Protein Interactions and Regulation of STIM and ORAI Genes

Dissertation

Zur Erlangung des Grades
des Doktors der Naturwissenshaften
der Naturwissenschaftlich-Technischen Fakultat III
der Universität des Saarlandes

von
Ruzianisra Binti Mohamed

Saarbrücken
2017

Tag des Kolloquiums:	22.02.2017
Dekan:	Prof. Dr. Guido Kickelbick
Perichterstatter:	Prof. Dr. Volkhard Helms
	Prof. Dr. Richard Zimmermann
Vorsitz:	Prof. Dr. Uli Müller
Akad. Mitarbeiter:	Dr. Jessica Hoppstädter

ABSTRACT

Protein interactions play major roles in many biological processes. This thesis is composed of three projects. Using five datasets, we explored the characteristics and composition of overlapping protein-protein (PP) and protein-ligand (PL) interfaces. Overall, characteristics of PP contacts and overlapping PL contacts are highly similar. Second study was designed to identify transcription factor binding site motifs in promoter regions of STIM and ORAI genes, to gain knowledge of their regulation and relation with breast cancer. Our findings form an important basis of predictive interactions between transcription factors targeting STIM and ORAI genes and underline roles of these genes in breast cancer. Thirdly, we evaluated the performance of seven protein prediction tools on a dataset of protein-ligand complexes. Although the tools predicted pockets of various sizes and shapes, we found comparable performance amongst the predictions of five tools. We trained a random forest model to output a list of suitable tools for a given protein structure. This classifier should be useful for prioritizing the tools to be used for unknown proteins.

ZUSAMMENFASSUNG

Wechselwirkungen zwischen Proteinen spielen in biologischen Prozessen eine wesentliche Rolle. Diese Dissertation ist in drei Projekte aufgegliedert. Unter Zuhilfenahme von fünf Datensätzen aus drei unterschiedlichen Datenbanken wurden die Eigenschaften und die Zusammensetzung von überlappenden Protein-Protein (PP) und Protein-Ligand (PL) Bindestellen untersucht. Statistisch gesehen sind sich die Eigenschaften von PP Kontakten und überlappenden PL Kontakten sehr ähnlich. Die zweite Studie diente dazu Motive von Transkriptionsfaktor-Bindestellen in bestimmten Promotor-Regionen der STIM und ORAI Gene zu identifizieren als auch Erkenntnisse über deren Regulation und Zusammenhang mit Brustkrebs zu gewinnen. Unsere Ergebnisse stellen eine wichtige Grundlage für die Vorhersage der Wechselwirkungen zwischen Transkriptionsfaktoren, welche an die beiden Gene STIM und ORAI binden, dar. Darüber hinaus konnten wir die Rolle dieser Gene in Zusammenhang mit Brustkrebs herausstellen. Im dritten Teil werteten wir die Performance von sieben verschiedenen Tools anhand eines Datensatzes von PL Komplexen aus. Obwohl die Tools von der Größe und Form her unterschiedliche Bindetaschen vorhersagten, konnten wir dennoch ein vergleichbares Verhalten zwischen fünf Tools feststellen. Wir trainierten ein Random-Forest Modell, welches, gegeben eine Proteinstruktur, eine Reihe von geeigneten Tools vorhersagt. Dieses Modell kann dazu dienen Vorhersagemodelle anhand unbekannter Proteine zu priorisieren.

ACKNOWLEDGEMENTS

Throughout my PhD study and research, it has been a pleasure working with my great supervisor and colleagues, this dissertation owe much to their sincere help and encouragement. First, I would like to thank my PhD supervisor Prof. Volkhard Helms, for his support and encouragement throughout the whole PhD work. His rational and rigorous approach in doing scientific research has been significantly influential and greatly helpful for this dissertation.

Secondly, I had a great fun and a good collaboration from my colleagues and their contributions should not be underestimated. I thank my collaborators Jennifer Degac, Zhao Yuan, Rahmad Akhbar, Riccha Sethi, and Dr. Mohamed Hamed. A big thank you also to Maryam Nazeriah, Thorsten Will, Kerstin Reuter, Dr. Michael Hutter, Dr. Ruslan Akulenko, Dr. Siba Shanak, Tran Vu Ha, Dr. Ahmad Bargash, Dr. Ozlem Ulucan, Daria Gadar, Dr. Christian Spaniol, Shimaa Khaled, Jan Riehm, Duy Nguyen and Hema as they gave me a lot of help and suggestions. I also appreciate the secretary of my supervisor Kerstin Pudelek for her big help in administration works.

I am also grateful to Ministry of Education of Malaysia for providing me a SLAI/SLAB scholarship for Phd. And big thanks to MARA University of Technology (UiTM) and all staffs of Faculty of Pharmacy, UiTM Puncak Alam for supporting me to study abroad.

Finally, I would like to mention my father, my siblings, nieces, nephews, best friends, my fiancé and the biggest appreciations to my late Bonda and my brother. I am very grateful for the supports from the whole family and friends. Without their supports and prayers, I could not finish my PhD. They bring me a lot of happiness and give me encouragement to overcome the difficulties during my scientific research.

LIST OF FIGURES

Figure 2.1 Classification of different types of protein interactions. 6

Figure 2.2 Pocket (orange sphere) identified by the protein pocket identification tool GHECOM that overlaps with the ligand dysiherbaine (red sticks) bound to human glutamate receptor, GluR5 (PDB ID: 3FV1).. 8

Figure 2.3 Example of formation of position weight matrices (PWM) and sequence motif. (A) Shown are eight known genomic binding sites in three *Saccharomyces cerevisiae* genes (HEM13, ANB1 and ROX1). (B) Frequencies matrix of nucleotides at each position. (C) Sequence logo used to visualize count at each position. (D) Sequence logo to represent the frequencies scaled relative using the information content at each position. (E) Energy normalized logo using relative entropy to adjust the GC content in *S.cerevisiae*.. 13

Figure 3.1 Schematic illustration of a PL complex illustrating the interface (black border) and the remaining surface regions 20

Figure 3.2 Flowchart summarizing the compilation of contacts between amino acids of the first protein (Pi1) and amino acids of the second protein (Pi2), atom contacts in PP and PL complexes, and the calculation of PR and IR 23

Figure 3.3 Percentage frequencies and propensities of amino acid residues at protein interfaces of PP complexes from the ABC dataset. (A) Percentage frequencies of amino acid residues at protein interfaces. (B) Propensities of amino acid residues at protein interfaces. 25

Figure 3.4 Percentage frequencies and propensities of amino acid residues at protein interfaces of PP complexes from the PIBASE dataset. (A) Percentage frequencies of amino acid residues at protein interfaces. (B) Propensities of amino acid residues at protein interfaces 26

Figure 3.5 Percentage frequencies and propensities of amino acid residues at protein interfaces of PL complexes from the ABC, PIBASE and Timbal datasets. (A) Percentage frequencies of amino acid residues at protein interfaces. (B) Propensities of amino acid residues at protein interfaces 29

Figure 3.6 Amino acid pairing propensities (in $\log_2$-format) for interfaces of PP complexes from the ABC dataset 30

Figure 3.7 Amino acid pairing propensities (in $\log_2$-format) for interfaces of PP complexes from the PIBASE dataset 31

Figure 4.1 Molecular structure of STIM1 (A) STIM1 domain organization. (B) Cartoon depicting a possible model of the STIM1 monomer in the resting state. 41

Figure 4.2 Molecular structure of STIM2 (A) STIM2 domain organization. (B) Cartoon depicting a possible model of the STIM2 monomer in the resting state. 41

Figure 4.3 Molecular structures of ORAI1, ORAI2, and ORAI3 (A) Domain organization of ORAI genes. (B) Cartoon depicting a possible model of ORAI in the resting state and sequence alignments 42

Figure 4.4 Flowchart summarizing the workflow of the analysis.....46

Figure 4.5 Dynamic graphical representations of 48 transcription factors targeting STIM and ORAI genes. The figure drawn using Cytoscape. 47

Figure 4.6 Schematic illustrations of transcription factors binding motifs in the promoters of the genes STIM1, STIM2, ORAI1, ORAI2 and ORAI3. (A) Transcription factor binding motifs in the promoters of genes STIM1 and STIM2. (B) Transcription factors binding motifs in the promoters of genes ORAI1, ORAI2 and ORAI3. .50

Figure 4.7 Network of 14 transcription factors targeting STIM and ORAI genes. Evidence view of the STRING database output depicting the transcription factors targeting STIM and ORAI genes obtained from http://string-db.org/.. 52

Figure 4.8 Plot of principal component analysis of 113 normal samples and 1102 tumor samples... 58

Figure 4.9 Topological overlap matrix (TOM) heatmap corresponding to the two co-expression modules in normal samples. Each row and column of the heatmap represents a single gene. Red indicates high levels of co-expression genes. The dendograms on the upper and left sides show the hierarchical clustering tree of genes... 59

Figure 4.10 Topological overlap matrix (TOM) heatmap corresponding to the three co-expression modules in tumor samples. Each row and column of the heatmap represents a single gene. Red indicates high levels of co-expression genes. The dendograms on the upper and left sides show the hierarchical clustering tree of genes... 60

