ from true ones.

To overcome the problems under those atomic-contact definitions in previous works, we propose a new definition called β atomic contacts. Our β contact definition has its foundation on β -skeletons and requires that there is no other atom in the contact sphere defined by two contact atoms i and j ; this sphere is similar to van der Waals spheres of atoms. To empirically quantify the importance of β contacts, we designed an SVM classifier β ACV with β contacts as input features to classify homodimers from crystal

packing. We found that our β ACV is able to achieve a superior classification performance to an SVM classifier with distance-based contacts as input, although β contacts are only a small fraction, about 10%, of conventional distance-based contacts. Our β ACV also outperforms several existing methods on four datasets. The promising performance suggests that β contacts can truly identify critical specific contacts in protein binding interfaces. β contacts thus provide a more precise model of atomic organization than distance-based contacts for 3D structures of biological molecules, and in the future, β contacts can greatly facilitate other structural analysis of biological molecules, such as protein folding and binding, and binding of proteins to RNA/DNA.

7.1 Introduction

Specific binding of proteins plays a fundamental role in molecular functions and biological processes. Discovery of governing principles behind specific protein interactions is thus an essential problem in proteomics. Protein interactions are generally considered to be governed by their binding interfaces which consist of interfacial residues/atoms and their contacts. In order to uncover contributing factors to specific interactions, binding interfaces and their contacts are firstly quantified. Several criteria are widely used to define interfacial contacts: some definitions consider atomic distance between atoms each from one protein [10, 46, 54, 55, 56, 57, 58], while another more complicated definitions take into account Voronoi diagrams of entire complexes [67, 68, 69, 70, 64, 136]; the change of solvent accessible surface area (Δ ASA) upon the formation of protein complexes is also used to define binding interfaces [66, 18, 22, 23, 30, 31, 65, 53]. Under these definitions, protein binding is considered to be driven by forces from those atomic contacts such as hydrogen bonds, electrostatic interactions, van der Waals forces, salt bridges, hydrophobic attractions and so on.

However, it is still very difficult, if not impossible, to pinpoint the governing principles according to these existing contact definitions above. This partially, if not all, results from non-specific contacts under these criteria which define an interface simply as a cluster of spatially close atoms and their contacts but neglect the atomic local organization of contacts. For example, the definitions based on spatially close distance [10, 46, 54, 55, 56, 57, 58] detects all contacts (dash lines) in Figure 7.1(b); but in fact some contacts are almost completely interrupted by other atoms as shown in Figure 7.1(a). Thus, these definitions pay little attention to local surroundings of its defined contacts. Another evidence for the non-specific contacts is that these definitions above will detect larger ‘binding interfaces’ in crystal packing, and it is very hard to distinguish these crystal-packing ‘binding interfaces’ from true ones. As discussed in Chapter 1, crystal packing is the artifact of the crystallographic packing environments and is randomly formed during

the crystallization process; but they do not occur in solution or in their physiological states [11]. With crystal packing as the reference state, a perfect contact definition is expected to satisfy that no or fewer contacts are detected in crystal packing; based on this definition, a simple learning algorithm should be able to easily distinguish crystal packing from specific biological binding of proteins.

In this work, we propose a new definition: β atomic contacts. A β atomic contact, c , of atoms i and j must satisfy β -skeletons [137] where there is no other atom in c 's forbidden region. This forbidden region is defined by the parameter β . In this work, β is set to 1, defining a sphere with the midpoint of i and j as the center and with the spatial distance between i and j as the diameter; this sphere is similar to van der Waals spheres of atoms. An example of c 's forbidden sphere in β -skeletons is shown in the gray region in Figure 7.2(b) and examples of β contacts are given in Figure 7.2(e) and in Figure 7.1(d). Obviously, this definition only detects “perceptually meaningful” contacts. We expect β atomic contacts can provide a more precise model of atomic organization in protein 3D structures than the previous definitions.

To demonstrate the efficacy of our β contacts in identifying critical atomic contacts in protein binding interfaces, we adopt β atomic contacts to define interfaces of proteins and then investigate the difference between homodimeric interfaces and crystal-packing interfaces. Several previous works also endeavored to detect distinguishing characteristics of crystal packing and specific biological binding. Some works have revealed a significant difference between protein surfaces and interfaces in amino acid composition, as well as a high similarity of protein surfaces to crystal packing [30, 83, 22, 121, 35]. Several other methods have been proposed to identify biological protein complexes from crystal packing, for example, ACV [26, 107], PITA [28], NOXclass [18] and DiMoVo [70]. However, almost all of them use knowledge extracted from the simple definitions, such as defining interfaces using ASA change or using a threshold of atomic distance for interfacial contacts.

In this work, we use β atomic contacts in interfaces to classify homodimers from crystal packing. In this classification, we represent an interface by an atomic contact vector (ACV) [26] based on β contacts, and then design a new classifier, called β atomic contact vector (β ACV). β ACV is a linear SVM classifier after using SVM-RFE to select distinguishable types of β atomic contacts as input features. Evaluated on four datasets, β ACV achieves better classification than the ACV classifier simply based on the distance-based contacts, although β atomic contacts are only a small fraction of the contacts under the latter definition. Our β ACV is also compared with several existing methods in literature, including PISA, DiMoVo and NOXclass. The results demonstrate that β contacts are greatly superior to these methods in most cases. All these comparisons suggest that β contacts are more capable in capturing specific binding contacts than the other definitions. A web server of β ACV is also available at <http://155.69.2>.

25/liuqian/bacv/index.py (visit from outside NTU) or at <http://172.21.100.239/liuqian/bacv/index.py> (visit from inside NTU), or <http://sunim1.sce.ntu.edu.sg/liuqian/bacv/index.py>.

β atomic contacts provide a new fundamental and precise unit for atomic organization in computational structural analysis. In the future, β atomic contacts can be applied in many other applications, such as for estimation of folding and binding free energy, prediction of binding hot spots, protein docking as well as other structural analyses for folding and binding of proteins and RNA/DNA.

7.2 β Atomic Contacts in Protein Interfaces

7.2.1 Various Definitions of Binding Interfaces and Their Atomic Contacts: A Background

Protein binding interfaces consist of interfacial residues. Further, residue contacts across interfaces are quantified as atomic contacts. Binding interfaces and their residue/atomic contacts are generally defined by several widely-used simple criteria in literature, as discussed in Chapter 1. Some works determined contact residues based on the change of solvent accessible surface area ΔASA upon the formation of protein complexes [18, 22, 23, 30, 31, 65, 53]. Another definition of contact residues is based on atomic distance [10, 46, 53, 54, 55, 56, 57, 58]. This kind of definitions consider how two residues/atoms from an interacted pair are spatially close. More complicated definitions of contact residues are according to Voronoi diagrams of entire complexes [67, 68, 69, 70, 64, 71]. Please refer to Section 1.1.4 for detail. We would like to note that the latter two kinds of definitions can be used to define both interfaces and residue/atomic contacts, while the first kind of the ΔASA criterion defines binding interfaces directly.

The simple definitions above have several limitations. (i) Tsai *et al.* compared the three kinds of definitions on 592 structures of protein-protein interfaces and found that the definition based on Voronoi diagrams provides a more accurate analysis of protein interfaces than the other two criteria [59]. However, they also argued that not all of these contacts based on Voronoi diagrams have a significant energetic contribution to protein binding. (ii) The second close-distance definition are commonly used, but it only considers spatial distance between two atoms, neglecting the local organization of their nearby atoms—whether there are other atoms between these two atoms. When the distance threshold becomes larger, the close-distance criterion will detect a huge number of contacts which are truly interrupted by other atoms. An intuitive example of such interrupted contacts is illustrated in Figure 7.1(a) where under the close-distance definition, OD1 in D53 has contacts with NH1 in R97, OD1 in D52 and other atoms; but most of them are truly interrupted by OD2 in D52 and NH2 in R97. Thus, it is doubtful

if these interrupted contacts have significant contribution to protein folding and binding. At least, these contacts should have greatly different importance to protein binding from the other “perceptually meaningful” contacts, such as the contact between OD1 in D52 and NH1 in R97 as shown in Figure 7.1(a) and Figure 7.1(d). (iii) The well-known driving forces of protein folding and binding, such as salt bridges and hydrogen bonds, suggest a completely different property from these simple definitions: they are quite “perceptually meaningful”—there are no other heavy atoms between them—instead just with close spatial distance. That is, these simple definitions above cannot exactly define hydrogen bonds and salt bridges.

Furthermore, atomic contacts are the basic unit in protein 3D structures. If a huge number of contacts are defined without considering their local organization, it will bring a great challenge for further structural analysis of protein folding and binding. For example, when the close-distance definition is used to define biological binding interfaces, it is still very difficult, if not impossible, to mine specific contacts from all atomic contacts defined. Another argument is that none of them exactly follows the biological meaning of protein interfaces. That is, these definitions will detect larger ‘binding interfaces’ in crystal packing, and what is even worse is that it is very hard to distinguish these crystal-packing ‘binding interfaces’ from true ones. To our best knowledge, few works in literature address the problems of the contacts under these simple definitions.

7.2.2 β Atomic Contacts: Our Definition

To tackle these problems of atomic contacts, we propose a new definition— β atomic contacts in this work. Given a 3D structure of a protein complex p , **an atomic contact, denoted as $c(i, j)$, between two atoms i and j is called a β contact if and only if $d(i, j) \leq T_d \wedge c(i, j) \in e(VD(p)) \wedge c(i, j) \in e(b(p))$** . Specifically, these three requirements are described as follows:

$d(i, j) \leq T_d$: This first requirement states that the surface distance between i and j , $d(i, j)$, must be less than or equal to T_d . Here, a surface distance of two heavy atoms is their Euclidian distance minus the sum of their van der Waals radii as defined in [12]. A similar surface distance definition is also used in [138, 56, 139]. For a simple description, the ‘distance’ in our definition and method is always calculated in this way except as otherwise specified. These contacts under this requirement are called distance-based(-cutoff) contacts.

$c(i, j) \in e(VD(p))$: For the second requirement, i and j must share a Voronoi facet in $VD(p)$ where $VD(p)$ is the Voronoi diagram of p and $e(VD(p))$ is a set of edges in $VD(p)$. These contacts under these two requirements above are called Voronoi-based contacts.

7.1.a Real atomic contacts

7.1.b Distance-based atomic contacts

7.1.c Voronoi-based atomic contacts

 7.1.d $\beta=1$ atomic contacts

Figure 7.1: An example of the difference of atomic contacts according to various definitions (better viewed in color). In (a), carbon, oxygen and nitrogen atoms are in green, red and blue respectively. In (a)-(d), the solid lines represent covalent bonds and the dash lines represent atomic contacts; XYZ or XY represents a specific atom of X type in the location YZ/Y in a specific residue; for example, OD1 specifies an oxygen atom in the location D1 for Asp52.

$c(i, j) \in e(b(p))$: The last requirement indicates that $c(i, j)$ cannot break β -skeletons [137]. That is, $c(i, j)$ is an edge of the β -skeleton $b(p)$ of p where $e(b(p))$ is a set of edges in $b(p)$. These contacts under these three requirements above are called β atomic contacts.

A β -skeleton of a discrete set p is an undirected graph in computational geometry where two points i and j have an edge if and only if any angle ikj are sharper than a threshold determined by β , $\forall k \in p, k \neq i, j$. This angle threshold is called $\angle\beta$. The relation of β and $\angle\beta$ is: $\angle\beta = \arcsin \frac{1}{\beta}$ if $\beta \geq 1$ and $\angle\beta = \pi - \arcsin \beta$ if $\beta \leq 1$. β or $\angle\beta$

actually defines a forbidden region, fr , for the contact between i and j , just like the gray regions in Figure 7.2. In Figure 7.2, two atoms i and j have an edge in β -skeletons if there is no other atoms k whose center is in their forbidden region. In other words, if there is any atom whose center is in the gray regions, the atomic contact between i and j is interrupted, and the contact should not exist in β -skeletons. For example when $\beta = 1$, that is, $\angle\beta = 90^\circ$, two atoms i' and j' do not have a β contact in Figure 7.2(d) because there is a k' in their forbidden region, while the contact between two atoms i and j is a β contact in Figure 7.2(e) since any atom k is outside the forbidden region. It is also interesting to note that in Figure 7.2(d), two atoms i' and j' have a Voronoi-based contact, but they only share a smaller-size facet (e.g., the dash gray line down in Figure 7.2(d)) which is also far away from the center region of the contact—the center region of the contact between two atoms i' and j' is a small arch region very close around the contact, that is, the magenta region in Figure 7.2(d). Compared to this contact in the Voronoi-based definition, our β criterion assumes that two atoms should have enough contact area in their center region to form an important interaction.

In β -skeletons, with different values of β from bigger to smaller, the forbidden gray regions decrease as shown in Figure 7.2 from (a) to (c) and the number of atomic contacts in β -skeletons increases. When β is small enough, the contacts defined on β -skeletons are similar to those on Voronoi diagrams or even to those on distance-based definitions. In this work, β is set to 1, and this β -skeleton is also called Gabriel graph [140, 141]. We would like to emphasize that β -skeletons are a totally different concept from β shape [142], a generalization of the α shape [143], which is also commonly used in analysis of protein structures, such as ASA calculation and protein shape detection.

7.2.3 The Difference of β Contacts from Other Contact Definitions

To show the difference of β atomic contacts from distance-based and Voronoi-based contacts, Figure 7.1 gives an example of real contacts. The real contacts in Figure 7.1(a) occur between three residues in the binding site of the PDB entry 3LZF [144]. Figures from 7.1(b) to 7.1(d) show the distance-based, Voronoi-based, and β atomic contacts, respectively. For simplicity, in Figures from 7.1(b) to 7.1(d), we only consider contacts between those atoms each from the following two groups: atoms CG, OD1 and OD2 from the residue Asp52, and atoms CZ, NH1 and NH2 from Arg97. The atomic contacts between these six atoms are also shown in a 2D space.

In Figure 7.1(b) under the distance-based definition, the two groups of atoms are fully connected, since they are spatially close to each other. Under the Voronoi-based definition, four distance-based atomic contacts are removed as shown in Figure 7.1(c). These

Figure 7.2: β skeletons and β contacts. Three points, i , j and k , represent atoms. The dash circles in blue represent van der Waals spheres in 2D space. The lines in yellow are of interest. In the first row, if i and j have a β contact, the gray regions are required to contain no other atom by β skeletons when $\beta > 1$ (left), $\beta = 1$ (center) and $\beta < 1$ (right), respectively. In (d), a region in magenta is the center area which is very close around the line in yellow, and the dash lines represent Voronoi facets.

removed atomic contacts, e.g., the contact of CG and CZ (CG.CZ for short), are interrupted by other atoms, such as NH2 (Please refer to Figure 7.1(a) for detail). However, some other interrupted contacts are still detected under the Voronoi-based definition, such as the OD2.NH1 contact (*In fact, these six atoms of the two groups are almost coplanar; this indicates that OD2 and NH1 should not have enough contact center region between them; this contact is similar to the contact between i' and j' in Figure 7.2(d).*). In contrast, these interrupted contacts are all filtered by our β contact definition as shown in Figure 7.1(d). Figure 7.1(d) also suggests three significant contacts: the contacts between OD1 and NH2, between OD2 and NH2, and between OD1 and NH1; these three β contacts are hydrogen bonds which are obviously observed in the real contacts in Figure 7.1(a). Thus, β atomic contacts are also “perceptually meaningful” contacts. We

believe that these “perceptually meaningful” atomic contacts can identify critical specific contacts in protein binding interfaces and aid the discovery of the principles governing protein binding.

Finally, we would like to make two remarks about the definition of β atomic contacts. The first remark is that β atomic contacts also implicitly require a large enough angle kij or kji besides the angle requirement of ikj in β -skeletons. This implicit requirement of kij can easily be demonstrated as follows. The requirement of kji can be demonstrated similarly.

Assumptions: (i) in a β contact between two atoms i and j and their close atoms k , atoms i , j and k define a triangle as shown in Figure 7.2(e); (ii) k is a covalent-bond atom of i . The second assumption is needed because covalent-bond atoms have smaller distance to atom i or j than those non-covalent-bond atoms do. That is, covalent-bond atoms usually constrain the contacts of atom i or j in β contacts. So, the closest neighbor atoms k of atom i or j are generally the covalent-bond atoms of atom i or j .

Demonstration: In this triangle, the sum of the angle ikj , kij and kji is π . Then if i and j have a β contact, the angle ikj must be sharper than a right angle here, that is, $ikj < \pi/2$. As a result, the sum of the angles kij and kji must be greater than $\pi/2$. Then the angle kij should be larger enough, since the covalent bond of k and i generally has much shorter length than the non-covalent contacts of i and j and of k and j in this triangle.

Although this implicit requirement on the angle kij in β contacts is not as large as hydrogen bonds do, β contacts share angle requirement which is required by hydrogen bonds and π -involving interactions: in a hydrogen bond, if a Nitrogen is considered to be k , a hydrogen to be i , and an Oxygen to be j , kij is generally required to be larger than $\pi/2$. Thus, β contacts can be considered as a generality extension of the angle requirement in hydrogen bonds and π -involving interactions. In literature, hydrogen bonds and π -involving interactions, which have angle requirements in contact definitions, both have been found to be important in protein binding and folding. Thus, whether it means that protein 3D structures are, instead of a cluster of spatially close atoms with no local structures, well organized according to some rules, for example, angle requirements here? The answer should be yes, but we have no idea about how atoms are well organized. In this work, β atomic contacts should be a promising attempt to detect well-organized protein 3D structures. We would like to find stronger evidence by using β atomic contacts for more applications in protein folding and binding in the future.

The other remark is that some features are not directly considered in β atomic contacts, e.g., van der Waals forces and atomic packing density. These features may be

somewhat reflected by distance-based atomic contacts if the distance threshold is large enough. In previous works, atomic packing density is observed to be an effective feature in protein characterization such as prediction of binding hot spots [66]. So, these two features are advantageous features in distance-based contacts to analyze protein 3D structures. However, both of them are necessary but not sufficient conditions of protein binding—they should not be critical specific patterns in protein binding either.

7.3 Materials and Methods

To evaluate the advantages of β contacts, we use them to design an SVM classifier β ACV to identify homodimers from crystal packing. Thus in this section, we firstly describe three datasets of crystal packing and biological homodimers in literature, which are also used in this work. We then present why and how we construct another new small dataset of HA1-2D1 binding for further evaluating β ACV and other existing classification methods. After that, we introduce a process to detect β atomic contacts from given protein 3D structures, and provide ACV representations for interfaces to design β ACV. Finally, we describe several common measures for evaluating classification performance. A brief introduction of machine-learning algorithms used here are given in Appendix B, including SVM, a feature-selection method SVM-RFE and decision tree.

7.3.1 Datasets

Three datasets in literature are used to comprehensively evaluate β atomic contacts.

The first *Bahadur* dataset contains 178 crystal packing and 113 biological homodimers from the previous works [31, 22]. This dataset has been used to develop DiMoVo [70].

The second *Ponstingl* dataset has 95 crystal packing and 76 homodimers [27]. This dataset has been used in several existing works [26, 107], including PITA [28] and PISA [145]. These two datasets are the same as those in Chapter 6.

The third *non-redundant* dataset is compiled by [70] from the Bahadur, Ponstingl and NOXclass datasets [18]. In this non-redundant dataset, two proteins have no more than 30% sequence identity. This dataset includes 314 crystal packing and 144 homodimers after preprocessing [70].

7.3.1.1 A HA1-2D1-binding Dataset for a Case Study.

For a case study, we also construct another small dataset, which has different properties from the three datasets above. In each of the above datasets, proteins generally must have lower sequence identity compared with each other, e.g., $< 30\%$. This is quite reasonable to avoid over-fitting in the following scenario when evaluating classification performance

of homodimers and crystal packing: if protein sequences in two complexes have higher identity, these two complexes generally have similar 3D structures; further if they belong to the same class (e.g., crystal packing), classifiers will overestimate them, thus resulting in evaluation bias.

However, protein binding is specific—a small fraction of mutations or difference of residues can result in dissociation of protein complexes. Therefore, there is another different scenario: two complexes have highly similar protein sequences, e.g., $> 30\%$ sequence identity, but they belong to different classes—for example, one is a true protein complex and the other is crystal packing. This scenario is common in real biological systems, but it is very little, if not never, considered in these three datasets above. Also, it is a difficult task for a classifier to produce correct predictions for those biological complexes and crystal packing with the properties in this scenario.

Hence, we want to construct a new dataset whose protein sequences in two complexes have the properties in the latter scenario. This dataset is used for test only. In this work, this dataset contains HA1 proteins in several different influenza viruses especially H1N1, such as A/South_Carolina/1/1918 (1918H1N1), A/California/04/2009, A/swine/Iowa/30, A/Brisbane/59/2007, A/Puerto Rico/8/34, A/x-31(H3N2), A/duck/Singapore/3/97(H5), and B/Hongkong/8/73. These HA1 proteins are highly similar in their sequences as shown in Table 7.1. But a specific antibody, for example the 2D1 antibody, can bind to some of them and cannot to the other due to residue mutations (Please refer to Table 7.1). Therefore, the main purpose here is to use these binding complexes to verify whether a method, which is constructed on either distance-based contacts or the proposed β contacts, can capture specific difference in these antigen-antibody complexes.

In antigen-antibody binding of influenza viruses, the 2D1 antibody is a monoclonal antibody from a survivor of the 1918 Spanish influenza [146], which is believed to bind to HA1s in the 1918, 1930 and 2009 H1N1 [147, 144] but not to HA1s in seasonal influenza viruses. HA1 is a major surface protein of influenza viruses. Its binding to the sialic-acid-containing receptors of target cells initiates virus infection to human. Inhibiting HA1 binding to receptors by antibodies is thus an important way against flu. Therefore, this dataset can also be used to examine whether a classifier for distinguishing biological binding from crystal packing can reproduce which HA1 proteins can bind to 2D1 and which cannot, and then help to understand antibody cross-reaction with different influenza viruses.

In these 2D1-HA1 binding in Table 7.1, only the 2D1 binding to the 1918H1N1 HA proteins has an available quaternary structure in the PDB entry 3LZF. Thus, we take the following computational steps to produce quaternary structures and their binding interfaces of the 2D1 binding to HA1 proteins in other influenza viruses (indicated by ‘pyMol’ in Table 7.1). For each of the other influenza viruses in Table 7.1, we first use PyMOL to align its HA protein structure onto the HA protein structure of the 1918H1N1

Table 7.1: The HA1-2D1-binding dataset.

H1N1 Virus	I_{HA1}	Tertiary	Quaternary		Bind to 2D1
			How	RMSD	
A/South_Carolina/1/1918	100	3LZF	3LZF		True [144, 147]
A/California/04/2009	83	3LZG	pyMOL	0.782	True [144, 147]
A/swine/Iowa/30	93	1RUY	pyMOL	0.442	True [146, 147]
A/Puerto Rico/8/34	85	1RU7	pyMOL	0.505	True [144, 147]
A/Brisbane/59/2007	80	ref. [148]	pyMOL	0.921	False [144]
A/x-31(H3N2)	38	1HGE	pyMOL	1.816	False [146]
A/duck/Singapore/3/97	58	1JSM	pyMOL	0.836	False [146]
B/Hongkong/8/73	22	3BT6	pyMOL	5.679	False [146]

‘Tertiary’ indicates where to find the tertiary structures for the HA1 proteins.

‘Quaternary’ means how to obtain the quaternary structures of the HA1-2D1 binding.

‘ I_{HA1} ’ provides sequence identity between the HA1 protein in 1918H1N1 and the HA1 protein in an influenza virus of each row.

‘pyMOL’ indicates that the quaternary structure for the virus in each row is produced by aligning its HA structure with the HA structure of 3LZF in pyMOL.

in the PDB 3LZF with the antibody coordinates. This PyMOL action will align the two HA protein structures to minimize RMSD (Root Mean Square Deviation) [115]. After that, we remove the HA coordinates of the 1918H1N1 and obtain the computational binding interface of the HA1 protein and the 2D1 antibody. As shown in Table 7.1, most of these HA1 proteins do not have much RMSD difference in their 3D structures from the 1918H1N1 HA1 protein. Finally, since HA1s in these computational binding interfaces remain in an unbound state with free side chains and may also have atomic clashes, we next use FoldX [75] to repair this interface when fixing the antibody binding site. The repaired interfaces are used for our subsequent analysis.

7.3.2 Detecting β Atomic Contacts in Protein 3D Structures

A protein 3D structure can be modeled as a β atomic contact graph $b(p)$ where heavy atoms are considered as points, β atomic contacts as edges. Given a protein 3D structure, its $b(p)$ can be produced by the following process.

First, Qhull is used to produce Delaunay triangulation [149] for all points. Second, a distance threshold T_d is used to remove those atomic contacts whose distances are too large. T_d is set to 3.3 Å (the diameter of a water molecule 2.8 Å plus 0.5 Å). T_d is the maximum distance between van der Waals spheres of two atoms, as discussed in Section 7.2.2. Thirdly, each atomic contact is checked to guarantee that it satisfies β

skeletons, that is, Gabriel graph here. Since we are interested in atomic contacts between proteins, all atoms which have no contact across binding interfaces are removed.

7.3.3 β Atomic Contact Vectors in Protein Interfaces

An atomic contact vector (ACV for short) [26] is adopted to represent an interface. In this vector representation, all heavy atoms of the twenty standard residues in proteins are grouped by twelve atomic types in Table 7.2. These atomic types are similar to those in [150]. So, the atomic contact vector for an interface has 78 atomic pairs ($78 = \frac{12 \times 11}{2} + 12$). The value for each pair is its occurrence in a β atomic contact graph when T_d is set to 3.3 Å. Since disulfide bonds are almost as strong as covalent bonds, disulfide bonds whose spatial distance of two sulfur atoms across interfaces is less than 2.6 Å is also considered as an atomic pair in the vector. Finally, the vector for a protein interface has 79 pairs, called β ACV1a (β atomic contact vectors) for short. Similarly, we also construct β ACV1 in which the distance threshold of two contact atoms is as small as 0.5 Å, that is, $T_d = 0.5$ Å.

Beside atomic types, a distance between contact atoms is also an important factor in protein binding. Given two atomic pairs with the same types, one pair has a small distance between atoms, while atoms in the other pair have much larger distance; the first pair generally has different importance to protein binding from the second pair. An instant example with this property is hydrogen bonds in interfaces: if a Nitrogen atom and an Oxygen have less than 3.5 Å spatial distance, their contact may be a hydrogen bond; but these two atoms cannot form a hydrogen bond directly if their spatial distance is too large, for example, more than 5 Å. Therefore, we take into account distance information of atomic contacts and split β ACV1a into three sub-vectors: each of them contains atomic contacts whose distance falls in one of the three regions: ≤ 0.5 , $(0.5, 1.9]$ and $(1.9, 3.3]$, and they are named as ≤ 0.5 contacts, $(0.5, 1.9]$ contacts and $(1.9, 3.3]$ contacts for short. Here $1.9 = 0.5 + 2.8/2$. This vector representation has 235 pairs ($235 = (\frac{12 \times 11}{2} + 12) \times 3 + 1$), which is referred to as β ACV3 for short.

Meanwhile, distance-based ACV1 (dACV1), ACV1a (dACV1a) and ACV3 (dACV3) are also constructed in a similar way, since we want to compare β atomic contacts with distance-based atomic contacts.

7.3.4 Our Proposed Classifier β ACV

We want to evaluate β contacts in classifying homodimers and crystal packing. In our classification task, a dataset of crystal packing and homodimers is represented by $D = \{(x_i, y_i) | y_i \in \{-1, 1\}_{i=1}^n\}$, where $y_i = -1$ indicates that this vector is from crystal packing or $y_i = 1$ indicates it is from homodimers; n is the total number of crystal packing and

Table 7.2: Atomic types.

Atomic group	Chemical formula	Valency ⁿ	Hydrogen ^m	Non-hydrogen ^l
N3H0	>N–	3	0	3
N3H1	>NH	3	1	2
N3H2	–NH ₂	3	2	1
N4(3)H3(2)	–NH ₄₍₃₎ ⁺	4(3)	3(2)	1
O1H0	=O	1	0	1
O1H0	–O [–]	1	0	1
O2H1	–OH	2	1	1
C3(S2)H0	–C<(–S–)	3(2)	0	3(2)
C3(S2)H1	–CH–(–SH)	3(2)	1	2(1)
C4H1	–CH<	4	1	3
C4H2	–CH ₂ –	4	2	2
C4H3	–CH ₃	4	3	1

ⁿ: Valency of X atom;

^m: The number of hydrogen atoms bonded to X;

^l: Non-hydrogen atoms bonded to X.

N, H, O, C and S represent Nitrogen, Hydrogen, Oxygen, Carbon and Sulfur atoms respectively.

homodimers; x_i is β atomic contact vector for interface i , no matter it is β ACV1, β ACV1a or β ACV3. Then, SVM with a linear kernel in LIBSVM [127] (the freely available SVM library) is employed to train our classifier for identifying homodimers from crystal packing. A short tutorial of SVM is provided in Appendix B.

In our β ACV classifier, RFE (discussed in Appendix B) is firstly used to find the feature set S_f with the best accuracy no matter the features are β ACV1a or β ACV3. This process uses a five-fold cross-validation. Then, two established ways are used to evaluate classification performance. One is feature-selection classification performance by using a leave-one-out cross-validation on the learning datasets. The other is the independent-dataset testing. That is, a β ACV classifier with features S_f is constructed for predicting those complexes whose proteins have low sequence similarity to the complexes in the dataset for feature selection. The independent-testing datasets for the Bahadur, Ponstingl and NOXclass datasets can be found in [70].

7.3.5 Evaluation Measures

In this chapter, homodimers are considered as the positive samples while crystal packing as the negative samples. Those measures in Equation 4.10, such as $recall(r.)$,

specificity(sp.), *accuracy(acc.)* and AUC, are adopted to evaluate the classification performance of β atomic contact vectors. Here, *recall(r.)* is the number of correct homodimer predictions divided by the number of homodimers; *specificity(sp.)* is the fraction of correct predictions of crystal packing over all crystal packing; *accuracy(acc.)* is the proportion of correct predictions in all homodimers and crystal packing; Matthew’s correlation coefficient (MCC), $MCC = \frac{TP \times TN - FP \times FN}{\sqrt{T \times F \times (TP + FP) \times (TN + FN)}}$, is a more meaningful measure in a dataset which has a great non-balance between the number of positive and of negative samples in those classifications of this chapter.

7.4 Results and Discussion

In this section, we use three ways to validate the advantages of β atomic contacts.

- We firstly compare the classification performance of β contacts and distance-based contacts on the three datasets. This comparison can show that β contacts are better than distance-based contacts to describe critical atomic contacts in protein binding interfaces. In this comparison with distance-based contacts, we also show that β contacts have more easily interpretable and significant features.
- We then substantiate the usefulness of β atomic contacts by comparing their classification performance against several existing prediction methods in literature on the three datasets. These existing methods include NOXclass [18], DiMoVo [70], PISA [145] and PITA [28].
- We further confirm that β contacts are powerful in identification of biological binding from crystal packing by testing classifiers of distance-based contacts and β contacts and the other existing prediction methods on the HA1-2D1-binding dataset. This dataset has proteins with higher sequence similarity in two protein complexes, while one complex is crystal packing and the other is biological binding. This property is not fully considered in these three dataset in the first two ways, and challenges the existing prediction methods.

7.4.1 Comparison of β Contacts with Distance-based Contacts

To demonstrate that β contacts are better than distance-based contacts to describe protein binding interfaces, we compare these two kinds of definitions from the following aspects. We first measure the numbers of distance-based and of β atomic contacts to see their difference. We then compare their prediction performance on the three datasets. After that, we provide a detailed comparative analysis of the selected features by RFE for distance-based contacts and β contacts, especially of the top 10 atomic contacts in Section 7.4.2.

7.4.1.1 β Atomic Contacts are a Small Fraction of Distance-Based Atomic Contacts.

We perform statistical analysis of the numbers of atomic contacts under the three different definitions: distance-based, Voronoi-based, and β contacts. The analysis is based on the non-redundant dataset when $T_d = 3.5$. The result is shown in Table 7.3 and Table 7.4.

From Table 7.3 and Table 7.4, we find that β atomic contacts are only a small fraction of distance-based atomic contacts, i.e., 9.4% of distance-based contacts in homodimers, and 10.1% in crystal packing.

We also show the comparison of the involving atoms in β contacts with those in the complement of β contacts with respect to distance-based contacts in Figure 7.3. These atoms and their ASA are from the non-redundant dataset. It seems that a lot of contact atoms in distance-based contacts are not defined to be in direct interfaces by β contacts; Figure 7.3 also clearly shows that atoms only in distance-based contacts mostly have small ASA change after complex formation, while β contacts contain most of atoms which have larger ASA change.

Table 7.3: The difference of the numbers of distance-based, Voronoi-based and β atomic contacts for 114 homodimers.

	Distance-based	Voronoi-based	β contacts
Voronoi-based	<i>587,052</i>	0	<i>73,572</i>
β contacts	<i>660,624</i>	<i>73,572</i>	0
Number of contacts	729,175	142,123	68,551

Italic numbers are the difference of the numbers of the contacts under the definitions of its column and row.

The numbers of the last row are the numbers of atomic contacts under the definition of their columns.

Table 7.4: The difference of the numbers of distance-based, Voronoi-based and β atomic contacts for 314 crystal packing.

	Distance-based	Voronoi-based	β contacts
Voronoi-based	<i>309,962</i>	0	<i>52,511</i>
β contacts	<i>362,473</i>	<i>52,511</i>	0
Number of contacts	403,321	93,359	40,848

These numbers has the similar meaning to those in Table 7.3.

Figure 7.3: The ASA change ($\text{\AA}$) of atoms in β contacts and in the complement of β contacts with respect to distance-based contacts (referred to as ‘c-beta contacts’ for short in the figure).

Table 7.5: The comparison of feature-selection classification performance achieved by distance-based and β atomic contacts and other methods in literature.

Dataset		distance-based contacts			β contacts			DiMoVo	
		dACV1	dACV3	dACV1a	β ACV1	β ACV3	β ACV1a		[107]
Ponstingl	r.	0.895	0.895	0.908	0.895	0.934	0.921	-	-
	sp.	0.947	0.947	0.947	0.937	0.968	0.989	-	-
	acc.	0.924	0.924	0.93	0.918	0.953	0.959	-	0.948
	MCC	0.846	0.846	0.858	0.834	0.905	0.918	-	-
	AUC	0.921	0.921	0.928	0.916	0.951	0.955	-	-
Bahadur	r.	0.84	0.877	0.821	0.887	0.925	0.877	0.89	-
	sp.	0.955	0.955	0.955	0.983	0.983	0.943	0.98	-
	acc.	0.911	0.926	0.904	0.947	0.961	0.918	0.945	-
	MCC	0.81	0.84	0.795	0.887	0.917	0.825	0.884	-
	AUC	0.898	0.916	0.888	0.935	0.954	0.91	0.935	-
Non-redundant	r.	0.882	0.917	0.861	0.896	0.958	0.868	-	-
	sp.	0.978	0.971	0.978	0.984	0.994	0.975	-	-
	acc.	0.948	0.954	0.941	0.956	0.983	0.941	-	-
	MCC	0.877	0.893	0.862	0.898	0.959	0.862	-	-
	AUC	0.93	0.944	0.920	0.94	0.976	0.923	-	-

r., sp., acc. and MCC represent recall, specificity, accuracy, and Matthew's correlation coefficient respectively. The **bold numbers** are the larger values in each of three pair-wise comparisons of β ACV1 vs dACV1, β ACV3 vs dACV3, and β ACV1a vs dACV1a. In this table, significant features are selected on a dataset, and then a method with the selected features is evaluated on this dataset under a leave-one-out cross-validation process, as discussed in Section 7.3.4.

Table 7.6: The comparison of classification performance on the independent datasets achieved by distance-based and β atomic contacts and other methods in literature.

Dataset		distance-based contacts			β contacts			Di-MoVo	PISA	PITA	NOX-class
		dACV1	dACV3	dACV1a	β ACV1	β ACV3	β ACV1a				
Training on Ponstingl	r.	0.908	0.934	0.934	0.947	0.947	0.921	0.71	0.92	0.84	-
	sp.	0.934	0.929	0.898	0.938	0.934	0.938	0.92	0.76	0.91	-
	acc.	0.927	0.93	0.907	0.94	0.937	0.934	0.868	0.802	0.894	-
	MCC	0.815	0.827	0.78	0.851	0.844	0.832	0.643	0.602	0.73	-
	AUC	0.921	0.932	0.916	0.943	0.941	0.930	0.815	0.84	0.875	-
Training on Bahadur	r.	0.868	0.974	0.947	0.842	0.947	0.974	0.84	-	-	-
	sp.	0.986	0.978	0.986	1.00	0.971	0.971	0.95	-	-	-
	acc.	0.96	0.977	0.977	0.966	0.966	0.972	0.929	-	-	-
	MCC	0.88	0.935	0.933	0.898	0.902	0.92	0.78	-	-	-
	AUC	0.927	0.976	0.967	0.921	0.959	0.973	0.895	-	-	-
Training on NOXclass	r.	0.861	0.924	0.937	0.911	0.937	0.81	0.79	-	-	0.95
	sp.	0.909	0.841	0.868	0.909	0.882	0.877	0.97	-	-	0.68
	acc.	0.896	0.863	0.886	0.91	0.896	0.86	0.92	-	-	0.751
	MCC	0.745	0.702	0.747	0.784	0.765	0.659	0.79	-	-	0.556
	AUC	0.885	0.883	0.903	0.91	0.91	0.844	0.88	-	-	0.815

r., sp., acc. and MCC represent recall, specificity, accuracy, and Matthew's correlation coefficient respectively. The **bold numbers** have the same meaning as those in Table 7.5. Here, significant features and a prediction method on them are trained on a dataset, and the evaluation is performed on another dataset, as discussed in Section 7.3.4.

It is also interesting to note that previous works have found a small fraction of residues in interfaces, named binding hot spots [36], which contribute most to protein binding free energy; several atomic contacts, such as hydrogen bonds, salt bridges and $\pi - \pi$ contacts, are also observed to be important features in hot spot prediction [79, 75, 76] or to closely relate to binding hot spots [66]. It seems that the distribution of binding free energy in interfaces is uneven not only on residues but also on atoms and their contacts. Thus, there should also be a small fraction of atoms and their contacts in binding interfaces which are closely related to hot spots; they can be referred to as hot atomic contacts for short. Although we cannot argue that β atomic contacts are hot atomic contacts, β contacts give a good criterion to exclude most contacts which should not be closely related to hot spots and to protein binding. The relation of β atomic contacts and hot atomic contacts is verified in Chapter 8 by using β atomic contacts to estimate the change of free energy of residue mutations and to identify hot spots in protein binding.

7.4.1.2 β Atomic Contacts and Distance-based Atomic Contacts in Classification of Crystal Packing and Homodimers.

We then compare β contacts with distance-based contacts in classifying crystal packing and homodimers. The results are shown in Table 7.5 and 7.6.

It is clearly seen from Table 7.5 that β contacts aggregately have better performance than distance-based contacts, although β contacts are only a small fraction of distance-based contacts. In the nine pair-wise comparisons in Table 7.5 (three pairs of β ACV1 vs dACV1, β ACV3 vs dACV3, and β ACV1a vs dACV1a on the three datasets), β contacts are superior in seven times, and distance-based contacts are superior only once (there is one tie). For example on Bahadur, β -contact β ACV3 has much better MCC, 7.7 percent points higher than distance-based dACV3.

Table 7.6 presents the classification performance of β contacts and distance-based contacts on the independent datasets. In these nine comparisons, β contacts perform better in six times. Again, β contacts demonstrate aggregately better classification performance than distance-based contacts.

The difference of β contacts and distance-based contacts are also evaluated by $c\beta$ ACV1 and $c\beta$ ACV3 which only use the corresponding complement of β contacts with respect to distance-based contacts, that is, those atomic contacts not in β ACV1 and β ACV3 but in dACV1 and dACV3, respectively. We find that $c\beta$ ACV1 has acc 0.726 and MCC 0.424 on the independent dataset of NOXclass, acc 0.848 and MCC 0.627 on the independent dataset of Ponstingl, and acc 0.960 and MCC 0.881 on the independent dataset of Bahadur; $c\beta$ ACV3 has acc 0.86 and MCC 0.697 on the independent dataset of NOXclass, acc 0.924 and MCC 0.808 on the independent dataset of Ponstingl, and acc 0.972 and MCC 0.915 on the independent dataset of Bahadur. β contacts still achieve better performance

on the independent datasets of NOXclass and Ponstingl and similar performance on the independent dataset of Bahadur. This similar performance should be due to that this independent dataset is easily to be distinguished with high classification performance for both distance-based and β contacts (shown in Table 7.6). Meanwhile, c β ACV3 have similar performances to dACV3, because the distance-based contacts predominate in their used contacts.

In conclusion, β contacts are generally more capable in capturing critical specific binding contacts than distance-based contacts.

7.4.2 Analysis of the Selected Features in the Classifications

7.4.2.1 Top 10 Selected Features in β Atomic Contacts and in Distance-Based Atomic Contacts.

To deepen understanding of the difference between β atomic contacts and distance-based atomic contacts, we show the top 10 selected features of β ACV3 and dACV3 in Table 7.7 when β ACV3 and dACV3 are trained on the non-redundant dataset.

Table 7.7: The top 10 features of distance-based and β atomic contacts when β ACV3 and dACV3 are trained on the non-redundant dataset.

Rank	distance-based contacts		β contacts	
	types of atomic contacts	T_d range	types of atomic contacts	T_d range
1st	N3H1_ C4H3	(1.9, 3.3]	C3(S2)H0_C4H3	≤ 0.5
2nd	C4H1_C4H1	(1.9, 3.3]	N3H2_ C3(S2)H1	≤ 0.5
3th	<i>O2H1_C4H2</i>	(1.9, 3.3]	C4H1_C4H3	≤ 0.5
4th	N3H1_ <i>O1H0</i> -	(1.9, 3.3]	<i>O1H0-O2H1</i>	(1.9, 3.3]
5th	C3(S2)H0_C3(S2)H1	(1.9, 3.3]	N4H3/2+_ C4H1	(0.5, 1.9]
6th	N3H1_ C4H1	(0.5, 1.9]	<i>O1H0-O1H0</i> -	(0.5, 1.9]
7th	<i>O1H0_C4H3</i>	(1.9, 3.3]	<i>O2H1_C4H2</i>	≤ 0.5
8th	<i>O1H0-O1H0</i>	(1.9, 3.3]	N3H1_ <i>O1H0</i>	≤ 0.5
9th	C3(S2)H1_C3(S2)H1	(0.5, 1.9]	C4H3_C4H3	(1.9, 3.3]
10th	C4H3_C4H3	(1.9, 3.3]	C4H1_C4H1	(0.5, 1.9]

X_Y means atomic contacts between X and Y, while X and Y are atomic types in Table 7.2. Carbon atoms are in **Bold**, while Oxygen atoms are in *italic*.

From Table 7.7, these top 10 features of β atomic contacts indicate two interesting phenomena. One is hydrophobic effect—the contacts among Carbon atoms are chosen to be significantly important to biological binding, although their atomic types are different and their surface distances are ≤ 0.5 Å, or in (0.5, 1.9] Å, or in (1.9, 3.3] Å. The other is

that hydrogen bonds, ≤ 0.5 contacts between N3H1 and O1H0 in β contacts, also play an important role in classifying biological binding from crystal packing.

The top 10 selected features of distance-based contacts capture contacts among Carbon atoms, but miss hydrogen bonds. Furthermore, most of the top features of distance-based contacts are those atomic contacts with relatively larger distance, in $(1.9, 3.3]$ Å in Table 7.7. The main reason is the spatial constraint: given a sphere with an atom as a center, the larger the radius is, i.e., the larger distance threshold T_d here, the more other atoms can be covered without atomic clashes. That is, in distance-based contacts, the number of contacts with the bigger distance $(1.9, 3.3]$ Å is generally much larger than the number of contacts with the smaller distance ≤ 0.5 Å. However, this does not hold in β contacts. For example on the non-redundant dataset, the number of ≤ 0.5 contacts in β contacts are 43,063, and the number of $(1.9, 3.3]$ contacts is 14,007. $(1.9, 3.3]$ contacts are about one third of ≤ 0.5 contacts in β contacts. However in distance-based contacts, the numbers of ≤ 0.5 contacts and of $(1.9, 3.3]$ contacts are 57,286 and 635,254; $(1.9, 3.3]$ contacts are over ten times more than ≤ 0.5 contacts in distance-based contacts. This misleads SVM and RFE to prefer the $(1.9, 3.3]$ contacts in distance-based contacts, since they have higher occurrence.

One argument of distance-based contacts here is why this misleading does not create a classifier with much worse performance. There are at least two helpful factors contributing to distance-based classifiers. One is that distance-based contacts can easily represent atomic density in interfaces; the other is that interface contact size can also greatly help the classification performance of distance-based contacts. Both atomic density and contact size should be distinguishable features and a possibly necessary condition for biological binding; they are easily but indirectly implied in distance-based ACV3 vector, although contacts are divided into different types. However, both of them should not be sufficient conditions for specific protein binding.

7.4.2.2 Decision Trees of Distinguishing Features in β Atomic Contacts and in Distance-Based Atomic Contacts.

To visualize the selected features by RFE and to provide some clues of governing principles behind protein binding, we show the decision trees in Figure 7.4 for β contacts and in Figure 7.5 for distance-based contacts when β ACV3 and dACV3 are trained on the non-redundant dataset with only these selected features. The detail of how to construct decision tree is provided in Appendix B. With these selected features, β ACV3 and dACV3 achieve accuracy 0.983 and 0.954; in decision tree with 5-fold cross-validation, β contacts have accuracy 0.891, and distance-based contacts have accuracy 0.915. Since the important features in SVM cannot to be guaranteed to be the same as those in decision tree, we do not pay more attention to the similar performances of the two trees. Instead,

we would like to see whether easily interpretable knowledge can be derived from these two decision trees, because SVM-based β ACV3 has much better classification performance but poor interpretability.

Figure 7.4 suggests three interesting rules. One is about contacts between two Carbon atoms, called R1 in the first line of Figure 7.4. R1 suggests that if an interface has more than four ≤ 0.5 contacts between the atomic types C3(S2)H0 and C4H3 (*C3(S2)H0-C4H3* for short), it has 98.8%(82/83) probability to be a biological binding. This rule is consistent with hydrophobic effect. The other two interesting rules are closely related to hydrogen bonds. One hydrogen-bond-involving rule is: given an interface with less ≤ 0.5 C3(S2)H0-C4H3, it can still be biological binding if this interface has: **(i)** more than four (0.5, 1.9] C3(S2)H1-C4H3 contacts, and **(ii)** more than three hydrogen bonds (≤ 0.5 N3H1-O1H0 contacts), and **(iii)** no (1.9, 3.3] O1H0-O2H1 contacts. In the non-redundant dataset, 20 biological binding interfaces and none of crystal packing satisfy this rule. In contrast, 251 crystal packing and only one biological interface satisfy the other hydrogen-bond-involving rule (called R2 for short). As shown in Figure 7.4, R2 requires an interface with **(i)** less than five ≤ 0.5 C3(S2)H0-C4H3 contacts, **(ii)** less than five (0.5, 1.9] C3(S2)H1-C4H3 contacts, **(iii)** less than four ≤ 0.5 C4H1-C4H3 contacts, **(iv)** less than eight (0.5, 1.9] O1H0-O1H0 contacts, and **(v)** no more than three hydrogen bonds (≤ 0.5 N3H1-O1H0 contacts). This rule is a little more complicated, but it is reasonable: there should be more than one type of specific important contacts to protein binding, and enough occurrence of several of them in an interface should produce a true biological binding; so, a false biological binding should not have enough critical specific contacts, which must exclude all potential combinations of specific important contacts. This makes a complicated rule unique to crystal packing.

According to these three rules of β contacts in Figure 7.4, we believe that the following contacts should be closely related to specific important contacts to protein binding: the ≤ 0.5 contacts of C4H3 with C3(S2)H0 and with C4H1, (0.5, 1.9] C4H3-C3(S2)H1 contacts, and ≤ 0.5 N3H1-O1H0 contacts. These contacts are consistent with previous observations: ≤ 0.5 N3H1-O1H0 contacts generally can be hydrogen bonds; C4H3 are mostly in the side chains of such hydrophobic residues as Val, Ile and Leu, indicating hydrophobic effect; C3(S2)H1 are almost in the aromatic side chains, providing π -involving interactions in binding interfaces. In contrast, it is crystal packing, not biological binding, which prefer (1.9, 3.3] O1H0-O2H1 contacts. Further in Figure 7.4, higher occurrence of contacts involving O1H0- almost suggests crystal packing prediction. O1H0- should play a destructive role in binding interfaces unless it can form salt bridges.

```

C3(S2)H0_C4H3 [0.5] > 4: Dimers (83.0/4+0)
C3(S2)H0_C4H3 [0.5] <= 4
| C3(S2)H1_C4H3 [0.5,1.9] <= 4
| C4H1_C4H3 [0.5] <= 3
| O1H0_O1H0 [0.5,1.9] <= 7
| N3H1_O1H0 [0.5] <= 3: CP (252.0/4+0)
| N3H1_O1H0 [0.5] > 3
| O2H1_C4H2 [0.5] <= 1: CP (17.0)
| O2H1_C4H2 [0.5] > 1
| N3H1_N4H3/2+ [0.5] <= 0
| O1H0_C4H2 [0.5] <= 3: Dimers (2.0)
| O1H0_C4H2 [0.5] > 3: CP (9.0/4+0)
| N3H1_N4H3/2+ [0.5] > 0: Dimers (2.0)
| O1H0_O1H0 [0.5,1.9] > 7
| C4H3_C4H3 [0.5,1.9] <= 0: CP (6.0)
| C4H3_C4H3 [0.5,1.9] > 0
| O1H0_O1H0- [0.5,1.9] <= 5
| C4H2_C4H2 [0.5] <= 3
| C4H1_C4H3 [0.5] <= 1: CP (4.0)
| C4H1_C4H3 [0.5] > 1: Dimers (3.0)
| C4H2_C4H2 [0.5] > 3: Dimers (8.0)
| O1H0_O1H0- [0.5,1.9] > 5: CP (3.0)
| C4H1_C4H3 [0.5] > 3
| N3H2_C3(S2)H1 [0.5] <= 1
| N3H2_N4H3/2+ [1.9,3.3] <= 0
| N3H1_C4H3 [0.5,1.9] <= 5
| O2H1_C4H2 [0.5] <= 3: CP (16.0/2+0)
| O2H1_C4H2 [0.5] > 3: Dimers (2.0)
| N3H1_C4H3 [0.5,1.9] > 5: Dimers (3.0)
| N3H2_N4H3/2+ [1.9,3.3] > 0: Dimers (2.0)
| N3H2_C3(S2)H1 [0.5] > 1: Dimers (4.0)
C3(S2)H1_C4H3 [0.5,1.9] > 4
| O1H0_O2H1 [1.9,3.3] <= 0
| N3H1_O1H0 [0.5] <= 3
| O1H0_O1H0- [0.5,1.9] <= 0
| N3H1_C4H3 [0.5,1.9] <= 2
| C3(S2)H1_C4H3 [0.5,1.9] <= 8: CP (4.0)
| C3(S2)H1_C4H3 [0.5,1.9] > 8: Dimers (4.0/4+0)
| N3H1_C4H3 [0.5,1.9] > 2: Dimers (8.0)
| O1H0_O1H0- [0.5,1.9] > 0: CP (2.0)
| N3H1_O1H0 [0.5] > 3: Dimers (20.0)
| O1H0_O2H1 [1.9,3.3] > 0: CP (4.0/4+0)

```

```

N3H1_C4H3 [1.9,3.3] <= 34
| C3(S2)H1_C3(S2)H1 [0.5,1.9] <= 8: CP (292.0/4+0)
| C3(S2)H1_C3(S2)H1 [0.5,1.9] > 8
| C4H3_C4H3 [0.5,1.9] <= 2
| C3(S2)H0_C3(S2)H1 [1.9,3.3] <= 122
| N3H1_C4H1 [0.5,1.9] <= 17: CP (20.0)
| N3H1_C4H1 [0.5,1.9] > 17: Dimers (3.0/4+0)
| C3(S2)H0_C3(S2)H1 [1.9,3.3] > 122: Dimers (2.0)
| C4H3_C4H3 [0.5,1.9] > 2
| O1H0_O1H0 [1.9,3.3] <= 20: Dimers (8.0)
| O1H0_O1H0 [1.9,3.3] > 20
| O1H0_O1H0 [1.9,3.3] <= 27: CP (2.0)
| O1H0_O1H0 [1.9,3.3] > 27: Dimers (3.0)
N3H1_C4H3 [1.9,3.3] > 34
| N3H1_C4H3 [1.9,3.3] <= 58
| C3(S2)H0_C3(S2)H1 [1.9,3.3] <= 54
| N3H1_C4H1 [0.5,1.9] <= 24
| N3H1_O1H0- [1.9,3.3] <= 5
| O2H1_C4H2 [1.9,3.3] <= 11: Dimers (8.0/4+0)
| O2H1_C4H2 [1.9,3.3] > 11
| N3H1_O1H0- [1.9,3.3] <= 1: Dimers (3.0/4+0)
| N3H1_O1H0- [1.9,3.3] > 1: CP (4.0)
| N3H1_O1H0- [1.9,3.3] > 5: CP (4.0)
| N3H1_C4H1 [0.5,1.9] > 24: Dimers (6.0)
| C3(S2)H0_C3(S2)H1 [1.9,3.3] > 54: Dimers (20.0)
| N3H1_C4H3 [1.9,3.3] > 58: Dimers (83.0)

```

Figure 7.5: Decision tree of distance-based contacts on the non-redundant dataset. Each line has the similar meaning to that in Figure 7.4; the format of a line is: types of atomic contacts with their distance information, followed by a splitting rule and a class label if possible. For example, the rule ‘C3(S2)H0_C4H3 [0.5] > 4: Dimers (83.0/4+0)’ in the first line of Figure 7.4 suggests the following prediction: 83 interfaces have more than four ≤ 0.5 contacts of C3(S2)H0 and C4H3 in the non-redundant dataset; among these interfaces, only one is crystal packing.

Figure 7.4: Decision tree of β contacts on the non-redundant dataset. Each line is a branch in decision trees; ‘|’ and an indent represent a sub-branch; ‘Dimers’ indicates a class label of biological binding, while ‘CP’ means a class label of crystal packing; ‘~~striketrough~~’ means the number of misclassified complexes in a branch; the format of each line is similar to that in Figure 7.5.

Similarly, distance-based contacts in Figure 7.5 also suggest two interesting rules. One is that only 83 biological binding have more than fifty-eight $(1.9, 3.3]$ N3H1_C4H3 contacts; the other is that in those interfaces which have no more than thirty-four $(1.9, 3.3]$ N3H1_C4H3 contacts, 281 crystal packing and 11 biological binding have no more than eight $(0.5, 1.9]$ C3(S2)H1_C3(S2)H1 contacts. These two rules are simple. However, the second rule has 11 false negative predictions, while the first rule is much harder to interpret according to our current biological knowledge.

Finally, we would like to note that β contacts are all in distance-based contacts, and then those contacts in the interesting rules of β contacts are in fact also in distance-based contacts. But distance-based contacts have so many non- β contacts, masking detection of critical specific atomic contacts in the interesting rules of β contacts.

7.4.2.3 Two Misclassified Examples in the Decision Tree of β Contacts

In the decision tree of β contacts, each of the two interesting rules, R1 and R2, has a misclassified interface. According to R1 in the first line of Figure 7.4, 82 biological interfaces have more than four ≤ 0.5 C3(S2)H0_C4H3 contacts, and only one out of 314 crystal packing interfaces satisfies R1 in the non-redundant dataset. This misclassified crystal packing interface is 1RB3 as shown in Figure 7.6(a). This interface has five ≤ 0.5 C3(S2)H0_C4H3 contacts. However, there are contradictory conclusions on this interface. On one hand, the authors who determined 1RB3 in PDB recommended it as a dimeric unit; it seems that the decision tree of β contacts shares the conclusion with the original authors of 1RB3. On the other hand, the non-redundant dataset labels 1RB3 as ‘crystal packing’; the existing classifiers in literature, such as NOXclass, DiMoVo and PISA, also predict it as a crystal packing. But 1RB3’s interface size is as small as $615 \text{ \AA}^2$, much smaller than the cut-off of interface ASA $856 \text{ \AA}^2$ suggested by the previous work [27] to distinguish crystal packing from biological binding. Meanwhile, almost all of the classifiers in literature have the same bias—they heavily rely on interface size. That is, 1RB3’s smaller interface size can easily mislead their predictions: these classifiers are more likely to have wrong predictions for 1RB3, and tend to consider 1RB3 as a crystal packing according to the same bias of 1RB3’s smaller interface size. Thus, these predictions provide nothing more than 1RB3’s smaller interface size. In summary, 1RB3 may be a potential dimer, just as R1 and the original authors suggest, but whether it is dimer is still an open problem until it can be verified in wet-lab experiments.

The second rule R2 is in the sixth line of Figure 7.4, which covers 251 crystal packing and only one biological binding in the non-redundant dataset. This biological binding is 3SDH whose interface is shown in Figure 7.6(b). This 3SDH interface has plenty of interfacial water molecules, and large-size non-standard residues as shown in Figure 7.6(b). These two factors are not evaluated in β contacts so far, which may be the reason why the decision tree of β contacts misclassifies 3SDH.

7.6.a 1RB3

7.6.b 3SDH

Figure 7.6: Two misclassified examples in the decision tree of β contacts (better viewed in color). (a) 1RB3 satisfies the biological rule R1. (b) the biological interface in 3SDH follows the crystal packing rule R2. In (a) and (b), residues labeled in black (chain a) and white (chain b) form an interface; interfacial waters whose spatial distance to both chains are less than 3.5 Å are in the red sphere view; non-standard residues are in the magenta sphere view; Carbon: green; Oxygen: red; Nitrogen: blue.

7.4.3 β Contacts Outperform Previous Methods in Classification of Crystal Contacts and Homodimers

7.4.3.1 Feature-Selection Classification Performance.

The feature-selection classification performance of β atomic contacts is evaluated against those achieved by DiMoVo [70], and Block’s method [107]. DiMoVo and the Block’s method used a similar backward-removal feature-selection method as described in Chapter 2; this process checks $\frac{\|F\|(\|F\|-1)}{2}$ potential subsets of features, while the feature-selection method in our approach, SVM-RFE, only checks $\|F\|$ potential subsets of features. The best classification performance of the Block’s method and the DiMoVo prediction are shown in Table 7.5.

In Table 7.5, our β contacts trained on the Ponstingl dataset have better classification performance than the best performance (accuracy 0.948) of the Block’s method. When the Block’s method uses SVM, its best accuracy is 0.919 with an RBF kernel, which is much less than our accuracy 0.959.

On the Bahadur dataset, DiMoVo’s performance is recalculated by using its recalls for homodimers and crystal packing in [70]. β contacts have at least comparable, if not better, performance with DiMoVo. However, our β ACV1a and β ACV3 on β contacts adopt a linear kernel which is simpler than the RBF kernel used in DiMoVo.

7.4.3.2 Classification Performance on the Independent Datasets.

Besides, the classification performance of β atomic contacts on the independent datasets is also compared with those achieved by other methods in literature, including DiMoVo, PISA, PITA [28], and NOXclass. Their classification results are shown in Table 7.6 where the performance of DiMoVo, PISA, PITA, and NOXclass on the independent datasets is recalculated by using the recall and specificity and the datasets in [70]. Here, the independent dataset has protein complexes whose proteins have less than 30% sequence similarity to those proteins of complexes in the training dataset.

Trained on the Ponstingl dataset, our β ACV3 of β contacts have accuracy 0.937, and MCC 0.844, which are much higher than those achieved by DiMoVo, PISA and PITA. For example, PITA achieves the best accuracy 0.894 and MCC 0.73 among DiMoVo, PISA and PITA; its accuracy is 4.3 percent points lower than β ACV3’s accuracy, and its MCC is 11.4 percent points lower than β ACV3’s MCC.

When β ACV1a and β ACV3 of β contacts are trained on the Bahadur dataset, they again achieve better performance than DiMoVo. Our β ACV3 has 0.966 accuracy, 3.7 percent points higher than DiMoVo’s, and β ACV3 has 0.902 MCC, 12.2 percent points higher than DiMoVo’s. MCC here is a better metric than accuracy to compare β contacts with DiMoVo, since the independent-testing dataset of Bahadur is quite unbalanced: crystal

packing is about four times larger than homodimers. Therefore, the great improvement of MCC suggests that β ACV3 is much better than DiMoVo to capture protein specific binding.

When the NOXclass dataset is the training dataset, β contacts have much better performance than NOXclass, although β contacts cannot achieve better performance than DiMoVo. A reason for this is that β contacts can easily distinguish crystal packing from homodimers in the NOXclass dataset. Removal of several features does not change training accuracy significantly, which misleads SVM and SVM-RFE into choosing all features or quite fewer features as the best feature set. But the samples in the NOXclass's independent dataset are much harder to distinguish.

In conclusion, β contacts demonstrate its superior classification power to the other methods in literature under non- β contact definitions. This partially, if not all, results from that the new β contact definition can capture specific binding patterns in homodimers and then benefit the classification of them from crystal packing.

7.4.4 β Atomic Contacts Can Identify Critical Specific Contacts: a Case Study on the HA1-2D1-Binding Dataset

Protein sequences in the datasets for the classification above have lower identity to avoid overfitting in performance evaluation. However in the HA1-2D1-binding dataset, proteins in two complexes are highly similar, but one belongs to crystal packing while the other is biological binding, clearly reflecting that biological interactions are specific. So, only a method which identifies specific binding in biological interactions can have a higher chance to correctly distinguish biological binding from false binding in the HA1-2D1-binding dataset. We use the HA1-2D1-binding dataset to further evaluate the classifiers, such as DiMoVo, NOXclass and PISA in literature, and our β ACV3 and dACV3 trained on the Ponstingl dataset. The prediction results are shown in Table 7.8.

In Table 7.8, it is not surprising that all of these methods except β atomic contacts misclassify almost all HA1-2D1-binding complexes as crystal packing. Two possible reasons can explain this result. One is that most of them use features extracted from distance-based contacts and do not highlight specific contacts in binding interfaces. The other is that most of them heavily rely on one distinguishing feature, interface size, to obtain correct predictions. However, two complexes of the HA1-2D1-binding dataset have similar protein sequences, resulting in quite similar size of binding interfaces as shown in Table 7.8; thus, these methods simply classify them into the same class, i.e., crystal packing here. But in fact, one of them is crystal packing and the other is biological binding.

In contrast, β atomic contacts have much better prediction as shown in Table 7.8. Our β ACV3 classifier does not classify the complexes in the HA1-2D1-binding dataset as

Table 7.8: Predictions on the HA1-2D1-binding dataset.

H1N1 Virus	Bind	Prediction					Interface (Å ²)
		dACV3	β ACV3	NOXclass	Dimovo	PISA	
A/South_Carolina/1/1918	True	False	True	False	False	False	900
A/California/04/2009	True	False	True	False	False	False	924
A/swine/Iowa/30	True	False	True	False	False	False	885
A/Puerto Rico/8/34	True	True	True	False	False	False	956
A/Brisbane/59/2007	False	False	False	False	False	False	814
A/x-31(H3N2)	False	False	False	False	False	False	787
A/duck/Singapore/3/97	False	False	False	False	False	False	913
B/Hongkong/8/73	False	False	False	False	False	False	849

dACV3 and β ACV3 are trained on Ponstingl.

the same type of binding, indicating that β contacts can capture the different preference of atomic contacts in false binding and true binding. Further, β contacts are able to correctly discriminate all of these complexes. This demonstrates that β atomic contacts can identify much more critical specific contacts in biological binding interfaces, although these interfaces have similar properties such as interface size to those false binding.

Finally, it is important to note that these complexes in the HA1-2D1-binding dataset are not homodimers, while all of the used classifiers are trained on homodimers and crystal packing. But both these biological complexes and homodimers belong to biological binding. A reliable classifier for identifying homodimers should capture some, if not all, sharing binding patterns between different types of biological binding. Meanwhile, whether a classifier can achieve correct predictions in Table 7.8 is also closely related to its prediction generalization—a classifier should not expect to have all possible scenarios in its training dataset. β ACV3’s performance in Table 7.8 indicates a good prediction generalization, demonstrating that our β contacts can more effectively highlight those common binding patterns in homodimers and other protein complexes than those non- β contacts.

In conclusion, β atomic contacts can identify critical specific contacts in binding interfaces which are hard to be detected by conventional methods based on distance-based atomic contacts.

7.5 Discussion and Future Improvement

7.5.1 Discussion on Infrequent and Redundant β Atomic Contact Patterns.

In the classification of β contacts, crystal packing is used as the contrast of biological interactions. Thus, it is natural to mine those specific contacts which frequently occur in biological interactions but infrequently, if not never, occur in crystal packing. These specific contacts give us an intuitive understanding of governing principles behind protein binding. If all specific contacts can be detected, these contacts as a whole should be a perfect definition for protein binding. In this work, β contacts do suggest several biologically meaningful rules in the decision tree—those hydrophobic contacts and hydrogen bonds truly rarely occur in crystal packing. Unfortunately, the decision tree only provides a few of those specific contacts. Many of important contacts to protein binding are not detected by machine-learning algorithms here, since detection of critical specific contacts is not a trivial task.

First of all, β atomic contacts only take up to 10% of distance-based contacts. Hence, β contacts have a fairly sparse matrix D where each row represents an interface while each column denotes an atomic pair. Here, ‘sparse’ means D has a lot of zero values. In the sparse matrix D , rarely-occurring atomic pairs are more likely to be considered as noise features by machine-learning algorithms—noise features are generally infrequent. One good example of these infrequent atomic pairs is disulfide bonds. Disulfide bonds cost a little less free energy than covalent bonds. The occurrence of disulfide bonds is closely related to biological binding: in our non-redundant dataset with 315 crystal packing and 144 homodimers, only 1 crystal packing and 9 dimeric interfaces have disulfide bonds (*this crystal packing with disulfide bonds is removed in our analysis, because there are contradictory conclusions on whether it is a homodimer*). So, the occurrence of disulfide bonds has high confidence to predict homodimers. However, disulfide bonds are much less frequent so that they are eliminated from the best feature set at the first several runs in SVM-RFE. This results from the facts that (i) disulfide bonds rarely occur and machine-learning algorithms, such as SVM, cannot distinguish this feature from noise features; and (ii) disulfide bonds may also be considered as redundant features. In other words, protein interactions which have disulfide bonds may also have other frequent atomic contacts, and then these interactions can be correctly distinguished by other more frequent features. Thus, disulfide bonds are not considered to be important in classification of homodimers and crystal packing. However, it does not mean disulfide bonds are not important in protein binding.

Similar to disulfide bonds, some other atomic contacts that appear frequently may also be considered as redundant features by machine-learning algorithms. The problem

here is that these redundant β atomic contact patterns may be of interest yet hardly be detected. This is due to that when β atomic contact patterns are used to identify homodimers from crystal packing: it is not necessary to detect all specific atomic contacts in homodimers to obtain a good prediction; employing only several important contacts is able to produce a fairly good classifier. This is supported by Figure 7.4 and 7.5 where most crystal packing and biological binding are grouped in several sub-branches. Meanwhile, some non-specific features in binding interfaces, such as interface size, are also helpful to distinguish homodimers from crystal packing. All of these indicate that only several of specific important contacts in protein binding, not all of them, are detected in this classification. In the future, we would like to evaluate β atomic contacts in other applications such as estimation of protein folding/binding free energy where critical specific contacts should be more important.

7.5.2 Future Improvement

Despite the promising results achieved by β ACV, one limitation of this representation is that it might not be able to discover complicated relation among atomic contacts, such as cooperativity. In fact, atomic contact graphs and their subgraphs are one of the best models of a protein tertiary/quaternary structure. An atomic contact subgraph can capture cooperativity among atomic contacts. For example, given two local structures: one has a hydrogen-bond network of hydrogen bonds between N_1 and O_1 as well as between N_1 and O_2 , while the other contains two individual hydrogen bonds between N_1 and O_1 as well as between N_2 and O_2 , where atoms N_1 and N_2 are hydrogen donors while atoms O_1 and O_2 are hydrogen acceptors; the hydrogen-bond network is believed to generate favorable electrostatic contribution to protein binding and to strongly stabilize protein complexes [151] than the two individual hydrogen bonds. Furthermore, an atomic contact subgraph can also find frequent atomic contact clusters which can contribute to protein binding greatly. For instance, several contacts involving C3H1 clearly demonstrate a π -involving contact. More cooperative atomic contact clusters can be found in atomic contact graphs. These cooperative clusters are hardly to be exactly detected in β atomic contact vectors of protein binding interfaces. But if atomic contact graphs are used to model protein 3D structures, at least two bottlenecks should be tackled in the future: (i) an efficient subgraph mining algorithm is necessary to find important frequent subgraphs—they may not be connected subgraphs; (ii) a much larger dataset is compulsory, since subgraph space is very large. If the learning dataset does not have enough protein structures, subgraphs generally have lower occurrence and become meaningless.

Meanwhile, β atomic contacts of water molecules and of non-standard amino acid residues are not considered in this work. This is due to the fact that crystal packing in the datasets generally contain no water information. These β atomic contact patterns will be evaluated in our future works.

7.6 Conclusion

The main contribution of this work is to propose the novel β atomic contacts to define critical specific contacts across protein binding interfaces (See Section 7.2.2). To evaluate the efficacy of the proposed β contacts, we design a classification scheme β ACV for classifying crystal packing and homodimers. We compare β ACV's classification performance with those achieved by the existing methods on the four datasets. The promising performance achieved by β ACV (See Subsection 7.4.3 and 7.4.4) demonstrates that β contacts can truly identify a compact set of critical specific contacts in protein binding interfaces which are only a small fraction of conventional distance-based contacts (See Subsection 7.4.1). In the future, β atomic contacts will also be used to define atomic contacts within protein tertiary/quaternary structures. The newly defined contacts can facilitate structural analysis on protein folding and binding.

Chapter 8

Prediction of Binding Free Energy and Binding Hot Spots by Integrating Water Exclusion into β Atomic Contacts

In Chapter 7, we propound the new definition, β contacts, to more precisely describe atomic organization of protein binding interfaces. After using β contacts to identify homodimers from crystal packing, we found that β contacts have better classification performance than distance-based contacts, although β contacts are only a small fraction of distance-based contacts.

Meanwhile, binding free energy and binding hot spots in protein interfaces provide basic knowledge to explore protein interactions and their functions in biology, and thus have attracted a lot of research. However, there are several limitations in existing prediction methods of binding free energy and binding hot spots, such as investigating ASA and atomic contacts separately, building a single prediction method for all 20 residues, and simply using distance-cutoff atomic contacts without considering their local atomic organization. Thus, the performance in existing works still has much room for improvement on the prediction of binding hot spots.

To tackle these above limitations, we would like to take both DWE and β contacts to define atomic contacts in protein binding interfaces. Then, we design $\beta\text{ACV}_{\text{ASA}}$ to linearly combine β contacts for the prediction of the change of binding free energy ($\Delta\Delta G$). $\beta\text{ACV}_{\text{ASA}}$ firstly uses atomic contact vector (ACV) of β contacts to represent an alanine mutation. As discussed in Chapter 7, β contacts can identify more critical atomic

contacts in protein binding interfaces than distance-cutoff contacts. Secondly, $\beta\text{ACV}_{\text{ASA}}$ is integrated by ASA under our DWE hypothesis in Chapters 3, and the contribution of a β contact of two atoms to protein binding is considered to be proportional to their buried ASA according to the water exclusion hypothesis of binding hot spots. Thirdly in particular, we build a $\beta\text{ACV}_{\text{ASA}}$ prediction for alanine mutations of polar residues and the other for non-polar residues, together called $\beta\text{ACV}_{\text{ASA}}^2$. Our empirical study on 396 mutations shows that $\beta\text{ACV}_{\text{ASA}}$ is superior to other methods, including Robetta, FoldX, HotPOINT, KFC2, an ACV method of β contacts without ASA integration, and an ACV_{ASA} method based on distance-cutoff contacts. This demonstrates that β contacts are more powerful than distance-cutoff contacts to predict $\Delta\Delta G$, although β contacts are only a small fraction of the latter contacts. It also reconfirms that water exclusion is a necessary condition of binding hot spots. Furthermore, $\beta\text{ACV}_{\text{ASA}}^2$ has better classification performance of binding hot spots with $\Delta\Delta G \geq 2$ kcal/mol, indicating that each residue should have their own way of contributing to protein binding.

8.1 Introduction

Protein binding free energy can precisely describe protein binding interfaces. Its accurate prediction is a fundamental problem in biology. It can serve as a cornerstone in many applications such as rational drug design and protein engineering. In a wet lab, residue contribution to binding free energy can be determined by mutation experiments. For example, alanine scanning mutagenesis [37] mutates each interfacial residue into an alanine, and measures the change of binding free energy, $\Delta\Delta G$, to quantify the contribution from the side chain of the mutated residue. As discussed in the previous chapters, it has been observed from such experiments that binding free energy is unevenly distributed in protein interfaces [38]. There are a small fraction of interfacial residues—the so-called binding hot spots—which contribute to protein binding greatly [38, 86], usually with $\Delta\Delta G \geq 2$ kcal/mol [36]. As conducting wet-lab experiments is time-consuming and expensive, reliable computational predictions have been actively explored as a complementary tool.

As discussed in Chapter 2, three kinds of existing computational works are closely related to the dissection of binding free energy and binding hot spots. One kind were designed to estimate $\Delta\Delta G$ after residue mutations [74, 75, 76, 77], another kind were proposed to qualitatively identify hot spots [79, 66], while the other kind endeavored to investigate the topological organization of binding hot spots [36, 52, 48].

Although extensive works have been done on the investigation of $\Delta\Delta G$ and binding hot spots, the prediction performance can be greatly improved. Those prediction methods have at least three major limitations. One is a single prediction method was developed

for all of 20 standard amino acids. However, different residues, such as polar and non-polar residues, usually have various contact preference; their contact behaviors should be unique. It is thus beneficial to have individual classifiers that are designed specifically for each type of residues [94]. Another limitation is that atomic contacts and their ASA, two distinguishing features for identifying hot spots, were mostly exploited separately. The other limitation is that the atomic contacts used were simply defined using a distance threshold without considering the local atomic organization of the contacts defined. As we discussed in Chapter 7, in this common distance-based definition, if the distance threshold is too large, e.g., larger than 6 Å, many atomic contacts defined between two atoms i and j actually accommodate an atom almost between them, resulting in smaller or no direct contact area between the two atoms i and j . It is highly questionable whether these interrupted contacts are important in protein binding, although they are predominant in the contacts defined according to previous works. Under this simple definition, many contacts are not fully understood yet, although hydrogen bonds, salt bridges and $\pi - \pi$ contacts are observed to be important features in hot spot prediction [79, 75, 76] or to closely relate to binding hot spots [66].