Figure 4.11 Differential Interactions Networks (A) Differential interactions network of normal-tumor. (B) Differential interactions network of tumor-normal. Red edges indicate interactions that found in differential interactions and black edges indicate the interactions which were not found in differential interactions. Blue nodes indicate the genes found in the blue module, brown (blue module), and turquoise (turquoise module), respectively……....62

Figure 5.1 Two dimensional illustrations of three possible ways to define the pocket volume of a pacman-shape surface cavity... 67

Figure 5.2 Pockets (blue) identified by DEPTH (lining residues) (top panel left), GHECOM (top panel middle), Fpocket (top panel right), DoGSiteScorer (middle panel left), PocketPicker (middle panel middle), IsoMif (middle panel right), and ProACT2 (bottom panel) that overlap with the ligand HNT (red sticks) bound to human phenylethanolamine N-methyltransferase, PNMT (PDB ID: 2G70). The figures were generated using PyMOL Molecular Graphics System.....67

Figure 5.3 Distribution of the number of predicted pockets per protein for Fpocket, DoGSiteScorer, and IsoMif, respectively. 72

Figure 5.4 Distribution of the number of pocket lining residues per pocket (top five) for Fpocket, DoGSiteScorer, IsoMif, GHECOM, and PocketPicker, respectively…….. 73

Figure 5.5 Cumulative density of Matthews correlation coefficient (MCC), precision (pre), recall, and residues overlap (overlap)... 74

Figure 5.6 Heatmap of TScores of DEPTH, GHECOM, Fpocket, DoGSiteScorer, PocketPicker, IsoMif, ProACT2, and maxTScore, respectively. The column maxTScore contains the maximum TScores among all methods... 76

LIST OF TABLES

Table 2.1 Classification of different types of protein interactions. 6

Table 2.2 Pocket (orange sphere) identified by the protein pocket identification tool GHECOM that overlaps with the ligand dysiherbaine (red sticks) bound to human glutamate receptor, GluR5 (PDB ID: 3FV1) ..…..8

Table 2.3 Example of formation of position weight matrices (PWM) and sequence motif. (A) Shown are eight known genomic binding sites in three *Saccharomyces cerevisiae* genes (HEM13, ANB1 and ROX1). (B) Frequencies matrix of nucleotides at each position. (C) Sequence logo used to visualize count at each osition. (D) Sequence logo to represent the frequencies scaled relative using the information content at each position. (E) Energy normalized logo using relative entropy to adjust the GC content in *S. cerevisiae.*. 13

Table 3.1 Schematic illustration of a PL complex illustrating the interface (black border) and the remaining surface regions.....20

Table 3.2 Flowchart summarizing the compilation of contacts between amino acids of the first protein (Pi1) and amino acids of the second protein (Pi2), atom contacts in PP and PL complexes, and the calculation of PR and IR 23

Table 3.3 Percentage frequencies and propensities of amino acid residues at protein interfaces of PP complexes from the ABC dataset. (A) Percentage frequencies of amino acid residues at protein interfaces. (B) Propensities of amino acid residues at protein interfaces. 25

Table 3.4 Percentage frequencies and propensities of amino acid residues at protein interfaces of PP complexes from the PIBASE dataset. (A) Percentage frequencies of amino acid residues at protein interfaces. (B) Propensities of amino acid residues at protein interfaces 26

Table 3.5 Percentage frequencies and propensities of amino acid residues at protein interfaces of PL complexes from the ABC, PIBASE and Timbal datasets. (A) Percentage frequencies of amino acid residues at protein interfaces. (B) Propensities of amino acid residues at protein interfaces.....29

Table 3.6 Amino acid pairing propensities (in $\log_2$-format) for interfaces of PP complexes from the ABC dataset 30

Table 3.7 Amino acid pairing propensities (in $\log_2$-format) for interfaces of PP complexes from the PIBASE dataset.....31

Table 4.1 Molecular structure of STIM1 (A) STIM1 domain organization. (B) Cartoon depicting a possible model of the STIM1 monomer in the resting state. 41

Table 4.2 Molecular structure of STIM2 (A) STIM2 domain organization. (B) Cartoon depicting a possible model of the STIM2 monomer in the resting state. 41

Table 4.3 Molecular structures of ORAI1, ORAI2, and ORAI3 (A) Domain organization of ORAI genes. (B) Cartoon depicting a possible model of ORAI in the resting state and sequence alignments 42

Table 4.4 Flowchart summarizing the workflow of the analysis.....46

Table 4.5 Dynamic graphical representations of 48 transcription factors targeting STIM and ORAI genes. The figure drawn using Cytoscape. 47

Table 4.6 Schematic illustrations of transcription factors binding motifs in the promoters of the genes STIM1, STIM2, ORAI1, ORAI2 and ORAI3. (A) Transcription factor binding motifs in the promoters of genes STIM1 and STIM2. (B) Transcription factors binding motifs in the promoters of genes ORAI1, ORAI2 and ORAI3. .50

Table 4.7 Network of 14 transcription factors targeting STIM and ORAI genes. Evidence view of the STRING database output depicting the transcription factors targeting STIM and ORAI genes obtained from http://string-db.org/.. 52

Table 4.8 Plot of principal component analysis of 113 normal samples and 1102 tumor samples... 58

Table 4.9 Topological overlap matrix (TOM) heatmap corresponding to the two co-expression modules in normal samples. Each row and column of the heatmap represents a single gene. Red indicates high levels of co-expression genes. The dendograms on the upper and left sides show the hierarchical clustering tree of genes... 59

Table 4.10 Topological overlap matrix (TOM) heatmap corresponding to the three co-expression modules in tumor samples. Each row and column of the heatmap represents a single gene. Red indicates high levels of co-expression genes. The dendograms on the upper and left sides show the hierarchical clustering tree of genes... 60

Table 4.11 Differential Interactions Networks (A) Differential interactions network of normal-tumor. (B) Differential interactions network of tumor-normal. Red edges indicate interactions that found in differential interactions and black edges indicate the interactions which were not found in differential interactions. Blue nodes indicate the genes found in the blue module, brown (blue module), and turquoise (turquoise module), respectively……....62

Table 5.1 Two dimensional illustrations of three possible ways to define the pocket volume of a pacman-shape surface cavity... 67

Table 5.2 Pockets (blue) identified by DEPTH (lining residues) (top panel left), GHECOM (top panel middle), Fpocket (top panel right), DoGSiteScorer (middle panel left), PocketPicker (middle panel middle), IsoMif (middle panel right), and ProACT2 (bottom panel) that overlap with the ligand HNT (red sticks) bound to human phenylethanolamine N-methyltransferase, PNMT (PDB ID: 2G70). The figures were generated using PyMOL Molecular Graphics System.....67

Table 5.3 Distribution of the number of predicted pockets per protein for Fpocket, DoGSiteScorer, and IsoMif, respectively. 72

Table 5.4 Distribution of the number of pocket lining residues per pocket (top five) for Fpocket, DoGSiteScorer, IsoMif, GHECOM, and PocketPicker, respectively…….. 73

Table 5.5 Cumulative density of Matthews correlation coefficient (MCC), precision (pre), recall, and residues overlap (overlap)... 74

Table 5.6 Heatmap of TScores of DEPTH, GHECOM, Fpocket, DoGSiteScorer, PocketPicker, IsoMif, ProACT2, and maxTScore, respectively. The column maxTScore contains the maximum TScores among all methods... 76

ABREVIATIONS

3D	Three dimensional
Å	Angstrom
aa	Amino acid
bp	Base pair
BRCA	Breast invasive carcinoma
$Ca^{2+`}$	Calcium
ChIP	Chromatin immuniprecipitation
ChIP-chip	Chromatin immuniprecipitation microarrays
CRAC	Calcium Release Activated Calcium
Da	Dalton
DESeq2	Differential Expression analysis for Sequence count data version 2
DNA	Deoxyribonucleic acid
edgeR	Empirical analysis of digital gene expression in R
ENCODE	Encyclopedia of DNA Elements
ER	Endoplasmic reticulum
FDR	False Discovery Rate
FN	False negative
FP	False positive
MD	Molecular Dynamics
MCC	Matthews coefficient correlation
ORAI1	Calcium release-activated calcium channel protein 1
ORAI2	Calcium release-activated calcium channel protein 2
ORAI3	Calcium release-activated calcium channel protein 3
PDB	Protein Data Bank
PL	Protein-ligand
PP	Protein-protein
PWM	Position weight matrices
SASA	Solvent Accessible Surface Area
SOCE	Store-Operated $Ca^{2+`}$entry
STIM	Stromal Interaction Molecules
STIM1	Stromal Interaction Molecule 1
STIM2	Stromal Interaction Molecule 2
SVM	Support Vector Machine
TCGA	The Cancer Genome Atlas
TF	Transcription factor
TFBS	Transcription factor binding site
TM	Transmembrane
TN	True negative
TOM	Topological overlap matrix
TP	True positive
TSS	Transcription start site
WGCNA	Weight Gene Co-expression Network Analysis

TABLE OF CONTENTS

Abstract i
Zusammenfassung ii
Acknowledgements iii
List of Figures iv
List of Tables vii
Abreviations x
Table of Contents xi

Chapter 1: Introduction 1
1.1 Introduction 1
1.2 Motivation 1
1.3 Contributions 2
1.4 Thesis Organization 3

Chapter 2: Biological Background and Computational Methods 4
2.1 The Nature of Protein 4
2.2 Protein-Protein Interactions, Protein-Ligand Interactions, and Protein Interfaces 4
2.3 Types of Protein Interactions 5
2.4 The Protein Data Bank (PDB) 6
2.5 Protein Pockets 7
2.6 Computational Methods for Identification of Protein Pockets 8
2.6.1 Geometry based Methods 8
2.6.1.1 Grid System Scanning 9
2.6.1.2 Probe Sphere Filling Methods 9
2.6.1.3 Alpha Shape Methods 10
2.6.2 Energy based Methods 10
2.6.3 Evolution based Methods 10
2.6.4 Blind Docking and Molecular Dynamic Methods 11
2.6.5 Combined Approaches 11
2.7 Transcription Factors, Promoters, Motifs, and Position Weight Matrix (PWM)…...12
2.7.1 Transcription Factors and Promoters 12
2.7.2 Motifs and position weight matrix (PWM) 12
2.8 Gene Expression, Differential Gene Expression, and Co-expression Analysis 13
2.8.1 Gene Expression 13
2.8.2 Differential Gene Expression Analysis 14
2.8.2.1 DESeq, DESeq2, and edgeR Packages 14
2.8.3 Co-expression Analysis 15

Chapter 3: Composition of Overlapping Protein-Protein and Protein-Ligand Interfaces 17
Abstract 17
3.1 Introduction 18

3.2 Material and Methods 19
3.2.1 Datasets 19
3.2.2 Surface and Interface Residues 20
3.2.3 Classification of the Amino Acids 21
3.2.4 Interface Residue Propensities 21
3.2.5 Surface and Interface Residues 21
3.2.6 Atom Contacts in Protein-protein and Protein-ligand Complexes 21
3.2.7 Calculation of Polarity Ratio and Interface Atom Ratio 22
3.3 Results and Discussion 22
3.3.1 Amino Acid Composition and Protein Interfaces Propensity 24
3.3.2 Amino Acid Contacts 29
3.3.3 Atomic Contacts in Protein-protein and Protein-ligand Complexes 32
3.3.4 Polarity Ratio and Interface Atom Ratio 37
3.4 Conclusions 38

Chapter 4: STIM and ORAI Genes, Interactions with Transcription Factors, Differential Gene Expression and Co-expression Analysis on Breast Invasive Carcinoma Dataset 39
Abstract 39
4.1 Introduction 40
4.2 Material and Methods 43
4.2.1 Transcription Factors Targeting STIM and ORAI Genes 43
4.2.2 Transcription Factors Binding Models 43
4.2.3 Sequence of Promoter Regions 44
4.2.4 Motif Over-representation Analysis 44
4.2.5Transcription Factors Predicted by CheA Mapped with the FIMO Results 44
4.2.6 Prediction of Physical Interactions of Protein-protein Complexes 45
4.2.7 TCGA BRCA Dataset 45
4.2.8 Computing Differential Analysis 45
4.2.9 Weight Gene Co-expression Network Analysis 45
4.2.10 Flowchart 46
4.3 Results and Discussion 47
4.3.1 Transcription Factors Targeting STIM and ORAI Genes 47
4.3.2 Promoter Sequences of STIM and ORAI Genes from the EPDnew Database 48
4.3.3 HOCOMOCO and FIMO Results 48
4.3.4 Physical Protein-protein Interactions 51
4.3.5 Differential Gene Expression Analysis 56
4.3.6 Results of Gene Co-expression Analysis 58
4.3.6.1 Normal samples 58
4.3.6.2 Tumor samples 59
4.3.7 Differential Interaction Networks 61
4.3.8 Regulation of STIM and ORAI Genes on Normal and Tumor Breast Cancer Tissue 63
4.4 Conclusions 63

Chapter 5: Evaluation of Protein Pocket Identification Tools on Protein-Ligand Complexes 65
Abstract 65
5.1 Introduction 66
5.2 Material and Methods 68
5.2.1 Dataset 68
5.2.2 Tools 68
5.2.3 Binding Site Definition 69
5.2.4 Model Evaluations 69
5.2.5 MCC 69
5.2.6 Precision 69
5.2.7 Recall 70
5.2.8 Overlap 70
5.2.9 Correlation between Features and Quality Metrics 70
5.2.10 TSore 70
5.2.11 TRatio 70
5.2.12 Classifiers 71
5.3 Results 71
5.3.1 Pocket Distributions 71
5.3.2 Size Distributions 72
5.3.3 Prediction versus Reality 73
5.3.4 Features and Quality Metrics 74
5.3.5 Tool Score 75
5.3.6 Identifying an Optimal Tool for a Protein 76
5.4 Discussion ..... .77
5.5 Conclusions ..... .78

Chapter 6: Conclusions and Future Works 79

7: References 81

8: Supplementary Information 99
8.1 Supplementary Information for Chapter 3 99
8.2 Supplementary Information for Chapter 4 133
8.3 Supplementary Information for Chapter 5 141

Chapter 1

Introduction

1.1 Introduction

Protein interactions play a major role in many biological processes. Numerous important applications benefit from the identification of protein-protein and protein-ligand interactions such as drug design, protein mimetics engineering, elucidation of molecular pathways, and understanding of disease mechanisms. Nowadays, a wide range of high-throughput experimental approaches are available for identification of protein-protein and protein-ligand interactions. For that reasons, many publicly accessible databases stores the high-quality information about protein-protein and protein-ligand complexes. Moreover, for almost three decades, gene expression was recognized to be mainly regulated at the transcriptional level and protein known as transcription factors functions in regulating the expression of a gene. Any changes of protein interactions can lead to mutations and diseases by affecting the functions of protein complexes or by affecting gene regulations.

1.2 Motivation

This thesis addresses three problems which related mostly to protein interactions and gene regulation. Nowadays, many efforts and studies have analyzed protein-protein and protein-ligand interactions, however a detailed of protein interfaces composition is remain elusive as targeting protein-protein interactions is challenging because usually no convenient natural substrates are available as starting point for small-molecule design. Hence, by using five datasets from three different databases, our aim was to explore the characteristics and composition of overlapping protein-protein and protein-ligand interfaces.

On the other hand, many studies have been done on STIM and ORAI genes as they play important roles in calcium signals and involved in store-operated Ca^{2+} entry (SOCE) in cells. STIM1 and STIM2 are needed for the development and functioning of various cell types such as lymphocytes, skeletal and smooth muscle myoblast, adipocytes and neurons [1,2]. In addition, previous studies stated that STIM and ORAI genes are associated with differentiation processes and linked to several diseases such as Alzheimers, Parkinsons diseases and cancer (e.g. breast cancer and prostate cancer) [3–6]. Today, cancer is a major public health problem worldwide which breast cancer is one of the most common and predominant cancer types that affects 1.3 million people and causes thousands of death annually [7,8]. In the United States of

America, 30.4% of women are diagnosed with breast cancer yearly [7]. However, only few studies addressed the involvements of STIM and ORAI genes to breast cancer. In addition, the discovery of transcription factor binding sites (TFBS) motifs in specific locations in the promoter regions of STIM and ORAI genes is remain elusive. Our aim was to find the interacting transcription factors bound to the particular promoter regions of STIM1, STIM2, ORAI1, ORAI2, and ORAI3 and their regulation and relation with normal and breast invasive carcinoma (BRCA) samples.

Identification of binding pockets is often a prerequisite step for structure-based drug design. However, the characterization of these sites is a main challenge in computational biology. Defining the correct pockets on protein surfaces is not an easy task. Recently, many protein pocket identification tools have been developed by using different strategies and approaches. So far apparently, none of the existing approaches performed in a way which can set as a standard or that is widely accepted as benchmarking method to identify protein pockets accurately. Hence, it is great of interest to evaluate and compare the performances of these tools. By using seven different protein pocket identification tools and tested on a dataset of protein-ligand complexes, our aim was to evaluate and compare their performances and suggest a set of the best tool which can identify protein pockets accurately.

1.3 Contributions

Most of the results chapters of this thesis are based on manuscripts that either published or ready for submissions as follow:

Chapter 3: Ruzianisra Mohamed, Jennifer Degac, and Volkhard Helms. Composition of Overlapping Protein-Protein and Protein-Ligand Interfaces. PloS One. 2015 Oct 30;10(10):e0140965. doi: 10.1371/journal.pone.0140965.

Chapter 4: Ruzianisra Mohamed, Riccha Sethi, Mohamed Hamed, and Volkhard Helms. STIM and ORAI Genes, Interactions with Transcription Factors, Differential Gene Expression and Co-expression Analysis on Breast Invasive Carcinoma Dataset (In preparation for submission to peer reviewed journal).

Chapter 5: Zhao Yuan, Rahmad Akbar, Volkhard Helms, and Ruzianisra Mohamed. Evaluation of Protein Pocket Identification Tools on Protein-Ligand Complexes (In preparation for submission to peer reviewed journal).