To overcome these above limitations, we propose $\beta\text{ACV}_{\text{ASA}}$ for the prediction of binding free energy and binding hot spots. $\beta\text{ACV}_{\text{ASA}}$ uses atomic contact vector (ACV) [26] and their surroundings to represent an alanine mutation. It has three main advantages. Firstly, $\beta\text{ACV}_{\text{ASA}}$ uses ACV of β contacts (called βACV vectors for short) instead of distance-based contacts. Secondly, βACV is integrated with ASA, named by $\beta\text{ACV}_{\text{ASA}}$, with characteristics of water exclusion hypothesis of binding hot spots. Thirdly, two $\beta\text{ACV}_{\text{ASA}}$ predictions are developed: one for mutations of polar residues and the other for mutations of non-polar residues. Finally, ridge regression is used to linearly combine these atomic contacts in $\beta\text{ACV}_{\text{ASA}}$ for predicting $\Delta\Delta G$ and binding hot spots.

In this work, $\beta\text{ACV}_{\text{ASA}}$ is evaluated on the dataset mainly from ASEdb with 396 alanine mutations. The evaluation includes comparisons of $\beta\text{ACV}_{\text{ASA}}$ with the following methods: i) an ACV method of distance-cutoff contacts to reveal the importance of β contacts to protein binding; (ii) a βACV method without ASA integration to verify whether water exclusion is necessary for binding hot spots; and (iii) several previous widely-used methods, such as Robetta, FoldX, HotPOINT and KFC2, to show the prediction capability of $\beta\text{ACV}_{\text{ASA}}$. The comparison results demonstrate that $\beta\text{ACV}_{\text{ASA}}$ achieves better performance than these other methods. This proves that (i) β contacts can provide a more precise model of atomic organization in protein binding than the definitions simply based on a distance cut-off, (ii) water exclusion is a necessary condition of a binding hot spot [36, 62], and (iii) polar residues and non-polar residues have their own unique contribution way to protein binding.

A web server of $\beta\text{ACV}_{\text{ASA}}$ is also available at http://155.69.2.25/liuqian/bacv_asa/index.py (visit from outside NTU), or at http://172.21.100.239/liuqian/bacv_asa/index.py (visit from inside NTU), or http://sunim1.sce.ntu.edu.sg/liuqian/bacv_asa/index.py.

8.2 Materials and Methods

8.2.1 A Dataset

A dataset, mostly collected from the ASEdb database [36], is used in the evaluation of $\beta\text{ACV}_{\text{ASA}}$ for the prediction of binding free energy of alanine mutations.

This dataset contains 22 protein complexes, as shown in Table C.1. All of them have quaternary structures in PDB and meet the following three requirements. Firstly, there is no nonredundancy among protein complexes. Given two interaction pairs with proteins A and B and with proteins C and D, the sequence identify is calculated by using BLAST with the default setting for A and C, A and D, B and C, and B and D, denoted by $S(A, C)$, $S(A, D)$, $S(B, C)$ and $S(B, D)$ respectively; these two interaction pairs are redundant if $S(A, C) \geq 40\%$ and $S(B, D) \geq 40\%$, or $S(A, D) \geq 40\%$ and $S(B, C) \geq 40\%$. In our dataset, two protein complexes are generally not redundant and, in redundant complexes, the similar proteins have different mutations. This requirement of sequence identity is enough, since we only use the knowledge from atomic contacts and ASA which are derived from a complex, instead of using sequence information such as conservation based on a single sequence. Secondly, only alanine mutations are considered. Thirdly, before mutations, the mutated atoms in a residue’s side chain should have at least one distance-cutoff atomic contact with the partner proteins. The mutated atoms are those atoms except N, CA, C, O and CB. Under the above requirements, our ASEdb dataset has 396 alanine mutations as shown in Table C.3; among them, 86 are binding hot spots with $\Delta\Delta G \geq 2$ kcal/mol.

8.2.2 β Atomic Contacts in Protein Binding Interfaces

We use a β atomic contact graph $b(p)$ to represent a given protein complex p , as defined in Chapter 7. To sharpen the difference of mutations with higher $\Delta\Delta G$ and with lower $\Delta\Delta G$, the angle threshold $\angle\beta$ in this study is set to 75° , whose forbidden region fr is larger than when $\angle\beta = 90^\circ$ as shown in Figure 7.2(e) and 7.2(f). Thus, β contacts under $\angle\beta = 75^\circ$ are stricter than those contacts under $\angle\beta = 90^\circ$.

We want note that in the above process, T_d is also set to $3.3 \text{ \AA}$. This threshold is an insensitive factor to β contacts in our dataset when it is large enough. For example, under $\angle\beta = 75^\circ$ and in the 396 mutations, there are 32 involving β contacts whose distances are between $2.8 \text{ \AA}$ and $3.3 \text{ \AA}$, 118 β contacts with distances between $1.9 \text{ \AA}$ and $2.8 \text{ \AA}$, and 5760 β contacts whose distance less than $1.9 \text{ \AA}$. Here, $1.9 \text{ \AA}$ is the radius of a water molecule plus $0.5 \text{ \AA}$. The latter contacts with small distance predominate in β contacts.

8.2.3 Our Prediction Method for Alanine Mutations

In our prediction method, we firstly use an atomic contact vector (ACV) [26] of β contacts to represent an alanine mutation in protein interfaces. Our proposed ACV is based on β contacts, called β ACV for short. Then, we integrate atomic ASAs into each atomic pair of this vector based on the observation of water exclusion, and design our method β ACV_{ASA}. Thirdly, we consider bound water molecules in interfaces and also add atomic environment of mutations into this vector for the prediction. Finally, we use ridge regression to learn the relation of β ACV_{ASAs} and their $\Delta\Delta G$.

8.2.3.1 β ACV Representation for Alanine Mutations in Binding Interfaces

To produce a β ACV vector for an alanine mutation mut_r^{ala} of residue r in a protein complex p , we first build the β atomic contact graph $b(p)$ for p . After that, we remove the coordinates of the mutated atoms in r and obtain a new quaternary structure after the mutation, called p_r^{mut} . Then, we produce another β atomic contact graph $b(p_r^{mut})$ for p_r^{mut} .

After this mut_r^{ala} mutation, some contacts in $b(p)$ may disappear in $b(p_r^{mut})$, called the mutated contacts, while another contacts not in $b(p)$ may newly form in $b(p_r^{mut})$, denoted as the new contacts. All of these contacts are represented by a β ACV vector for mut_r^{ala} . In a β ACV vector, all heavy atoms from 20 standard residues are grouped into 8 atomic types, as shown in Table C.2. A β ACV vector then has 36 possible pairs of atomic types. The value $v(Ti, Tj)$ of each element in β ACV with atomic types Ti and Tj is calculated using Equation 8.1.

$$v(Ti, Tj) = \sum_{m \in M, t_m = Ti} \sum_{n \in C1, t_n = Tj} (-1)^s \frac{1}{2 \times d_{(m,n)}^2} + \sum_{l \in M, t_l = Tj} \sum_{k \in C2, t_k = Ti} (-1)^s \frac{1}{2 \times d_{(l,k)}^2} \quad (8.1)$$

where m, n, l and k are atoms, t_* is the type of atom $*$, M is a set of atoms which have mutated or new contacts after this mut_r^{ala} mutation, $C1/C2$ is a set of atoms which have mutated or new contacts with atom m/l , $d(*, *)$ is the spatial distance between two atoms, and $s = 0$ if the contact is a mutated contact and $s = 1$ if the contact is a new contact. That is, other contacts are not used in β ACV if they are common in $b(p)$ and $b(p_r^{mut})$. Furthermore, we would like to note that alanine mutations of Ala here are assumed to have insignificant $\Delta\Delta G$, and alanine mutations of Gly are not considered here.

In this work, β ACV considers all mutated contacts and new contacts, including both across-interface contacts and those contacts from the same proteins or same biological

units. However, atomic contacts between nearby atoms are not used in Equation 8.1. The nearby atoms of a given atom i are those atoms that have no more than three covalent-bond steps from i . For example, suppose $i - j - k - l - m$, where $-$ indicates a covalent bond. From i , the covalent-bond step is 0 to i , is 1 to j , is 2 to k , is 3 to l and is 4 to m , respectively. Thus, i, j, k and l are nearby atoms of i , while m is not. In producing β ACV, even if i and its nearby atoms, for example k , have mutated or new contacts, these contacts are excluded from C in Equation 8.1. This is reasonable, since i and its nearby atoms have spatially close distance more likely due to the rigidity of their covalent bonds.

8.2.3.2 Integrating the Water Exclusion Hypothesis into β ACV

The solvent provides a compulsory surrounding of protein binding, but water exclusion is a necessary condition of binding hot spots [36, 62]. Few highly exposed residues contribute greatly to protein binding. Thus, we integrate ASA into β ACV and design β ACV_{ASA}.

Given a protein complex p , we use NACCESS [12] to produce ASA for atoms and residues both in a bound state and in an unbound state. For p in the bound state, we then define several ASA terms for an atom i using Equation 8.2, and for a residue R_i using Equation 8.3 and Equation 8.4.

$$E_i = \sqrt{\frac{ASA_i}{50}} \quad B_i = \max(0, 1 - E_i) \quad (8.2)$$

$$E_{R_i}^{bb} = \sqrt{\frac{ASA_{R_i}^{bb}}{\max(ASA_{R_i}^{bb})}} \quad B_{R_i}^{bb} = \max(0, 1 - E_{R_i}^{bb}) \quad (8.3)$$

$$E_{R_i}^{sc} = \sqrt{\frac{ASA_{R_i}^{sc}}{\max(ASA_{R_i}^{sc})}} \quad B_{R_i}^{sc} = \max(0, 1 - E_{R_i}^{sc}) \quad (8.4)$$

In Equation 8.2, ASA_i is accessible surface area of atom i , while E_i is its relative ASA and B_i is its relative buried ASA; in Equation 8.3 and Equation 8.4, $ASA_{R_i}^{bb}$ and $ASA_{R_i}^{sc}$ are accessible surface area of backbone atoms (i.e., bb) and of its side-chain atoms (i.e., sc) for a residue R_i , while $E_{R_i}^*$ and $B_{R_i}^*$ are the relative ASA and the relative buried ASA of $* \in \{bb, sc\}$. $\max(ASA_{R_i}^{bb})$ and $\max(ASA_{R_i}^{sc})$ are maximum ASA of backbone atoms and of side-chain atoms for the residue type of R_i ; they are calculated in a triplet of ALA- R_i -ALA by NACCESS. Backbone atoms and side-chain atoms are defined in the same way as those in [12]. In Equation 8.2, 50 is roughly half of NACCESS-calculated ASA of a single water molecule without any neighbor atoms and hydrogens; the ASA of this molecule is actually the surface area of a sphere with a radius 2.8 Å, that is, $98.47 = 4 \times 3.14 \times 2.8^2 \text{ Å}^2$.

Then given an atom i , its local ASA E_i^{loc} and local buried ASA B_i^{loc} are defined using Equation 8.5 where a cluster of nearby atoms together, instead of a single atom, are assumed to cooperatively contribute to protein binding.

$$\begin{aligned} E_i^{loc} &= \begin{cases} E_i \times E_{R_i}^{bb} & \text{if } i \text{ is a backbone atom of a residue } R_i \\ E_i \times E_{R_i}^{sc} & \text{if } i \text{ is a side-chain atom of a residue } R_i \end{cases} \\ B_i^{loc} &= \begin{cases} B_i \times B_{R_i}^{bb} & \text{if } i \text{ is a backbone atom of a residue } R_i \\ B_i \times B_{R_i}^{sc} & \text{if } i \text{ is a side-chain atom of a residue } R_i \end{cases} \end{aligned} \quad (8.5)$$

With the mutated contacts and new contacts of an alanine mutation mut_r^{ala} , we use Equation 8.6 instead of Equation 8.1 to calculate the value $v(Ti, Tj)$ of each element in βACV_{ASA} for the water exclusion integration.

$$\begin{aligned} v(Ti, Tj) = & \sum_{m \in M, t_m = Ti} \sum_{n \in C1, t_n = Tj} (-1)^s \frac{B_m^{loc} \times B_n^{loc}}{2 \times d_{(m,n)}^2} + \\ & \sum_{l \in M, t_l = Tj} \sum_{k \in C2, t_k = Ti} (-1)^s \frac{B_l^{loc} \times B_k^{loc}}{2 \times d_{(l,k)}^2} \end{aligned} \quad (8.6)$$

where t_* , $C1/C2$ and s have the same meaning as those in Equation 8.1.

8.2.3.3 Two Additional Factors in Protein Binding Interfaces

Bound Water Molecules in Protein Interfaces Protein folding and binding occur in a solvent environment *in vivo*. It is clear that water molecules are involved in protein binding and form a compulsory part of protein interfaces. In this work, a water molecule is considered as a part of a binding interface if (i) it has at least 3 potential hydrogen-bond contacts, or (ii) it has 2 potential hydrogen-bond contacts and also has at least another 2 contacts with spatial distances less than 4 Å. A potential hydrogen-bond contact should have a spatial distance less than 3.2 Å between a nitrogen atom and an oxygen atom. Under this requirement, these water molecules are heavily involved in protein folding and binding such that they are an integral part of proteins, especially called bound water molecules. In βACV , bound water molecules are grouped into the Oxygen type with more than one hydrogen as shown in Table C.2. But the contacts between any two water molecules are not considered.

The Surrounding of Mutated Residues Finally, the surrounding information of mut_r^{ala} is appended to the βACV representation. The surrounding information of mut_r^{ala} are from those atoms which are close to the mutated atoms of mut_r^{ala} . Assume that S is a set of atoms which have β contacts with the mutated atoms when $\angle\beta = 90^\circ$. For each atom in S including the mutated atoms, its nearby atoms are added into S . Then,

an atomic vector is used to represent those atoms in S in the bound state. This atomic vector has 8 atomic types, similar to β ACV. The value of its element Tk is calculated using $v(Tk)^b$ in Equation 8.7. Similarly, $v(Tk)^u$ in Equation 8.7 is used to calculate another atomic vector for representing the atoms in S in the unbound state.

$$\begin{aligned} v(Tk)^b &= \sum_{j \in S}^{t_j=Tk} E_j^{loc} \\ v(Tk)^u &= \sum_{j \in S}^{t_j=Tk} E_j^{loc}(u) \end{aligned} \quad (8.7)$$

where $E_j^{loc}(u)$ is the relative local ASA of atom j in the unbound state. Water molecules are not considered here. Thus, each β ACV vector is added by another 16 atomic types for representing surrounding information of the mutated contacts, and both β ACV and β ACV_{ASA} have a vector representation with 52 elements.

Finally, to compare β contacts with distance-based contacts in the prediction of binding free energy, ACV_{ASA} is produced based on distance-based atomic contacts in a similar way to β ACV_{ASA}. To further show the importance of β contacts in protein binding interfaces, non β ACV_{ASA} is also constructed for alanine mutations when $\angle\beta = 90^\circ$; in non β ACV_{ASA}, the values of its elements are the difference of the values of the 52 corresponding elements in ACV_{ASA} and in β ACV_{ASA}.

8.2.4 Separate Prediction Methods for Polar Residues and Non-Polar Residues

In [94], Lise *et al.* built two individual hot spot classifiers for those mutations involving either an arginine or a glutamic residue, and improved their overall classification performance. It seems that the 20 standard amino acids have different physicochemical properties, and they may have their unique way of contributing to protein binding. Hence, it may be better to construct individual hot spot classifiers for each type of residues when enough mutations are available.

However, we do not have enough binding hot spot information for most types of amino acids. Thus, we aggregate the 20 standard amino acids into polar residues and non-polar residues, and build two prediction methods, one for mutations of polar residues and the other for mutations of non-polar residues. In this work, polar residues are Thr, Ser, His, Glu, Gln, Asp, Asn, Lys, Arg, and Pro, while non-polar residues are Ile, Val, Leu, Phe, Cys, Met, Trp, and Tyr.

Specifically, β ACV_{ASA} with the above setting is used for predicting mutation effect of polar residues. Another β ACV_{ASA} method for mutations of non-polar residues only has

two differences. One is that the angle threshold of β contacts is relaxed from $\angle\beta = 75^\circ$ to $\angle\beta = 90^\circ$. This is because heavy atoms of non-polar residues are mostly Carbon atoms, and their contacts are not as specific as contacts between polar atoms such as Oxygen and Nitrogen. The other difference is that for an alanine mutation of non-polar residues, β contacts within protein tertiary structures are used to produce a βACV vector, and those contacts across interfaces to build another βACV vector. Thus, $\beta\text{ACV}_{\text{ASA}}$ for mutations of non-polar residues has 72 pairs of atomic types. The reason for this is that the side chains of non-polar residues tend to repel water molecules, and their carbon atoms prefer to well-organized clusters to free movement in their tertiary structures. Thus, their mutated contacts within tertiary structures will have different contact behaviors from across-interface contacts. This is different from polar residues: polar residues have the side chains with Oxygen and Nitrogen which prefer free movement with bulk solvent if exposed in tertiary structures; thus, all of the contacts with their side chains should form together upon the formation of protein complexes, no matter across-interface contacts or contacts within a single biological unit.

Finally, these two $\beta\text{ACV}_{\text{ASA}}$ predictions together are collectively called $\beta\text{ACV}_{\text{ASA}}^2$ for short. For a better review, the evaluated methods, $\beta\text{ACV}_{\text{ASA}}$ and its compared methods, are summarized in Table 8.1.

Table 8.1: **Summary of the evaluated methods: $\beta\text{ACV}_{\text{ASA}}$ and its compared methods.**

Methods	Description (of representation for an alanine mutation)
$\beta\text{ACV}_{\text{ASA}}$	An ACV of β contacts with ASA integration
ACV_{ASA}	An ACV of distance-based contacts with ASA integration
$\text{non}\beta\text{ACV}_{\text{ASA}}$	The difference of $\beta\text{ACV}_{\text{ASA}}$ and ACV_{ASA}
βACV	An ACV of β contacts without ASA integration
$\beta\text{ACV}_{\text{ASA}}^2$	two $\beta\text{ACV}_{\text{ASA}}$, one for mutations of polar residues and the other for non-polar residues

8.2.5 Ridge Regression of $\beta\text{ACV}_{\text{ASA}}$ and Evaluation Measures

Ridge regression in Matlab is used to build the relation of βACV and $\Delta\Delta G$. In this regression, values in each column are normalized on a training data. Ridge regression minimizes average square error SE between the experimental binding free energy ($\Delta\Delta G_i^e$) and the predicted binding free energy $\Delta\Delta G_i^p$ in a training data with N mutations. SE is calculated by $SE = \frac{\sum_i (\Delta\Delta G_i^e - \Delta\Delta G_i^p)^2}{N-1}$.

In evaluation with ASEdb, we use leave-one-out cross-validation for performance evaluation, and then calculate correlation coefficient R and average standard deviation $\delta = \sqrt{SE}$. In this cross-validation, there are 1 outlier prediction by $\beta\text{ACV}_{\text{ASA}}$, 1 outlier

by β ACV, and 9 outliers by β ACV $^2_{ASA}$ in the whole ASEdb dataset with 396 mutations. These outliers have less than -3 kcal/mol predicted $\Delta\Delta G$, or more than 11 kcal/mol predicted $\Delta\Delta G$, as shown in Table C.3. This may be due to limited alanine mutations with higher $\Delta\Delta G$ in ASEdb. These outlier predictions are excluded in the regression evaluation.

After regression, β ACV $_{ASA}$ is also evaluated using the classification performance for binding hot spots. The predicted hot spots are defined using ≥ 2 kcal/mol as the threshold, same as the true hot spot definition. After taking hot spots as the positive samples while non-hot spots as the negative samples, classification performance is measured by using the metrics in Equation 4.10, such as *precision*, *recall*, *accuracy*, *F1* and *AUC*.

8.3 Results and Discussion

8.3.1 β Contacts are Better than Distance-Based Contacts to Predict Binding Free Energy

First of all, the prediction performance of β ACV $_{ASA}$ is compared with ACV $_{ASA}$ and non β ACV $_{ASA}$ to show the important role of β contacts in the prediction of binding free energy for alanine mutations. The prediction results are shown in Figure 8.1.a, 8.1.b and 8.1.c, respectively.

Table 8.2: Predictions of binding hot spots by different methods. **AUC in bold** is calculated using ‘perfcurve’ function in MATLAB

Methods	Precision	Recall	F1	Accuracy	AUC
β ACV $_{ASA}$	0.615	0.593	0.604	0.830	0.852
ACV $_{ASA}$	0.526	0.477	0.500	0.793	0.773
non β ACV $_{ASA}$	0.513	0.454	0.482	0.788	0.772
β ACV	0.564	0.616	0.589	0.813	0.834
β ACV $^2_{ASA}$	0.663	0.679	0.671	0.855	0.821
FoldX	0.400	0.488	0.440	0.730	0.706
Robetta	0.526	0.465	0.494	0.793	0.775
HotPOINT	0.439	0.547	0.487	0.750	0.676
KFC2a	0.563	0.419	0.480	0.803	0.664
KFC2b	0.522	0.547	0.534	0.793	0.7

It can be seen from Figure 8.1.a and 8.1.b that β ACV $_{ASA}$ has better $\Delta\Delta G$ prediction according to both correlation coefficient R and average standard deviation δ , although β contacts are only a small fraction of distance-based contacts used in ACV $_{ASA}$. For example, in the 22 ASEdb protein complexes, there are 54,286 distance-based contacts,

9,830 β contacts when $\angle\beta = 90^\circ$, and 4,096 β contacts when $\angle\beta = 75^\circ$ which are used in βACV_{ASA} . That is, βACV_{ASA} only uses 7.55% atomic contacts of distance-based contacts, but it achieves better prediction performance.

The comparison between βACV_{ASA} and $\text{non}\beta\text{ACV}_{ASA}$ can further suggests that β contacts are crucial in $\Delta\Delta G$ prediction in Figure 8.1.a and 8.1.c. In Figure 8.1.c, $\text{non}\beta\text{ACV}_{ASA}$ has much lower R , 0.469, and higher δ , 1.486, than βACV_{ASA} . Please note that $\text{non}\beta\text{ACV}_{ASA}$ uses non- β contacts when $\angle\beta = 90^\circ$, which are 81.8% distance-based contacts.

In ASEdb, a lot of alanine mutations are not binding hot spots with small $\Delta\Delta G$, i.e., <2 kcal/mol; these mutations greatly affect the calculation of R and δ . Meanwhile, the prediction of residue mutations with high $\Delta\Delta G$ is more important. Thus, the prediction performance of binding hot spots is also calculated. The results are shown in Table 8.2. We would like to note that F1 is not the objective function in the regression process.

In Table 8.2 among βACV_{ASA} , ACV_{ASA} and $\text{non}\beta\text{ACV}_{ASA}$, βACV_{ASA} has the highest precision, recall, F1 and accuracy, while $\text{non}\beta\text{ACV}_{ASA}$ has the lowest. For example, βACV_{ASA} 's F1 is 0.604, 0.122 higher than $\text{non}\beta\text{ACV}_{ASA}$'s F1. However, ACV_{ASA} and $\text{non}\beta\text{ACV}_{ASA}$ show quite similar performances. This should be because non- β contacts generally are dominant in distance-cutoff contacts used in ACV_{ASA} and $\text{non}\beta\text{ACV}_{ASA}$.

Paired t-test in MATLAB is also used to test the difference among Figure 2(a)-2(c) for the predictions of true hot spots. This test shows that the p-value of the difference between Figure 2(a) and 2(b), between Figure 2(a) and 2(c), and between Figure 2(b) and 2(c) are 0.016, 0.108 and 0.798, respectively. Figure 2(a) of βACV_{ASA} predictions is more significantly different from the other two figures, while Figure 2(b) of ACV_{ASA} predictions and Figure 2(c) of $\text{non}\beta\text{ACV}_{ASA}$ predictions have insignificant difference.

8.3.2 Water Exclusion is a Necessary Condition of Binding Hot Spots

Previous works have reported that water exclusion is a necessary condition of binding hot spots [36, 62]. In βACV_{ASA} , ASA is integrated based on this observation. To investigate the role of water exclusion in the prediction of binding free energy, βACV without ASA integration is also designed. Its prediction performance is shown in Figure 8.1.d and in Table 8.2. It can be seen from Figure 8.1.d and Figure 8.1.a that βACV_{ASA} has slightly better performance than βACV , which is also supported by the worse F1 of βACV in Table 8.2. This verifies that water exclusion is a necessary condition of binding hot spots.

8.3.3 Prediction Results for Polar and Non-Polar Residues Separately

In protein folding and binding, polar and non-polar residues should have different contact behaviors. Thus, we build two individual βACV_{ASA} prediction methods, each for one of them. These two methods together are called βACV_{ASA}^2 whose prediction performance is shown in Table 8.2 and Figure 8.2.a.

It can be seen from Table 8.2 that βACV_{ASA}^2 achieves much better prediction performance of binding hot spots—its F1 is 0.671, much higher than the other methods. However, compared with βACV_{ASA} , Figure 8.2.a indicates that βACV_{ASA}^2 has worse R and δ . The worse R and δ is due to the limited mutations compared with the number of elements in βACV_{ASA} . This dilemma becomes worse after ASEdb is split into two groups for polar residues and non-polar residues, especially for non-polar residues which are represented by ‘o’ in Figure 8.2.a.

8.3.4 Our Method βACV_{ASA} is Superior to Several Widely-used Methods

The prediction ability of our βACV_{ASA} is finally evaluated against two gold standard methods FoldX [74, 75] and Robetta [76]. Among 396 mutations on ASEdb, FoldX only considers 378 mutations, and Robetta only uses 338 mutations. The prediction performance of FoldX and Robetta are much worse than our βACV_{ASA} according to Figure 8.2.b and 8.2.c. These two methods have R 0.324 and 0.485, much smaller than βACV_{ASA} ’s R 0.569; their δ are 1.788 and 1.554, much larger than βACV_{ASA} ’s δ 1.349. We want to note that the non-detected mutations by FoldX and Robetta still can significantly affect protein binding. For example, Robetta has no prediction for both Pro mutation in position 306 of Chain B in 1A22 and Glu mutation in position 80 of Chain M in 1DX5, and these two mutations can be considered to have small predicted $\Delta\Delta G$. But Pro mutation has $\Delta\Delta G$ 3.31kcal/mol, while Glu mutation has $\Delta\Delta G$ 3.4 kcal/mol. Hence, Robetta and FoldX will have lower regression performances if evaluated on 396 mutations.

Evaluated on the 396 mutations in ASEdb, the classification performance of binding hot spots is shown in Table 8.2 by βACV_{ASA} , FoldX and Robetta. FoldX’s F1 is 0.44, while Robetta’s F1 is 0.494; both are much worse than βACV_{ASA} ’s F1 0.604 and βACV_{ASA}^2 ’s F1 0.671. Again, βACV_{ASA} is proved to be superior to FoldX and Robetta. Also from Table 8.2, our method has better classification performance than HotPOINT [82] and KFC2 [152].

We also measure the classification performance of βACV_{ASA} on the 265 mutations used by Z-score in Chapter 4; the result is shown in Table 8.3. According to F1 in Table 8.3, βACV_{ASA} outperforms Z-score.

Table 8.3: Predictions of binding hot spots by βACV_{ASA} and Z-score

	Methods	Precision	Recall	F1
265 mutations	βACV_{ASA}	0.667	0.603	0.633
	Z-score	0.470	0.820	0.590

8.3.4.1 An example of βACV_{ASA} Predictions

The hot spot predictions in 3HFM by βACV_{ASA} , FoldX and Robetta are presented as a case study. This 3HFM complex is about an antibody-antigen binding of HyHEL-10 and hen egg white lysozyme. According to ASEdb, this complex has 25 alanine mutations where 11 have $\Delta\Delta G$ more than 2 kcal/mol.

As shown in Figure 8.3, our βACV_{ASA} correctly identifies 9 binding hot spots with 0.818 recall, and has 3 false positive predictions with 0.75 precision; this gives 0.783 F1. In contrast, FoldX has only one hot spot prediction which is correct with 0.091 recall, while Robetta has 0.455 recall (5 out of 11) and 0.833 precision (5 out of 6), resulting in 0.588 F1. Both have lower prediction performance than our βACV_{ASA} . Meanwhile, the four uniquely correct predictions by βACV_{ASA} also have larger $\Delta\Delta G$, such as Trp in position 98 of Chain H and Tyr in position 50 of Chain L. Please refer to Table C.3 for detail.

8.4 Conclusion

In this chapter, we have proposed βACV_{ASA} to predict binding free energy and binding hot spots, which integrates water exclusion into β contacts (See Subsection 8.2.3). In particular, we proposed two βACV_{ASA} s, one for mutations of polar residues and the other for mutations of non-polar residues (See Subsection 8.2.4). In the evaluation on 396 alanine mutations in ASEdb, βACV_{ASA} outperforms the other methods (See Subsection 8.3). This shows that β contacts are truly better than distance-based contacts to model protein quaternary structures for protein binding and water exclusion is a necessary condition of binding hot spots.

8.1.a $\Delta\Delta G$ prediction by βACV_{ASA}

8.1.b $\Delta\Delta G$ prediction by ACV_{ASA} based on distance-based contacts

8.1.c $\Delta\Delta G$ prediction by $\text{non}\beta\text{ACV}_{ASA}$ based on non- β contacts

8.1.d $\Delta\Delta G$ prediction by βACV without ASA integration

Figure 8.1: $\Delta\Delta G$ predictions by different ACV methods. In (a)-(d), ‘o’ represents non-polar residues, while ‘x’ represents polar residues. R is correlation coefficient, while δ is average standard deviation. Two diagonal red lines represent $y = x \pm 1.5$.

8.2.a $\Delta\Delta G$ prediction by βACV^2_{ASA}

8.2.b $\Delta\Delta G$ prediction by FoldX

8.2.c $\Delta\Delta G$ prediction by Robetta

Figure 8.2: $\Delta\Delta G$ predictions by βACV^2_{ASA} , FoldX and Robetta. In (a)-(c), 'o', 'x', R , δ and two diagonal red lines have the same meaning as those in Figure 8.1.

8.3.a Prediction for chain H and L

8.3.b Prediction for chain Y

Figure 8.3: **An example of βACV_{ASA} predictions in 3HFM.** In (a) and (b), true predicted hot spots are in magenta, false predicted non-hot spots in yellow, false predicted hot spots in orange, while true predicted non-hot spots in cyan; Other residues: Carbon in green, Oxygen in red, and Nitrogen in blue.

Chapter 9

Structural Analysis of the Hot Spots in the binding between H1N1 HA and the 2D1 Antibody: Do Mutations of H1N1 from 1918 to 2009 Affect Much on This Binding?

In 2009, the H1N1 influenza A caused worldwide mortality which has stimulated a new surge of research on H1N1 viruses. An epitope conservation has been learned in the HA1 protein of 2009 H1N1 that allows antibodies to cross-neutralize both 1918 H1N1 and 2009 H1N1. However, few works have thoroughly studied the binding hot spots in these two antigen-antibody interfaces which are responsible for the antibody cross neutralization.

In this thesis, from the chapters 3 to 8, we propose the DWE hypothesis to model and investigate protein binding interfaces and their binding hot spots at both residue and atomic levels. In particular, we focus on the analysis and prediction of binding hot spots and the discrimination of biological protein interactions from crystal packing.

Thus in this chapter, we would like to take predictive methods to understand the binding behaviors between the 2D1 antibody and H1N1 influenza A viruses in 1918, 2007 and 2009. Besides several methods in literature, these used methods include the Z-score method in Chapter 4, the β ACV classifier in Chapter 7 and the β ACV_{ASA} estimation method of binding free energy in Chapter 8. In particular, we identify binding hot spots at the epitope sites of the HA1 proteins for 1918 H1N1 and 2009 H1N1, and hot spots at the paratope sites of the 2D1 antibody. We find that the six mutations at the HA1's

epitope from 1918 H1N1 to 2009 H1N1 should not harm their binding to 2D1. Instead, the change of binding free energy on the whole exhibits an increased tendency after these mutations, making the binding stronger. This is consistent with the observation that the 1918 H1N1 neutralizing antibody can cross-react with 2009 H1N1.

Further, we identified three distinguished hot spot residues, including Lys¹⁶⁶, common between the two epitopes of 1918 H1N1 and 2009 H1N1. These common hot spots again can explain why 2D1 cross-reacted. We believe that these hot spot residues are mutation candidates which may help H1N1 viruses evade the immune system. We also identified eight residues at the paratope site of 2D1, five from its heavy chain and three from its light chain, that are predicted to be energetically important in the HA1 recognition.

Meanwhile, 2D1 binding to 1918 H1N1 and 2009 H1N1 is predicted as biological interactions by β ACV in Chapter 7, while 2D1-2007 H1N1 binding is classified as a false interaction; these classifications are consistent with the observation that 2D1 can neutralize 1918 H1N1 and 2009 H1N1 and cannot identify 2007 H1N1. Thus, these structural analyses and the identification of these hot spot residues are potentially useful to fight against H1N1 viruses.

9.1 Introduction

The H1N1 influenza A caused two notable pandemics with substantial mortality in 1918 and 2009. Fortunately, it has been found that some antibodies can work against the Hemagglutinin (HA) proteins in these two pandemics [144]. HA is a homotrimeric glycoprotein. HA monomers are synthesized as precursors that are then cleaved into two proteins, HA1 and HA2, which form the major surface proteins of influenza A viruses. The infection is started by the binding of HA proteins to the sialic-acid-containing receptors of target cells and by fusing the viral membrane with the endosomal membrane of the target cells. The viral genome enters and infects the target cells after the binding. So, inhibiting this binding by antibodies is an important way against flu. Previous works have learned that there is an epitope (binding site) conservation that exists between the 1918 and 2009 H1N1 HA proteins [153, 144]. Such epitope conservation enables the older population to avoid infection from 2009 H1N1 because their pre-existing immunity against 1918 H1N1 can neutralize the 2009 H1N1 HA proteins. Thus, studies on these antibody-HA binding interfaces are crucial to understand how the antibodies recognize the antigens. However, there are few studies on the energetic importance of the binding residues in the HA1 protein in complex with the 2D1 antibody.

We take predictive and comparative methods to examine the interfaces between the 2D1 antibody and the HA1 proteins of 1918 and 2009 H1N1, and to investigate an assumed 2D1 binding to the seasonal influenza virus A/Brisbane/59/2007 to understand

why 2D1 did not bind to the 2007 strain [147, 144]. This 2D1 antibody is a monoclonal antibody from a survivor of the 1918 Spanish influenza [146], which is believed to bind to HA1s in both of 1918 and 2009 H1N1. Of particular interests, we identify *binding hot spot residues* from the above mentioned three antibody-antigen interfaces, since binding hot spots contribute most to binding free energy [38, 36] and their mutations of hot spots—e.g., alanine mutations—can reduce binding affinity remarkably [38].

We address the problems whether the interfacial mutations from 1918 H1N1’s HA1 to 2009’s are hot spot residues and whether these mutations make the binding stronger with 2D1. We explain how the computational methods find those antigenic residues that are energetically important in the antibody binding, such as Asn¹²⁹, Lys¹⁵⁷ and Lys¹⁶⁶. These three hot spot residues are actually common between the 1918 and 2009 epitopes of HA1. Their mutations may make the 2D1-antibody binding ineffective. Therefore, they are mutation candidates which may help H1N1 viruses evade the immune system. We also describe and characterize hot spot residues at the paratope site of the 2D1 antibody—e.g., Asp⁵² and Arg⁹⁷ from the heavy chain and Asn³¹, Trp⁹¹ and Asp⁹³ from the light chain. Knowledge gained from these binding hot spot studies can be useful to fight against H1N1 viruses in future.

9.2 Methods

9.2.1 The HA-2D1 Binding Structures

We retrieve from the PDB entry 3LZF the crystal structure of the 2D1 antibody binding to the HA proteins of 1918 H1N1 (1918HA1), and from 3LZG the atomic coordinates of the HA proteins of the 2009 H1N1 A/California/04/2009 [144] (2009HA1). The structure information of the HA1 proteins of A/Brisbane/59/2007 (2007HA1) is taken from [148]. Our comparative analysis is on the three interfaces: the interface between 2D1 and 1918HA1 (2D1-1918HA1), the interface between 2D1 and 2009HA1 (2D1-2009HA1), and an assumed artificial interface between 2D1 and 2007HA1 (2D1-2007HA1).

We use MAFFT to align the three HA1 sequences to examine the specific mutations among 1918HA1, 2009HA1 and 2007HA1. We also take the steps in Section 7.3.1.1 of Chapter 7 to produce the structures of the 2D1-2009HA1 (and 2007HA1) binding.

9.2.2 Computational Methods for Predicting Hot Spots

Binding hot spot residues can be predicted by computational methods such as by Robetta [76], FoldX [75], KFC [79], GCR [64] and a Z-score method in Chapter 4. These methods are applied in the three interfaces—2D1-1918HA1, 2D1-2007HA1 and 2D1-2009HA1. Default parameters are used for the Robetta and KFC web servers and for the FoldX software. Since Robetta and FoldX estimate $\Delta\Delta G$, we are interested in those residues whose alanine mutation results in $\Delta\Delta G \geq 1$ kcal/mol.

9.2.2.1 A Meta-learning Approach to Combine the Computational Methods for Predicting Hot Spots

Then, we apply a meta-learning approach [126] by combining the Z-score method with the other methods. The reason is that the Z-score method has a very high recall with low precision rate; however, the other methods generally have a low recall but a high precision rate. Therefore, in this work, we are interested in those hot spot residues that are predicted by Z-score and are also confirmed by at least one of the other methods (Robetta, FoldX or KFC). Meanwhile, we also trust with high confidence that non-hot spot residues predicted by Z-score generally have insignificant contribution to the binding. The hot spot residues which are predicted by a single method only are considered having intermediate contribution to binding.

9.3 Results

Firstly, we use β ACV in Chapter 7 to identify the interactions types of the 2D1 binding to 1918HA1, 2009HA1 and 2007HA1. As shown in Table 7.8, our β ACV3 classifies the first two interactions as true biological binding and the last as false binding; this is consistent with the observation that 2D1 can cross-react with 1918 H1N1 and 2009 H1N1 [144, 147], and cannot with 2007 H1N1 [144].

Secondly, we show the sequence alignment among 1918HA1, 2007HA1 and 2009HA1 in Figure 9.1(a). There are a total of six interfacial mutations between 1918HA1 and 2009HA1, namely E131D, T133N, S159N, V169I, N171D, and T242K. The structural alignment of 1918HA1 and 2009HA1 between their interfacial segments is shown in Figure 9.1(c) where the structural match—based on the C α s of these interfacial residues—has a RMSD of 0.725 Å. Previous works have reported that protein sequences with about 50% identity or above in crystallographic models can differ by about 1 Å RMSD, while proteins in NMR models can have even larger deviations [154, 155]. In some cases, sequences with more than 95% identity can also have an interface RMSD up to about 1.2 Å [156]. Thus the small 0.725 Å RMSD suggests that the interfacial segments of the two HA1 proteins have a very good match. This indicates that the mutations from 1918HA1 to 2009HA1 result in little structural change at the binding site, and that the computationally produced binding structure of 2009HA1 and the 2D1 antibody may not have any big deviation from the real one.

9.3.1 Energy Change Tendency of the Six Mutations

	126	157	195
3LZF1918	FPKTS ^W PNHETTKGVTAACSYAGASSFYRNLLWLT ^K GGSSYPKLSKSYVNNKGKEVLVL	WGVHHPPTGTDQQSL ^Y QNA	DAYVSVGSSKYNRRFTPEIARPKVRD
3LZG2009	FPKTS ^W PNHDSNKGVTAAACPHAGAKSFYKNLIWL ^V KKGNSYPKLSKSYINDKGKEVLVL	WGIHHPSTSADQQSL ^Y QNA	DYVVFVGSSRYSKKPKPEIAIRPKVRD
Bris2007	FPKES ^S WPNHTVT-GVSASCSHNGESSFYRNLLWLT ^G KNGLYPNLSKYANNKEKEVLVL	WGVHHPNIGNQKAL ^Y HT	ENAYVSVVSSHYSRKFTPEIAKRPKVRD
	242		
3LZF1918	QAGRMNYYWTLLEPGDTITFEATGNLIAPWYAFALNRGSGGIITSDAPVHDCNTKCQTP	HGAINSSLPFQNIH	PVTIGECPKYVRSTKLRMATGLRNIPSIQSR
3LZG2009	QEGRMNYWTLVEPGDKITFEATGNLVVPYAFAMERNAGSGIISDTPVHDCNTTCQTP	KGAINSSLPFQNIH	PITIGCKPKYVKSTKLRATGLRNIPSIQSR
Bris2007	QEGRINYYWTLLEPGDTIIFANGNLIAPRYAFALSRGFGSGIINSNAPMDKCDACQTP	QGAINSSLPFQNVH	PVTIGECPKYVRSKLRMTGLRNIPSIQSR

9.1.a

9.1.b

9.1.c

Figure 9.1: The epitopes of HA1 in 1918, in 2009 and in 2007, and their sequence and structural alignment (better viewed in color). (a) The sequence alignment between 1918HA1(3LZF1918), 2009HA1(3LZG2009) and 2007HA1(Bris2007). Interface residues are in yellow, and the positions are in accordance to the 1918HA1 numbering. (b) The binding interface of 2D1-1918HA1. The HA1 epitope is in cyan and in sphere view, and the antibody paratope is in yellow and in stick view. (c) The aligned structure of the epitopes between 1918HA1 (cyan) and 2009HA1 (green). In (b) and (c), residues in magenta are the mutations from 1918HA1 to 2009HA1.

Using the Z-score method, three of the six mutations—T133N, S159N and N171D—are predicted as non-hot spot residues in both 1918HA1 and 2009HA1. Two mutations—V169I, and T242K—are believed to contribute, though slightly, to the antibody binding in 2009HA1 only after the mutations. They may be newly formed hot spots in the epitope of 2009HA1 after the mutations. The remaining one of the six mutations—E131D—is predicted to contribute to the binding free energy both before and after the mutation. Robetta, FoldX and KFC predict all of the mutations as non-hot spot residues in both 1918HA1 and 2009HA1. However, $\Delta\Delta G$ of V169I is predicted to increase from 0.4 to 0.76 kcal/mol by Robetta and from 0.07 to 0.83 kcal/mol by FoldX. Hence, the six mutations from 1918HA1 to 2009HA1 do not appear to adversely affect the binding between the 2D1 antibody and the two HA1s. Instead, on the whole, the change of binding free energy exhibits a possible increased tendency after the mutation, making the binding stronger. This is consistent with the result that the 1918HA1 neutralizing 2D1 antibody can cross-react with 2009HA1 [144].

Geometrically, the six mutations are located at the rim of the binding interface (Figure 9.1(b) and (c)), forming a part of an O-ring structure [36, 62]. Most of them have a large exposed portion to water and are not deeply buried. Their absolute and relative ASA information and burial levels are presented in Table 9.1. Only the residue at position 131 is buried with little exposure to water. The influential O-ring theory [36, 62] stated that residues on an O-ring, though of structural importance, are usually not energetically important. Therefore, these six mutations can only slightly destruct the antibody binding of 2009HA1 energetically if they are adverse.

We closely examined the S159N mutation from 1918HA1 to 2009HA1. It was reported to have a high $\Delta\Delta G$ about 2 kcal/mol [144]. However, we believe that this residue *does not* contribute greatly to the antibody binding either before or after the mutation. First, Ser¹⁵⁹ has no significant contact with the antibody; see Figure 9.3(a) where the closest contact has a 3.5 Å distance for the side-chain atom OG of Ser¹⁵⁹ and CB of Thr³¹ from the antibody heavy chain. Second, the S159N mutation makes the backbone deviate far away—the C α deviation is about 1.1 Å. This results from an increased flexibility for Asn¹⁵⁹ side chain. At the side chain of Ser¹⁵⁹, OG has a hydrogen bond with its backbone O. The distance between O and OG is 2.2 Å, and the angle of OG-H $\cdots$ O is 151.0 degree. Therefore, it is this hydrogen bond that confines Ser¹⁵⁹ side chain. But the mutation makes Asn¹⁵⁹ side chain free to contact water molecules, which can drag and affect the backbone structure a lot. Third, the mutation breaks the hydrogen bond of OG and the backbone O at Ser¹⁵⁹, changing the binding free energy. In fact, at least half of the 2 kcal/mol $\Delta\Delta G$ of the mutation should come from this hydrogen-bond break. This can be seen from the mutation of Ser¹⁵⁹ to Gly¹⁵⁹ (no side chain in Gly), whose $\Delta\Delta G$ is bigger than 1 kcal/mol [144]. Therefore, the S159N mutation from 1918HA1 to 2009HA1 did not greatly destroy the binding of the 2D1 antibody to 2009HA1.

As the Z-score method has a high negative precision value for predicting non-hot spot residues, the above residues predicted as non-hot spot residues can be considered as energetically unimportant to the antibody binding with high confidence. The additional twelve non-hot spot predictions in the HA1 epitopes are Pro(¹²⁵ in 1918HA1, ¹²² in 2009HA1), Thr(^{125B} in 1918HA1, ¹²⁴ in 2009HA1), Ser(^{125C} in 1918HA1, ¹²⁵ in 2009HA1), Ser¹²⁶, His¹³⁰, Gly¹⁵⁸, Ser¹⁶⁰, Leu¹⁶⁴, Ser¹⁶⁵, Ser¹⁶⁷, Tyr¹⁶⁸ and Thr²⁴⁸. Some of these non-hot spot predictions can be verified by past non-alanine mutation experiments [144]. For example, G158E/D, S160L or S165K cause only a small $\Delta\Delta G$ (less than 1 kcal/mol). They also happened between different residue-type groups, e.g., a mutation from a polar uncharged residue Ser¹⁶⁰ to a hydrophobic residue Leu¹⁶⁰. These suggest that these non-hot spot residues have little contribution to the binding to either 1918HA1 or 2009HA1, just as Z-score predicts. So, mutating these predicted non-hot spot residues provides little chance for H1N1 to evade the capture by the 2D1 antibody.

9.3.2 Hot Spot Residues at the Epitopes of the Two HA1s

The hot spot residues in 2D1-1918HA1 or 2D1-2009HA1 predicted by Robetta, FoldX, KFC and Z-score are shown in Figure 9.2. All of them are considered to have potential contribution to the antibody binding. In particular, Pro¹²⁸, Asn¹²⁹, Lys¹⁵⁷, Pro¹⁶², Lys¹⁶³, and Lys¹⁶⁶ are confirmed as hot spot residues at 2D1-1918HA1 or at 2D1-2009HA1 by Z-score and at least one of the other methods; see Table 9.2 and Figure 9.2(a) and (b). Three of them are common in both 2D1-1918HA1 and 2D1-2009HA1. For the other three residues, two of them (Pro¹²⁸ and Pro¹⁶²) are hot spot residues for 2D1-1918HA1 but also with not low $\Delta\Delta G$ in 2D1-2009HA1 (Table 9.2). Similar observation can be found for the remaining one (Lys¹⁶³) in 2D1-2009HA1. These residues all have a very small ASA, and are buried with a burial level up to 2.0 (Table 9.2). These double-confirmed hot spot residues are believed to contribute *greatly* to the antibody binding, as the combined prediction by Z-score and the other computational methods has a much higher precision. So, they are positions for mutations that can lead to H1N1's escape from 2D1's neutralization.

Figure 9.2: The binding hot spots in the HA-2D1 binding interfaces (better viewed in color). (a) The hot spots in the 1918HA1 epitope. (b) The hot spots in the 2009HA1 epitope. (c) The artificial hot spots in the 2007HA1 epitope and their glycosylation sites. (d) The hot spots in the antibody paratope of 2D1-1918HA1. (e) The hot spots in the antibody paratope of 2D1-2009HA1. (f) The hot spots in the artificial paratope of 2D1-2007HA1. (a) and (d) form 2D1-1918HA1, (b) and (e) form 2D1-2009HA1, and (c) and (f) form artificial 2D1-2007HA1. In (a)-(f), binding sites are in sphere view; residues in red (with yellow labels) are predicted by Z-score to be hot spot residues and confirmed by at least one of other computational methods; residues in green (with lightpink labels) are predicted only by one method; residues in blue are predicted as non-hot spots by all methods.

Some of these double-confirmed hot spot residues at 2D1-1918HA1 or 2D1-2009HA1 are also supported by wet-lab experiments. For example, the mutation of Asn to Lys at position 129 and the mutation of Lys to Asn at position 163 in wet-lab experiments resulted in more than 1 kcal/mol $\Delta\Delta G$ [144]. This fact indicates that Asn¹²⁹ and Lys¹⁶³ are truly energetically important although the mutations are non-alanine mutations.

Lys¹⁶⁶ has been comprehensively studied in the past by wet-lab experiments. It was found that this residue contributes greatly to this antibody binding: its mutations to residue types such as its similar hydrophilic residues Glu and Gln, or Pro resulted in more than 3 kcal/mol $\Delta\Delta G$ [144]. As can be seen in Table 9.2, Lys¹⁶⁶ is predicted by three computational methods (Z-score, Robetta and FoldX) as a hot spot residue in both 1918HA1 and 2009HA1 epitopes. To investigate why this residue is energetically so important, we examine its contacts using Figure 9.3(b). First, this residue is deeply buried with a small ASA and high burial level. Second, Lys¹⁶⁶ has several hydrogen bonds with its NZ atom as the donor: one hydrogen bond forms with the backbone O of the Ser¹²⁶ from the same chain as Lys¹⁶⁶, and the other two form with the side chain O of Asp⁹³ and of Asn³¹ from the light chain of the antibody. Third, CE of Lys¹⁶⁶ has a π -involving contact with Trp¹²⁷ from the same chain as Lys¹⁶⁶. These contacts suggest that this residue contributes greatly to the antibody binding and to the antigen folding by the π -involving contact with Trp¹²⁷ and the hydrogen bond with the Ser¹²⁶. In fact, Ser¹²⁶ is also at the epitope site, and it is predicted as a hot spot residue by FoldX and Robetta in the 2009HA1 epitope site but as a non-hot spot residue in the 1918HA1 by all the methods; so this hydrogen bond contributes to the antibody binding indirectly.

The residue Lys¹⁶⁶ was also reported as a selected escape mutation at 2D1-antibody by several viruses including 2009 H1N1, 1918 H1N1, and the 1930 swine viruses [146, 147, 144]. What is more interesting is that Lys¹⁶⁶'s two hydrogen-bond contact residues, Asp⁹³ and Asn³¹, are all predicted to be hot spot residues by Z-score and more than one other methods, instantiating the hot spot coupling property [48].

Although we have not found wet-lab evidence and report for Pro¹²⁸, Lys¹⁵⁷ or Pro¹⁶², we suggest that all of the six double-confirmed hot spot residues—Pro¹²⁸, Asn¹²⁹, Lys¹⁵⁷, Pro¹⁶², Lys¹⁶³, and Lys¹⁶⁶—are potential escape mutations for H1N1 to elude the 2D1 antibody.

9.3.3 Hot Spot Residues at the 2D1's Paratope

Table 9.3: Binding hot spots in the antibody paratopes predicted by Z-score and confirmed by other computational methods

Residues	Robetta		FoldX		KFC		p-value ^z		abs ASA (Å)		rel ASA (%)		burial level	
	<i>a</i> 2D1	<i>b</i> 2D1	<i>a</i> 2D1	<i>b</i> 2D1	<i>a</i> 2D1	<i>b</i> 2D1	<i>a</i> 2D1	<i>b</i> 2D1	<i>a</i> 2D1	<i>b</i> 2D1	<i>a</i> 2D1	<i>b</i> 2D1	<i>a</i> 2D1	<i>b</i> 2D1
<i>Asp</i> ⁵²	3.93	2.85	1.06			No	0.2522	0.1735	2.44	2.06	1.7	1.5	1.88	1.88
<i>Tyr</i> ⁵⁸	0.96	1.35				No	0.0585	0.1691	23.57	26.04	11.1	12.2	1.33	1.33
<i>Arg</i> ⁹⁷	4.5	4.57	1.32	3.53		Yes ^g	0.0075	0.0106	0.71	0.69	0.3	0.3	2.00	2.00
<i>Tyr</i> ^{100B}	2.28	1.14				Yes ^g	0.0178	0.015	26.35	33.65	12.4	15.8	1.17	1.08
<i>c Val</i> ^{100C}			1.02	1.16		No ^g	0.2539	0.3271	16.97	16.5	11.2	10.9	1.57	1.57
^d Asn ³¹	2.44	1.56	1.68	1.86		No	0.3236	0.3251	11.30	10.62	7.9	7.4	1.5	1.88
Trp ⁹¹	3.05	3.02	2.96	2.7	✓	Yes	<1E-324	<1E-324	12.39	10.99	5.0	4.4	1.29	2.14
Asp ⁹³	3.08	3.32		2.78	✓	No	0.016	0.0065	13.34	11.13	9.5	7.9	1.25	1.25

^a is for residues in 2D1-1918HA1, and ^b is for residues in 2D1-2009HA1.

^c Z-score failed to identify this hot spot in 2D1-2009HA1.

Italic are for the antibody heavy chain, and **bold** for the light chain.

^z is for p-values of Z-score.

^g Paratope residues which are close to potential glycosylation sites in 2D1-2007HA1.

^d Z-score failed to identify this hot spot.

We also studied those residues in the paratope (the antigen binding site) of the 2D1 antibody that can contribute greatly to the binding. These antibody hot spot residues can uncover how the 2D1 antibody captures the H1N1 viruses.

Using the Z-score method, all and only six hot spot residues are predicted in the antibody light chain which are also common between 2D1-1918HA1 and 2D1-2009HA1. Meanwhile, eight hot spot residues in the antibody heavy chain are identified in 2D1-1918HA1, and seven in 2D1-2009HA1. These predicted paratope hot spots are depicted in Figure 9.2(d) and (e). Among them, five from the heavy chain and three from the light chain are confirmed by more than one existing computational methods (Table 9.3). In the antibody light chain, the hot spot residues Asp⁹³ and Asn³¹ have significant contacts with the antigen hot spot residue Lys¹⁶⁶ as we discussed above. We believe that they are mainly responsible for the binding to the antigen.

In the antibody heavy chain, we are interested in the predicted hot spot residue Arg⁹⁷, as it is predicted to be energetically important by three methods in both 2D1-1918HA1 and 2D1-2009HA1 (Table 9.3). Its close contact with Asp⁵² is shown in Figure 9.3(c). Asp⁵² is also from the antibody heavy chain and confirmed as a hot spot residue by Robetta and FoldX. As seen in Figure 9.3(c), Arg⁹⁷ has two hydrogen bonds with Asp⁵² with NH1 and NH2 as donors and OD1 and OD2 as acceptors. Furthermore, Arg⁹⁷'s NH2 has a hydrogen bond with OD1 from Asp⁵³, and Arg⁹⁷'s NH1 has another hydrogen bond with OD1 from the antigen residue Asn¹²⁹. Asn¹²⁹ is considered to contribute greatly to the antibody binding by the two existing methods (Table 9.2). These contacts form a hydrogen-bond network which is believed to generate a favorable electrostatic contribution to the protein binding that can strongly stabilize the protein complexes [151]. Another finding is that the side chain of Arg⁹⁷ confines at least three side chains: the side chain of Arg⁹⁷ which is positively electrically charged, and the side chains of Asp⁵² and Asp⁵³ which are negatively charged. These side chains should prefer solvent water molecules if they are free. So, confining by the side chain of Arg⁹⁷ can make remarkable contribution to the antibody binding by removing the freedom of these three charged side chains.

Our investigation also finds a large cavity at the core of the binding interface, as seen in Figure 9.3(c) and (d). This cavity has a surface surrounded by the binding residues. Its narrowest part is more than 9 Å wide, which is equivalent to more than three water molecule diameters (2.75 Å). What is more important is that the side chains of Arg⁹⁷ and Asp⁵² are in the rim of the cavity, and some side chain atoms of Asp⁵³ contact the solvent as seen in Figure 9.3(c) and (d). Thus, removing the side chains of Arg⁹⁷ and/or Asp⁵² would increase the chance that this cavity is open to contact the solvent. In other words, the binding interface would have a larger open empty or solvent core. This makes it impossible to have stable binding. Hence, we believe that both Arg⁹⁷ and Asp⁵² are energetically significant in the binding.

9.3.4 Analysis on the Assumed 2D1-2007HA1 Binding

The hot spot prediction results on the assumed artificial 2D1-2007HA1 interface are also presented in Table 9.2, Table 9.3 and Figure 9.2(c) and (f). We find that more than half of the predicted hot spots in the binding 2D1-1918HA1 or 2D1-2009HA1 are not predicted to make contribution to the artificial binding 2D1-2007HA1, although Leu¹⁵⁶ is newly predicted as a hot spot residue at the epitope of 2D1-2007HA1.

We have two interesting remarks about this assumed binding. One remark is about Lys¹⁶² in 2007HA1 (Lys¹⁶⁶ in 1918HA1 and 2009HA1) which is predicted as a non-hot spot residue. First, this residue is conserved in the three HA1s. Second, it is predicted as a hot spot residue in 1918HA1 and 2009HA1 by all of the computational methods here. Third, its binding importance in 2D1-1918HA1 and 2D1-2009HA1 has been demonstrated by [144]. So, the reduced contribution of this Lys in 2D1-2007HA1 indicates an escape of 2007HA1 from 2D1. This together with the less number of predicted hot spots in 2D1-2007HA1 suggests a very small occurrence probability of 2D1-2007HA1. The other remark is that although Asn¹⁵⁹ in 2007HA1 (Lys¹⁶³ in 1918HA1 and 2009HA1) is predicted to contribute to the artificial 2D1-2007HA1 binding, this mutation creates a potential N-glycosylation site [144] as shown in Figure 9.2(c). 2007HA1 contains another glycosylation site Asn¹²⁵ [144]; see Figure 9.2(c). To better understand the assumed binding, we use Figure 9.2(f) to depict the binding region of 2D1's paratope to the glycosylation sites, and this region covers two of the three predicted hot spot residues of the paratope. It can be observed that the glycosylation sites mask the surface of 2007HA1 to block the cross neutralization by 2D1 [144]. The computational methods did make some predictions of hot spots in 2D1-2007HA1, because all of them only the residue information and none of them considers other important factors, such as the potential glycosylation sites here. Hence, these predicted binding hot spots are false positive, and they should be not true if the glycosylation sites are considered. In summary, 2D1 cannot recognize 2007HA1 for neutralization.

9.3.5 Another Two Evidences

9.3.5.1 $\Delta\Delta G$ Prediction by βACV_{ASA}

We also use βACV_{ASA} in Chapter 8 to predict $\Delta\Delta G$ for contact residues in the three binding interfaces of 2D1-1918HA1, 2D1-2009HA1 and 2D1-2007HA1. βACV_{ASA} is proved by the evaluation in Chapter 8 to have better prediction performance than other methods including Z-score, Robetta and FoldX. To help understanding of the 2D1 binding behaviors, we show βACV_{ASA} predictions for the six mutations in Table 9.4, and present binding hot spot predictions in Table 9.5. Since the βACV_{ASA} -predicted $\Delta\Delta G$ of binding hot spots is generally lower than the true $\Delta\Delta G$, we are interested in those hot spot predictions whose $\Delta\Delta G \geq 1.5$ kcal/mol.

Table 9.4: Predictions by βACV_{ASA} for the six mutations from 1918HA1 to 2009HA1.

1918HA1	βACV_{ASA} -predicted $\Delta\Delta G$		2009HA1
Glu ¹³¹	-0.03	0.18	Asp ¹³¹
Thr ¹³³	-0.48	0.09	Asn ¹³³
Ser ¹⁵⁹	0.61	0.20	Asn ¹⁵⁹
Val ¹⁶⁹	-0.44	0.21	Ile ¹⁶⁹
Asn ¹⁷¹	0.18	-0.06	Asp ¹⁷¹
Thr ²⁴²	0.35	0.59	Lys ²⁴²

Firstly, it can be clearly seen from Table 9.4 that all of these mutations from 1918HA1 to 2009HA1 are again confirmed to have lower $\Delta\Delta G$ by βACV_{ASA} . Secondly, in Table 9.5, βACV_{ASA} computationally reconfirmed several hot spots predicted by the meta-learning approach above, such as Pro¹⁶², Lys¹⁶³ and Lys¹⁶⁶ for 1918HA1 or for 2009HA1, and Arg⁹⁷ and Asp⁵² for the antibody heavy chain. In contrast, βACV_{ASA} only predicted one antigenic hot spot for 2007HA1, i.e., Phe¹¹⁷ (Phe¹²⁴ in 1918HA1 and Phe¹²¹ in 2009HA1), although βACV_{ASA} also predicted several hot spots in the paratope of 2D1-2007HA1, such as Asp⁵², Ser⁹⁹, Asp^{100A}, Tyr^{100B}, Asn³¹, Trp⁹¹ and Asp⁹³. These predicted hot spots in the paratope in 2D1-2007HA1 are partially due to that βACV_{ASA} considers atomic contacts within a biological unit. But these predictions still suggest that this assumed binding 2D1-2007HA1 cannot happen *in vivo*, because according to these βACV_{ASA} predictions, although the 2D1 paratope is ready to bind, 2007HA1 would not like to (because it has only one predicted antigenic binding hot spot). In conclusion, Table 9.4 and Table 9.5 again indicate that 2D1 can bind to 1918HA1 and 2009HA1, and cannot bind to 2007HA1.

9.3.5.2 Eight Different 2009 H1N1 Strains

More 2009 H1N1 strains are investigated; see Table 9.6. From their sequence alignment in Figure 9.4, it can be seen that there are only seven mutations and none of them is in the binding site of 2009HA1. Thus, the 2D1 antibody should be capable of neutralizing all of them.

9.4 Conclusion

We have done a structural analysis on the interfaces between the 2D1 antibody and the HA1 proteins of 2009 H1N1 and 1918 H1N1. The cross neutralization of this antibody is clearly demonstrated by the hot spot residues common in the two binding interfaces. Our comprehensive investigation suggests that there are six outstanding epitope residues whose mutations will help H1N1 evade capture by this antibody. We further pinpointed

Table 9.5: Binding hot spot predictions by $\beta\text{ACV}_{\text{ASA}}$ for 2D1-1918HA1 and 2D1-2009HA1 interfaces.

1918HA1	β ACV _{ASA} -predicted $\Delta\Delta G$		2009HA1
epitope			
Phe ¹²⁴	1.57	1.47	Phe ¹²¹
Pro ¹²⁸	1.00	1.18	Pro ¹²⁸
Asn ¹²⁹	0.88	1.35	Asn ¹²⁹
Lys ¹⁵⁷	1.11	1.38	Lys ¹⁵⁷
Ser ¹⁶⁰	1.86	0.66	Ser ¹⁶⁰
Pro ¹⁶²	1.72	1.00	Pro ¹⁶²
Lys ¹⁶³	0.73	1.56	Lys ¹⁶³
Lys ¹⁶⁶	2.83	1.66	Lys ¹⁶⁶
Asn ¹⁹⁷	0.99	2.05	Asn ¹⁹⁷
paratope of heavy chain			
Asp ⁵²	3.40	3.35	Asp ⁵²
Tyr ⁵⁸	1.54	0.97	Tyr ⁵⁸
Arg ⁹⁷	2.45	2.71	Arg ⁹⁷
Asp ^{100A}	1.88	1.87	Asp ^{100A}
Tyr ^{100B}	1.55	1.15	Tyr ^{100B}
paratope of light chain			
Asn ³¹	2.14	2.30	Asn ³¹
Trp ⁹¹	2.15	2.51	Trp ⁹¹
Asp ⁹³	1.91	0.71	Asp ⁹³
Trp ⁹³	2.20	2.09	Trp ⁹³

0-9 indicates predictions which are not binding hot spots, but some have higher $\Delta\Delta G$.

Table 9.6: HA proteins and their corresponding H1N1 viruses in 2009.

Name	Accession	Virus Name
3LZG2009	PDBID 3LZG	A/California/04/2009
Mxic2009	ADG27988	A/Mexico city/CIA6/2009 2009/06/14
Sing2009	ADI24621	A/Singapore/TT017/2010 2010/01/08
HK2009	ADM32929	A/Hong Kong/WRAIR0114T/2009 2009/08/27
Beij2009	ACR32998	A/Beijing/01/2009 2009/05/16
UK2009	ADU76756	A/England/00380004/2009 2009/11/10
Isra2009	ACT21572	A/Israel/276/2009 2009/04/
Agnt2009	ADA83603	A/Argentina/HNRG18/2009 2009/07/01

the hot spot residues at the paratope site of the 2D1 antibody which are responsible for the antigen recognition. The understanding of these hot spot residues can potentially facilitate drug design to neutralize influenza viruses.

3LZG2009	PGDTLCIGYHANNSTDTVDTVLEKNVTVTHSVNLEDKHNGKLCRLGVAPLHLGKCNIA	GWILGNPECESLSTASSWSYIVETSSDNGTCYPGDFIDYEELREQLSSVSSFERFEI
Mxic2009	--DTLCIGYHANNSTDTVDTVLEKNVTVTHSVNLEDKHNGKLCRLGVAPLHLGKCNIA	GWILGNPECESLSTASSWSYIVETSSDNGTCYPGDFIDYEELREQLSSVSSFERFEI
Beij2009	--DTLCIGYHANNSTDTVDTVLEKNVTVTHSVNLEDKHNGKLCRLGVAPLHLGKCNIA	GWILGNPECESLSTASSWSYIVETSSDNGTCYPGDFIDYEELREQLSSVSSFERFEI
HK2009	--DTLCIGYHANNSTDTVDTVLEKNVTVTHSVNLEDKHNGKLCRLGVAPLHLGKCNIA	GWILGNPECESLSTASSWSYIVETSSDNGTCYPGDFIDYEELREQLSSVSSFERFEI
UK2009	--DTLCIGYHANNSTDTVDTVLEKNVTVTHSVNLEDKHNGKLCRLGVAPLHLGKCNIA	GWILGNPECESLSTASSWSYIVETSSDNGTCYPGDFIDYEELREQLSSVSSFERFEI
Agnt2009	--DTLCIGYHANNSTDTVDTVLEKNVTVTHSVNLEDKHNGKLCRLGVAPLHLGKCNIA	GWILGNPECESLSTASSWSYIVETSSDNGTCYPGDFIDYEELREQLSSVSSFERFEI
Sing2009	--DTLCIGYHANNSTDTVDTVLEKNVTVTHSVNLEDKHNGKLCRLGVAPLHLGKCNIA	GWILGNPECESLSTASSWSYIVETSSDNGTCYPGDFIDYEELREQLSSVSSFERFEI
Isra2009	--DTLCIGYHANNSTDTVDTVLEKNVTVTHSVNLEDKHNGKLCRLGVAPLHLGKCNIA	GWILGNPECESLSTASSWSYIVETSSDNGTCYPGDFIDYEELREQLSSVSSFERFEI
3LZG2009	KTSSWPNHDSNKGVTAAACPHAGAKSFYKNLIWLVKKGSYYPKLSKSYINDKGKEVLVLWG	IHPSTADQOSLYQNADYVVFVGSRYSKKFKPEIAIRPKVRDQEGRMNYWTLVEPGD
Mxic2009	KTSSWPNHDSNKGVTAAACPHAGAKSFYKNLIWLVKKGSYYPKLSKSYINDKGKEVLVLWG	IHPSTADQOSLYQNADYVVFVGSRYSKKFKPEIAIRPKVRDQEGRMNYWTLVEPGD
Beij2009	KTSSWPNHDSNKGVTAAACPHAGAKSFYKNLIWLVKKGSYYPKLSKSYINDKGKEVLVLWG	IHPSTADQOSLYQNADYVVFVGSRYSKKFKPEIAIRPKVRDQEGRMNYWTLVEPGD
HK2009	KTSSWPNHDSNKGVTAAACPHAGAKSFYKNLIWLVKKGSYYPKLSKSYINDKGKEVLVLWG	IHPSTADQOSLYQNADYVVFVGSRYSKKFKPEIAIRPKVRDQEGRMNYWTLVEPGD
UK2009	KTSSWPNHDSNKGVTAAACPHAGAKSFYKNLIWLVKKGSYYPKLSKSYINDKGKEVLVLWG	IHPSTADQOSLYQNADYVVFVGSRYSKKFKPEIAIRPKVRDQEGRMNYWTLVEPGD
Agnt2009	KTSSWPNHDSNKGVTAAACPHAGAKSFYKNLIWLVKKGSYYPKLSKSYINDKGKEVLVLWG	IHPSTADQOSLYQNADYVVFVGSRYSKKFKPEIAIRPKVRDQEGRMNYWTLVEPGD
Sing2009	KTSSWPNHDSNKGVTAAACPHAGAKSFYKNLIWLVKKGSYYPKLSKSYINDKGKEVLVLWG	IHPSTADQOSLYQNADYVVFVGSRYSKKFKPEIAIRPKVRDQEGRMNYWTLVEPGD
Isra2009	KTSSWPNHDSNKGVTAAACPHAGAKSFYKNLIWLVKKGSYYPKLSKSYINDKGKEVLVLWG	IHPSTADQOSLYQNADYVVFVGSRYSKKFKPEIAIRPKVRDQEGRMNYWTLVEPGD
3LZG2009	KITFEATGNLVVPRYAFAMERNAGSGIIISDTPVHDCNTTCQTPKGAINTSLPFQNIHPI	TIGKCPYVKSTKLRLATGLRNVPSIQSR
Mxic2009	KITFEATGNLVVPRYAFAMERNAGSGIIISDTPVHDCNTTCQTPKGAINTSLPFQNIHPI	TIGKCPYVKSTKLRLATGLRNVPSIQSR
Beij2009	KITFEATGNLVVPRYAFAMERNAGSGIIISDTPVHDCNTTCQTPKGAINTSLPFQNIHPI	TIGKCPYVKSTKLRLATGLRNVPSIQSR
HK2009	KITFEATGNLVVPRYAFAMERNAGSGIIISDTPVHDCNTTCQTPKGAINTSLPFQNIHPI	TIGKCPYVKSTKLRLATGLRNVPSIQSR
UK2009	KITFEATGNLVVPRYAFAMERNAGSGIIISDTPVHDCNTTCQTPKGAINTSLPFQNIHPI	TIGKCPYVKSTKLRLATGLRNVPSIQSR
Agnt2009	KITFEATGNLVVPRYAFAMERNAGSGIIISDTPVHDCNTTCQTPKGAINTSLPFQNIHPI	TIGKCPYVKSTKLRLATGLRNVPSIQSR
Sing2009	KITFEATGNLVVPRYAFAMERNAGSGIIISDTPVHDCNTTCQTPKGAINTSLPFQNIHPI	TIGKCPYVKSTKLRLATGLRNVPSIQSR
Isra2009	KITFEATGNLVVPRYAFAMERNAGSGIIISDTPVHDCNTTCQTPKGAINTSLPFQNIHPI	TIGKCPYVKSTKLRLATGLRNVPSIQSR

Figure 9.4: The sequence alignment of HA proteins in different 2009 H1N1 viruses (better viewed in color). The binding site is in yellow and the mutations are in red.

Chapter 10

Summary and Future Work

In post-genomic biology, accurate understanding of protein binding becomes the key to uncovering the principles underlying living cells. However, protein interactions have great functional and structural diversity, which makes effective analysis of protein interactions a challenging topic.

In this thesis, we propound a new DWE hypothesis to model binding hot spots in protein interfaces. Under this hypothesis, we analyze protein binding interfaces from the following two levels. At the first residue level, we use bicliques under DWE to dissect protein interfaces based on their physicochemical properties and their different interaction types, such as obligate or non-obligate interactions, or crystal packing; at the second atomic level, we propose β atomic contacts for modeling atomic organization of protein binding interfaces. β contacts can overcome the problems in conventional computational definitions of protein interfaces and their defined contacts.

10.1 Summary

In detail, we propound DWE (Double Water Exclusion) to characterize the topology of binding hot spots in protein interfaces. DWE refines and integrates the three insightful propositions of binding hot spots. To apply DWE to a protein quaternary structure of two chains, each chain is considered as a partite, each residue as a vertex and each residue contact as an edge. Then, the protein quaternary structure is converted into a bipartite graph. The DWE hot spots are modeled by the corresponding maximal bicliques with the constraint of residues' ASA. The verification with ASEdb indicates that these DWE bicliques are truly rich in binding hot spots and DWE is a nice model of binding hot spots in protein interfaces.

However, there is no statistical way to filter biclique noises which should have little biological significance. To detect biologically significant DWE bicliques and binding hot spots, we introduce the notion of Z-score of a contact residue for capturing the probability of the residue occurring in or contributing to protein binding interfaces. This probability is calculated by using pairwise potential with crystal packing as the reference state. On one hand, crystal packing is used to calculate pairwise occurrence in a non-biological environment, in contrast with pairwise occurrence in a biological environment for residue pairwise potentials. On the other hand, a null distribution of this potential score is generated from artificial crystal packing contacts. Biological significance of DWE bicliques and binding hot spots is then estimated by using their pairwise potentials according to this null distribution. On two benchmark datasets, ASEdb and BID, this Z-score method is mostly superior to earlier methods to classify binding hot spots.

To understand the physicochemical properties of interfaces and their hot spots, we detect DWE bicliques from a non-redundant dataset with the three types of interactions: crystal packing, non-obligate and obligate interactions. DWE bicliques of these three types of protein interactions are investigated separately, as different types of interactions possess distinct behaviors in biological functions and processes. Our comparative investigation of DWE bicliques between biological interactions (i.e. obligate and non-obligate interactions) and crystal packing suggests that there are abundant unique bicliques to biological interactions, indicating the specificity of protein binding behaviors. Further, the two different types of biological interactions also have their own signature bicliques, emphasizing different specificity of binding behaviors in these two types of biological interactions. In detail, we also analyze the composition of residues and their pairs in DWE bicliques, and find interesting patterns including interaction-preferred residues (ipRs) and interaction-preferred residue pairs (ipRPs). For example, in obligate interactions, hydrophobic ipRs and ipRP contacts only involved by hydrophobic residues are frequent, while none of anti-ipRs are hydrophobic and the contacts involved by hydrophilic residues are also depleted in ipRPs; besides, aromatic ipRs and their involved ipRPs much prefer biological interactions, especially non-obligate interactions; in contrast, hydrophilic ipRs and ipRPs involved by hydrophilic residues seem to be richer in crystal packing; the anti-ipRs of crystal packing, such as Met, Trp and Cys, are just the ipRs of biological interactions. These patterns demonstrate the diverse binding features among the three types of interactions. Thus, DWE biclique patterns provide a new insight into protein interfaces.

Based on the above knowledge of DWE bicliques for the three types of protein interfaces, we design a new classification method OringPV to identify the types of protein interactions. OringPV is an SVM approach with propensity vector of low-ASA residue pairs as input. Low-ASA residue pairs are from O-ring-surrounded regions in protein interfaces under the O-ring and DWE hypotheses. The main part of our propensity vector

is 210-dimensional, consisting of all possible low-ASA residue pairs; the value of every feature is determined by a propensity measure. OringPV then uses propensity vectors of protein interactions for SVM. Finally, OringPV is tested on three benchmark datasets for a variety of classification tasks such as the distinction between crystal packing and biological interactions, the distinction between two different types of biological interactions, etc. The evaluation frameworks include within-dataset, cross-dataset comparison, and leave-one-out cross-validation. The results show that low-ASA residue pairs and the propensity vector description of protein interactions are truly strong in the distinction. In particular, the generalization capability tests of OringPV have achieved excellent recalls and overall accuracies on many cross datasets, much outperforming existing benchmark methods, such as NOXclass and DiMoVo.