1.4 Thesis Organization

The structure of the thesis is as follows:

- Chapter 2 provides general introduction to biological background and computational methods such as proteins, protein-protein and protein-ligand interactions, protein pocket identification algorithms, transcription factors, promoters, gene expressions, differential gene expression analysis and co-expression analysis.
- Chapter 3 will discuss the topic on compositions of overlapping protein-protein and protein-ligand interfaces.
- Chapter 4 will show the study on STIM and ORAI genes, interactions with transcription factors and differential gene expression and co-expression analysis on a dataset of breast invasive carcinoma.
- Chapter 5 will discuss the topic of evaluation of seven protein pocket identification tools on a dataset of 167 protein-ligand complexes.
- Chapter 6 summarizes the results of the projects and provides conclusions with regard to the aims of the studies and contributions made and present the future works.

Chapter 2

Biological Background and Computational Methods

In this chapter, we present the background on the biological concepts and computational methods relevant to this thesis. We also briefly describe the publicly accessible biological databases, tools, and packages that were used in this thesis.

2.1 The Nature of Protein

Proteins rarely act by themselves and in most cases in order to function in biological systems, they work in groups which known as protein complexes. Protein complexes perform many important tasks within biological cells such as catalyzing metabolic reactions, replicating DNA, and transporting molecules from one place to another. Being workhorses that assist many biological processes, their detection is needed to increase our knowledge about cellular organizing and function. Generally, a protein consists of a sequence of 20 amino acids (also known as residues) which are linked together through peptide bonds, forming a polypeptide chain known as the primary structure. The primary structure forms secondary structural elements, for example alpha helices and beta sheets through hydrogen bonds, which interact to form the tertiary structure via protein folding. The tertiary structure may combine with another to construct a quaternary structure.

2.2 Protein-Protein Interactions, Protein-Ligand Interactions, and Protein Interfaces

As mentioned above, proteins play major roles in almost all biological functions. Most of the molecular processes are based on molecular machines which are composed of a large number of proteins and bind to each other through protein-protein interactions. The interactions of proteins are determined and usually mediated by their interfaces [9]. Protein interfaces are defined as binding sites of certain patches on each of the two protein's surface, which enables the interaction between two proteins. The characteristics of the protein-protein and protein-ligand interfaces are important to describe binding principles and provide clues about algorithms for binding site prediction. One of the most important aims in interface studies is to identify properties which may distinguish residues at binding sites from the rest of the protein surfaces. A few studies showed that interfaces are rather large, planar or well packed depending on the type of the interactions [10,11].

Due to the importance of protein interfaces in mediating protein-protein and protein-ligand interactions, various studies have analyzed protein interfaces. The analysis of three dimensional (3D) protein structures revealed that protein interfaces are composed of buried cores which are surrounded by partially accessible rims [12,13]. Additionally, the size of the interface patches ranges between 400 and 1600Å^2 [12].

Interestingly, PPI interfaces contain small regions called “hot spots” that contribute the most to the total binding energy [10,14,15]. A powerful experimental approach named alanine scanning mutagenesis is used to identify hot spot regions by successively mutating each interface residue to alanine. Those residues are termed hot spots where mutation into alanine leads to a decrease in binding free energy by more than 2.0 kcal/mol [16]. Since hot spots by definition make up the largest contribution to the binding energy, the ability to predict hot spots is important to identify, analyze and lead to targets for drug binding sites. To circumvent the required time and costs and the experimental uncertainty whether protein mutants will properly express and purify, several computational methods have been designed and used to predict hot spots. They typically either use energy based methods to calculate the energetic contributions of residues [17,18], solvent accessible surface area (SASA) [19], and sequence conservation based approaches [20].

2.3 Types of Protein Interactions

Studies showed that many factors are influencing the classification of protein interactions into different types of interactions [10,21]. The most basic classification of interactions is based on the composition of the protein complex. For example a complex consisting of identical proteins is termed homo-oligomer (e.g. homodimer complex). Usually, homo-oligomers are very stable permanent protein structures such as in the case of homodimer (two proteins) complexes [10]. On the other hand, a complex made of non-identical proteins is defined as hetero-oligomers (e.g. heterodimer complex) [22]. Then, protein interactions can be further grouped into obligate or non-obligate depending on the lifetime of the interactions [22,23]. Obligate complexes only exist in the bound form. Components of non-obligate complexes such as antibodies are also stable in the unbound form. Based on their binding affinity, the interaction can be considered to be permanent or transient. Generally, all obligate interactions are permanent, however not all permanent interactions are obligate. Non-obligate and transient protein interactions usually have lower binding affinities. The transient interactions are subdivided into weak and strong transient. The strong transient ones include protein interactions that shift from an unbound or weakly bound state to a strongly bound state. This is frequently stimulated by an effector

molecule. Figure 2.1 shows the classification of protein interactions. The protein interaction types depend on the situation and cellular process. Therefore, it is importance to understand protein-protein interaction types and the effects they have on certain biological processes.

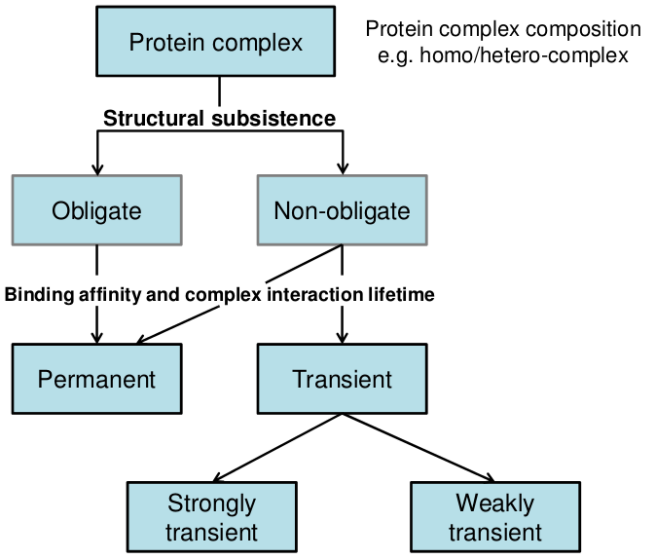

Figure 2.1 Classification of different types of protein interactions.

2.4 The Protein Data Bank (PDB)

The 3D coordinates of protein structures are generally deposited in The Protein Data Bank (PDB) [24], the globally recognized primary depository for experimentally determined atomistic structure of 3D biological macromolecules. The PDB was developed in 1971 at the Brookhaven National Laboratories containing a set of seven protein structures. Then, in 1998, the PDB moved under management of the Research Collaboratory for Structural Bioinformatics (RCSB) at the Rutgers University, New Jersey (http://www.rcsb.org/pdb/). Since early 1990's, the number of 3D structures deposited in the PDB has been increasing exponentially. As for 16 May 2016, the PDB stores 118748 structure of biological macromolecules of which more than 92% are proteins, 2% nucleic acid complexes, 4% protein-nucleic acid complexes and 0.02% other complexes. 99% of the structures were determined by X-ray crystallography (89%) and by Nuclear Magnetic Resonance (NMR) Spectroscopy (10%). 1022 structures were solved with electron microscopy, 95 hybrid and 196 by other methods. The PDB also stores 20955 ligands which interact with the proteins (protein-ligand complexes).

The PDB is the primary resource to study the diversity of protein-protein (PP) and protein-ligand (PL) interactions. Thus, some secondary databases have been developed to

assist the research on PP interactions such as the Biological General Repository for Interaction Datasets BioGRID (thebiogrid.org) [25,26] which stores data of small molecules modulating protein–protein complexes, the Molecular INTeraction database (MINT) (mint.bio.uniroma2.it/mint/Welcome.do) [27], the Biomolecular Interaction Network Database (BIND) (http://bind.ca) [28], the Database of Interacting Proteins (DIP: http://dip.doe-mbi.ucla.edu) [29], the IntAct molecular interaction (IntAct) database (www.ebi.ac.uk/intact/) [30], the ABC database (http://service.bioinformatik.uni-saarland.de/ABCSquareWeb/) [31], and the database of structurally defined protein interfaces named PIBASE (http://pibase.janelia.org/pibase2010/queries.html) [32].

On the other hand, the high-quality databases of PL interactions can assist the field of structure-based drug design to develop the best computational tools. There are several publicly available databases such as the Timbal database (http://mordred.bioc.cam.ac.uk/timbal) [33], the Mother of All Database (MOAD) (http://bindingmoad.org) [34,35], the 2P2I database (http://2p2idb.cnrs-mrs.fr) [36], the PDBbind (http://sw16.im.med.umich.edu/databases/pdbbind/index.jsp) [37], the BindingDB database (www.bindingdb.org) [38,39], and Relibase (http://relibase.ccdc.cam.ac.uk/account_utilities/login_form.php) [40]. The ABC and PIBASE databases also stored data on PL interactions.

2.5 Protein Pockets

The accurate prediction of protein pockets from a 3D protein structure is an important issue for structure-based drug design [41,42] which can elucidate protein functions [43]. The basic principle of molecular interactions was proposed by Emil Fischer in 1894, who stated that ligand binding to proteins follows a “lock and key” mechanism [44]. These interactions often occur in particular sites on protein surfaces (binding sites). These binding sites can be distinguished from other parts of the protein surface by their unique characteristics as such the electrostatic properties and the size of a cavity on the protein surface [45]. Additionally, in a single protein there can be many pockets and the sizes of ligands are varying. They can be as small as ions or large polymers. Figure 2.2 shows an example of a pocket identified by the protein pocket identification tool GHECOM, that overlap with the ligand dysiherbaine bound to human glutamate receptor, GluR5 (PDB ID: 3FV1).