These above analyses of protein interactions are at a residue level, where each residue is considered as a point although it is composed of a group of heavy atoms. To dissect atomic detail of residues and their contacts, we propose β contacts at an atomic level to identify critical contacts in binding interfaces. A β atomic contact cannot be interrupted by other atoms and should have enough directly contact area. To show the advantages of β atomic contacts, we design β ACV to classify homodimers from crystal packing. In this classification, each protein binding interface is represented by a β atomic contact vector (a β ACV vector), after grouping all heavy atoms of 20 standard residues into 12 atomic types; this vector is used as input of an SVM to build our classifier called β ACV. On four datasets of homodimers and crystal packing, β ACV shows better classification performance than existing methods in previous works and is also superior to the ACV methods of distance-based contact definitions. This demonstrates that β atomic contacts can truly identify critical specific contacts in protein binding interfaces.

Using β contacts, we construct a simple method, β ACV_{ASA}, to predict binding free energy and binding hot spots. In this method, an alanine mutation of interfacial residues is represented by an ACV based on β contacts and under our DWE hypothesis. Then, on the ASEdb dataset, β ACV_{ASA} is compared with ACV_{ASA} based on distance-based contacts, with ACV of β contacts but having no ASA integration, and with other methods in previous works. This comparison suggests the best prediction performance of β ACV_{ASA}, proving that both DWE and β contacts provide nice models for protein binding.

Finally, we apply the Z-score methods in Chapter 4, the β ACV in Chapter 7 and β ACV_{ASA} in Chapter 8 together with several existing methods to understanding of why and how 2D1 can neutralize 1918 H1N1 and 2009 H1N1 and cannot react with 2007 seasonal H1N1. After detecting predicted binding hot spots of the two binding interfaces between 2D1 and HA1 of 1918 H1N1 and between 2D1 and HA1 of 2009 H1N1, we found that the mutations from 1918 HA1 to 2009 HA1 do not destruct 2D1 binding; several common binding hot spots are also found in the two interfaces, suggesting that 2D1 can have enough high binding affinity for these two HA1s. In contrast, more than half of

binding hot spots in these two 2D1 interfaces are not predicted to greatly contribute to the assumed 2D1 binding to 2007 H1N1. This prediction together with two potential glycosylation sites in HA1 of 2007 H1N1 indicates that the assumed 2D1-2007HA1 binding truly never happen. Meanwhile, β ACV also classifies the 2D1 binding to HA1 both in 1918 H1N1 and in 2009 H1N1 as true biological binding and the 2D1 binding to the HA1 protein in 2007 H1N1 as false binding, which are consistent with our current knowledge about these three interactions [144, 147].

10.2 Future Directions

In this thesis, we propose DWE and β contacts to computationally analyze protein interactions and their binding hot spots at both residue and atomic levels. The achievements are interesting and promising. The deep understanding of protein binding interfaces can further improve other computational problems in Figure 1.7, as we discussed in Chapter 1. Therefore in the future, our works can be extended in a lot of fields.

10.2.1 Atomic Pairwise Potentials of β Contacts

β atomic contacts provide a new fundamental unit in protein 3D structures. Therefore, a lot of structural analyses based on distance-based contacts can be redone to get new results, such as atomic pairwise potentials. Atomic pairwise potentials are commonly used in previous structural bioinformatics, but they are obtained according to distance-based contacts. In the future, β contacts can improve atomic pairwise potentials from the following two aspects.

10.2.1.1 β Atomic Pairwise Potentials with Crystal Packing as the Reference State

We can calculate β atomic pairwise potentials using a large set of known protein quaternary structures of biological interactions as the preference state and the 3D structures of crystal packing as the reference state (similar to Z-score in Chapter 4). In this calculation, we use the process in Section 7.3.2 to produce atomic graphs of β contacts for all 3D structures. In these graphs, each node represents an atom, while each edge is a β contact. Then, the similar formula of Equation 4.4 in Chapter 4 is used to derive β atomic pairwise potentials. We expect that atomic pairwise potentials of β contacts possess more biological significance than the conventional contacts, such as distance-based contacts.

10.2.1.2 β Atomic Pairwise Potentials with 3D Structures Determined by NMR

In previous works, atomic pairwise potentials are mostly calculated using the 3D structures determined by X-ray where 3D structures are in a static model. Clearly, protein binding is a dynamic process. In a real-cell environment, vibrating protein interactions provide different binding interfaces according to the computational definitions as discussed in Section 1.1.4. It is reasonable to assume that important atomic contacts to protein binding should exist in most of the lifetime of protein complexes. 3D structures determined by NMR (the NMR structures for short) can help determine such contacts, since NMR provides a series of 3D structures (each corresponds to a model in PDB) of a protein or a protein complex. On a NMR 3D structure, the atomic graph of β contacts has a weight for each edge to indicate the percentage of this contact occur in all models of the 3D structure. Then, atomic pairwise potentials can be obtained on these atomic graphs.

In this calculation of atomic pairwise potentials, β contacts are better than distance-based contacts, since the former provide a more precise model of atomic contacts than the latter as discussed in Chapter 7. In distance-based criterion, two models of a protein 3D structure will have very similar defined interfaces, although atomic organization of their interfaces in the NMR structures differ greatly. This is due to that distance-based contacts are insensitive to the change of atomic organization. β contacts can capture such kind of properties in NMR structures. Both protein folding and binding can use this idea to obtain their own β atomic pairwise potentials.

10.2.2 A Probability Graph Model of Protein Quaternary Structures

After representing a protein 3D structure by an atomic graph of β contacts, a probability graph model can be used to describe protein binding. In this model, each edge of two atoms has a weight to denote the pairwise potential of the two atoms. Since interfacial atomic contacts are observed to be cooperative [52], those potentials should propagate using a Bayesian network or random walk method to optimize the contribution of local structures in protein binding. In this process of potential propagation, the co-occurrence probability of atomic pairs will be beneficial if available. Similar to atomic pairwise potentials, the co-occurrence probability of atomic pairs estimates the times of two atomic pairs which are in close vicinity in a protein binding interface. Thus, the co-occurrence knowledge needs much larger structure datasets than the potentials. In this model, propagated potentials can be used in many applications of structural bioinformatics, such as binding free energy prediction, binding hot spot prediction, binding type prediction and so on. Also, a similar probability graph model can be used for protein folding.

10.2.3 Unique Atomic Contacts in Binding Interfaces

In Chapter 5, we found that there are unique DWE bicliques to different types of biological interactions and to false interactions crystal packing. However, DWE bicliques are based on residues; their atomic contact definition is also simply based on distance. The atomic organization is not well described. Thus, residue-based DWE bicliques in Chapter 5 lose atomic-contact information. For example, given a DWE biclique $\mathcal{G}=\{\{\text{SER}, \text{LYS}\}, \{\text{GLU}, \text{THR}\}\}$, two residues, LYS and GLU, may only have one atomic pair with 6 Å in a protein binding interface, while they can have many close atomic contacts with small enough distance, such as less than 4 Å, in another interface. Residue-based contacts cannot describe these details.

In Chapter 7, we propose β contacts to nicely model atomic organization in protein 3D structures. We can then extend the work in Chapter 5 by using β atomic contact model for each interface, and define atomic graphs of β contacts for those interfaces. After that, we can use a subgraph mining algorithm to detect frequent subgraphs (may not be cliques or bicliques with full connection between atoms). These subgraphs, especially those subgraphs which are unique to biological binding or to crystal packing, can provide more detail about why biological binding is specific. For example, we want to see whether it is possible that crystal packing rarely have those contacts, such as the β contacts between the two groups CG, OD1 and OD2 from Asp52, and CZ, NH1 and NH2 from Arg97 as shown in Figure 7.1(a)).

10.2.4 Analysis of Protein-Small Molecule Binding

Protein-ligand binding is a critical interaction in biological systems. Small molecules, often called ligands, include drugs, metals, sugars, enzyme substrates and inhibitors, cofactors and so on. Knowledge of protein-ligand interactions is important to understand proteins' functions and activity. For example, hemoglobin proteins in the blood bind oxygen in respiratory organs, and release the oxygen in the rest of the body, such as tissues, where hemoglobin bring the resultant carbon dioxide back to the respiratory organs. In this process, hemoglobin should be able to bind to the small molecules, oxygen and carbon dioxide. Meanwhile, some small molecules can modulate protein-protein binding; identifying the binding of such small molecules is a major challenge for drug discovery.

In protein-ligand complexes, binding affinity is a key descriptor. Its prediction is challenging but a crucial issue of the determination of protein-ligand binding in a lot of fields. For example, scoring protein-ligand binding affinity is a compulsory step but a major obstacle in protein-ligand docking after identifying binding mode of ligands to target receptors. Thus, accurate prediction of protein-ligand binding affinity plays a fundamental role in structure-based drug design.

Our β contacts can be used to produce an atomic graph under the DWE hypothesis for an interface of a protein binding to a small molecule. Based on those interfacial graphs, on one hand, frequent β -contact subgraphs are able to indicate what kinds of atoms and atomic contacts are more likely conserved (i.e., with high frequency) in binding of small molecule to proteins; on the other hand, β contacts can be used to design a better prediction method for protein-ligand binding affinity or docking solutions, since in Chapter 8, βACV_{ASA} has shown a better estimation accuracy for $\Delta\Delta G$ than a prediction based on conventional distance-based contacts. Based on the probability graph model in Section 10.2.2, protein-ligand quaternary structures can be decomposed into β contacts with propagated weights of atomic pairwise potentials, where the potentials should be learned from a large set of known protein-ligand structure databases based on β contacts. And then the sum of propagated potentials between protein-ligand interfaces can be used to predict protein-ligand binding affinity, as an extension of βACV_{ASA} in Chapter 8.

In particular, the prediction method for protein-ligand binding affinity, integrated into protein-ligand docking, can also be used to estimate the stability of short, linear epitopes binding to Major histocompatibility complex (MHC) in T cells or antibodies in B cells. This estimation can greatly speed up drug design by removing non-binder peptides which have very low binding affinity.

10.2.5 Analysis of Protein Binding to DNA/RNA

Protein binding to both DNA and RNA play a crucial role in the maintenance and the biological processing of genetic information. DNA-binding proteins regulate DNA transcription, replication, repair, packaging and rearrangements, while RNA binding proteins participate post-transcriptional regulation, including splicing, mRNA transport and localization, mRNA stability and translation.

In protein-DNA/RNA complexes, it is not yet fully understood how proteins recognize DNA/RNA and what is binding specificity at the molecular basis [157]. In particular, DNA/RNA involving interactions are dominated by charged and polar groups [157]. In Chapter 8, βACV_{ASA} for alanine mutations of polar residues demonstrates high classification performance of binding hot spots. Thus, our DWE and β contacts should be able to model protein binding to DNA and to RNA. A prediction method based on β contacts can be expected to more accurately identify binding hot spot residues in protein-DNA/RNA complexes.

Meanwhile, base-pair rules provide a simple and fundamental knowledge for DNA in genetics. If protein-DNA/RNA complexes are modeled by β contacts and DWE, it is interesting to know whether there are atomic complementary pairs between nucleotides and amino acids.

10.2.6 Protein Docking

As discussed in subsection 1.2.1 of Chapter 1, one essential step in protein docking is to estimate the score of predicted docking interfaces. This is exactly a classification problem of false interactions from true interactions. In Chapter 7, β contacts already prove to be powerful in distinguishing crystal packing from homodimers. β contacts can also be used to find potential biological binding from decays in docking.

Furthermore, it should be also helpful to use β contacts to estimate residue contribution to docking interfaces. It is reasonable that a true docking interface should have several interfacial residues which have high $\Delta\Delta G$. But a necessary caution should be taken that side chains in docking interfaces may be in an unbound state. Thus, it is better to optimize the conformation of those side chains before using β contacts.

10.2.7 A More Accurate Estimation Method for $\Delta\Delta G$

In Chapter 8, we use β contacts to design βACV_{ASA} for predicting $\Delta\Delta G$ and binding hot spots. But βACV_{ASA} is naive, since it converts the complicated subgraph of an alanine mutation into a simple atomic contact vector; this vector representation loses cooperative contribution among a cluster of atoms.

To more precisely characterize contribution from interfacial residues, at least two factors as follows should be considered in the future. One is that after mutations (maybe not alanine mutations), mutated side chains and surrounding side chains should be repacked to build a new optimal local structure; in this repacking, residues' potential rotamers can be used to reduce search space for the conformation of these side chains. The other is that a probability graph model in Section 10.2.2 should be built for a protein binding interface to consider the cooperativity of atomic contacts. For a mutation, the difference of the two local structures in this model before and after mutations can be used as a measure of $\Delta\Delta G$ prediction.

Appendix A

A Short Tutorial on Bipartite Graphs and Their Maximal Bicliques

A bipartite graph is a special graph which comprises two disjoint sets of vertices as its partites, and whose edges are only available between two vertices each from one partite. Generally, bipartite graphs are preferable to general homogeneous graphs when modeling real-life applications. This is due to that vertices in many applications may have different types of objects, and bipartite graphs can easily capture this property. Thus, bipartite graphs are widely used in many fields, such as data mining, Web mining, bioinformatics, and business.

A.1 Definitions of Bipartite Graphs and of Maximal Bicliques

Mathematically, a bipartite graph is defined as $G = \langle \{V_1, V_2, E\} \rangle$ [112], where V_i (a partite) is a set of vertices, $i=1,2$, $V_1 \cap V_2 = \emptyset$, and E is a set of edges between V_1 and V_2 .

A bipartite graph is a biclique if and only if $E = V_1 \times V_2$. It is also simply denoted as $\mathcal{G} = \langle \{V_1, V_2\} \rangle$ without edges due to their full interactions. Given a biclique $\mathcal{G}$, if there is no any another proper biclique $\mathcal{G}'$ satisfying $\mathcal{G} \subseteq \mathcal{G}'$, $\mathcal{G}$ is a maximal biclique.

It should be noted that in a bipartite graph, there may exist other useful subgraphs, such as quasi-bicliques [158]. In quasi-bicliques, it is allowed to miss a certain number of or a percentage of edges compared to bicliques. These missing edges are considered to tolerate noises in data. However, we only focus on bicliques here due to that the

algorithm to mine maximal bicliques is highly efficient and there is less noise in our applications of protein 3D structures.

A.2 Mining Maximal Bicliques

The problem of mining maximal (quasi-)bicliques from a graph has been long studied, and several algorithms are available. Among them, LCM-MBC is a highly efficient method. LCM-MBC originally transforms this problem into a long-and-well-studied problem in data mining [112]: enumerating closed patterns from a transactional database. The key finding in LCM-MBC is that there is one-to-one correspondence between maximal bicliques of a bipartite graph and pairs of closed patterns in the adjacency matrix of the graph. The algorithms for enumeration of closed patterns, such as LCM [159], are so well-designed that it can be solved in a linear time. Computationally, LCM-MBC firstly transforms a bipartite graph into its adjacency matrix, and then uses LCM to mine closed patterns. Finally, the combination of these closed patterns and their occurrence set are maximal bicliques.

In applications, the LCM-MBC algorithm needs two parameters: p and q , $p \leq q$ to control partite size of maximal bicliques. Supposed that V_i is with less vertices than V_j in a maximal biclique $\mathcal{G}$, $i, j=1$ or 2 and $i \neq j$, p is the minimum size of V_i and q is the minimum size of V_j . That is, the LCM-MBC algorithm filters $\mathcal{G}$ where minimum size of V_i is less than p or minimum size of V_j is less than q . Small maximal bicliques will not be enumerated. The parameter setting of p and q is up to domain knowledge.

Appendix B

A Short Tutorial on Several Machine-Learning Algorithms Used in This Thesis

B.1 SVM.

SVM [160] is a state-of-the-art machine-learning algorithm. It usually achieves better performance than traditional machine-learning algorithms and has been successfully applied in a wide range of applications, including in bioinformatics [66, 18, 107, 70, 61].

Given a binary-class dataset D , SVM constructs an optimal hyperplane based on the maximum margin principle, that is, a linear classifier $f(x) = w \cdot x - b$ where x is any input vector from D , $\cdot$ denotes dot product, and w is a normal vector perpendicular to the hyperplane. When the dataset D is not linearly separable, SVM typically uses kernel functions to map data to a high dimensional vector space and then finds an optimal hyperplane in the new space.

B.2 RFE: a Feature-Selection Method.

Feature selection is an important pre-processing step for a classification task, since noise or irrelevant features often significantly hinder classifier construction. In β contacts in Chapter 7, not all types of atomic pairs are specific to protein binding; some types of atomic contacts may just happen by chance in crystal packing and homodimers. Feature-selection algorithms are able to remove noise or irrelevant features and identify relevant features in a given dataset D , which helps to construct efficient and robust classification

methods. Two common feature-selection approaches are filter and wrapper. In this work, a wrapper-based approach is adopted to select best β atomic contacts.

Wrapper-based feature selection approaches employ some specific learning algorithms (e.g., SVM) to identify which features are more discriminative for classification task. However, the full space of feature subsets is often huge, i.e., 2^m by an exhaustive search where m is the number of features. Several heuristic ways are commonly used to select feature sets efficiently, including forward selection and backward elimination. Among them, SVM-RFE (SVM-Recursive Feature Elimination) [161] is one of the most successful wrappers. SVM-RFE first uses SVM learning to obtain feature weight w_i for each feature (a type of atomic contacts here). Then, it removes features (one feature in this work) with the lowest value $\|w_i\|^2$. This removal process is repeated until predefined criteria are satisfied. We implemented SVM-RFE in Algorithm 2 to find a subset of features which can produce the best empirical classification performance on datasets.

Algorithm 2 SVM-RFE in β atomic contact vectors

Input: a given data D

Output: a subset of features S_f with the best accuracy

Description:

```

1: set  $S_f = \{\}$ , and the best accuracy  $A_b = 0$ 
2: while more than five features left in  $D$  do
3:   calculate classification accuracy  $A_c$  by using  $k$ -fold cross-validation
4:   if  $A_b < A_c$  then
5:     set  $A_b = A_c$ 
6:     set  $S_f$  to all features in  $D$ 
7:   end if
8:   use the linear LIBSVM to train an SVM model
9:   calculate weights for all features in  $D$  based on the training model
10:  rank the features in a descending order according to the square of their weights
     $\|w_i\|^2$ .
11:  remove the last feature from  $D$ 
12: end while

```

B.3 Decision Tree.

Decision tree is an easily interpretable machine-learning algorithm. It uses a recursive divide-and-conquer strategy to organize data into a tree structure. Each leaf in this tree represents a prediction by a rule from the root node to this leaf. So, given a data D , decision tree chooses a best discriminating feature to split D into several partitions. For

each partition and its sub-partitions, decision tree repeats the split until predefined criteria are satisfied. Here, we use one decision-tree implementation C4.5 [162] in Weka [163] to visualize the selected features by RFE in Chapter 7.

Appendix C

Supplementary Files for Chapter 4 and 8

Table C.1: The ASEdb dataset used in Chapter 8.

PDB ID	Biological Unit 1			Biological Unit 2		
	Chain Name(s)	mutations	≥ 2 kcal/mol	Chain Name(s)	mutations	≥ 2 kcal/mol
1A22 ¹	growth hormone			growth hormone receptor		
	A	30	3	B	32	5
1A4Y ⁴	ribonuclease inhibitor			angiogenin		
	A	13	2	B	11	1
1AHW ⁵	immunoglobulin fab 5g9			tissue factor		
	AB	0	0	C	8	1
1BRS	barnase			barstar		
	A	7	6	D	5	3
1BXI	colicin e9 immunity			colicin e9		
	A	17	7	B	0	0
1CBW ⁶	bovine chymotrypsin			BPTI		
	GH	0	0	I	8	0
1DAN ⁵	blood coagulation factor VIIA			soluble tissue factor		
	HL	25	0	TU	35	2
1DFJ ⁴	ribonuclease			ribonuclease inhibitor		
	E	0	0	I	13	4
1DVF ³	FV D1.3			FV E5.2		
Continued on next page						

Table C.1 – continued from previous page

PDB ID	Biological Unit 1			Biological Unit 2		
	Chain Name(s)	mutations	≥ 2 kcal/mol	Chain Name(s)	mutations	≥ 2 kcal/mol
	AB	16	6	CD	8	3
1DX5	thrombomodulin			thrombin		
	I	0	0	M	16	5
1F47	cell division protein ftsz			cell division protein zipa		
	A	9	3	B	0	0
1FC2	protein A			immunoglobulin fc		
	C	3	1	D	0	0
1FCC	IGG1 MO61 FC			streptococcal protein G		
	A	0	0	C	8	4
1GC1	CD4			envelope protein GP120		
	C	17	0	G	0	0
1JCK	T cell antigen receptor			staphylococcal enterotoxin C3		
	A	0	0	B	9	4
1JRH	antibody A6			interferon-gamma receptor alpha chain		
	HL	17	3	I	12	5
1JTG	beta-lactamase TEM			beta-lactamase inhibitory protein		
	A	6	0	B	4	2
1NMB	Fab NC10			N9 neuraminidase		
	HL	1	0	N	0	0
1VFB ^{2,3}	IGG1-KAPPA D1.3 FV			hen egg white lysozyme		
	AB	14	2	C	12	1
2PTC ⁶	beta-trypsin			trypsin inhibitor		
	E	0	0	I	2	2
3HFM ²	Hyhel-10 IGG1			hen egg white lysozyme		
	HL	12	8	Y	13	3
3HHR ¹	human growth hormone			human growth hormone receptor		
	A	10	0	BC	3	0

¹: These two complexes have identical protein sequences, but they have different alanine mutations. Please refer to Table C.3 for detail.

²: Chain C in 1VFB and Chain Y in 3HFM are identical, but they have different binding sites in 1VFB and in 3HFM.

³: Chain A and B in 1VFB and in 1DVF are identical, but their partner proteins Chain C in 1VFB and Chain C and D in 1DVF are dissimilar, indicating that 1VFB and 1DVF possess different interfaces in 3D structures. $\Delta\Delta G$ of alanine mutations for the same specific residues in Chain A and B are almost different. In detail, $\Delta\Delta G$ for the same mutations in Chain A are: His¹³⁰:0.8(1.7), Tyr³²:1.3(2), Tyr⁴⁹:0.8(1.7), Tyr⁵⁰:0.4(0.7), Trp⁹²:1.71(0.3) and Ser⁹³:0.11(1.2); $\Delta\Delta G$ for the same mutations in Chain B are:

Continued on next page

Table C.1 – continued from previous page

PDB ID	Biological Unit 1			Biological Unit 2		
	Chain Name(s)	mutations	≥ 2 kcal/mol	Chain Name(s)	mutations	≥ 2 kcal/mol
<p>Tyr³²:0.5(1.8), Trp⁵²:1.23(4.2), Asp⁵⁴:1.95(4.3), Asp⁵⁸:-0.2(1.6), Arg⁹⁹:0.47(1.9), Asp¹⁰⁰:3.1(2.8) and Tyr¹⁰¹:4(4); the numbers after ‘:’ are $\Delta\Delta G$ in 1VFB, and the numbers in brackets are $\Delta\Delta G$ in 1DVF.</p> <p>4: Chain B in 1A4Y has dissimilar sequence to Chain E in 1DFJ, and Chain A in 1A4Y is similar to Chain I in 1DFJ. The latter two similar sequences have several same mutations but with different $\Delta\Delta G$ as follows: Trp²⁶¹:0.1(²⁵⁷:1.3), Trp²⁶³:1.2(²⁵⁹:2.2), Ser²⁸⁹:0(²⁸⁵:0.8)), Trp³¹⁸:1.5(³¹⁴:1), Lys³²⁰:-0.3(³¹⁶:1.3), Glu⁴⁰¹:0.9(³⁹⁷:1.3)), Tyr⁴³⁴:3.3(⁴³⁰:5.9)), Asp⁴³⁵:3.5(⁴³¹:3.6)), Tyr⁴³⁷:0.8(⁴³³:2.6)) and Ile⁴⁵⁹:0.7(⁴⁵⁵:0.3)). The numbers after ‘:’ are $\Delta\Delta G$ in 1A4Y and those in brackets are $\Delta\Delta G$ in 1DFJ.</p> <p>5: Chain T and U in 1DAN constitute Chain C in 1AHW, but Chain A and B in 1AHW have different sequences from Chain H and L in 1DAN. Chain C in 1AHW (Chain T and U in 1DAN) also has different binding sites in these two complexes.</p> <p>6: Chain I in 2PTC and in 1CBW are identical, but Chain E in 2PTC has 42% sequence identity to Chain G in 1CBW and 44% to Chain H in 1CBW. Further, Chain I in 2PTC only has two mutations: Lys¹⁵ and Ile¹⁸ which has $\Delta\Delta G$ 10 kcal/mol and 4.97 kcal/mol. In contrast, their mutations have much smaller $\Delta\Delta G$ in 1CBW, i.e., 1.99 kcal/mol and 1.4 kcal/mol.</p>						

Table C.2: Eight atomic types used in Chapter 8.

Short Name	Atomic names
N_H+	ILE_N_ VAL_N_ LEU_N_ PHE_N_ CYS_N_ MET_N_ ALA_N_ GLY_N_ THR_N_ SER_N_ TRP_N_ TYR_N_ HIS_N_ GLU_N_ GLN_N_ ASP_N_ ASN_N_ LYS_N_ ARG_N_ ARG_NE_ ARG_NH1 ASN_ND2 GLN_NE2 HIS_NE2 TRP_NE1
N+H+	ARG_NH2 LYS_NZ_
O_H0	ILE_O_ VAL_O_ LEU_O_ PHE_O_ CYS_O_ MET_O_ ALA_O_ GLY_O_ THR_O_ SER_O_ TRP_O_ TYR_O_ PRO_O_ HIS_O_ GLU_O_ GLN_O_ ASP_O_ ASN_O_ LYS_O_ ARG_O_ ASP_OD1 GLU_OE1 ASN_OD1 GLN_OE1 PRO_N_ HIS_ND1
O-H0	ASP_OD2 GLU_OE2 ILE_OXT VAL_OXT LEU_OXT PHE_OXT CYS_OXT MET_OXT ALA_OXT GLY_OXT THR_OXT SER_OXT TRP_OXT TYR_OXT PRO_OXT HIS_OXT GLU_OXT GLN_OXT ASP_OXT ASN_OXT LYS_OXT ARG_OXT
O_H+	SER_OG_ THR_OG1 TYR_OH_ HOH_O_
C_NO	ILE_C_ VAL_C_ LEU_C_ PHE_C_ CYS_C_ MET_C_ ALA_C_ GLY_C_ THR_C_ SER_C_ TRP_C_ TYR_C_ PRO_C_ HIS_C_ GLU_C_ GLN_C_ ASP_C_ ASN_C_ LYS_C_ ARG_C_ ILE_CA_ VAL_CA_ LEU_CA_ PHE_CA_ CYS_CA_ MET_CA_ ALA_CA_ GLY_CA_ THR_CA_ SER_CA_ TRP_CA_ TYR_CA_ PRO_CA_ HIS_CA_ GLU_CA_ GLN_CA_ ASP_CA_ ASN_CA_ LYS_CA_ ARG_CA_ ARG_CZ_ ARG_CD_ LYS_CE_ PRO_CD_ ASP_CG_ GLU_CD_ SER_CB_ THR_CB_ ASN_CG_ GLN_CD_
C_NON	MET_SD_ CYS_SG_ LEU_CG_ ILE_CB_ VAL_CB_ LEU_CB_ PHE_CB_ CYS_CB_ MET_CB_ ALA_CB_ TRP_CB_ TYR_CB_ PRO_CB_ HIS_CB_ GLU_CB_ GLN_CB_ ASP_CB_ ASN_CB_ LYS_CB_ ARG_CB_ ARG_CG_ GLU_CG_ GLN_CG_ LYS_CG_ LYS_CD_ PRO_CG_ ILE_CG1 MET_CG_ MET_CE_ ILE_CD1 THR_CG2 ILE_CG2 LEU_CD1 LEU_CD2 VAL_CG1 VAL_CG2
ARC	HIS_CD2 HIS_CE1 HIS_CG_ PHE_CG_ PHE_CD1 PHE_CD2 PHE_CE1 PHE_CE2 PHE_CZ_ TRP_CD1 TRP_CE2 TRP_CH2 TRP_CE3 TRP_CZ3 TRP_CG_ TRP_CD2 TRP_CZ2 TYR_CZ_ TYR_CG_ TYR_CD1 TYR_CD2 TYR_CE1 TYR_CE2

‘H+’: more than one hydrogen

‘H0’: without any hydrogens

‘N/O_’: nitrogen/oxygen atoms without charged

‘N+’: nitrogen atoms with positively charged

‘O-’: oxygen atoms with negatively charged

‘C_NO(N)’: carbon atoms with(out) covalent-bond nitrogen or oxygen

‘ARC’: carbon atoms in an aromatic ring

For ‘XXXYYZZ’ or ‘XXX_YZZ’ or ‘XXX.YZ_’ in the second column, XXX represents a residue type, and _Y or YY denotes an atomic name, while Z or Z_ indicates a specific position of _Y or YY in XXX.

Table C.3: Alanine mutations in the ASEdb dataset for Chapter 8. ‘Inter’ denotes the interaction of two biological units, and ‘C’ indicates protein chain names. a and sc represent the ASA change of all atoms from a residue or of side-chain atoms, while ad and m denote the distance d of those contacts of all atoms from a residue or of the mutated atoms; d is the Euclidian distance of two atoms minus the sum of their van der Waals radii as defined in [12]; $\leq$ or $\bar{\Delta}$ indicates an outlier prediction with predicted $\Delta\Delta G$ less than -3 kcal/mol or more than 11 kcal/mol.

PDB	Inter	C	Pos	AA	$\Delta\Delta G_{exp}$	ASA a	ASA sc	d^{ad}	d^m	FoldX	Robetta	βACV_{ASA}	βACV_{ASA}^2
1A22	A-B	A	18	HIS	-0.5	71.67	71.66	-0.47	-0.47	-0.220	2.290	0.402	-0.420
	A-B	A	21	HIS	0.2	31.11	31.11	-0.15	-0.15	0.190	0.830	-0.284	-0.756
	A-B	A	22	GLN	-0.2	16.35	15.83	0.13	0.40	-0.060	0.050	0.372	0.410
	A-B	A	25	PHE	-0.4	47.89	47.89	0.12	0.12	0.920	1.290	1.005	-0.861
	A-B	A	26	ASP	-0.2	0.00	0.00	2.93	2.93	-0.040	-	0.379	0.610
	A-B	A	42	TYR	0.2	78.51	78.51	-0.05	-0.05	0.940	2.020	1.016	1.297
	A-B	A	45	LEU	1.2	68.67	52.10	-0.40	0.04	1.430	1.150	1.762	1.860
	A-B	A	46	GLN	0.1	76.77	71.59	-0.30	-0.08	-0.220	1.040	0.608	0.649
	A-B	A	48	PRO	0.4	57.51	57.50	-0.04	-0.04	1.090	-	0.485	1.433
	A-B	A	51	SER	0.3	10.78	8.02	1.24	1.34	-0.360	-0.020	1.081	0.436
	A-B	A	56	GLU	0.4	20.98	20.98	-0.17	-0.17	1.120	0.970	2.796	3.076
	A-B	A	61	PRO	1.2	6.41	0.00	0.51	2.15	0.290	-	0.409	0.686
	A-B	A	62	SER	0.1	74.12	42.31	-0.34	1.02	0.190	-0.170	1.004	1.052
	A-B	A	63	ASN	0.3	41.34	38.72	0.15	0.95	0.240	0.390	1.135	1.782
	A-B	A	64	ARG	1.6	122.01	115.92	-0.15	-0.15	1.280	2.000	1.181	1.631
	A-B	A	65	GLU	-0.5	23.68	23.68	1.01	1.01	0.010	-0.110	-0.657	-0.582
Continued on next page													

Table C.3 – continued from previous page

PDB	Inter	C	Pos	AA	$\Delta\Delta G_{exp}$	ASA ^a	ASA ^{sc}	d ^{ad}	d ^m	FoldX	Robetta	βACV_{ASA}	βACV^2_{ASA}
A-B	A	66	GLU	0.43	0.00	37.83	0.00	2.76	2.76	0.890	-	0.851	0.991
A-B	A	68	GLN	0.6	37.83	37.83	0.00	-0.07	-0.07	1.620	1.810	1.684	1.403
A-B	A	164	TYR	0.3	22.41	22.41	22.41	1.14	1.14	1.780	-	2.022	-0.745
A-B	A	167	ARG	0.3	34.74	34.74	34.74	-0.01	-0.01	1.190	0.420	1.724	1.331
A-B	A	168	LYS	-0.2	50.24	49.83	49.83	0.10	0.10	2.090	2.120	4.543	4.814
A-B	A	171	ASP	0.8	58.04	57.55	57.55	0.14	0.14	1.490	9.000	4.044	3.666
A-B	A	172	LYS	2	30.23	30.23	30.23	0.21	0.21	2.540	0.740	1.227	1.490
A-B	A	174	GLU	-0.9	11.59	11.59	11.59	0.60	0.60	1.790	1.380	1.057	0.362
A-B	A	175	THR	2	48.25	48.11	48.11	-0.40	-0.40	3.570	2.310	3.198	2.812
A-B	A	176	PHE	1.9	6.85	6.85	6.85	0.52	0.52	1.860	-	2.174	3.074
A-B	A	178	ARG	2.4	61.29	61.29	61.29	-0.07	-0.07	1.900	0.380	1.307	0.957
A-B	A	179	ILE	0.8	22.30	22.14	22.14	0.21	0.21	2.800	1.000	1.557	0.589
A-B	A	182	CYS	1.01	32.49	32.27	32.27	-0.19	-0.19	1.490	-0.070	1.267	0.779
A-B	A	183	ARG	0.5	2.50	2.50	2.50	2.32	2.32	1.960	-	-0.756	-0.521
A-B	B	243	ARG	2.12	30.86	30.48	30.48	-0.40	-0.40	5.930	6.540	2.179	2.179
A-B	B	244	GLU	1.69	32.18	32.18	32.18	0.01	0.01	2.050	0.490	0.941	0.736
A-B	B	270	ARG	0.69	0.40	0.40	0.40	2.31	2.31	2.120	-	0.099	0.043
A-B	B	271	ARG	0.54	75.59	75.59	75.59	-0.31	-0.31	2.290	1.310	$\underline{\vee}$	$\underline{\vee}$
A-B	B	274	GLN	0	30.70	14.25	14.25	0.56	1.68	2.100	0.030	0.143	0.200
A-B	B	275	GLU	-0.1	21.98	21.98	21.98	0.11	0.11	1.990	0.090	-0.242	0.074
A-B	B	276	TRP	0.51	56.88	56.87	56.87	-0.40	-0.40	5.220	2.860	2.572	$\underline{\vee}$
A-B	B	280	TRP	-0.02	3.85	3.85	3.85	1.88	1.88	2.000	-	-0.015	0.015
A-B	B	298	SER	-0.05	22.82	12.65	12.65	1.01	2.55	2.000	-	0.298	0.353
A-B	B	301	THR	1.76	12.52	0.00	0.00	0.66	2.75	-	-	0.031	-1.290
A-B	B	302	SER	-0.2	24.90	24.90	24.90	0.23	0.95	1.960	-0.110	0.944	1.107
A-B	B	303	ILE	1.61	18.56	7.78	7.78	-0.34	0.36	2.860	0.310	2.207	$\underline{\vee}$
A-B	B	304	TRP	4.5	154.87	135.22	135.22	0.10	0.15	5.960	5.380	3.626	$\underline{\vee}$

Continued on next page

Table C.3 – continued from previous page

PDB	Inter	C	Pos	AA	$\Delta\Delta G_{exp}$	ASA ^a	ASA ^{sc}	d ^{ad}	d ^m	FoldX	Robetta	βACV_{ASA}	βACV_{ASA}^2
	A-B	B	305	ILE	1.94	13.13	13.13	0.50	0.50	2.560	0.130	1.020	-1.275
	A-B	B	306	PRO	3.31	36.26	36.26	0.04	0.49	2.710	-	2.263	3.022
	A-B	B	320	GLU	-0.19	39.73	31.76	-0.00	-0.00	1.910	0.670	0.726	0.691
	A-B	B	321	LYS	0.08	16.18	16.18	0.45	0.79	2.440	-	1.265	1.408
	A-B	B	324	SER	0.28	11.38	11.38	1.36	1.67	2.370	-0.060	0.485	1.073
	A-B	B	326	ASP	0.99	14.49	14.49	0.61	1.41	2.050	-0.310	1.446	1.220
	A-B	B	327	GLU	0.97	63.40	63.40	-0.01	-0.01	1.860	1.110	1.719	0.430
	A-B	B	364	ASP	1.49	19.02	19.02	-0.15	-0.15	5.770	1.370	1.598	0.923
	A-B	B	365	ILE	2.13	29.47	17.32	0.44	1.30	2.750	0.090	0.537	0.716
	A-B	B	366	GLN	0.02	35.46	17.83	0.50	2.12	2.800	0.680	0.333	0.246
	A-B	B	367	LYS	-0.02	66.29	35.17	-0.11	-0.11	1.780	0.220	0.462	0.732
	A-B	B	369	TRP	4.5	92.93	91.27	-0.05	-0.05	6.670	3.560	2.497	4.597
	A-B	B	371	VAL	-0.64	36.84	36.84	0.17	0.17	2.910	0.260	0.459	-0.028
	A-B	B	394	THR	0.2	0.00	0.00	3.23	3.23	-	-	-0.875	-0.840
	A-B	B	395	THR	-0.09	4.77	4.77	2.25	2.25	2.640	-	-0.129	-0.111
	A-B	B	416	GLN	0.89	0.08	0.07	2.75	2.75	2.840	-	0.092	-0.582
	A-B	B	417	ARG	0.28	47.44	44.15	-0.47	-0.47	2.260	0.230	0.451	-0.296
	A-B	B	418	ASN	0.3	129.65	105.00	-0.17	-0.17	1.610	1.840	0.921	0.929
	A-B	B	419	SER	0.03	25.37	24.84	0.47	0.47	3.100	0.980	0.945	1.106
1A4Y	A-B	A	261	TRP	0.1	32.00	32.00	0.35	0.35	1.760	1.060	1.382	-0.069
	A-B	A	263	TRP	1.2	65.96	65.96	-0.29	-0.29	2.200	2.270	4.137	9.377
	A-B	A	289	SER	0	7.54	7.54	-0.05	-0.05	1.760	0.600	1.291	0.627
	A-B	A	318	TRP	1.5	42.53	42.53	-0.08	-0.08	2.440	2.190	1.490	2.289
	A-B	A	320	LYS	-0.3	39.47	39.47	0.12	0.12	1.760	-0.210	0.793	1.001
	A-B	A	344	GLU	0.2	16.01	16.01	0.14	0.14	1.790	1.370	0.590	1.937
	A-B	A	375	TRP	1	56.59	56.59	-0.35	-0.35	2.820	2.830	2.684	0.022
	A-B	A	401	GLU	0.9	19.59	19.59	0.44	0.44	1.760	0.020	0.565	0.352
Continued on next page													