Generally, clinically approved drugs are classified into two broad classes (i) small molecules which typically comprise of <100 atoms with a molecular mass of <1,000 Da and (ii) biologics such as antibodies, modified nucleic acids, peptides, and vaccines [46]. The small

molecules usually traverse cellular membranes before they can reach intracellular target molecules and work on a relatively limited surface area. On the other hand, biologics generally interact with a large surface area that contains multiple interaction sites. They are also larger in sizes which restrict their mode of delivery and pose more challenges in drug design.

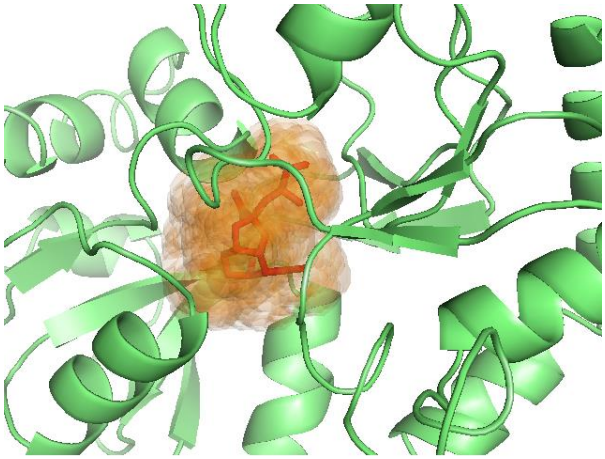

Figure 2.2 Pocket (orange sphere) identified by the protein pocket identification tool GHECOM that overlaps with the ligand dysiherbaine (red sticks) bound to human glutamate receptor, GluR5 (PDB ID: 3FV1). The figure was generated using PyMOL Molecular Graphics System [47].

2.6 Computational Methods for Identification of Protein Pockets

As the experimental identification of a binding site is not always feasible, as alternative computational protein pocket identification methods are required. Nowadays, many computational approaches for protein pocket identification have been developed which greatly accelerate drug discovery and protein designs. These protein pocket identification tools and algorithms fall into five different categories according to the methods applied (i) geometry based, (ii) energy based, (iii) evolution based, (iv) blind docking, and (v) combined approaches [48]. Geometry based methods can be grouped into three subcategories (i) grid system scanning, (ii) probe sphere filling, and (iii) based on the alpha shape theory [49,50]. In the following sections these methods are introduced in more details.

2.6.1 Geometry based Methods

Geometry based methods aim to identify solvent accessible regions that are located in surface cavities and clefts by the analysis of geometric criteria [48]. These approaches have been popular for years because they appear to exhibit good performance in the identification protein pockets. They were reported to predict almost 95% of the known binding sites. At a fast computational speed and they are robust in handling cases of structural variations or missing atoms/residues in the protein complexes [51]. Generally, geometry based methods are based

on the hypothesis that the ligand binding site are represented by the largest pocket [52–54]. Although these methods only considered the largest pocket, in reality this is not always the cases. Therefore, these methods have been further improved and new algorithms of prediction have been developed. In the following subsections the subcategories of geometry based methods are introduced in more detail.

2.6.1.1 Grid System Scanning

Grid based methods are a subcategory of geometry based methods that use a grid to define the molecular surface. These methods typically focus on the buriedness of grid points and the protein surface [54,55]. For example, DogSiteScorer is a grid based method. When tested on a dataset of 1069 structures, it achieved prediction accuracies more than 80% [56]. Generally, the initial step in the DoGSiteScorer process is the prediction of pockets on the protein surface based on the coordinates of the protein heavy atoms. A grid is spanned around the protein and grid points are labelled according to their spatial overlap with protein atoms. Then, the difference of Gaussian (DoG) filter is applied to the grid to identify the position of a cavity on the protein surface. Next, these positions are clustered to potential subpockets based on a density threshold. Finally, pockets are identified as the collection of merged neighbouring subpockets [48,56]. On the other hand, DEPTH [57,58], GHECOM [49], PocketDepth [59], and PocketPicker [55] also implement grid system scanning methods.

2.6.1.2 Probe Sphere Filling Methods

Probe sphere filling methods identify protein pockets by generating a set of probe spheres to fill cavities in a protein. Those regions containing the largest number of spheres are defined as pockets. In addition, these approaches use different types of probes such as (i) gap sphere [60], (ii) rotating probe, (iii) multiscale probe [49], (iv) the combination of big and small probes [61], and (v) probes placed tangential to triplets of protein atoms [62]. For example, SURFNET [60] and IsoMif [63] are representatives of these methods.

2.6.1.3 Alpha Shape Methods

In the1990s, the studies by Edelsbrunner and co-workers led to the development of new methods based on the alpha-shape theory [64,65]. The accuracy of the alpha-shape methods is influenced by the alpha values. The Automatic Protein Pocket Search (APROPOS) [66], the Computed Atlas of Surface Topology (CAST) [67], and Fpocket [68] are tools that implement alpha shape methods. For instance, CAST applied alpha shape methods to compute a triangulation of the protein surface atoms. This tool uses the discrete flow concept by allowing the small triangle flow to the neighbouring bigger triangles which act as "sinks" that later collect excess flow from neighbouring empty triangles. The collection of empty triangles is defined as the protein pocket [67].

2.6.2 Energy based Methods

Energy based methods rely on the assumption that the potential binding sites are characterized by binding energies which are different from the rest of the protein surface. Usually, these methods uses simple van der Waals (vdW) probes to locate the grid points around the protein surfaces and calculate interaction energies between the probe and a protein. The protein pockets are defined only by the energetic features. For examples, ProACT2 [69,70] and Q-SiteFinder [71] are tools that apply energy based methods to predict protein pockets.

2.6.3 Evolution based Methods

Evolution based methods also known as sequence based evolutionary conservation methods, are based on the assumption that functionally important residues of proteins are typically conserved during the evolution because of natural selection and these functional areas are mainly the protein pockets. Generally, the degree of conservation at each amino acid site is divided into two categories (i) slowly evolving sites which refer to evolutionarily conserved and (ii) rapidly evolving sites [72]. Based on this, several tools have been developed such as ConSeq [73], ConSurf [72], and a new version of ConSurf called Consurf 2010 [74] to identify the functional areas of unknown proteins by comparing their amino acid sequences to the already known amino acid sequences of proteins. These methods are fast, robust, and require only a protein sequence to predict the binding sites. Exceptions to this are the methods by de Rinaldis et al. 1998 [75] and siteFINDER|3D [76].

2.6.4 Blind Docking and Molecular Dynamic Methods

In addition to methods mentioned above, molecular dynamics (MD) and blind docking algorithms are also considered as useful methods to identify protein pockets. These approaches are most practical in cases where the ligand that binds in the target protein is known, but the binding site is unknown and requires the 3D-structure of the protein. Generally, MD calculations use two different approaches which are based on two situations (i) mobility of water molecules and (ii) long scale MD simulations to find the correct binding site of an already known active ligand. For example, a representative for this methods is molecular-docking binding site finding (MolSite) [77]. On the other hand, blind docking methods work by scanning the entire surface of the protein target to identify binding sites and modes of peptide ligands [78]. AnchorDock [79] is a representative of blind docking methods.

2.6.5 Combined Approaches

So far, apparently none of the methods described above performed in a way which can set as a standard or that is widely accepted as benchmarking method to identify protein pockets accurately. Several of the existing methods are often unsuccessful in certain types of cases, in which the algorithms are not able to take into account correctly or adequately all properties of the target site and did not work for all the entries in the dataset [80,81]. Furthermore, improving the existing algorithms does not necessarily can produce better predictions results. Thus the idea of extending the methods in other ways which aim to reduce the weaknesses of other algorithm appears promising. Consequently, the combination of two or more methods seems to be a good solution to improve the identification of protein pockets.

The first attempt of combined approaches was started by a study of Del Sol Mesa and co-workers (2003) [82]. They combined three separate evolution based methods but did not combine different types of approaches. Later, a study by Huang B et al., 2009 [83] introduced the MetaPocket tool which uses and combines the four methods LIGSITE^{cs} [84], PASS [62], Q-SiteFinder [71], and SURFNET [60] to predict protein pocket binding sites. The method was tested on two different datasets of 48 unbound/bound structures and 210 bound structures. The results showed a success rate of 70 to 75%. Two years later, Zhang et al. and co-workers introduced MetaPocket2.0 [85]. By applying these combined approaches to the same datasets, the results increased by 5% compared to the previous study.

2.7 Transcription Factors, Promoters, Motifs, and Position Weight Matrix (PWM)

2.7.1 Transcription Factors and Promoters

The information within a gene is expressed by the cellular processes of transcription and translation. Transcription factors are regulatory DNA-binding proteins which play major roles in the regulation of gene expression. Previous studies noted that the larger total number of transcription factors reflects the larger number of genes in the genome of an organism which results in a larger size of the genome [86,87]. Single transcription factor or transcription factor complexes bind to the promoter region of the coding sequence which influencing RNA transcription. In addition, transcription factors can regulate gene expression by either activating or repressing gene transcription.