Table C.3 – continued from previous page

PDB	Inter	C	Pos	AA	$\Delta\Delta G_{exp}$	ASA ^a	ASA ^{sc}	d ^{ad}	d ^m	FoldX	Robetta	βACV_{ASA}	βACV_{ASA}^2
	A-B	A	434	TYR	3.3	98.47	98.47	0.10	0.10	2.850	3.000	2.386	3.957
	A-B	A	435	ASP	3.5	90.55	80.28	-0.10	-0.10	3.410	0.570	3.436	3.239
	A-B	A	437	TYR	0.8	146.14	143.38	0.11	0.11	4.920	3.130	2.068	2.028
	A-B	A	457	ARG	-0.2	2.78	2.78	2.62	2.62	1.470	-	-0.664	-0.506
	A-B	A	459	ILE	0.7	23.86	23.86	-0.38	-0.38	2.270	0.620	0.937	1.656
	A-B	B	5	ARG	2.3	100.13	100.13	-0.43	-0.43	7.310	2.540	1.712	1.430
	A-B	B	8	HIS	0.9	24.05	24.05	-0.23	-0.23	1.960	0.850	0.882	0.875
	A-B	B	12	GLN	0.3	22.77	22.77	-0.21	-0.21	2.040	0.870	2.100	1.803
	A-B	B	13	HIS	-0.3	4.70	4.70	1.11	1.11	2.310	-	-1.476	-0.737
	A-B	B	31	ARG	0.2	125.12	119.92	-0.58	-0.58	1.760	2.760	1.830	1.885
	A-B	B	32	ARG	0.9	64.35	62.16	-0.57	-0.57	2.660	0.180	0.103	0.405
	A-B	B	68	ASN	0.2	10.66	10.67	0.64	0.64	1.740	0.520	1.337	1.065
	A-B	B	84	HIS	0.2	63.98	63.97	0.03	0.03	-	1.050	1.030	1.201
	A-B	B	89	TRP	0.2	84.78	84.77	-0.08	-0.08	4.390	2.710	1.343	1.539
	A-B	B	108	GLU	-0.3	43.90	43.90	0.29	0.29	3.820	1.730	1.500	1.502
	A-B	B	114	HIS	0.65	62.62	62.62	0.17	0.33	1.760	1.690	1.155	1.302
1AHW	AB-C	C	156	TYR	4	65.88	65.88	-0.22	-0.22	2.950	4.530	2.203	-2.288
	AB-C	C	157	TYR	-1.9	5.75	5.75	1.83	1.83	0.090	-	1.441	-0.417
	AB-C	C	167	THR	0	70.63	41.50	-0.23	-0.20	0.440	-0.240	1.710	0.044
	AB-C	C	170	THR	1	11.52	11.29	0.35	1.83	-0.080	-0.070	0.125	0.915
	AB-C	C	178	ASP	-0.5	14.32	14.32	1.11	1.11	-0.010	-0.090	0.311	0.648
	AB-C	C	197	THR	1.3	22.21	8.87	0.80	1.46	-	-0.020	-0.190	-0.138
	AB-C	C	198	VAL	-0.3	25.99	7.15	-0.35	1.64	-0.010	-0.010	-0.226	-0.087
	AB-C	C	199	ASN	1.1	0.68	0.68	1.33	3.08	-	-	-0.037	-0.003
1BRS	A-D	A	27	LYS	5.4	56.69	56.69	0.05	0.05	1.590	1.880	2.947	3.827
	A-D	A	58	ASN	3.1	1.85	1.69	0.47	2.28	1.590	-	1.006	1.111
	A-D	A	59	ARG	5.2	159.13	147.78	-0.18	-0.18	2.050	3.010	1.516	1.080

Continued on next page

Table C.3 – continued from previous page

PDB	Inter	C	Pos	AA	$\Delta\Delta G_{exp}$	ASA ^a	ASA ^{sc}	d ^{ad}	d ^m	FoldX	Robetta	βACV_{ASA}	βACV_{ASA}^2	
	A-D	A	60	GLU	-0.2	69.39	67.45	-0.21	-0.21	1.590	1.410	1.388	1.203	
	A-D	A	73	GLU	2.8	10.14	10.14	1.69	1.69	1.590	-	0.934	-0.302	
	A-D	A	87	ARG	5.5	3.84	3.84	-0.12	-0.12	3.350	4.440	2.736	3.460	
	A-D	A	102	HIS	6	107.74	96.86	-0.27	-0.27	-	5.080	4.587	5.036	
	A-D	D	29	TYR	3.4	96.78	86.59	-0.15	-0.15	2.470	3.130	1.899	3.092	
	A-D	D	35	ASP	4.5	115.75	111.62	-0.17	-0.17	1.820	1.420	1.147	1.012	
	A-D	D	39	ASP	7.7	84.15	84.06	-0.55	-0.55	6.490	9.400	3.579	4.689	
	A-D	D	42	THR	1.8	39.91	18.77	0.05	0.05	2.950	1.660	1.868	1.326	
	A-D	D	76	GLU	1.3	24.09	24.10	-0.18	-0.18	2.670	1.540	0.588	1.554	
	1BXI	A-B	A	23	CYS	0.92	38.11	10.61	-0.10	0.08	0.370	-0.080	1.454	1.923
	A-B	A	24	ASN	0.14	20.14	6.41	0.04	1.84	0.060	0.000	0.404	0.313	
	A-B	A	27	THR	0.73	12.57	5.33	0.08	0.08	1.390	0.600	1.623	0.850	
	A-B	A	29	SER	0.96	9.10	6.56	-0.05	2.87	0.020	-	-0.310	0.242	
	A-B	A	30	GLU	1.41	106.09	100.51	-0.35	-0.35	1.350	2.970	2.348	1.612	
	A-B	A	33	LEU	3.42	33.40	33.40	0.13	0.13	1.190	1.020	1.068	2.473	
	A-B	A	34	VAL	2.58	61.92	61.50	0.33	0.33	1.220	0.980	0.838	2.032	
	A-B	A	37	VAL	1.66	16.20	16.20	-0.08	-0.08	0.810	0.500	1.878	3.018	
	A-B	A	38	THR	0.9	19.24	19.24	0.46	0.46	1.650	1.350	0.986	0.454	
	A-B	A	41	GLU	2.08	32.84	32.84	-0.30	-0.30	0.730	-0.080	3.708	3.903	
	A-B	A	48	SER	0.01	9.38	2.14	-0.05	3.00	0.360	-	0.341	0.257	
	A-B	A	50	SER	2.19	35.71	33.03	-0.25	-0.25	0.890	5.400	1.338	0.985	
	A-B	A	51	ASP	5.92	36.70	36.63	-0.39	-0.39	2.160	0.820	1.701	0.973	
	A-B	A	53	ILE	0.85	12.73	2.21	-0.49	-0.15	0.750	0.170	2.218	0.953	
	A-B	A	54	TYR	4.83	98.94	80.73	-0.25	-0.25	3.250	2.860	1.943	4.323	
	A-B	A	55	TYR	4.63	90.64	90.63	-0.25	-0.25	3.600	3.490	1.826	0.651	
	A-B	A	56	PRO	1.24	5.40	5.41	0.53	0.53	1.410	-	1.256	1.589	
	1CBW	GH-I	I	11	THR	0.22	39.50	39.50	0.08	0.08	0.550	0.180	0.272	-0.307
	Continued on next page													

Table C.3 – continued from previous page

PDB	Inter	C	Pos	AA	$\Delta\Delta G_{exp}$	ASA ^a	ASA ^{sc}	d ^{ad}	d ^m	FoldX	Robetta	βACV_{ASA}	βACV_{ASA}^2
	GH-I	I	13	PRO	-0.06	59.11	47.88	-0.10	0.62	0.610	-	0.373	0.308
	GH-I	I	15	LYS	1.99	176.99	143.34	-0.65	-0.19	0.190	1.540	3.070	3.960
	GH-I	I	17	ARG	0.55	171.37	146.39	-0.42	-0.42	-0.010	1.820	1.267	1.501
	GH-I	I	18	ILE	1.40	28.33	28.33	1.04	1.17	1.100	0.730	0.603	1.308
	GH-I	I	19	ILE	0.14	28.17	27.90	-0.23	-0.23	0.730	0.540	-0.943	-0.903
	GH-I	I	34	VAL	0.05	8.69	8.69	1.06	1.06	0.450	0.260	0.909	0.176
	GH-I	I	39	ARG	0.22	67.80	62.57	0.31	0.31	0.490	1.330	0.120	-0.169
1DAN	HL-TU	H	129B	ARG	0	1.34	1.33	2.63	2.63	-	-	-0.619	-0.867
	HL-TU	H	129F	PHE	0	96.60	71.15	0.01	0.01	-	1.660	2.372	-0.729
	HL-TU	H	134	ARG	0.51	126.88	120.63	-0.20	-0.20	-0.770	3.090	1.891	1.662
	HL-TU	H	135	PHE	0	26.38	25.83	0.59	0.59	0.380	0.530	0.597	0.573
	HL-TU	H	162	ARG	0.65	4.26	2.85	1.74	1.74	0.340	-	1.711	2.990
	HL-TU	H	164	MET	0.5	111.85	110.80	-0.17	-0.17	1.160	1.030	1.552	0.975
	HL-TU	H	165	THR	0	8.75	6.03	-0.03	0.85	0.300	0.270	0.723	1.041
	HL-TU	H	166	GLN	0	73.86	72.81	-0.14	-0.03	1.230	2.600	1.071	0.733
	HL-TU	H	167	ASP	0.41	42.46	42.46	-0.12	-0.12	1.290	1.340	2.204	1.176
	HL-TU	H	170A	GLN	0	0.00	0.00	2.76	2.76	-	-	0.014	0.270
	HL-TU	H	170	GLN	0	4.96	4.96	1.66	1.66	0.670	-	0.452	0.740
	HL-TU	H	230	ARG	0.51	30.87	30.87	-0.01	-0.01	1.180	1.330	0.186	-0.572
	HL-TU	L	39	LEU	0	107.16	103.32	0.01	0.01	0.960	1.240	0.464	-0.288
	HL-TU	L	42	ILE	0	3.54	3.53	2.00	2.00	0.250	0.080	0.212	0.748
	HL-TU	L	62	LYS	0	71.99	71.99	-0.25	-0.25	-0.010	-0.270	0.677	0.905
	HL-TU	L	64	GLN	0.8	62.34	60.52	-0.10	-0.10	0.250	2.980	1.158	0.636
	HL-TU	L	69	ILE	1.9	74.69	74.69	-0.15	-0.15	1.140	1.570	1.510	2.549
	HL-TU	L	71	PHE	1.2	107.57	107.49	-0.49	-0.49	2.370	3.070	1.170	0.392
	HL-TU	L	73	LEU	0	1.04	0.00	-0.17	2.98	1.050	-	0.162	0.190
	HL-TU	L	77	GLU	0	60.26	60.26	0.09	0.09	0.280	0.360	1.236	1.591
Continued on next page													

Table C.3 – continued from previous page

PDB	Inter	C	Pos	AA	$\Delta\Delta G_{exp}$	ASA ^a	ASA ^{sc}	d ^{ad}	d ^m	FoldX	Robetta	βACV_{ASA}	βACV_{ASA}^2
HL-TU	L	79	ARG		1.2	119.07	110.44	-0.31	-0.31	2.060	2.450	-0.388	-0.227
HL-TU	L	88	GLN		0	29.41	29.41	0.29	0.29	1.020	0.350	0.102	-0.092
HL-TU	L	92	VAL		0	79.29	65.59	-0.10	-0.10	1.990	1.050	0.625	0.332
HL-TU	L	93	ASN		0	29.29	29.28	-0.24	-0.10	1.360	0.120	0.867	0.368
HL-TU	L	94	GLU		0	39.69	38.31	0.51	0.51	0.800	-	0.149	0.258
HL-TU	T	15	LYS		-0.4	3.32	3.32	2.23	2.29	0.970	-	-0.420	-0.158
HL-TU	T	17	THR		0.1	28.20	25.69	-0.15	-0.15	1.130	0.130	0.372	0.250
HL-TU	T	18	ASN		0.2	11.27	11.27	0.50	0.97	1.090	0.040	0.916	1.048
HL-TU	T	20	LYS		2.6	75.90	75.89	0.05	0.05	2.810	1.500	1.714	2.736
HL-TU	T	22	ILE		0.7	27.72	27.72	0.24	0.24	2.640	0.650	1.182	-0.595
HL-TU	T	24	GLU		0.7	14.75	14.74	-0.31	-0.31	2.230	0.640	0.007	0.407
HL-TU	T	37	GLN		0.55	27.30	27.30	0.14	0.14	3.370	1.410	1.758	2.220
HL-TU	T	41	LYS		0.35	22.37	3.21	0.39	2.87	1.420	-	-0.139	0.087
HL-TU	T	42	SER		-0.1	35.00	18.10	0.34	2.40	1.420	-0.050	-0.114	-0.040
HL-TU	T	44	ASP		0.7	51.16	51.16	-0.18	-0.18	2.520	0.890	1.016	1.071
HL-TU	T	45	TRP		1.6	61.09	49.10	0.08	0.20	1.420	1.050	1.355	$\bar{\Delta}$
HL-TU	T	46	LYS		0.25	39.11	39.11	0.51	0.51	1.550	-	0.574	0.594
HL-TU	T	47	SER		0.05	26.81	22.94	0.31	0.83	1.160	0.600	0.842	0.407
HL-TU	T	48	LYS		0.4	22.83	22.83	0.09	0.09	1.840	0.430	1.018	1.442
HL-TU	T	50	PHE		0.4	101.62	101.18	-0.24	-0.24	4.060	2.610	1.229	-0.395
HL-TU	T	51	TYR		-0.1	34.10	34.11	0.43	0.43	2.110	-	-0.365	0.523
HL-TU	T	58	ASP		2.18	48.82	43.19	-0.24	-0.23	2.560	1.090	3.011	2.302
HL-TU	T	61	ASP		0.24	50.59	50.26	-0.16	-0.16	0.640	0.010	0.713	0.688
HL-TU	T	62	GLU		0	1.07	1.08	2.38	2.38	1.260	-	1.263	1.866
HL-TU	T	72	LEU		-0.06	0.24	0.24	2.65	3.11	1.430	-	0.591	0.148
HL-TU	T	76	PHE		1.2	15.98	15.98	0.22	0.22	1.700	0.620	2.068	1.469
HL-TU	T	78	TYR		0.7	0.26	0.26	2.78	2.78	1.400	-	-0.133	1.281
Continued on next page													

Table C.3 – continued from previous page

PDB	Inter	C	Pos	AA	$\Delta\Delta G_{exp}$	ASA ^a	ASA ^{sc}	d ^{ad}	d ^m	FoldX	Robetta	βACV_{ASA}	βACV_{ASA}^2
	HL-TU	U	94	TYR	1	91.98	76.28	-0.14	-0.14	2.140	2.700	0.709	1.709
	HL-TU	U	110	GLN	1.4	69.94	62.32	-0.17	-0.17	0.750	1.690	0.675	0.209
	HL-TU	U	128	GLU	0.1	14.49	14.49	0.79	0.79	0.750	-0.110	0.148	0.358
	HL-TU	U	131	ARG	0	66.22	59.92	-0.49	-0.49	0.750	0.280	0.438	-0.091
	HL-TU	U	132	THR	0	1.04	0.00	0.25	3.05	0.750	-	0.069	-0.337
	HL-TU	U	133	LEU	0	41.56	41.56	-0.05	-0.05	1.320	1.620	2.187	5.061
	HL-TU	U	135	ARG	0.55	69.95	69.96	-0.48	-0.48	0.750	0.940	0.992	1.791
	HL-TU	U	138	ASN	0	18.61	15.84	1.61	1.61	0.750	-	-0.306	-0.061
	HL-TU	U	140	PHE	1.5	23.34	23.33	0.42	0.42	0.910	1.540	1.072	1.393
	HL-TU	U	163	SER	0	28.06	24.60	-0.36	-0.36	0.750	0.420	0.232	0.005
	HL-TU	U	203	THR	0.1	25.60	25.61	0.19	0.19	0.750	0.220	0.757	0.594
	HL-TU	U	207	VAL	-0.2	73.65	54.98	0.02	0.02	1.960	1.120	1.122	-0.554
	HL-TU	U	208	GLU	0	25.82	25.82	0.33	0.33	0.750	0.280	0.111	0.376
1DFJ	E-I	I	202	GLU	1	24.08	24.08	0.14	0.14	-0.480	0.770	1.388	0.072
	E-I	I	257	TRP	1.3	39.18	39.18	0.14	0.14	1.030	1.570	2.607	0.254
	E-I	I	259	TRP	2.2	43.45	43.45	-0.02	-0.02	0.910	3.090	2.788	2.373
	E-I	I	283	GLU	1.3	23.90	23.90	0.01	0.01	0.290	-0.100	-0.307	-0.188
	E-I	I	285	SER	0.8	2.18	2.18	1.72	1.72	0.410	0.000	0.704	-0.048
	E-I	I	314	TRP	1	54.72	54.72	0.22	0.22	1.150	1.270	0.857	1.983
	E-I	I	316	LYS	1.3	42.17	42.17	0.32	0.32	-0.150	0.380	0.355	0.570
	E-I	I	397	GLU	1.3	26.34	26.34	-0.31	-0.31	-0.100	0.420	0.509	0.509
	E-I	I	430	TYR	5.9	107.72	107.72	-0.11	-0.11	3.750	2.720	2.247	5.802
	E-I	I	431	ASP	3.6	77.71	62.73	0.36	0.36	0.090	0.160	-0.126	0.121
	E-I	I	433	TYR	2.6	163.26	158.27	-0.34	-0.34	3.190	3.620	2.237	1.940
	E-I	I	453	ARG	0.8	28.63	28.62	-0.23	-0.23	0.810	0.210	-0.208	-0.224
	E-I	I	455	ILE	0.3	35.29	35.29	0.11	0.11	2.010	0.500	1.189	1.254
1DVF	AB-CD	A	30	HIS	1.7	18.00	18.00	0.35	0.35	0.140	0.260	0.277	0.426

Continued on next page

Table C.3 – continued from previous page

PDB	Inter	C	Pos	AA	$\Delta\Delta G_{exp}$	ASA ^a	ASA ^{sc}	d ^{ad}	d ^m	FoldX	Robetta	βACV_{ASA}	βACV_{ASA}^2
	AB-CD	A	32	TYR	2	44.27	44.28	0.16	0.16	0.060	0.430	2.385	1.193
	AB-CD	A	49	TYR	1.7	23.07	23.07	-0.23	-0.23	1.040	0.870	1.476	0.670
	AB-CD	A	50	TYR	0.7	53.01	53.01	-0.12	-0.12	0.060	0.550	1.014	2.055
	AB-CD	A	92	TRP	0.3	74.62	55.78	-0.20	0.01	0.470	1.550	1.834	0.216
	AB-CD	A	93	SER	1.2	11.38	11.38	0.99	3.22	0.130	-0.020	-0.173	-0.304
	AB-CD	B	30	THR	0.9	38.44	18.03	-0.12	1.32	-0.080	0.010	0.232	0.244
	AB-CD	B	32	TYR	1.8	24.36	22.14	0.16	1.29	0.840	0.120	1.184	0.853
	AB-CD	B	52	TRP	4.2	65.45	65.46	-0.13	-0.13	2.390	3.260	3.575	1.540
	AB-CD	B	54	ASP	4.3	68.05	65.23	-0.24	-0.24	2.130	0.480	1.953	2.479
	AB-CD	B	56	ASN	1.2	53.22	53.22	-0.29	-0.15	0.890	1.060	0.735	1.287
	AB-CD	B	58	ASP	1.6	14.14	14.14	0.14	0.14	1.270	0.710	1.012	1.234
	AB-CD	B	98	GLU	4.2	10.64	7.42	-0.26	-0.26	2.930	1.150	5.462	4.141
	AB-CD	B	99	ARG	1.9	59.96	59.96	0.16	0.16	-0.160	0.220	0.526	0.892
	AB-CD	B	100	ASP	2.8	56.32	46.95	0.00	0.00	-0.340	0.430	1.122	0.930
	AB-CD	B	101	TYR	4	69.34	67.93	-0.04	-0.04	1.840	2.190	4.101	5.855
	AB-CD	C	49	TYR	1.9	15.95	15.95	-0.24	-0.24	1.640	0.040	0.479	0.530
	AB-CD	D	33	HIS	1.9	26.99	26.99	0.00	0.00	0.650	0.790	-0.203	-0.131
	AB-CD	D	52	ASP	1.7	15.78	15.78	0.48	0.48	0.230	-0.230	1.005	0.034
	AB-CD	D	54	ASN	1.9	46.18	46.18	-0.23	-0.23	1.140	2.090	1.004	0.917
	AB-CD	D	97	ILE	2.7	49.54	49.55	-0.10	-0.10	1.900	1.070	0.715	0.182
	AB-CD	D	98	TYR	4.7	188.42	165.71	-0.26	-0.26	5.780	5.710	2.351	4.645
	AB-CD	D	100B	ARG	4.1	92.07	92.07	-0.20	-0.20	-	1.660	2.809	2.605
	AB-CD	D	100	GLN	1.6	96.01	90.58	-0.29	-0.29	1.540	1.440	0.959	0.730
1DX5	I-M	M	34	PHE	2.6	15.87	15.87	0.32	0.32	3.400	1.010	1.088	0.981
	I-M	M	36A	SER	-0.2	17.80	17.79	-0.05	-0.05	-	0.570	0.477	0.523
	I-M	M	37	PRO	1.2	53.02	53.02	0.12	0.73	2.410	-	0.315	0.462
	I-M	M	38	GLN	1.4	116.04	92.86	-0.24	-0.23	2.870	2.310	1.804	1.454
Continued on next page													

Table C.3 – continued from previous page

PDB	Inter	C	Pos	AA	$\Delta\Delta G_{exp}$	ASA ^a	ASA ^{sc}	d ^{ad}	d ^m	FoldX	Robetta	βACV_{ASA}	βACV_{ASA}^2
	I-M	M	39	GLU	-0.2	6.82	6.82	2.00	2.39	1.610	-0.090	0.511	0.622
	I-M	M	65	LEU	1	21.78	21.78	0.49	0.49	2.060	-	0.890	1.725
	I-M	M	67	ARG	3.4	14.96	14.96	-0.30	-0.30	3.220	-0.010	6.244	6.003
	I-M	M	74	THR	0.8	42.62	22.45	-0.23	-0.23	1.610	0.200	0.272	0.236
	I-M	M	75	ARG	0.7	32.74	32.74	0.05	0.13	1.900	0.050	0.044	0.478
	I-M	M	76	TYR	3	87.26	81.60	-0.23	-0.15	3.790	2.030	2.617	3.018
	I-M	M	77A	ARG	1.5	86.70	86.70	-0.25	-0.25	-	2.070	1.006	0.365
	I-M	M	80	GLU	3.4	2.26	0.00	1.52	1.52	1.610	-	3.078	3.294
	I-M	M	81	LYS	1	44.74	44.74	0.32	0.42	2.890	0.800	1.105	1.600
	I-M	M	82	ILE	2.6	71.53	47.74	-0.22	0.15	3.690	1.070	1.848	2.790
	I-M	M	84	MET	0.3	54.10	54.10	-0.14	-0.14	3.060	0.770	0.498	0.898
	I-M	M	110	LYS	0	52.26	52.27	0.18	0.18	1.910	0.170	0.434	0.544
1F47	A-B	A	4	ASP	0.7	56.93	32.43	-0.26	0.07	0.110	-0.150	0.508	0.877
	A-B	A	5	TYR	0.88	105.84	96.36	-0.31	-0.31	0.620	1.650	0.820	1.050
	A-B	A	6	LEU	0.94	96.95	89.61	-0.21	-0.01	1.440	1.940	1.360	2.096
	A-B	A	7	ASP	1.76	15.11	12.44	0.91	1.82	0.280	-	-0.340	-0.610
	A-B	A	8	ILE	2.56	85.67	85.67	0.08	0.08	2.760	2.070	1.129	2.451
	A-B	A	9	PRO	-0.06	4.74	4.74	1.80	1.80	0.230	-	0.222	-0.031
	A-B	A	11	PHE	2.49	91.36	90.43	-0.27	-0.27	3.180	2.430	1.813	2.142
	A-B	A	12	LEU	2.33	67.29	67.29	0.08	0.08	1.650	1.180	1.108	1.623
	A-B	A	15	GLN	-0.05	55.53	55.21	-0.80	-0.56	-0.340	0.040	0.828	0.683
1FC2	C-D	C	147	ASN	0.6	47.72	46.00	-0.30	-0.30	0.880	0.290	0.483	0.468
	C-D	C	150	ILE	2.2	26.04	25.51	0.27	0.27	0.450	0.810	1.519	0.642
	C-D	C	154	LYS	1.2	59.74	59.74	0.12	0.12	-0.310	0.180	1.043	0.594
1FCC	A-C	C	25	THR	0.24	27.59	27.59	1.10	1.10	-0.020	0.100	0.664	0.581
	A-C	C	27	GLU	4.9	59.61	59.61	-0.56	-0.56	4.060	3.140	3.872	3.590
	A-C	C	28	LYS	1.3	141.49	134.63	-0.21	-0.21	0.390	0.990	1.451	1.782
Continued on next page													

Table C.3 – continued from previous page

PDB	Inter	C	Pos	AA	$\Delta\Delta G_{exp}$	ASA ^a	ASA ^{sc}	d ^{ad}	d ^m	FoldX	Robetta	βACV_{ASA}	βACV_{ASA}^2
	A-C	C	31	LYS	3.5	73.61	72.13	-0.21	-0.21	0.290	1.910	3.865	4.237
	A-C	C	35	ASN	2.4	80.44	80.44	-0.61	-0.61	-1.620	1.180	2.048	2.426
	A-C	C	40	ASP	0.3	41.07	41.06	0.10	0.10	-0.310	-0.150	0.294	0.797
	A-C	C	42	GLU	0.4	38.85	38.85	-0.37	-0.37	-0.490	0.030	0.141	0.725
	A-C	C	43	TRP	3.8	55.53	30.91	-0.26	-0.24	1.250	2.710	2.033	8.300
1GC1	C-G	C	23	SER	0.29	4.22	4.22	1.42	1.42	0.290	-	0.509	0.656
	C-G	C	25	GLN	0.03	20.08	20.08	-0.01	-0.01	0.030	0.420	0.176	0.942
	C-G	C	27	HIS	0.28	38.62	38.62	-0.39	-0.39	1.010	0.800	0.323	0.527
	C-G	C	29	LYS	0.59	24.95	24.95	-0.24	-0.24	2.430	2.460	1.462	1.484
	C-G	C	32	ASN	0.18	15.44	3.01	-0.18	2.44	0.540	0.000	0.211	0.233
	C-G	C	33	GLN	0.1	28.86	17.33	-0.34	1.38	0.690	0.030	0.703	0.726
	C-G	C	35	LYS	0.32	84.45	83.89	-0.28	-0.28	-0.210	0.550	1.459	0.968
	C-G	C	40	GLN	-0.41	87.86	78.97	-0.16	0.03	-0.700	1.660	0.622	0.138
	C-G	C	42	SER	0	75.02	67.54	-0.14	-0.12	-0.800	0.030	1.237	0.393
	C-G	C	44	LEU	1.04	29.13	7.83	0.23	1.40	0.690	0.070	0.463	-0.835
	C-G	C	45	THR	-0.15	23.68	23.69	0.05	0.05	0.750	0.320	1.273	1.314
	C-G	C	52	ASN	0.7	23.31	23.31	-0.08	-0.08	0.640	1.010	1.349	1.325
	C-G	C	59	ARG	1.16	56.12	47.19	-0.57	-0.57	1.440	1.020	1.302	0.237
	C-G	C	60	SER	-0.09	54.63	39.70	-0.38	1.26	-	0.140	0.476	0.254
	C-G	C	63	ASP	-0.32	36.61	36.61	1.05	1.05	0.070	-0.050	0.719	0.582
	C-G	C	64	GLN	0.44	29.57	29.56	0.31	0.31	0.680	1.020	0.267	0.402
	C-G	C	85	GLU	1.31	1.77	1.77	1.30	1.30	0.770	-	-0.114	0.415
1JCK	A-B	B	20	THR	1.4	59.98	59.68	0.03	0.03	-0.370	1.260	1.075	0.317
	A-B	B	23	ASN	2.5	49.05	49.05	-0.38	-0.38	-1.120	1.960	2.483	2.346
	A-B	B	26	TYR	1.7	38.74	38.73	-0.33	-0.33	0.260	0.920	1.604	1.888
	A-B	B	60	ASN	1.3	42.33	42.33	0.29	0.29	-0.370	0.820	1.160	0.691
	A-B	B	90	TYR	2.5	32.16	31.73	-0.17	-0.13	0.710	1.000	1.012	$\bar{\lambda}$
													Continued on next page

Table C.3 – continued from previous page

PDB	Inter	C	Pos	AA	$\Delta\Delta G_{exp}$	ASA ^a	ASA ^{sc}	d ^{ad}	d ^m	FoldX	Robetta	βACV_{ASA}	βACV_{ASA}^2
	A-B	B	91	VAL	2.1	71.86	71.68	-0.16	-0.16	0.220	1.040	0.590	-2.089
	A-B	B	103	LYS	0.4	35.30	19.12	-0.54	1.44	-0.170	-0.260	2.121	2.663
	A-B	B	176	PHE	1.9	44.52	44.52	0.27	0.27	-0.270	0.630	-0.126	0.627
	A-B	B	210	GLN	2.5	14.05	14.05	-0.30	-0.30	-0.140	1.120	1.550	2.431
1JRH	HL-I	H	32	TYR	1.4	60.47	60.48	0.10	0.10	2.090	1.630	0.874	1.295
	HL-I	H	52	TRP	2.7	18.74	18.74	0.17	0.17	1.310	1.550	5.463	3.700
	HL-I	H	53	TRP	2.4	27.95	27.96	-0.13	-0.13	1.730	0.730	1.713	3.145
	HL-I	H	54	ASP	1.9	11.50	11.50	-0.08	-0.08	0.750	1.640	0.577	0.975
	HL-I	H	56	ASP	1.8	18.11	18.11	-0.08	-0.08	0.420	0.490	1.178	1.044
	HL-I	H	58	TYR	1.2	48.16	48.16	0.18	0.18	0.940	2.050	1.664	1.114
	HL-I	H	95	ARG	0.54	12.49	12.49	-0.12	-0.12	1.850	1.480	-0.036	-0.271
	HL-I	H	99	TYR	1.1	87.97	87.98	-0.35	-0.35	1.610	1.980	1.023	1.871
	HL-I	H	100B	HIS	1.7	49.52	49.52	-0.13	-0.13	-	3.380	1.685	2.977
	HL-I	I	47	LYS	3.6	36.96	36.96	-0.08	-0.08	2.550	1.400	4.665	4.751
	HL-I	I	48	ASN	-0.3	0.00	0.00	2.16	2.16	0.280	-	0.729	0.605
	HL-I	I	49	TYR	3.4	113.37	112.48	-0.26	-0.26	3.520	3.870	3.959	3.799
	HL-I	I	51	VAL	1.9	93.80	68.43	-0.13	-0.13	1.750	1.000	1.570	0.505
	HL-I	I	52	LYS	3	105.93	104.00	-0.08	-0.08	1.530	2.380	1.228	0.654
	HL-I	I	53	ASN	3.9	114.90	95.90	-0.17	-0.17	0.810	2.720	2.246	3.766
	HL-I	I	54	SER	0.3	21.04	14.86	0.19	2.50	0.180	-0.030	0.286	0.415
	HL-I	I	55	GLU	-0.4	60.34	60.28	-0.23	0.08	0.250	-0.100	-0.320	-0.618
	HL-I	I	79	ASN	-0.4	0.00	0.00	0.63	1.99	0.230	-	1.887	1.956
	HL-I	I	82	TRP	4.5	45.68	45.68	0.25	0.25	1.740	1.350	1.087	1.247
	HL-I	I	84	ARG	-0.3	19.08	19.08	0.60	0.60	0.650	0.370	0.217	0.600
	HL-I	I	98	LYS	0	64.43	64.43	-0.12	-0.12	-0.260	0.050	0.349	0.245
	HL-I	L	27	GLU	0.54	39.52	39.52	0.48	0.48	0.040	0.670	0.588	0.549
	HL-I	L	28	ASP	0.44	27.51	23.86	-0.12	0.55	0.540	0.670	0.770	0.356
Continued on next page													

Table C.3 – continued from previous page

PDB	Inter	C	Pos	AA	$\Delta\Delta G_{exp}$	ASA ^a	ASA ^{sc}	d ^{ad}	d ^m	FoldX	Robetta	βACV_{ASA}	βACV_{ASA}^2	
	HL-I	L	30	TYR	1.1	67.76	67.76	-0.10	-0.10	0.960	1.090	1.685	1.132	
	HL-I	L	91	TYR	0.58	5.32	0.06	0.18	0.28	0.260	0.280	1.072	1.659	
	HL-I	L	92	TRP	2.8	114.91	90.24	-0.08	-0.03	2.790	3.010	3.398	3.989	
	HL-I	L	93	SER	-0.65	46.74	46.72	-0.23	-0.23	-0.190	0.540	0.625	0.347	
	HL-I	L	94	THR	0.38	37.77	31.33	-0.17	0.08	0.230	0.360	0.715	0.563	
	HL-I	L	96	TRP	1.7	2.98	2.99	0.26	0.26	0.470	0.590	4.821	1.387	
1JTG	A-B	A	104	GLU	1.55	98.59	94.89	-0.22	-0.22	3.650	2.100	1.752	-0.953	
	A-B	A	105	TYR	-0.17	154.63	145.98	-0.27	-0.27	3.930	4.490	2.988	2.601	
	A-B	A	130	SER	0.33	21.34	20.74	-0.23	-0.23	3.020	-0.180	-0.208	-1.082	
	A-B	A	234	LYS	1	0.00	0.00	0.23	0.23	4.270	0.600	4.279	4.501	
	A-B	A	235	SER	1.3	4.21	3.95	-0.01	-0.01	3.820	2.330	0.678	-0.629	
	A-B	A	243	ARG	1.27	13.40	13.40	-0.30	-0.30	3.960	3.760	3.934	3.485	
	A-B	B	49	ASP	1.8	138.08	115.15	-0.30	-0.30	8.530	6.420	2.355	3.392	
	A-B	B	74	LYS	3.56	26.41	26.41	-0.27	-0.27	6.420	0.680	6.571	5.091	
	A-B	B	142	PHE	2.1	142.58	135.37	-0.06	-0.01	5.350	3.230	2.122	3.348	
	A-B	B	143	TYR	0.38	28.88	23.82	-0.11	-0.09	4.810	0.830	0.866	∞	
	1NMB	HL-N	H	99	TYR	1.5	59.68	59.69	-0.01	-0.01	0.320	1.280	-0.238	0.821
	1VFB	AB-C	A	30	HIS	0.8	28.66	28.67	0.33	0.33	0.400	0.260	0.486	0.684
	AB-C	A	32	TYR	1.3	54.48	54.49	0.14	0.14	2.630	1.310	2.206	1.466	
	AB-C	A	49	TYR	0.8	21.06	21.06	0.38	0.38	0.610	0.490	1.022	1.417	
	AB-C	A	50	TYR	0.4	66.18	66.18	-0.06	-0.06	2.610	1.520	1.760	4.406	
	AB-C	A	53	THR	-0.23	17.71	17.71	-0.25	-0.25	1.170	0.640	0.538	0.558	
	AB-C	A	92	TRP	1.71	60.86	60.86	0.15	0.15	3.290	2.060	1.965	-1.066	
	AB-C	A	93	SER	0.11	42.22	35.45	-0.16	0.63	0.050	0.000	0.487	0.672	
	AB-C	B	32	TYR	0.5	30.94	30.52	0.37	0.37	0.510	0.660	0.917	2.080	
	AB-C	B	52	TRP	1.23	53.78	53.78	-0.12	-0.12	1.520	1.750	2.530	2.592	
	AB-C	B	54	ASP	1.95	36.05	32.45	0.29	0.29	-0.500	-0.140	0.598	1.140	
	Continued on next page													