On the other hand, promoters are defined as the genomic regions that are located 5'upstream of the transcription start site (TSS) of genes. Promoters are the important element of expression vectors as they control the attachment of RNA polymerase and required to recruit the transcription initiation complexes and initiate transcription. There is no accurate definition of promoter length. Promoter binding is different in bacteria compared to eukaryotes as in bacteria RNA polymerase only needs the association of protein sigma factor to bind to the promoter. In contrast, eukaryotes require several transcription factors for the binding of RNA polymerase II to the promoter. Usually the complex consists of specific TFs, general TFs, co-factors, and RNA polymerase II [88]. Transcription does not only depend on co-factors and TFs and their ability to work together, but also on the structure of the chromatin. Additional mechanisms that also control gene expression are RNA interference, methylation and acetylation. The EPDnew database is one of the several publicly available resources of species-specific databases of experimentally validated promoters [89].

2.7.2 Motifs and position weight matrix (PWM)

Transcription factor binding sites (TFBSs) are the specific recognition sites in the DNA sequence for a given transcription factor. These binding sites are often referred to as occurrences of the motif for the corresponding TF. Nowadays, there are several approaches for the genome-wide detection of TFBSs such as (i) computational approaches which are based on the DNA sequence-based analysis and (ii) experimental approaches, which are nowadays based on chromatin immunoprecipitation (ChIP) and DNaseI HS-based technologies.

A position weight matrix (PWM) is a common way of representing patterns in biological sequences. It consists of a stack of letters representing each nucleotide at each position and the height of each letter is proportional to its value in the PWM. In addition, the

sequence logos are regularly used to visualize count or frequency matrices. Figure 2.2 shows an example formation of PWM and the resulting sequence motif for eight known genomic binding sites in three genes HEM13, ANB1 and ROX1 from *Saccharomyces cerevisiae*. PWMs and sequence logos are available in several databases such as the JASPAR [90], HOCOMOCO [91], and TRANSFAC [92]. In this thesis, we identified the putative TFBS in the promoter regions of human STIM and ORAI genes. The description of STIM and ORAI genes are discussed in the Introduction section (4.1) of Chapter 4.

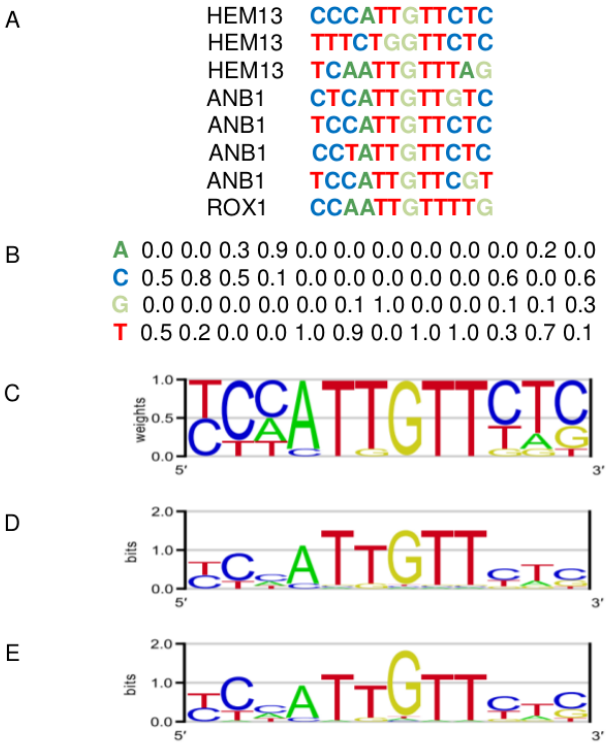

Figure 2.3 Example of formation of position weight matrix (PWM) and sequence motif. (A) Shown are eight known genomic binding sites in three *Saccharomyces cerevisiae* genes (HEM13, ANB1 and ROX1). (B) Frequencies matrix of nucleotides at each position. (C) Sequence logo used to visualize count at each position. (D) Sequence logo to represent the frequencies scaled relative using the information content at each position. (E) Energy normalized logo using relative entropy to adjust the GC content in *S.cerevisiae*. The figures were taken from [93,94].
GC content, guanine-cytosine content.

2.8 Gene Expression, Differential Gene Expression, and Co-expression Analysis

2.8.1 Gene Expression

The high-throughput sequencing technologies (HTS) are the most common approaches used in genomic studies which later involve statistical analysis to measure quantitative differences between experiments. It is important to analyze RNA expression levels (also known as RNA-seq data) to detect which genes that are differentially expressed across a group of samples.

Nowadays, there have been active efforts due to the advent of sequencing technologies with reduced costs that produce detailed profiling of gene expression levels, which are important in life sciences fields and clinical use [95]. Furthermore, the RNA-seq technology has been used to study complex transcriptomes and has assisted identification of levels of transcripts and isoforms, translocation events, sequence variations (for example SNPs) in the transcribed regions and post-transcriptional base modification.

Hence, various statistical approaches and tools have been developed such as differential expression analysis, random effects, gene set enrichment, gene set testing, and co-expression analysis to analyze the large datasets of genome-wide gene expression experiments. In this thesis, the differential gene expression analysis, co-expression analysis and differential interactions networks was performed using a dataset of breast invasive carcinoma (BRCA) obtained from the online data portal, The Cancer Genome Research (TCGA) (see Chapter 4).

2.8.2 Differential Gene Expression Analysis

The rapid growth of high-throughput technologies and publicly available datasets of RNA expression levels has motivated many studies to develop statistical algorithms that implement various approaches for normalization and differential gene expression analysis such as edgeR [96], DESeq [97], DESeq2 [98], PoisssonSeq, baySeq, and linear models for microarray data (limma). Generally, differential gene expression analysis of RNA-Seq data consists of three steps such as normalization of counts, parameter estimation of the statistical model and statistical tests for different expression.

2.8.2.1 DESeq, DESeq2, and edgeR Packages

DESeq, DESeq2 the extendable version of successful DESeq method and digital gene expression in R (EdgeR) are widely used Bioconductor packages for differential expression analysis of RNA-Seq and ChIP-Seq count data. These packages are very popular among user with biological background because they are easy to use, well documented and perform the best in replicated experiments. Generally, these statistical methods are based on the negative binomial distribution. Moreover, these packages have similar steps to perform differential analysis for count data. However they differ in several aspects such as (i) their look and feel, (ii) default normalization for example, edgeR applies the trimmed mean of M values and DESeq applies a relative log expression approach, and (iii) the application of dispersion estimate [99]. Typically, edgeR uses moderates individual dispersion estimates toward a trended-by-mean estimation [96]. A study by Dillies et al. 2013 [100] showed that the

normalization methods of edgeR and DESeq2 packages outperforms other approaches either in the case of expressed RNA repertoires that vary across biological conditions or in the presence of highly expressed genes. On the other hand, for the quality control step, clustering and Principal Component Analysis (PCA) can be used to assess the structure of the dataset.

In differential gene expression analysis, the False Discovery Rate (FDR) was used as a multiple testing correction approach. This was introduced by Benjamini and Hochberg (BH) in 1995 [101]. Multiple testing corrections adjust the p-values obtained from a large number of hypotheses testing to correct the occurrence of false positives. The FDR is defined as the expected proportion of falsely rejected null hypotheses among all rejected null hypotheses.

$$FDR = E\left(\frac{\text{number of falsely rejected null hypotheses}}{\text{number of rejected null hypotheses}}\right) \quad (2.1)$$

For example, a threshold of 0.02 FDR indicates that only two false positives are expected in 100 predictions. DESeq2 computes a q-value (FDR adjusted p-value) for each gene and uses it as the threshold to identify differentially expressed genes.

2.8.3 Co-expression Analysis

Generally, gene co-expression analysis is used to detect gene pairs that are coordinated in their expression profiles and to explore network characteristics of complex traits. In addition, gene co-expression network differential analysis is designed to assist biologists in many applications such as discovering protein-protein interaction relationships, predicting new gene functions, pathways, and identifying disease biomarkers or genes. In this network nodes represent genes and edges link two genes to show to what degree this pair of genes is co-expressed across several samples. The edges are based on correlation coefficients between each gene pair, where higher correlation means a higher probability of existing co-functionality between them. Recently, several computational approaches have been developed for the co-expression analysis and networks such as the Average Specific Connection, Differential Coexpression profile (DCp) [102], Differential Coexpression enrichment (DCe) [102], Differential Correlation in Expression for meta-module Recovery (DICER) [103], DiffCoEx [104], Log Ratio of Connections [105], and the Weighted Gene Co-Expression Network Analysis (WGCNA) [106]. For example, the popular R-package WGCNA has been used to analyze high dimensional data, distinct biological states and complex diseases. This approach identifies candidate genes relevant to a particular process of interest, construction of networks, module identifications, calculation of topological properties and visualizations [106]. The WGCNA

package does not only concentrate on the individual gene expression, but it also focuses on modules of the genes which provide the relationships between modules [107]. In this thesis, we applied the WGCNA package for our co-expression analysis and differential interaction networks analysis.