Table C.3 – continued from previous page

PDB	Inter	C	Pos	AA	$\Delta\Delta G_{exp}$	ASA ^a	ASA ^{sc}	d ^{ad}	d ^m	FoldX	Robetta	βACV_{ASA}	βACV_{ASA}^2
	AB-C	B	58	ASP	-0.2	0.00	0.00	2.98	2.98	0.040	-	0.540	0.887
	AB-C	B	99	ARG	0.47	48.98	48.98	-0.37	-0.37	-0.100	0.730	1.640	2.829
	AB-C	B	100	ASP	3.1	61.62	60.34	-0.14	-0.14	2.600	3.030	3.747	4.773
	AB-C	B	101	TYR	4	68.40	68.40	-0.13	-0.13	4.890	3.230	2.997	6.263
	AB-C	C	18	ASP	0.3	32.14	31.89	-0.06	-0.06	2.310	0.550	0.704	0.406
	AB-C	C	19	ASN	0.3	69.81	69.81	-0.25	-0.25	1.130	1.000	1.924	1.396
	AB-C	C	23	TYR	0.4	9.27	9.27	0.23	1.41	0.660	-	0.008	—
	AB-C	C	24	SER	0.8	43.53	41.63	-0.14	0.04	0.540	0.710	1.088	0.866
	AB-C	C	116	LYS	0.7	61.49	52.60	0.31	0.31	0.680	0.850	1.225	0.575
	AB-C	C	118	THR	0.8	44.15	33.90	0.04	0.29	0.900	0.100	0.567	0.816
	AB-C	C	119	ASP	1	44.87	40.94	-0.12	-0.10	1.570	1.670	1.505	1.386
	AB-C	C	120	VAL	0.9	13.75	13.27	0.12	0.12	1.370	0.230	1.630	6.105
	AB-C	C	121	GLN	2.9	120.40	119.81	-0.23	-0.23	5.330	4.200	3.120	2.683
	AB-C	C	124	ILE	1.2	22.89	21.77	0.49	0.49	1.580	-	1.203	1.759
	AB-C	C	125	ARG	1.8	56.47	56.47	0.12	0.12	2.450	2.220	0.608	1.174
	AB-C	C	129	LEU	0.2	28.38	0.00	0.33	0.33	0.840	-	0.081	—
2PTC	E-I	I	15	LYS	10.05	201.11	165.20	-0.48	0.06	1.560	4.160	4.115	2.994
	E-I	I	18	ILE	4.97	56.56	56.56	0.54	0.54	1.510	0.940	1.260	2.177
3HFM	HL-Y	H	31	SER	0.18	40.97	40.12	-0.59	-0.59	-0.140	0.380	0.795	0.572
	HL-Y	H	32	ASP	1.93	36.26	31.97	0.49	0.71	0.610	1.100	0.462	0.833
	HL-Y	H	33	TYR	6	74.92	74.92	-0.05	-0.05	2.770	2.900	3.916	7.609
	HL-Y	H	50	TYR	7.4	16.75	16.75	-0.54	-0.54	0.460	2.960	5.108	3.166
	HL-Y	H	53	TYR	3.3	91.75	89.66	-0.49	-0.26	1.510	1.660	2.160	4.738
	HL-Y	H	58	TYR	1.7	45.89	45.89	-0.50	-0.50	-0.630	1.770	1.195	3.219
	HL-Y	H	98	TRP	5.5	52.02	52.02	0.12	0.12	0.870	0.990	3.252	4.288
	HL-Y	L	31	ASN	5.2	53.35	53.35	-0.20	-0.16	1.450	1.860	1.159	1.470
	HL-Y	L	32	ASN	5.13	29.83	29.83	0.19	0.19	1.320	2.480	3.055	3.477
Continued on next page													

Table C.3 – continued from previous page

PDB	Inter	C	Pos	AA	$\Delta\Delta G_{exp}$	ASA ^a	ASA ^{sc}	d ^{ad}	d ^m	FoldX	Robetta	βACV_{ASA}	βACV_{ASA}^2
	HL-Y	L	50	TYR	4.6	57.84	57.84	-0.39	-0.39	1.090	1.400	2.913	-1.497
	HL-Y	L	53	GLN	0.96	53.23	53.23	-0.45	-0.45	1.640	0.830	1.079	0.828
	HL-Y	L	96	TYR	2.73	16.30	16.31	-0.16	-0.16	-0.730	0.450	3.018	2.763
	HL-Y	Y	15	HIS	-0.44	16.36	12.91	-0.09	0.98	-	0.100	-0.195	-0.039
	HL-Y	Y	20	TYR	4.9	66.59	66.59	0.02	0.02	0.800	2.230	3.631	8.701
	HL-Y	Y	21	ARG	1.07	121.93	118.40	-0.54	-0.54	-0.790	3.420	2.229	1.363
	HL-Y	Y	63	TRP	0.31	22.90	22.90	-0.17	-0.17	1.190	0.830	1.938	0.828
	HL-Y	Y	73	ARG	-0.33	67.88	67.88	-0.60	-0.60	-0.460	0.620	0.886	0.594
	HL-Y	Y	75	LEU	0.69	61.94	58.98	-0.06	-0.06	1.140	1.310	1.829	1.060
	HL-Y	Y	89	THR	0	15.37	15.37	0.36	0.36	0.300	0.170	0.009	0.268
	HL-Y	Y	93	ASN	0.21	48.27	44.04	-0.45	-0.45	1.650	1.530	1.045	0.988
	HL-Y	Y	96	LYS	7	53.30	51.52	-0.39	-0.39	1.200	2.130	3.384	4.369
	HL-Y	Y	97	LYS	6.2	77.64	72.16	-0.05	0.14	0.910	1.420	1.359	1.244
	HL-Y	Y	98	ILE	0	2.53	2.10	0.20	1.59	0.080	-	0.843	0.531
	HL-Y	Y	100	SER	0.26	41.61	23.24	0.01	0.16	-0.150	0.840	2.056	1.886
	HL-Y	Y	101	ASP	0.94	93.94	84.66	-0.49	-0.49	0.070	0.320	2.788	2.799
3HHR	A-BC	A	2	PRO	-0.05	86.40	75.92	0.09	0.09	1.420	-	0.160	0.135
	A-BC	A	3	THR	-0.05	4.59	0.77	1.72	3.16	-	0.010	-0.211	-0.476
	A-BC	A	4	ILE	0.41	69.54	68.36	0.40	0.40	0.330	1.050	0.699	0.103
	A-BC	A	5	PRO	0.43	0.00	0.00	2.86	2.86	-	-	0.255	0.400
	A-BC	A	8	ARG	0.2	88.45	88.45	0.14	0.14	-2.140	1.360	1.030	1.384
	A-BC	A	9	LEU	-0.04	18.77	14.72	-0.02	-0.02	0.320	0.440	1.031	1.934
	A-BC	A	12	ASN	0.1	59.83	59.14	-0.33	-0.33	0.020	3.020	2.881	1.872
	A-BC	A	15	LEU	0.15	75.00	70.45	-0.19	0.18	0.280	1.260	1.429	0.918
	A-BC	A	16	ARG	0.24	80.77	80.77	-0.21	-0.21	0.960	2.780	3.592	4.375
	A-BC	A	19	ARG	0.05	44.37	44.37	0.74	1.02	0.120	0.230	-0.059	-0.670
	A-BC	B	108	CYS	0	9.97	9.97	0.47	0.47	0.130	-0.040	1.002	0.941

Continued on next page

Table C.3 – continued from previous page

PDB	Inter	C	Pos	AA	$\Delta\Delta G_{exp}$	ASA ^a	ASA ^{sc}	d ^{ad}	d ^m	FoldX	Robetta	βACV_{ASA}	βACV_{ASA}^2
	A-BC	B	122	CYS	0	71.60	37.92	-0.09	1.08	0.210	-0.200	1.341	1.774
	A-BC	B	173	GLU	0.08	0.00	0.00	2.84	2.84	-	-	-0.139	0.068

Table C.4: Alanine mutations in ASEdb and their predictions for Chapter 4. 1 in columns KFC, Minerva², Minerva¹ and HotPoint means the mutated residue is predicted as a hot spot residue. The dataset and the prediction of Robetta, FoldX, KFC and Minerva are from [66].

PDB	Chain	Pos	Amino acid	$\Delta\Delta G_{exp}$	Robetta	FoldX	KFC	Minerva ²	Minerva ¹	HotPoint	Z-score
1a4y	A	261	TRP	0.1	1.05	0.6	-	-	-	-	3.024
	A	263	TRP	1.2	2.27	2.2	-	-	1	1	5.346
	A	289	SER	0	0.55	0.84	-	-	1	-	0.788
	A	318	TRP	1.5	2.19	2.7	-	-	1	1	4.186
	A	320	LYS	-0.3	-0.21	-0.75	-	-	-	-	-3.641
	A	344	GLU	0.2	1.37	-0.44	-	-	-	1	-1.334
	A	375	TRP	1	2.83	2.51	-	-	1	1	3.282
	A	401	GLU	0.9	0.02	-0.2	-	-	-	1	-1.640
	A	434	TYR	3.3	3.03	1.93	-	-	1	1	2.460
	A	435	ASP	3.5	0.58	3.2	1	-	1	-	-2.885
	A	437	TYR	0.8	3.13	1.07	-	-	-	1	2.656
	A	457	ARG	-0.2	-0.04	0.32	-	-	-	-	-0.061
	A	459	ILE	0.7	0.63	0.41	-	-	-	1	1.013
	B	5	ARG	2.3	2.54	2.09	-	-	1	1	-0.596
	B	8	HIS	0.9	0.84	-0.37	-	-	-	-	-0.013
	B	12	GLN	0.3	1.07	0.4	-	-	-	1	-1.706
	B	13	HIS	-0.3	0.02	0.19	-	-	-	-	0.299
	B	31	ARG	0.2	2.77	-0.83	-	-	-	-	-1.876
	B	32	ARG	0.9	0.18	0.06	-	-	-	-	-0.419
	B	68	ASN	0.2	0.52	0.05	-	-	-	-	-0.175
	B	84	HIS	0.2	1.01	0.25	-	-	-	1	-0.379
	B	89	TRP	0.2	2.71	0.88	-	-	-	-	1.658
	B	108	GLU	-0.3	1.73	-0.15	-	-	-	-	-1.857

Continued on next page

Table C.4 – continued from previous page

PDB	Chain	Pos	Amino acid	$\Delta\Delta G_{exp}$	Robetta	FoldX	KFC	Minerva ²	Minerva ¹	HotPoint	Z-score
1a22	B	114	HIS	0.65	1.69	0.49	-	-	-	-	0.045
	A	18	HIS	-0.5	1.84	-1.32	-	-	-	-	0.331
	A	21	HIS	0.2	0.78	0.2	-	-	-	-	0.205
	A	22	GLN	-0.2	0.07	0.36	-	-	-	-	-1.003
	A	25	PHE	-0.4	1.3	1.08	-	-	-	1	1.496
	A	42	TYR	0.2	2	1.34	1	-	1	-	1.486
	A	45	LEU	1.2	1.15	1.36	-	-	-	1	3.402
	A	46	GLN	0.1	1.04	0.17	-	-	-	-	-1.793
	A	51	SER	0.3	-0.02	0	-	-	-	-	0.259
	A	56	GLU	0.4	0.97	0.21	-	-	-	-	-1.379
	A	62	SER	0.1	-0.06	1.03	-	-	-	-	-0.439
	A	63	ASN	0.3	0.39	-0.25	-	-	-	-	-2.033
	A	64	ARG	1.6	2	1.42	1	-	1	-	-1.430
	A	65	GLU	-0.5	-0.11	-0.16	-	-	-	-	-0.558
	A	68	GLN	0.6	1.81	-0.22	-	-	-	-	-1.378
	A	164	TYR	0.3	0.96	-1.08	1	-	1	-	2.594
	A	167	ARG	0.3	0.42	-0.64	1	-	-	1	-1.465
	A	168	LYS	-0.2	2.1	-2.41	-	1	-	-	-5.873
	A	171	ASP	0.8	8.73	3.94	1	-	-	-	-3.326
	A	172	LYS	2	0.74	0.77	1	-	-	-	-3.523
	A	174	GLU	-0.9	0.85	1.66	1	-	-	-	-3.006
	A	175	THR	2	2.04	1.23	1	-	-	-	0.436
	A	176	PHE	1.9	0.6	0.48	-	-	-	-	4.004
	A	178	ARG	2.4	0.38	0.69	-	-	1	1	-2.041
	A	179	ILE	0.8	1	0.83	1	-	-	1	1.613
	A	183	ARG	0.5	0.07	-0.08	-	-	-	-	-0.166
	B	243	ARG	2.12	6.54	2.67	1	1	-	1	-1.246
	B	244	GLU	1.69	0.49	0.02	-	-	-	-	-1.454

Continued on next page

Table C.4 – continued from previous page

PDB Chain	Pos	Amino acid	$\Delta\Delta G_{exp}$	Robetta	FoldX	KFC	Minerva ²	Minerva ¹	HotPoint	Z-score
B	271	ARG	0.54	1.31	3.84	-	-	1	1	-2.737
B	273	THR	0.11	0	0	-	-	-	-	0.173
B	274	GLN	0	0.03	-0.01	-	-	-	-	0.003
B	275	GLU	-0.1	0.09	0.46	-	-	-	-	-0.708
B	276	TRP	0.51	2.86	2.49	-	-	-	1	3.155
B	280	TRP	-0.02	0.04	0.03	-	-	-	-	0.354
B	298	SER	-0.05	0.29	-0.01	-	-	-	-	0.224
B	302	SER	-0.2	-0.11	0.96	-	-	-	-	0.201
B	303	ILE	1.61	0.31	0.9	-	-	-	1	2.833
B	304	TRP	4.5	5.38	5.17	1	1	1	1	9.443
B	305	ILE	1.94	0.13	0.4	-	-	-	-	2.505
B	320	GLU	-0.19	0.86	-0.29	-	-	1	-	-1.266
B	321	LYS	0.08	0.05	0.24	-	-	-	-	-2.941
B	324	SER	0.28	-0.06	-0.96	-	-	-	-	0.461
B	326	ASP	0.99	-0.31	-0.95	1	-	-	-	-2.069
B	327	GLU	0.97	1.11	-1.54	1	-	1	-	-2.409
B	364	ASP	1.49	1.37	0.85	1	-	-	-	-0.938
B	365	ILE	2.13	0.09	0.15	-	-	-	1	2.554
B	366	GLN	0.02	0.68	0.29	-	-	-	-	-0.620
B	367	LYS	-0.02	0.22	0.07	-	-	-	-	-3.217
B	369	TRP	4.5	3.63	4.44	1	1	1	1	6.535
B	371	VAL	-0.64	0.26	0.02	1	-	-	-	1.164
B	416	GLN	0.89	-0.03	0.02	-	-	-	-	-1.021
B	417	ARG	0.28	0.23	-0.7	-	-	-	-	-0.820
B	418	ASN	0.3	0.79	0	-	-	-	-	-1.000
B	419	SER	0.03	0.98	0.74	-	-	-	-	-0.044
1ahw	C 156	TYR	4	4.6	2	1	1	1	1	1.502
C	167	THR	0	-0.03	0.05	-	-	1	-	-0.423

Continued on next page

Table C.4 – continued from previous page

PDB Chain	Pos	Amino acid	$\Delta\Delta G_{exp}$	Robetta	FoldX	KFC	Minerva ²	Minerva ¹	HotPoint	Z-score
1brs	C 170	THR	1	-0.06	0.45	-	-	-	-	0.357
	C 176	LEU	1	0.01	0	-	-	-	-	0.418
	C 178	ASP	-0.5	-0.09	0.15	-	-	-	-	-0.042
	C 197	THR	1.3	-0.02	0.01	-	-	-	-	0.016
	C 198	VAL	-0.3	-0.01	-0.03	-	-	-	-	0.397
	A 27	LYS	5.4	1.91	2.58	-	1	1	-	-2.520
	A 59	ARG	5.2	3.01	0.83	-	-	1	-	-1.344
	A 60	GLU	-0.2	1.65	0.43	1	-	-	-	-1.693
	A 73	GLU	2.8	-0.2	-0.59	-	-	-	-	-1.621
	A 87	ARG	5.5	4.44	-0.12	1	-	-	1	-1.209
1bxi	A 102	HIS	6	5.08	1.81	-	1	1	1	1.145
	D 29	TYR	3.4	3.13	1.86	1	1	1	-	1.640
	D 35	ASP	4.5	1.42	0.54	1	1	1	1	-3.096
	D 39	ASP	7.7	9.4	3.07	1	1	1	1	-3.457
	D 42	THR	1.8	1.69	2.01	-	-	-	-	0.870
	D 76	GLU	1.3	1.54	0.12	-	-	-	-	-0.359
	A 23	CYS	0.92	-0.08	-0.1	-	-	1	1	3.540
	A 24	ASN	0.14	0	-0.05	-	-	-	-	-0.622
	A 27	THR	0.73	0.59	0.9	1	-	-	-	0.547
	A 28	SER	0.17	0	0.04	-	-	-	-	0.200
A 29	SER	0.96	0.35	-0.04	-	-	-	-	0.174	
A 30	GLU	1.41	2.97	0.94	1	-	-	1	-2.855	
A 33	LEU	3.42	1.02	1.07	1	1	1	1	1.544	
A 34	VAL	2.58	0.98	0.91	-	-	1	1	0.909	
A 37	VAL	1.66	0.5	0.44	-	-	1	1	1.599	
A 38	THR	0.9	1.34	1	-	-	-	-	0.181	
A 41	GLU	2.08	-0.08	0.58	-	-	-	-	-2.430	
A 48	SER	0.01	-0.01	-0.05	-	-	-	-	0.259	
Continued on next page										

Table C.4 – continued from previous page

PDB Chain Pos			Amino acid	$\Delta\Delta G_{exp}$	Robetta	FoldX	KFC	Minerva ²	Minerva ¹	HotPoint	Z-score
1cbw	A	49	GLY	1.49	-0.43	-0.52	-	-	-	-	0.198
	A	50	SER	2.19	5.4	1.39	1	1	-	1	0.352
	A	51	ASP	5.92	0.82	1.07	1	-	1	-	-1.827
	A	55	TYR	4.63	3.31	3.31	1	1	1	-	2.201
	A	56	PRO	1.24	0.56	0.13	-	-	-	-	-0.814
	D	11	THR	0.2	0.18	0.06	1	-	-	-	-0.631
1dan	D	15	LYS	2	-0.03	-0.74	1	-	1	-	-13.034
	D	17	ARG	0.5	1.53	1.22	1	-	-	-	-3.710
	D	19	ILE	0.1	0.68	0.4	-	-	-	-	0.869
	D	34	VAL	0	0.3	0.4	1	-	-	-	0.668
	D	39	ARG	0.2	1.58	1.17	-	-	-	-	-1.509
	T	15	LYS	-0.4	-0.02	-0.08	-	-	-	-	-0.981
1dvf	T	17	THR	0.1	0.13	0.22	1	-	-	-	0.287
	T	18	ASN	0.2	0.04	0.19	-	-	-	-	-1.866
	T	20	LYS	2.6	1.5	0.16	1	1	1	-	-7.374
	T	21	THR	-0.2	0	0.5	-	-	1	-	0.204
	T	22	ILE	0.7	0.65	0.64	1	-	-	1	1.940
	T	24	GLU	0.7	0.64	-0.44	-	-	1	-	-0.409
	T	41	LYS	-0.04	-0.04	-0.07	-	-	-	-	-1.235
	T	42	SER	-0.05	-0.05	-0.08	-	-	-	-	0.153
	T	44	ASP	0.7	0.89	0.89	-	-	-	-	-0.739
	T	46	LYS	0.25	0.1	0.22	-	-	-	-	-2.902
	T	47	SER	0.05	0.56	-0.19	-	-	-	-	0.080
	T	48	LYS	0.4	0.43	0.04	-	-	-	-	-4.798
	T	50	PHE	0.4	2.61	3.09	-	-	1	1	2.946
	T	58	ASP	2.18	1.24	1.13	1	-	-	-	-1.652
	T	68	LYS	-0.1	0	1	-	-	-	-	0.204
	A	30	HIS	1.7	0.26	0.2	-	-	-	-	0.076
Continued on next page											

Table C.4 – continued from previous page

PDB	Chain	Pos	Amino acid	$\Delta\Delta G_{exp}$	Robetta	FoldX	KFC	Minerva ²	Minerva ¹	HotPoint	Z-score
1f47	A	32	TYR	2	0.47	0.24	1	1	1	-	1.627
	A	49	TYR	1.7	0.87	0.43	-	1	1	-	0.996
	A	50	TYR	0.7	0.55	-0.11	-	-	-	-	1.192
	A	92	TRP	0.3	1.56	0.68	-	-	1	-	2.961
	A	93	SER	1.2	-0.02	0.01	-	-	-	-	0.285
	B	30	THR	0.9	0.01	0.18	-	-	-	-	0.042
	B	32	TYR	1.8	0.11	0.19	-	-	1	1	1.304
	B	52	TRP	4.2	3.26	3.52	1	1	1	1	5.626
	B	54	ASP	4.3	0.72	0.86	1	-	-	-	-1.315
	B	56	ASN	1.2	1.07	-0.68	-	-	-	-	-0.172
1fc2	B	58	ASP	1.6	0.71	0.28	-	-	-	-	-0.445
	B	98	GLU	4.2	0.05	1.76	1	-	-	1	-2.310
	B	99	ARG	1.9	0.22	0.94	-	-	1	-	-1.566
	B	100	ASP	2.8	0.43	0.69	1	-	-	-	-1.771
	B	101	TYR	4	2.13	1.62	1	1	-	-	1.613
	D	33	HIS	1.9	0.79	0.46	-	-	-	-	0.437
	D	52	ASP	1.7	-0.18	-0.29	-	-	1	-	-2.057
	A	4	ASP	0.7	-0.15	0.4	-	-	-	-	-0.251
	A	5	TYR	0.9	1.65	1.26	-	-	-	-	1.289
	A	6	LEU	0.9	1.94	2.27	1	-	-	1	2.194
1fc2	A	7	ASP	1.8	-0.08	0.09	-	-	-	-	-0.533
	A	8	ILE	2.5	2.07	2.39	-	1	1	1	2.054
	A	11	PHE	2.5	2.43	2.26	1	1	1	1	4.933
	A	12	LEU	2.3	1.13	1.37	-	-	1	1	1.153
	A	15	GLN	0	0.03	0.08	-	-	-	-	-0.655
	C	147	ASN	0.6	0.61	0.92	-	-	-	-	-0.639
	C	150	ILE	2.2	0.78	0.32	-	-	-	1	2.537
	C	154	LYS	1.2	0.16	-0.13	-	-	-	-	-2.574
	Continued on next page										

Table C.4 – continued from previous page

PDB Chain Pos			Amino acid	$\Delta\Delta G_{exp}$	Robetta	FoldX	KFC	Minerva ²	Minerva ¹	HotPoint	Z-score
1fcc	C	25	THR	0.24	0.1	0.21	-	-	-	-	0.779
	C	27	GLU	4.9	3.21	1.23	1	1	1	1	-4.397
	C	28	LYS	1.3	0.99	0.55	-	1	1	-	-6.289
	C	31	LYS	3.5	1.91	1.55	-	1	1	1	-6.033
1gc1	C	35	ASN	2.4	1.18	1.69	-	1	1	-	-1.499
	C	40	ASP	0.3	-0.15	0.22	-	-	-	-	-0.743
	C	42	GLU	0.4	0.03	0.02	-	-	-	-	-0.566
	C	43	TRP	3.8	2.71	1.01	1	1	1	1	2.525
	C	23	SER	0.29	0.27	-0.15	-	-	-	-	0.459
	C	25	GLN	0.03	0.42	2.82	-	-	-	-	-0.666
	C	27	HIS	0.28	0.8	0.4	-	-	-	-	0.645
	C	29	LYS	0.59	2.46	-0.07	-	-	-	-	-3.885
	C	32	ASN	0.18	0	0.02	-	-	-	-	-0.118
	C	33	GLN	0.1	0.03	-0.05	-	-	-	-	-0.567
1jrh	C	35	LYS	0.32	0.56	-1.19	1	-	-	-	-7.313
	C	40	GLN	-0.41	1.66	4.17	1	1	1	1	-2.880
	C	42	SER	0	0.2	-1.06	-	-	-	-	0.431
	C	44	LEU	1.04	0.07	0.43	-	-	1	1	1.051
	C	45	THR	-0.15	0.32	0.65	-	-	-	-	0.250
	C	52	ASN	0.7	1.01	-0.04	-	-	-	-	-0.592
	C	59	ARG	1.16	1.02	0.45	-	-	1	-	-1.495
	C	60	SER	-0.09	0.13	-0.36	-	-	-	-	0.387
	C	63	ASP	-0.32	-0.05	0.49	-	-	-	-	-0.906
	C	64	GLN	0.44	1.16	-0.47	-	-	-	-	-0.575
	C	85	GLU	1.31	2.1	1.02	-	-	-	-	-1.335
	L	27	GLU	0.54	0.67	0.31	-	-	-	-	-0.868
	L	28	ASP	0.44	0.67	0.29	-	-	-	-	-0.393
	L	30	TYR	1.1	1.08	1.11	-	1	1	-	1.134
Continued on next page											

Table C.4 – continued from previous page

PDB	Chain	Pos	Amino acid	$\Delta\Delta G_{exp}$	Robetta	FoldX	KFC	Minerva ²	Minerva ¹	HotPoint	Z-score
L	91	TYR	0.58	0.28	0.13	-	-	-	-	-	1.149
L	92	TRP	2.8	3.01	3.45	1	1	1	1	1	4.484
L	93	SER	0.54	0.54	-0.18	-	-	-	-	-	0.565
L	94	THR	0.36	0.36	0.25	-	-	-	-	-	-0.065
L	96	TRP	0.59	0.59	0.17	-	-	-	-	-	2.341
H	32	TYR	1.4	1.63	-0.22	-	-	-	-	-	0.682
H	52	TRP	2.7	1.55	1.88	-	-	-	-	1	3.500
H	53	TRP	2.4	0.73	0.87	-	-	-	-	-	2.588
H	54	ASP	1.9	1.64	0.24	-	-	-	-	-	-0.293
H	56	ASP	1.8	0.49	0.09	-	-	-	1	-	-0.518
H	58	TYR	1.2	2.49	0.87	-	1	1	1	-	1.590
H	95	ARG	0.54	1.49	0.16	-	-	-	-	-	-1.201
H	99	TYR	1.1	1.98	2.81	-	-	-	1	-	0.592
H	100	GLY	1.7	3.38	-0.23	-	-	-	-	-	-0.562
I	47	LYS	3.6	1.4	3.55	-	1	1	1	-	-4.739
I	49	TYR	3.4	3.87	1.73	1	1	1	1	-	2.958
I	51	VAL	1.9	1	1.38	1	-	-	-	1	1.007
I	52	LYS	3	2.36	2.21	1	1	1	1	-	-5.757
I	53	ASN	3.9	2.74	0.99	-	1	1	1	-	-3.014
I	54	SER	0.3	-0.03	-0.14	-	-	-	-	-	-0.325
I	55	GLU	-0.4	-0.1	0.08	-	-	-	-	-	-1.231
I	82	TRP	4.5	1.35	1.85	-	1	-	-	-	2.619
I	84	ARG	-0.3	0.37	0.16	-	-	-	-	-	-0.228
I	98	LYS	0	0.04	0.42	-	-	-	-	-	-1.433
1nmb	H	99	TYR	1.5	1.27	1.03	-	-	-	-	0.758
1vfb	A	30	HIS	0.8	0.26	0.21	-	-	1	-	0.488
A	32	TYR	1.3	1.3	1	-	1	1	1	1	1.654
A	49	TYR	0.8	0.42	0.5	-	1	1	1	-	0.987

Continued on next page

Table C.4 – continued from previous page

PDB	Chain	Pos	Amino acid	$\Delta\Delta G_{exp}$	Robetta	FoldX	KFC	Minerva ²	Minerva ¹	HotPoint	Z-score
2ptc 3hfm	A	50	TYR	0.4	1.54	1.28	-	-	1	-	1.363
	A	53	THR	-0.23	0.58	0.66	-	-	-	-	-0.172
	A	92	TRP	1.71	2.06	3.01	-	1	1	1	2.992
	A	93	SER	0.11	0	-0.29	-	-	-	-	0.473
	B	30	THR	0.09	1.22	-0.01	-	-	-	-	0.273
	B	32	TYR	0.5	0.64	-0.09	-	1	-	1	1.429
	B	52	TRP	1.23	0.23	1.65	-	-	-	1	3.527
	B	99	ARG	0.47	0.73	0.37	-	-	-	-	-1.237
	B	100	ASP	3.1	3.29	2.75	-	-	1	-	-1.851
	B	101	TYR	4	2.63	3.04	1	1	1	-	2.544
	C	18	ASP	0.3	0.55	0.58	-	-	-	-	-1.374
	C	19	ASN	0.3	1.04	0.84	-	-	1	-	-1.390
	C	23	TYR	0.4	0	0.15	-	1	1	1	2.301
	C	24	SER	0.8	0.97	-0.16	-	-	1	-	0.599
	C	116	LYS	0.7	0.83	-0.01	-	-	1	-	-3.338
	C	118	THR	0.8	0.1	0.05	-	-	-	-	0.215
	C	119	ASP	1	-0.2	0.35	-	-	-	-	-1.729
	C	120	VAL	0.9	0.23	0.44	-	-	-	1	1.635
	C	121	GLN	2.9	4.2	2.58	-	1	1	-	-3.168
	C	124	ILE	1.2	0.46	0.4	-	1	1	1	1.187
	C	125	ARG	1.8	2.22	0.65	-	-	1	-	-1.146
	C	129	LEU	0.2	0.06	-0.11	-	-	-	-	0.315
	I	15	LYS	10	4.16	0.33	1	1	1	1	-16.562
	H	31	SER	0.2	0.45	0.29	-	-	-	-	-0.105
	H	32	ASP	2	1.1	0.03	-	-	-	-	-1.910
	H	33	TYR	6	2.9	2.53	1	1	1	1	2.894
	H	50	TYR	7.5	2.92	1.97	-	1	1	1	1.710
	H	53	TYR	3.29	2.04	0.9	-	1	1	1	2.149

Continued on next page

Table C.4 – continued from previous page

PDB Chain	Pos	Amino acid	$\Delta\Delta G_{exp}$	Robetta	FoldX	KFC	Minerva ²	Minerva ¹	HotPoint	Z-score
H	58	TYR	1.7	1.77	0.33	-	-	-	-	1.403
Y	20	TYR	5	2.72	1.46	1	1	1	1	1.831
Y	21	ARG	1	3.47	-1.32	1	1	1	-	-1.797
Y	63	TRP	0.3	0.83	0.75	-	1	1	1	2.885
Y	73	ARG	-0.2	0.62	-0.4	-	-	1	-	-0.814
Y	75	LEU	1.25	0.69	0.89	-	-	1	1	1.384
Y	89	THR	0	0.17	0.08	-	-	-	-	0.228
Y	93	ASN	0.6	1.56	2.01	-	-	1	-	-0.233
Y	96	LYS	7	2.13	-1.36	1	1	1	1	-6.429
Y	97	LYS	6	1.42	0.32	1	1	1	-	-4.465
Y	100	SER	0.25	0.95	-0.11	1	-	1	-	-0.369
Y	101	ASP	1.02	0.94	-2.39	1	1	1	-	-2.519
Y	15	HIS	-0.5	0.71	0.02	-	-	-	-	0.850
L	31	ASN	5.25	1.86	1.2	-	-	1	-	-1.282
L	32	ASN	5.2	1.27	2.41	-	-	1	-	-1.744
L	50	TYR	4.6	1.42	1.82	-	-	1	1	1.956
L	53	GLN	1	0.83	1.2	-	-	-	-	-0.870
L	96	TYR	2.8	0.55	0.91	-	1	1	1	1.223

Table C.5: Alanine mutations in BID and their predictions for Chapter 4. 1 in columns KFC, Minerva², Minerva¹ and HotPoint means the mutated residue is predicted as a hot spot residue. The dataset and the prediction of Robetta, FoldX, KFC and Minerva are from [66].

PDB	Chain	Pos	Amino acid	$\Delta\Delta G_{exp}$	Robetta	FoldX	KFC	Minerva ²	Minerva ¹	HotPoint	Z-score	
1cdl	A	12	PHE	insignificant	0.7	0.63	-	1	-	1	3.455	
	A	19	PHE	weak	1.79	1.71	1	-	1	1	7.364	
	A	92	PHE	strong	1.51	1.65	1	1	1	1	4.674	
	E	799	LYS	insignificant	0.61	2.52	-	-	-	-	-3.931	
	E	800	TRP	strong	5.18	4.4	1	1	1	1	7.517	
	E	802	LYS	intermediate	1.53	2.03	-	-	-	-	-5.531	
	E	804	GLY	strong	-3.4	-6.57	-	-	-	1	-1.101	
	E	808	ARG	intermediate	2.7	0.69	1	-	-	1	-2.241	
	E	810	ILE	strong	1.92	4.49	1	1	-	1	2.614	
	E	811	GLY	intermediate	-3.78	-6.69	-	1	-	-	0.103	
	E	812	ARG	strong	1.24	0.51	-	1	1	1	-0.806	
	E	813	LEU	strong	2.01	4.61	1	-	1	1	2.282	
	1dva	H	38	GLY	intermediate	-0.43	-1.2	-	-	1	-	0.016
H		65	ILE	insignificant	0.73	0.82	-	-	-	-	1.595	
H		67	VAL	insignificant	0.7	1.13	-	-	1	1	2.151	
H		70	GLU	weak	1.02	0.01	-	-	-	1	-2.664	
H		73	LEU	insignificant	0.53	1.17	1	1	1	1	1.780	
H		74	SER	insignificant	0.11	-0.01	-	-	-	-	0.283	
H		75	GLU	insignificant	0	-0.08	-	-	-	-	-0.732	
H		76	HIS	strong	0.43	1.51	-	-	-	-	0.210	
H		80	GLU	insignificant	0.01	-1.57	-	-	-	-	-2.772	
H		82	SER	insignificant	-0.01	-0.5	-	-	-	-	-0.063	
H		144	LEU	insignificant	0.28	0.23	-	-	-	1	0.564	
Continued on next page												

Table C.5 – continued from previous page

PDB	Chain	Pos	Amino acid	$\Delta\Delta G_{exp}$	Robertta	FoldX	KFC	Minerva ²	Minerva ¹	HotPoint	Z-score
	H	153	LEU	weak	0.82	1.1	-	-	1	1	1.168
	X	1	ALA	insignificant	0.02	0	-	-	-	-	0.204
	X	2	LEU	strong	2.31	2.27	-	1	1	1	1.206
	X	5	ASP	weak	1.65	0	-	-	-	-	-1.666
	X	7	ARG	weak	4.4	1.45	-	-	1	-	-1.380
	X	8	VAL	intermediate	0.57	0.91	-	-	-	1	1.345
	X	9	ASP	intermediate	0.66	-2.25	-	1	1	-	-3.421
	X	11	TRP	strong	2.61	2.45	-	-	1	1	5.439
	X	12	TYR	strong	3.16	1.92	1	-	1	1	3.318
	X	14	GLN	insignificant	0.1	-0.11	-	-	-	-	-0.388
	X	15	PHE	strong	1.58	1.98	1	1	1	1	3.668
	X	16	VAL	insignificant	0.22	0.23	-	-	-	-	0.507
1dx5BN	N	24	ILE	insignificant	0.16	0.02	-	-	-	-	0.732
	N	235	LYS	insignificant	-0.07	-0.06	-	-	-	-	-2.092
1dx5NJ	N	34	PHE	intermediate	0.98	1.47	1	-	-	1	2.784
	N	36	LYS	weak	0.01	-0.09	-	-	-	-	-1.918
	N	37	PRO	weak	0.9	0.06	-	-	-	-	-0.432
	N	38	GLN	weak	2.36	1.44	-	1	1	1	-2.892
	N	39	GLU	insignificant	-0.08	-0.1	-	-	-	-	-1.737
	N	65	LEU	weak	0.5	0.48	-	-	-	-	1.395
	N	67	ARG	strong	0.02	0.56	-	-	-	1	-1.681
	N	74	THR	weak	0.15	0.29	-	-	-	-	0.074
	N	75	ARG	weak	-0.01	0.67	-	-	1	-	-0.656
	N	76	TYR	strong	1.99	1.79	1	-	1	1	2.109
	N	80	GLU	strong	2.26	-0.21	-	-	-	-	-3.181
	N	81	LYS	weak	0.85	1.2	-	-	-	-	-3.184
	N	82	ILE	intermediate	1.05	1.48	-	1	1	1	2.162
	N	84	MET	insignificant	0.74	0.62	-	-	-	-	0.704

Continued on next page

Table C.5 – continued from previous page

PDB	Chain	Pos	Amino acid	$\Delta\Delta G_{exp}$	Robetta	FoldX	KFC	Minerva ²	Minerva ¹	HotPoint	Z-score
1ebpAC	N	110	LYS	insignificant	0.15	0.11	-	-	-	-	-1.496
	A	93	PHE	strong	0.75	0.75	-	-	-	-	3.683
	A	150	MET	strong	1.56	2.14	-	1	1	1	1.651
	A	151	THR	weak	-0.03	0.61	-	-	-	-	-0.055
	A	205	PHE	strong	0.71	1.06	-	-	1	1	1.575
	C	9	GLY	intermediate	0.08	-0.06	-	-	-	-	-0.256
	C	10	PRO	intermediate	0.31	0.28	-	-	1	-	-0.209
	C	11	LEU	insignificant	0.48	0.57	-	-	1	-	0.912
	C	12	THR	weak	-0.24	-0.12	-	-	1	-	-0.042
	C	13	TRP	strong	1.29	1.66	-	1	1	-	2.827
1es7AB	A	49	PHE	intermediate	2.78	2.8	1	1	-	1	7.036
	A	50	PRO	intermediate	2.08	2.15	1	1	-	-	-2.124
1es7AC	A	26	VAL	intermediate	0.4	0.59	-	-	-	-	1.186
1es7AD	A	31	TRP	strong	2.09	2.55	-	1	1	-	3.292
1fakHT	T	37	GLN	weak	0.64	0.23	1	-	-	-	-2.289
	T	41	LYS	insignificant	-0.01	-0.07	-	-	-	-	-0.874
	T	42	SER	insignificant	-0.07	0.03	-	-	-	-	0.134
	T	44	ASP	weak	0.89	3.06	-	-	-	-	-0.819
	T	94	TYR	weak	1.77	1.31	-	-	1	1	1.344
1fakLT	T	15	LYS	insignificant	-0.03	-0.17	-	-	-	-	-1.130
	T	17	THR	insignificant	0.35	0.21	-	-	-	-	0.263
	T	18	ASN	insignificant	0.98	0.99	-	-	-	1	-1.827
	T	20	LYS	strong	1.18	-0.26	-	1	1	-	-7.119
	T	22	ILE	weak	0.68	0.6	-	-	-	1	2.110
	T	24	GLU	weak	0.43	-0.43	-	-	-	-	-0.441
	T	47	SER	insignificant	-0.09	-0.09	-	-	-	-	0.080
	T	48	LYS	insignificant	0.74	0.4	1	-	-	-	-3.293
	T	50	PHE	insignificant	2.15	2.87	-	-	-	-	2.433