Chapter 3

Composition of Overlapping Protein-Protein and Protein-Ligand Interfaces

This chapter is based on the following publication:
Ruzianisra Mohamed, Jennifer Degac, and Volkhard Helms. Composition of Overlapping Protein-Protein and Protein-Ligand Interfaces. PloS One. 2015 Oct 30;10 (10):e0140965. doi: 10.1371/journal.pone.0140965 [108].

My contribution was to write the manuscript, designed the research project and analyzed the results together with the co-authors Jennifer Degac and Volkhard Helms. The calculations were performed by me and Jennifer Degac.

Abstract

Protein-protein interactions (PPIs) play a major role in many biological processes and they represent an important class of targets for therapeutic intervention. However, targeting PPIs is challenging because often no convenient natural substrates are available as starting point for small-molecule design. Here, we explored the characteristics of protein interfaces in five non-redundant datasets of 174 protein-protein (PP) complexes, and 161 protein-ligand (PL) complexes from the ABC database, 436 PP complexes, and 196 PL complexes from the PIBASE database and a dataset of 89 PL complexes from the Timbal database. In all cases, the small molecule ligands must bind at the respective PP interface. We observed similar amino acid frequencies in all three datasets. Remarkably, also the characteristics of PP contacts and overlapping PL contacts are highly similar.

Keywords: Atomic contact, protein interface, protein-ligand interaction, protein-protein interaction, and amino acid composition.

3.1 Introduction

Protein-protein interactions (PPIs) play major roles in many biological processes such as bioenergetics, immune response, signal transduction, structural organization, and apoptosis [10,109]. Recently, PPIs also became a promising new target for therapeutic intervention. Unlike established pharmaceutical efforts that are directed, for example, at enzymes, G-protein coupled receptors (GPCR), or ion-channels, PPIs are challenging subjects because there are usually no convenient natural substrates that can be exploited as starting points for small-molecule design. Moreover, the lack of information about particular interface residues determining the affinities and specificities at such interfaces makes it quite hard to design compounds that are capable of interfering with PPIs. Hence, there is a strong need to characterize the properties of protein interfaces that may also bind small-molecule ligands and the underlying molecular principles of contacts they are involved in.

The Protein Data Bank (PDB) [24] is the primary resource for elucidating the diversity of atomic contacts in protein-protein (PP) and protein-ligand (PL) interactions. Many statistical analyses of molecular interactions have been done based on this resource [1, 4–6]. Furthermore, some secondary databases that are derived from the PDB have been created to assist the integrated research on PP and PL interactions. Examples for this are the Timbal database (http://mordred.bioc.cam.ac.uk/timbal) which stores data of small molecules modulating protein–protein complexes [33], the Mother of All Database (MOAD) which contains data on ligand-protein binding (http://bindingmoad.org) [8-9], the 2P2I database of structures of PP complexes with known small molecule inhibitors (http://2p2idb.cnrs-mrs.fr) [36], the Analysing Biomolecular Contacts (ABC) database (http://service.bioinformatik.uni-saarland.de/ABCSquareWeb/) [31], and the database of structurally defined protein interfaces named PIBASE (http://pibase.janelia.org/pibase2010/queries.html) [32]. One important aim in interface analysis is to identify properties which may distinguish binding residues from the rest of the protein surfaces.

Although protein-protein interfaces are rather large, planar and well packed depending [10,11], some parts of these interfaces termed overlap or bifunctional regions may bind both to small-molecule ligands and to proteins. The remaining regions of the interface which bind only to either protein or ligand are called non-overlap or monofunctional regions. Davis and Sali [113] found that bifunctional regions were enriched in tyrosine and tryptophan residues and depleted from alanine, isoleucine, leucine and valine when compared to monofunctional positions. Walter et al. [114] found for a different dataset that the overlap regions were mostly

found in pockets and some of their surfaces were exposed to the solvent. Koes and Camacho [115] used Small Molecular Inhibitor Starting Points (SMISPs) from PL and PP complexes in the PDB to train statistical classifiers for predicting such SMISPs.

In this study, we analyzed the residue-residue and atomic contact frequencies and propensities of five non-redundant datasets i) 174 protein-protein and ii) 161 protein-ligand complexes from Walter [114], iii) 436 protein-protein and iv) 196 protein-ligand complexes from the PIBASE database [32], and v) a dataset of 89 protein-ligand complexes from the Timbal database [33]. Our main research question was to find out whether small molecule ligands have similar physio-chemical features as protein binding interfaces when they bind at overlapping PP/PL binding interfaces and this was indeed found to be the case.

3.2 Material and Methods

3.2.1 Datasets

Non-redundant datasets from three different databases were used to investigate the composition of protein interfaces. The first pair of datasets consists of 174 PP complexes and 161 PL complexes compiled by Walter et al. [114] from the ABC database [31] (see Tables A and B in Supplementary Information Table 3.1). 25 entries of this PL dataset had been updated in the PDB in the meantime. We changed 22 previous ligand names to the current ligand names in the PDB files and removed 14 PDB files because they contain modified residues that were wrongly recognized as ligands before [114]. As described by Walter et al. [114], these complementary PP and PL datasets fulfill the following criteria: (i) PP: PL pairs represent pairs of complexes, where one protein may bind either a second protein or a small molecule ligand at the same interface, (ii) every pair of the dataset is represented as (Pi1, Pi2): (Pi3, Lj), where Pi1, Pi2 and Pi3 are three proteins and Lj is a small molecule ligand, (iii) Pi1 and Pi3 share at least 40% sequence identity, and (iv) the aligned positions in the binding interfaces of Pi1–Pi2 and Pi3 – Lj have at least two residues in common.

The same criteria of (Pi1, Pi2):(Pi3, Lj) pairs of PP and PL complexes from Walter et al., were then applied to the datasets of PP and PL complexes from the PIBASE database [32]. To avoid redundancy among these complexes, we clustered the PL complexes using the CD-Hit program [116,117] with the same sequence identity cut-off of 40%. Within a cluster, we selected the representative PP:PL pair with the highest identity score of the interface residues. Additionally, we discarded clusters which contained only sequences with fewer than 40 amino acids. The final pair of datasets comprises 436 PP complexes (Table C in Supplementary

Information Table 3.1) and 196 PL complexes (Table D in Supplementary Information Table 3.1).

Interactions where both interacting chains have > 90% sequence identity are defined as homodimer complexes and the remainder as heterodimer complexes. As a result, the PP complexes from the ABC dataset comprised 94 homodimer complexes and 80 heterodimer complexes (see Tables A and B in Supplementary Information Table 3.2). The PP complexes from the PIBASE dataset were grouped into 335 homodimer complexes and 101 heterodimer complexes (see Tables A and B in Supplementary Information Table 3.3).

The fifth dataset was extracted from the table of PDB entries in the Timbal database (see Table E in Supplementary Information Table 3.1). First, the 1695 entries in the current version of the Timbal database were filtered by removing complexes containing ligands that are annotated to act as stabilizers. Then, the CD-Hit program was applied to remove redundancy among the protein chains of the complexes with the sequence identity cut-off of 40%. We also eliminated clusters of proteins with fewer than 40 amino acids. This gave a final dataset of 89 protein-small molecule complexes.

Data from the ABC, PIBASE, and Timbal databases was retrieved by using MySQL queries, Java, Biojava [118] and analyzed with the R software (http://www.R-project.org).

3.2.2 Surface and Interface Residues

The solvent accessible surface area (SASA) was calculated using the NACCESS program [119]. As surface residues we considered those residues with a SASA value larger than zero. Labeled as interface residues were those residues that are within a radius of either 3 Å, 4 Å or 5 Å of any residue of the binding partner. Figure 3.1 shows a schematic diagram how we determined the interface and the remaining surface of PL complexes.

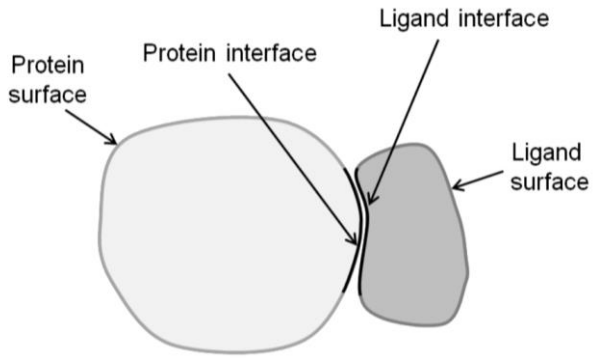

Figure 3.1 Schematic illustration of a PL complex illustrating the interface (black border) and the remaining surface regions. PL, protein-ligand.

3.2.3 Classification of the Amino Acids

The standard classification according to the Eisenberg hydrophobicity scale [120] was used to classify amino acids into four categories: hydrophobic (Ala, Ile, Leu, Met, Phe, Pro, Val), charged (Arg, Asp, Glu, Lys), polar (Cys, Asn, Gln, His, Ser, Thr, Trp, Tyr), and Gly.