Continued on next page

Table C.5 – continued from previous page

PDB	Chain	Pos	Amino acid	$\Delta\Delta G_{exp}$	Robetta	FoldX	KFC	Minerva ²	Minerva ¹	HotPoint	Z-score
	T	58	ASP	strong	0.98	0.62	-	1	-	-	-1.210
	T	128	GLU	insignificant	-0.08	-0.1	-	-	-	-	-0.226
	T	133	LEU	insignificant	2.04	2.4	-	-	-	1	2.082
	T	135	ARG	insignificant	1.57	-0.02	-	-	-	-	-0.959
	T	140	PHE	intermediate	1.62	1.02	-	-	-	-	2.099
	T	203	THR	insignificant	0.82	1.5	1	-	-	-	0.230
	T	207	VAL	insignificant	0.32	0.43	-	-	-	-	0.802
1fe8AH	A	987	GLU	insignificant	0.06	0.3	-	-	-	-	-4.193
	A	990	HIS	insignificant	1.74	1.76	1	-	1	1	0.112
1fe8AL	A	963	ARG	insignificant	2.94	1.88	-	-	-	-	-1.186
	A	987	GLU	insignificant	0.35	-0.29	-	-	-	-	-4.193
	A	1023	HIS	insignificant	0.03	-0.07	-	-	-	-	0.306
1foeAB	B	41	SER	intermediate	1.85	0.34	-	-	-	-	-0.051
	B	54	GLY	strong	-0.32	-0.43	-	-	-	-	-0.308
1g3iAG	A	438	ASP	strong	-0.37	-0.17	-	-	1	-	-1.626
	A	439	LEU	strong	0.43	0.84	-	-	-	-	1.011
	A	441	ARG	strong	0.05	0.07	-	-	-	-	-2.031
	A	442	PHE	strong	0.39	0.1	-	1	1	-	5.879
	A	443	ILE	strong	1.68	2.18	-	1	1	1	4.132
	A	444	LEU	strong	0.43	-1.83	-	1	1	-	2.896
1gl4AB	A	403	ARG	intermediate	1.63	-0.02	-	-	-	-	-1.821
	A	427	ASP	strong	0.77	-0.18	-	-	-	-	-1.863
	A	429	HIS	strong	3.23	1.41	-	1	1	1	1.221
	A	431	TYR	strong	1.69	1.81	1	-	-	1	1.803
	A	440	TYR	intermediate	3.32	2.69	1	1	1	-	2.670
	A	616	GLU	strong	0.91	0.67	1	-	-	-	-0.394
	A	620	ARG	strong	2.07	1.31	1	1	1	1	-2.994
1ihbAB	B	101	ASN	insignificant	-0.01	-0.57	-	-	-	-	0.113

Continued on next page

Table C.5 – continued from previous page

PDB	Chain	Pos	Amino acid	$\Delta\Delta G_{exp}$	Robetta	FoldX	KFC	Minerva ²	Minerva ¹	HotPoint	Z-score
	B	133	ARG	weak	2.08	-0.67	-	-	-	1	-2.871
	B	135	HIS	weak	0.69	-1.96	-	-	-	1	0.411
	B	136	LYS	intermediate	0.28	0.11	-	-	-	-	-1.356
1jatAB	A	55	GLU	strong	2.39	-0.06	-	-	1	1	-4.334
	B	8	PHE	strong	4.45	4.25	1	1	1	1	5.455
1jppBD	B	345	LYS	strong	0.09	-0.25	-	-	-	-	-1.172
	B	354	LYS	insignificant	0.19	-0.04	-	-	-	-	-2.515
	B	383	TRP	strong	0.93	-0.72	-	-	-	1	1.492
	B	386	ARG	intermediate	0.5	-0.26	-	-	1	-	-1.455
	B	435	LYS	intermediate	0.5	2.92	-	-	-	-	-2.706
	B	469	ARG	intermediate	-0.07	-0.07	-	-	-	-	-0.604
	B	470	HIS	intermediate	0.47	0.49	-	-	-	1	-0.408
1mq8AB	B	206	THR	strong	3.43	0.98	1	-	-	1	1.253
1nfiAF	F	215	CYS	insignificant	-0.08	-0.02	-	-	-	-	6.828
1nfiBF	F	181	TYR	strong	2.35	3.27	-	1	1	1	1.906
1nunAB	A	76	ASP	intermediate	2.51	-0.98	1	-	-	1	-2.923
	A	78	ARG	intermediate	2.74	2.29	-	-	1	-	-2.251
	A	155	ARG	intermediate	0.83	0.84	1	-	-	-	-0.475
1ub4AC	C	453	PHE	insignificant	1.89	2.74	-	-	-	-	3.153
2hbbAB	B	35	TYR	insignificant	0.82	0.09	-	-	-	1	2.381

References

- [1] J. Kyte and R. F. Doolittle. A simple method for displaying the hydropathic character of a protein. *J Mol Biol*, 157(1):105–132, May 1982.
- [2] anonymous. All structures. http://share.chuagh.net/science/Share_AminoAcids.php?structure=1, Last accessed time: Sep, 2009.
- [3] anonymous. Amino acid hydrophobicity. <http://www.cgl.ucsf.edu/chimera/docs/UsersGuide/midas/hydrophob.html>, Last accessed time: Oct, 2011.
- [4] Wikipedia the free encyclopedia. Protein folding. http://en.wikipedia.org/wiki/Protein_folding, Last accessed time: Oct, 2011.
- [5] Wikipedia the free encyclopedia. X-ray crystallography. http://en.wikipedia.org/wiki/X-ray_crystallography, Last accessed time: Oct, 2011.
- [6] A. Murzin, S. Brenner, T. Hubbard, and C. Chothia. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol*, 247:536–540, April 1995.
- [7] C. A. Orengo, A. D. Michie, S. Jones, D. T. Jones, M. B. Swindells, and J. M. Thornton. CATH: a hierarchic classification of protein domain structures. *Structure (London, England: 1993)*, 5(8):1093–1108, August 1997.
- [8] Irene M.A. Nooren and Janet M. Thornton. Diversity of protein-protein interactions. *EMBO J*, 22(14):3486–3492, July 15 2003.
- [9] C. J. Tsai, D. Xu, and R. Nussinov. Protein folding via binding and vice versa. *Fold Des*, 3(4):R71–R80, 1998.
- [10] Y Ofra and B Rost. Analysing six types of protein-protein interfaces. *J Mol Biol*, 325(2):377–387, Jan 10 2003.
- [11] Nurcan Tuncbag, Gozde Kar, Ozlem Keskin, Attila Gursoy, and Ruth Nussinov. A survey of available tools and web servers for analysis of protein-protein interactions and interfaces. *Brief Bioinform*, 10:217–232, February 2009.

REFERENCES

- [12] S. J. Hubbard and J. M. Thornton. ‘NACCESS’, computer program. Technical report, Department of Biochemistry Molecular Biology, University College London, 1993.
- [13] Peter Brick, Peter Briggs, and Ian Tickle. AREAIMOL (CCP4: Supported program). <http://www.ccp4.ac.uk/dist/html/areaimol.html>, Last accessed time: Sep, 2009.
- [14] Wikipedia. Delaunay triangulation. http://en.wikipedia.org/wiki/Delaunay_triangulation, 2011.
- [15] S. Jones and J. M. Thornton. Principles of protein-protein interactions. *Proc Natl Acad Sci USA*, 93(1):13–20, January 1996.
- [16] H. Neuvirth, R. Raz, and G. Schreiber. ProMate: a structure based prediction program to identify the location of protein-protein binding sites. *J Mol Biol*, 338(1):181–199, April 2004.
- [17] William S. J. Valdar and Janet M. Thornton. Conservation helps to identify biologically relevant crystal contacts. *J Mol Biol*, 313(2):399–416, October 2001.
- [18] H. Zhu, F. S. Domingues, I. Sommer, and T. Lengauer. NOXclass: prediction of protein-protein interaction types. *BMC Bioinformatics*, 7:27, 2006.
- [19] S. De, O. Krishnadev, N. Srinivasan, and N. Rekha. Interaction preferences across protein-protein interfaces of obligatory and non-obligatory components are different. *BMC Struct Biol*, 5:15, August 2005.
- [20] Rudra Prasad Saha, Ranjit Prasad Bahadur, and Pinak Chakrabarti. Interresidue contacts in proteins and protein-protein interfaces and their use in characterizing the homodimeric interface. *J Proteome Res*, 4:1600–1609, August 2005.
- [21] Suryani Lukman, Kelvin Sim, Jinyan Li, and Yi-Ping Phoebe Chen. Interacting amino acid preferences of 3D pattern pairs at the binding sites of transient and obligate protein complexes. In *APBC*, pages 69–78, 2008.
- [22] R. P. Bahadur, P. Chakrabarti, F. Rodier, and J. Janin. A dissection of specific and non-specific protein-protein interfaces. *J Mol Biol*, 336(4):943–955, February 2004.
- [23] Fabian Glaser, David M. Steinberg, Ilya A. Vakser, and Nir Ben-Tal. Residue frequencies and pairing preferences at protein-protein interfaces. *Proteins*, 43(2):89–102, 2001.

REFERENCES

- [24] Sanzo Miyazawa and Robert L. Jernigan. Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *J Mol Biol*, 256:623–644, 1996.
- [25] G. Moont, H. A. Gabb, and M. J. Sternberg. Use of pair potentials across protein interfaces in screening predicted docked complexes. *Proteins*, 35(3):364–373, May 1999.
- [26] J. Mintseris and Z. Weng. Atomic contact vectors in protein-protein recognition. *Proteins*, 53(3):629–639, November 2003.
- [27] Hannes Ponstingl, Kim Henrick, and Janet M. Thornton. Discriminating between homodimeric and monomeric proteins in the crystalline state. *Proteins*, 41(1):47–57, July 2000.
- [28] Hannes Ponstingl, Thomas Kabir, and Janet M. Thornton. Automatic inference of protein quaternary structure from crystals. *Journal of Applied Crystallography*, 36(5):1116–1122, Oct 2003.
- [29] C. Zhang, G. Vasmatzis, J.L. Cornette, and C. DeLisi. Determination of atomic desolvation energies from the structures of crystallized proteins. *J Mol Biol*, 267:707–726(20), April 1997.
- [30] S. Jones and J. M. Thornton. Analysis of protein-protein interaction sites using surface patches. *J Mol Biol*, 272(1):121–132, September 1997.
- [31] R. P. Bahadur, P. Chakrabarti, F. Rodier, and J. Janin. Dissecting subunit interfaces in homodimeric proteins. *Proteins*, 53(3):708–719, November 2003.
- [32] L. Young, R. L. Jernigan, and D. G. Covell. A role for surface hydrophobicity in protein-protein recognition. *Protein Sci*, 3(5):717–729, May 1994.
- [33] Joel Janin and Francis Rodier. Protein-protein interaction at crystal contacts. *Proteins*, 23(4):580–587, December 1995.
- [34] Joel Janin. Specific versus non-specific contacts in protein crystals. *Nature Structural Biology*, 4:973–974, December 1997.
- [35] O. Carugo and P. Argos. Protein-protein crystal-packing contacts. *Protein sci*, 6(10):2261–3, 1997.
- [36] Andrew A. Bogan and Kurt S. Thorn. Anatomy of hot spots in protein interfaces. *J Mol Biol*, 280(1):1–9, July 1998.

REFERENCES

- [37] JA Wells. Systematic mutational analyses of protein-protein interfaces. *Methods Enzymol*, 202:390–411, 1991.
- [38] T. Clackson and J. Wells. A hot spot of binding energy in a hormone-receptor interface. *Science*, 267:383–386, January 1995.
- [39] BC Cunningham and JA Wells. High-resolution epitope mapping of hGH-receptor interactions by alanine-scanning mutagenesis. *Science*, 244(4908):1081–1085, 1989.
- [40] F Lefevre, MH Remy, and JM Masson. Alanine-stretch scanning mutagenesis: a simple and efficient method to probe protein structure and function. *Nucl Acids Res*, 25(2):447–448, 1997.
- [41] Solene Grosdidier and Juan F. Recio. Identification of hot-spot residues in protein-protein interactions by computational docking. *BMC Bioinformatics*, 9(1):447, 2008.
- [42] Maria José M. Castro and Stephen Anderson. Alanine point-mutations in the reactive region of bovine pancreatic trypsin inhibitor: effects on the kinetics and thermodynamics of binding to β -trypsin and α -chymotrypsin. *Biochemistry*, pages 11435–11446, 1996.
- [43] James R. Bradford and David R. Westhead. Improved prediction of protein-protein binding sites using a support vector machines approach. *Bioinformatics*, 21(8):1487–1494, April 2005.
- [44] J. Bernauer, A. Poupon, J. Azé, and J. Janin. A docking analysis of the statistical physics of protein-protein recognition. *Phys Biol*, 2(1-2), June 2005.
- [45] I O Daar, P J Artymiuk, D C Phillips, and L E Maquat. Human triose-phosphate isomerase deficiency: a single amino acid substitution results in a thermolabile enzyme. *Proc Natl Acad Sci USA*, 83(20):7903–7907, 1986.
- [46] C. J. Tsai, S. L. Lin, H. J. Wolfson, and R. Nussinov. A dataset of protein-protein interfaces generated with a sequence-order-independent comparison technique. *J Mol Biol*, 260(4):604–620, July 1996.
- [47] C. J. Tsai, S. L. Lin, H. J. Wolfson, and R. Nussinov. Studies of protein-protein interfaces: a statistical analysis of the hydrophobic effect. *Protein Sci*, 6(1):53–64, January 1997.
- [48] Inbal Halperin, Haim Wolfson, and Ruth Nussinov. Protein-protein interactions: coupling of structurally conserved residues and of hot spots across interfaces. implications for docking. *Structure*, 12(6):1027–1038, June 2004.

REFERENCES

- [49] X. Li, O. Keskin, B. Ma, R. Nussinov, and J. Liang. 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(3):781–795, November 2004.
- [50] O. Keskin, C. J. Tsai, H. Wolfson, and R. Nussinov. A new, structurally nonredundant, diverse data set of protein-protein interfaces and its implications. *Protein Sci*, 13(4):1043–1055, April 2004.
- [51] Shira Mintz, Alexandra S. Peleg, Haim J. Wolfson, and Ruth Nussinov. Generation and analysis of a protein-protein interface data set with similar chemical and spatial patterns of interactions. *Proteins*, 61(1):6–20, Oct 2005.
- [52] O. Keskin, B. Ma, and R. Nussinov. Hot regions in protein-protein interactions: the organization and contribution of structurally conserved hot spot residues. *J Mol Biol*, 345(5):1281–1294, February 2005.
- [53] Sungsam Gong, Changbum Park, Hansol Choi, Junsu Ko, Insoo Jang, Jungsul Lee, Dan M. Bolser, Donghoon Oh, Deok-Soo Kim, and Jong Bhak. A protein domain interaction interface database: InterPare. *BMC Bioinformatics*, 6:207, 2005.
- [54] M. C. Lawrence and P. M. Colman. Shape complementarity at protein/protein interfaces. *J Mol Biol*, 234(4):946–950, December 1993.
- [55] T. A. Larsen, A. J. Olson, and D. S. Goodsell. Morphology of protein-protein interfaces. *Structure*, 6(4):421–427, Apr 1998.
- [56] R. Preissner, A. Goede, and C. Frommel. Dictionary of interfaces in proteins (DIP) data bank of complementary molecular surface patches. *J Mol Biol*, 280:535–550, 1998.
- [57] D. Korkin, F. P. Davis, and A. Sali. Localization of protein-binding sites within families of proteins. *Protein Sci*, 14(9):2350–2360, September 2005.
- [58] Fred P. Davis and Andrej Sali. PIBASE: a comprehensive database of structurally defined protein interfaces. *Bioinformatics*, 21(9):1901–1907, May 2005.
- [59] T. B. Fischer, J. B. Holmes, I. R. Miller, J. R. Parsons, L. Tung, J. C. Hu, and J. Tsai. Assessing methods for identifying pair-wise atomic contacts across binding interfaces. *J Struct Biol*, 153(2):103–112, Feb 2006.
- [60] Emre Guney, Nurcan Tuncbag, Ozlem Keskin, and Attila Grsoy. HotSprint: database of computational hot spots in protein interfaces. *Nucl Acids Res*, 36(Database-Issue):662–666, 2008.

REFERENCES

- [61] Qian Liu and Jinyan Li. Propensity vectors of low-ASA residue pairs in the distinction of protein interactions. *Proteins*, 78(3):589–602, August 2009.
- [62] Jinyan Li and Qian Liu. ‘Double water exclusion’: a hypothesis refining the O-ring theory for the hot spots at protein interfaces. *Bioinformatics*, 25(6):743–750, January 2009.
- [63] Qian Liu and Jinyan Li. Protein binding hot spots and the residue-residue pairing preference: a water exclusion perspective. *BMC Bioinformatics*, 11(1):244, May 2010.
- [64] Zhenhua Li and Jinyan Li. Geometrically centered region: a “wet” model of protein binding hot spots not excluding water molecules. *Proteins*, 78(16):3304–3316, 2010.
- [65] Pinak Chakrabarti and Joël Janin. Dissecting protein-protein recognition sites. *Proteins*, 47(3):334–343, 2002.
- [66] Kyu-Il Cho, Dongsup Kim, and Doheon Lee. A feature-based approach to modeling protein-protein interaction hot spots. *Nucl Acids Res*, 37(8):2672–2687, May 2009.
- [67] J. J. Headd, Y. E. A. Ban, P. Brown, H. Edelsbrunner, M. Vaidya, and J. Rudolph. Protein-protein interfaces: properties, preferences, and projections. *J Proteome Res*, 6(7):2576–2586, July 2007.
- [68] F. Cazals, F. Proust, R. P. Bahadur, and J. Janin. Revisiting the Voronoi description of protein-protein interfaces. *Protein Sci*, 15(9):2082–2092, 2006.
- [69] Benjamin Bouvier, Raik Grünberg, Michael Nilges, and Frédéric Cazals. Shelling the Voronoi interface of protein-protein complexes reveals patterns of residue conservation, dynamics, and composition. *Proteins*, 76(3):677–692, 2009.
- [70] Julie Bernauer, Ranjit Prasad P. Bahadur, Francis Rodier, Joël Janin, and Anne Poupon. DiMoVo: a Voronoi tessellation-based method for discriminating crystallographic and biological protein-protein interactions. *Bioinformatics*, 24:652–8, January 2008.
- [71] Anne Poupon. Voronoi and Voronoi-related tessellations in studies of protein structure and interaction. *Current Opinion in Structural Biology*, 14(2):233–241, 2004.
- [72] J. Janin. Elusive affinities. *Proteins*, pages 30–39, 1995.
- [73] C. Chothia and J. Janin. Principles of protein-protein recognition. *Nature*, 256:705–708, 1975.

REFERENCES

- [74] R. Guerois, J. E. Nielsen, and L. Serrano. Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. *J Mol Biol*, 320(2):369–387, July 2002.
- [75] J. Schymkowitz, J. Borg, F. Stricher, R. Nys, F. Rousseau, and L. Serrano. The FoldX web server: an online force field. *Nucl Acids Res*, 33(Web Server issue), July 2005.
- [76] T. Kortemme and D. Baker. A simple physical model for binding energy hot spots in protein-protein complexes. *Proc Natl Acad Sci USA*, 99(22):14116–14121, October 2002.
- [77] T. Kortemme, D. E. Kim, and D. Baker. Computational alanine scanning of protein-protein interfaces. *Sci STKE*, 2004(219), February 2004.
- [78] Ying Gao, Renxiao Wang, and Luhua Lai. Structure-based method for analyzing protein-protein interfaces. *J Mol Model*, 10:44–54, 2004.
- [79] Steven J J. Darnell, Laura Legault, and Julie C C. Mitchell. KFC server: interactive forecasting of protein interaction hot spots. *Nucl Acids Res*, June 2008.
- [80] Jun-Feng Xia, Xing-Ming Zhao, Jiangning Song, and De-Shuang Huang. APIS: accurate prediction of hot spots in protein interfaces by combining protrusion index with solvent accessibility. *BMC Bioinformatics*, 11(1):174, April 2010.
- [81] Yanay Ofran and Burkhard Rost. Protein-protein interaction hotspots carved into sequences. *PLoS Comput Biol*, 3(7), July 2007.
- [82] N Tuncbag, A Gursoy, and O Keskin. Identification of computational hot spots in protein interfaces: combining solvent accessibility and inter-residue potentials improves the accuracy. *Bioinformatics*, 25(12):1513–1520, April 2009.
- [83] Loredana Lo Conte, Cyrus Chothia, and Joel Janin. The atomic structure of protein-protein recognition sites. *J Mol Biol*, 285(5):2177–2198, 1999.
- [84] C Yan, F Wu, RL Jernigan, D Dobbs, and V. Honavar. Characterization of protein-protein interfaces. *Protein J*, 27(1):59–70, January 2008.
- [85] A. Ben-Naim. On the driving forces for protein-protein association. *J Chem Phys.*, 125(2):24901, July 2006.
- [86] K. S. Thorn and A. A. Bogan. ASEdb: a database of alanine mutations and their effects on the free energy of binding in protein interactions. *Bioinformatics*, 17(3):284–285, March 2001.

REFERENCES

- [87] Irina S S. Moreira, Pedro A A. Fernandes, and Maria J J. Ramos. Hot spots-a review of the protein-protein interface determinant amino-acid residues. *Proteins*, 68(4):803–812, June 2007.
- [88] Juan Fernandez-Recio. Prediction of protein binding sites and hot spots. *Wiley Interdisciplinary Reviews Computational Molecular Science*, (February):1–19, 2011.
- [89] Zengjian Hu, Buyong Ma, Haim Wolfson, and Ruth Nussinov. Conservation of polar residues as hot spots at protein interfaces. *Proteins: Structure, Function, and Bioinformatics*, 39(4):331–342, 2000.
- [90] Buyong Ma, Tal Elkayam, Haim Wolfson, and Ruth Nussinov. Protein-protein interactions: structurally conserved residues distinguish between binding sites and exposed protein surfaces. *Proceedings of the National Academy of Sciences*, 100(10):5772–5777, 2003.
- [91] Alexander Benedix, Caroline M. Becker, Bert L. de Groot, Amedeo Caflisch, and Rainer A. Bockmann. Predicting free energy changes using structural ensembles. *Nature Methods*, 6(1):3–4, Jan 2009.
- [92] Steven J J. Darnell, David Page, and Julie C C. Mitchell. An automated decision-tree approach to predicting protein interaction hot spots. *Proteins*, 36:W265–W269, June 2007.
- [93] Stefano Lise, Cedric Archambeau, Massimiliano Pontil, and David Jones. Prediction of hot spot residues at protein-protein interfaces by combining machine learning and energy-based methods. *BMC Bioinformatics*, 10(1):365, 2009.
- [94] Stefano Lise, Daniel Buchan, Massimiliano Pontil, and David T. Jones. Predictions of hot spot residues at protein-protein interfaces using support vector machines. *PLoS ONE*, 6(2):e16774, 02 2011.
- [95] Salam A. Assi, Tomoyuki Tanaka, Terence H. Rabbitts, and Narcis Fernandez-Fuentes. PCRPI: presaging critical residues in protein interfaces, a new computational tool to chart hot spots in protein interfaces. *Nucl Acids Res*, 38(6), April 2010.
- [96] DEL SOL Antonio and O’MEARA Paul. Small-world network approach to identify key residues in protein-protein interaction. *Proteins*, 58:672–682, 2005.
- [97] Nurcan Tuncbag, F. Sibel Salman, Ozlem Keskin, and Attila Gursoy. Analysis and network representation of hotspots in protein interfaces using minimum cut trees. *Proteins: Structure, Function, and Bioinformatics*, 78(10):2283–2294, 2010.

REFERENCES

- [98] T. B. Fischer, K. V. Arunachalam, D. Bailey, V. Mangual, S. Bakhru, R. Russo, D. Huang, M. Paczkowski, V. Lalchandani, C. Ramachandra, B. Ellison, S. Galer, J. Shapley, E. Fuentes, and J. Tsai. The binding interface database (BID): a compilation of amino acid hot spots in protein interfaces. *Bioinformatics*, 19(11):1453–1454, July 2003.
- [99] Peter Kollman. Free energy calculations: applications to chemical and biochemical phenomena. *Chemical Reviews*, 93(7):2395–2417, November 1993.
- [100] Hiroaki Gouda, Irwin D. Kuntz, David A. Case, and Peter A. Kollman. Free energy calculations for theophylline binding to an RNA aptamer: Comparison of MM-PBSA and thermodynamic integration methods. *Biopolymers*, 68(1):16–34, 2003.
- [101] Shuanghong Huo, Irina Massova, and Peter A. Kollman. Computational alanine scanning of the 1:1 human growth hormone-receptor complex. *Journal of Computational Chemistry*, 23:15–27, November 2001.
- [102] B. de Groot, D. van Aalten, R. Scheek, A. Amadei, G. Vriend, and H. Berendsen. Prediction of protein conformational freedom from distance constraints. *Proteins*, 29(2):240–251, Dec 1997.
- [103] Shiyong Liu, Qingliang Li, and Luhua Lai. A combinatorial score to distinguish biological and nonbiological protein-protein interfaces. *Proteins*, 64:68–78, 2006.
- [104] J. Bernauer, J. Azé, J. Janin, and A. Poupon. A new protein-protein docking scoring function based on interface residue properties. *Bioinformatics*, 5(23):555–62, 2007.
- [105] KI Cho, K Lee, KH Lee, D Kim, and D Lee. Specificity of molecular interactions in transient protein-protein interaction interfaces. *Proteins*, 65(3):593–606, 2006.
- [106] Sung H. Park, Jose Reyes, David Gilbert, Ji W. Kim, and Sangsoo Kim. Prediction of protein-protein interaction types using association rule based classification. *BMC Bioinformatics*, 10(1):36, 2009.
- [107] Peter Block, Juri Paern, Eyke Hullermeier, Paul Sanschagrin, Christoph A. Sotriffer, and Gerhard Klebe. Physicochemical descriptors to discriminate protein-protein interactions in permanent and transient complexes selected by means of machine learning algorithms. *Proteins*, 65(3):607–622, 2006.
- [108] W. L. DeLano. Unraveling hot spots in binding interfaces: progress and challenges. *Curr Opin Struct Biol*, 12(1):14–20, February 2002.

REFERENCES

- [109] A. Fernández and R. Scott. Dehydron: a structurally encoded signal for protein interaction. *Biophys J*, 85(3):1914–1928, September 2003.
- [110] PL Privalov, AI Dragan, C Crane-Robinson, KJ Breslauer, DP Remeta, and CA Minetti. What drives proteins into the major or minor grooves of DNA? *J Mol Biol*, 365(1):1–9, 2007.
- [111] David Eppstein. Arboricity and bipartite subgraph listing algorithms. *Inf Process Lett*, 51(4):207–211, 1994.
- [112] Jinyan Li, Guimei Liu, Haiquan Li, and Limsoon Wong. Maximal biclique subgraphs and closed pattern pairs of the adjacency matrix: a one-to-one correspondence and mining algorithms. *IEEE TKDE*, 19(12):1625–1637, 2007.
- [113] H. Li, J. Li, and L. Wong. Discovering motif pairs at interaction sites from protein sequences on a proteome-wide scale. *Bioinformatics (Oxford, England)*, 22(8):989–996, April 2006.
- [114] Julie L. Morrison, Rainer Breitling, Desmond J. Higham, and David R. Gilbert. A lock-and-key model for protein-protein interactions. *Bioinformatics*, 22(16):2012–2019, August 2006.
- [115] Schrödinger, LLC. The PyMOL molecular graphics system, version 1.3r1. August 2010.
- [116] Deepa Rajamani, Spencer Thiel, Sandor Vajda, and Carlos J. Camacho. Anchor residues in protein-protein interactions. *Proc Natl Acad Sci USA*, 101(31):11287–11292, August 2004.
- [117] L. Jin and JA Wells. Dissecting the energetics of an antibody-antigen interface by alanine shaving and molecular grafting. *Protein Sci*, 3(12):2351–2357, September 1994.
- [118] L Li, B Zhao, Z Cui, J Gan, MK Sakharkar, and P Kanguane. Identification of hot spot residues at protein-protein interface. *Bioinformation*, 1(4):121–126, April 2006.
- [119] D. Kozakov, R. Brenke, S. R. Comeau, and S. Vajda. PIPER: an FFT-based protein docking program with pairwise potentials. *Proteins*, 65(2):392–406, November 2006.
- [120] Julian Mintseris, Brian Pierce, Kevin Wiehe, Robert Anderson, Rong Chen, and Zhiping Weng. Integrating statistical pair potentials into protein complex prediction. *Proteins*, 69(3):511–520, July 2007.

REFERENCES

- [121] J. Janin, S. Miller, and C. Chothia. Surface, subunit interfaces and interior of oligomeric proteins. *J Mol Biol*, 204(1):155–164, November 1988.
- [122] Julian Mintseris and Zhiping Weng. Structure, function, and evolution of transient and obligate protein-protein interactions. *Proc Natl Acad Sci USA*, July 2005.
- [123] N. V. Dokholyan, L. Li, F. Ding, and E. I. Shakhnovich. Topological determinants of protein folding. *Proc Natl Acad Sci USA*, 99(13):8637–8641, June 2002.
- [124] Eric Chea and Dennis Livesay. How accurate and statistically robust are catalytic site predictions based on closeness centrality? *BMC Bioinformatics*, 8(1):153, May 2007.
- [125] SB Cantor and MW Kattan. Determining the area under the ROC curve for a binary diagnostic test. *Med Decis Making*, 20(4):468–70, 2000.
- [126] Ricardo Vilalta and Youssef Drissi. A perspective view and survey of Meta-Learning. *Artificial Intelligence Review*, 18:77–95, 2002.
- [127] Chih C. Chang and Chih J. Lin. LIBSVM: a library for support vector machines, 2001.
- [128] T. Pupko, R. E. Bell, I. Mayrose, F. Glaser, and N. Ben-Tal. Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. *Bioinformatics*, 18 Suppl 1, 2002.
- [129] Juliette Martin, Leslie Regad, Catherine Etchebest, and Anne-Claude Camproux. Taking advantage of local structure descriptors to analyze interresidue contacts in protein structures and protein complexes. *Proteins*, 73(3):672–689, May 19 2008.
- [130] Gianluca Pollastri, Pierre Baldi, Pietro Fariselli, and Rita Casadio. Prediction of coordination number and relative solvent accessibility in proteins. *Proteins*, 47(2):142–153, 2002.
- [131] Nello Cristianini and John Shawe-Taylor. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press, March 2000.
- [132] Chihwei Hsu, Chihchung Chang, and Chihjen Lin. A practical guide to support vector classification. Technical report, Department of Computer Science and Information Engineering, National Taiwan University, 2003.

REFERENCES

- [133] John A. Rice. *Mathematical Statistics and Data Analysis*. Duxbury Press, April 2001.
- [134] Vladimir Sobolev, Rebecca C. Wade, Gert Vriend, and Marvin Edelman. Molecular docking using surface complementarity. *Proteins*, 25:120–129, 1996.
- [135] Nathalie Japkowicz. *Concept learning in the absence of counterexamples: an autoassociation-based approach to classification*. PhD thesis, New Brunswick, NJ, USA, 1999. Director-Hanson, Jose and Director-Kulikowski, Casimir.
- [136] B.J. McConkey, V. Sobolev, and M. Edelman. Quantification of protein surfaces, volumes and atom-atom contacts using a constrained Voronoi procedure. *Bioinformatics*, 18(10):1365–1373, 2002.
- [137] David G. Kirkpatrick and John D. Radke. A framework for computational morphology. *Computational Geometry, Machine Intelligence and Pattern Recognition*, 2, pages 217–48, 1985.
- [138] Ajay N. Jain. Scoring noncovalent protein-ligand interactions: A continuous differentiable function tuned to compute binding affinities. *Journal of Computer-Aided Molecular Design*, 10(5):427–440, 1996.
- [139] Oleg Trott and Arthur J. Olson. AutoDock vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem.*, 31(2):455–461, Jan 2010.
- [140] K. R. Gabriel and R. R. Sokal. A new statistical approach to geographic variation analysis. *Systematic Zoology (Society of Systematic Biologists)*, 18(3):259–270, 1969.
- [141] D. W. Matula and R. R. Sokal. Properties of Gabriel graphs relevant to geographic variation research and clustering of points in the plane. *Geogr. Anal.*, 12:205–222, 1980.
- [142] Deok-Soo Kim, Jeongyeon Seo, Donguk Kim, Joonghyun Ryu, and Cheol-Hyung Cho. Three-dimensional beta shapes. *Computer-Aided Design*, 38(11):1179–1191, 2006.
- [143] Herbert Edelsbrunner and Ernst P. Mücke. Three-dimensional alpha shapes. *ACM Trans. Graph.*, 13:43–72, 1994.
- [144] Rui Xu, Damian C. Ekiert, Jens C. Krause, Rong Hai, James E. Crowe, and Ian A. Wilson. Structural Basis of Preexisting Immunity to the 2009 H1N1 Pandemic Influenza Virus. *Science*, 328(5976):357–360, 2010.

REFERENCES

- [145] E. Krissinel and K. Henrick. Inference of macromolecular assemblies from crystalline state. *J Mol Biol*, 372(3):774–797, 2007.
- [146] Xiacong Yu, Tshidi Tsibane, Patricia A. McGraw, Frances S. House, Christopher J. Keefer, Mark D. Hicar, Terrence M. Tumpey, Claudia Pappas, Lucy A. Perrone, Osvaldo Martinez, James Stevens, Ian A. Wilson, Patricia V. Aguilar, Eric L. Altschuler, Christopher F. Basler, and James E. Crowe. Neutralizing antibodies derived from the B cells of 1918 influenza pandemic survivors. *Nature*, August 2008.
- [147] Jens C. Krause, Terrence M. Tumpey, Chelsey J. Huffman, Patricia A. McGraw, Melissa B. Pearce, Tshidi Tsibane, Rong Hai, Christopher F. Basler, and Jr. Crowe, James E. Naturally occurring human monoclonal antibodies neutralize both 1918 and 2009 pandemic influenza a (H1N1) viruses. *J. Virol.*, 84(6):3127–3130, 2010.
- [148] Manabu Igarashi, Kimihito Ito, Reiko Yoshida, Daisuke Tomabechi, Hiroshi Kida, and Ayato Takada. Predicting the antigenic structure of the pandemic (H1N1) 2009 influenza virus hemagglutinin. *PLoS ONE*, 5(1):e8553, 01 2010.
- [149] Bradford C. Barber, David P. Dobkin, and Hannu Huhdanpaa. The Quickhull algorithm for convex hulls. *ACM Transactions on Mathematical Software*, 22(4):469–483, 1996.
- [150] Jerry Tsai, Robin Taylor, Cyrus Chothia, and Mark Gerstein. The packing density in proteins: standard radii and volumes. *J Mol Biol*, 290(1):253–266, 1999.
- [151] Felix B. Sheinerman and Barry Honig. On the role of electrostatic interactions in the design of protein-protein interfaces. *J Mol Biol*, 318(1):161–177, 2002.
- [152] Xiaolei Zhu and Julie C. Mitchell. KFC2: A knowledge-based hot spot prediction method based on interface solvation, atomic density and plasticity features. *Proteins*, pages 2671–2683, 2011.
- [153] Damian C. Ekiert, Gira Bhabha, Marc-André A. Elsliger, Robert H. Friesen, Mandy Jongeneelen, Mark Throsby, Jaap Goudsmit, and Ian A. Wilson. Antibody recognition of a highly conserved influenza virus epitope. *Science (New York, N.Y.)*, 324(5924):246–251, April 2009.
- [154] C. Chothia and A. M. Lesk. The relation between the divergence of sequence and structure in proteins. *The EMBO journal*, 5(4):823–826, April 1986.
- [155] T Schwede, A Diemand, N Guex, and MC. Peitsch. Protein structure computing in the genomic era. *Res Microbiol*, 151(2):107–12, March 2000.

REFERENCES

- [156] Akira R. Kinjo and Haruki Nakamura. Geometric similarities of protein-protein interfaces at atomic resolution are only observed within homologous families: an exhaustive structural classification study. *J Mol Biol*, 399(3):526–540, June 2010.
- [157] Alpay N. Temiz, Panayiotis V. Benos, and Carlos J. Camacho. Electrostatic hot spot on DNA-binding domains mediates phosphate desolvation and the pre-organization of specificity determinant side chains. *Nucleic Acids Research*, 38(7):2134–2144, 2010.
- [158] Jinyan Li, Kelvin Sim, Guimei Liu, and Limsoon Wong. Maximal quasi-bicliques with balanced noise tolerance: concepts and co-clustering applications. In *SDM*, pages 72–83, 2008.
- [159] Takeaki Uno, Tatsuya Asai, Yuzo Uchida, and Hiroki Arimura. Lcm: An efficient algorithm for enumerating frequent closed item sets. In Bart Goethals and Mohammed Javeed Zaki, editors, *FIMI*, volume 90 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2003.
- [160] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.
- [161] Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46(1-3):389–422, 2002.
- [162] J. Ross Quinlan. *C4.5: Programs for Machine Learning (Morgan Kaufmann Series in Machine Learning)*. Morgan Kaufmann, 1 edition, 1993.
- [163] Weka machine learning project. Weka. URL <http://www.cs.waikato.ac.nz/~ml/weka>.