3.2.4 Interface Residue Propensities

Residue interface propensities were calculated for the homodimeric and heterodimeric protein-protein complexes of the ABC and PIBASE datasets and for the protein-ligand complexes of the ABC, PIBASE and Timbal datasets. These propensities give a measure of the relative importance of different amino acid residues in the interface, compared with the surface as a whole. The propensities were calculated with the following formula:

$$Interface\ residue\ propensity\ AAj = \left(\frac{\sum interface\ residues\ of\ type\ j}{\sum all\ interface\ residues}\right) / \left(\frac{\sum surface\ residues\ of\ type\ j}{\sum all\ surface\ residues}\right) \qquad (3.1)$$

An interface residue propensity of >1.0 indicates that a residue type occurs more frequently in interfaces than on the protein surface in general.

3.2.5 Contacts between Amino Acids of the Two Proteins

For every PP complex, we counted the observed number of contacts between amino acids of the first protein and amino acids of the second protein. A contact exists between two residues of these proteins if any residue of the first protein is within a distance threshold of 5.0 Å from the other protein. This was represented in a 20 x 20 table. From the 400 observed counts of amino acid pairs in the two datasets of protein-protein complexes, we derived normalized pair frequencies with the following formula:

$$Normalization = \frac{\left(\frac{\sum contacts\ of\ residue\ pair\ XY}{\sum all\ residue\ contacts}\right)}{\left(\frac{\sum observed\ X\ on\ surface}{\sum all\ surface\ residues\ of\ first\ protein}\right)\left(\frac{\sum observed\ Y\ on\ surface}{\sum all\ surface\ residues\ of\ second\ protein}\right)} \qquad (3.2)$$

Here, XY is the number of observed contact pairs between residues X and Y across the interface, X is the count of amino acid X in the first protein and Y is the count of amino acid Y in the second protein.

3.2.6 Atom Contacts in Protein-protein and Protein-ligand Complexes

In protein-protein and protein-ligand complexes, we considered two surface atoms belonging to separate molecules to be in contact and labeled them as interface atoms if the distance between them is less than 5.0 Å. We counted contacts between all pairs of carbon (C), fluorine (F), nitrogen (N), oxygen (O), phosphorus (P), and sulfur (S) atoms resulting in 36 contact pairs. Then, the absolute counts were normalized as follows:

$$Normalization = \frac{\left(\frac{\sum contacts\ of\ atom\ pair\ AB}{\sum all\ atom\ contacts}\right)}{\left(\frac{\sum observed\ A\ on\ surface}{\sum all\ surface\ atom\ of\ ligand\ or\ protein}\right)\left(\frac{\sum observed\ B\ on\ surface}{\sum all\ surface\ atom\ of\ protein}\right)} \quad (3.3)$$

where A is the count of atom type A in the first protein (PP complexes) or protein (PL complexes), B is the count of atom type B in the second protein (PP complexes) or ligand (PL complexes) and, AB is the number of observed contact pairs between atom types A and B across the interface.

According to Higueruelo et al. [121], atom type contacts were grouped into polar and apolar contacts as follows: For protein-protein complexes, apolar contacts exist between C…C, C…S and S…S (not in Cys-Cys bridges). Polar contacts involve the pairs N…O, O…O, N…N, O…S and N…S (from Cys). For protein-ligand complexes, apolar contacts are C…C, and C…S pairs whereas polar contacts are formed by the pairs N…O, O…O, N…N, O…S, N…S, N…F, O…F, and S…F (from Cys).

3.2.7 Calculation of Polarity Ratio and Interface Atom Ratio

The polarity ratio (PR) is a simple measure of the polarity of the interface [122]. It was defined as the ratio of the number of polar atoms N, O, S at the interface to the sum of all C, N, O, S at the interface.

The interface atom ratio (IR) is a measure for the fraction of surface atoms that are located at the interface. It was calculated for the interfaces of protein-protein and protein-ligand complexes. Only the six heavy atom types C, N, O, S, P and F were considered in the calculation. IR is the ratio of the sum of all atoms at the interface to the sum of all atoms at the surface.

3.3 Results and Discussion

PPI interfaces are known to possess particular geometric and physicochemical characteristics, see e.g. [10,123–125]. Comparing these features of protein interfaces to those of overlapping protein-ligand interfaces should aid in targeting protein-protein interaction sites. Here, we used the ABC, PIBASE and Timbal databases as data sources for protein interfaces and surfaces. All three databases are secondary database that are derived from the PDB. However, due to the different way of identifying overlapping PP/PL pairs, the direct overlap between the three non-redundant datasets derived from them is fairly small. We believe that this may have resulted from the clustering with the CD-Hit program that selected different cluster representatives in each case. We found only the following redundant PP complexes 1AB8 (B-A), 1AZZ (A-C), 1BMF (F-B), 1EYS (H-M), 1RQ8 (A-E), 1SGF (G-B) from the ABC dataset and 1AB8 (A-B), 1AZZ (C-A), 1BMF (C-D), 1EYS (M-C), 1RQ8 (E-A), 1SGF (G-Z) from the PIBASE dataset. Furthermore, both datasets share the following lists of PDB IDs with same chain interactions 1DPJ (A-B), 1P0S (H-E), and 2G2U (A-B). Similarly, there are few redundancies between datasets of PL complexes from ABC and PIBASE, namely 1C50 (A-CHI), 1KYN (A-KTP), 1LBC (A-CYZ) and 1M2Z (A-BOG), respectively. There is also one overlapping member between the datasets of PL complexes from PIBASE and Timbal, namely the PDB ID 1AB8 (A-FOK). Figure 3.2 summarizes the workflow of the analysis of the five datasets. The fraction of homodimers and heterodimers in the datasets derived from ABC and PIBASE are 54%: 46% and 77%: 23%, respectively.

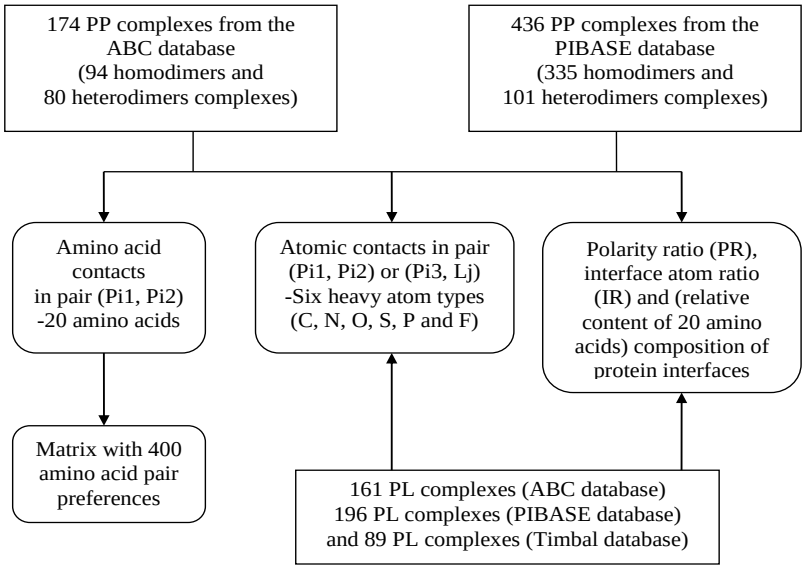

Figure 3.2 Flow chart summarizing the compilation of contacts between amino acids of the first protein (Pi1) and amino acids of the second protein (Pi2), atom contacts in PP and PL complexes, and the calculation of PR and IR.
PP, protein-protein; PL, protein-ligand; PR, polarity ratio; IR, interface atom ratio.

3.3.1 Amino Acid Composition and Protein Interfaces Propensity

Figures 3.3 and 3.4 show the percentage frequencies and propensities of amino acids at the interfaces of homodimeric and heterodimeric PP complexes from the ABC and PIBASE datasets, respectively. Figure 3.5 shows the percentage frequencies and propensities of amino acids at the protein interfaces of the PL complexes from the ABC, PIBASE and Timbal datasets, respectively. Previous studies showed that protein-protein interfaces have unique characteristics that distinguish them from non-interface portions of protein surfaces [123,126,127].

By grouping the amino acids according to the Eisenberg hydrophobicity scale (see methods) we found that, hydrophobic amino acids account for 38.06% (ABC-P1-homo), 38.87% (ABC-P2-homo), 38.81% (PIB-P1-homo) and 38.75% (PIB-P2-homo) at interfaces of homodimeric PP complexes compared to 35.60% (ABC-P1-hetero), 36.11% (ABC-P2-hetero), 37.94% (PIB-P1-hetero) and 36.38% (PIB-P2-hetero) at interfaces of heterodimeric PP complexes (Figures 3.3A and 3.4A). This matches the general finding e.g. of Jones and Thornton who stated that homodimer complexes are more hydrophobic [10].

At interfaces of both homodimeric and heterodimeric PP complexes from the ABC and PIBASE datasets, alanine, valine, and lysine residues are underrepresented with propensities lower than 1.0 (Figures 3B and 4B). One hydrophobic amino acid (leucine), one charged amino acid (lysine) and two polar amino acids (glutamine and threonine) have higher propensities at interfaces of homodimer complexes than at interfaces of heterodimer complexes of the ABC dataset. In the PIBASE dataset, four hydrophobic amino acids (alanine, leucine, proline and valine), one polar amino acid (threonine) and glycine have higher propensities in homodimer complexes than in heterodimer complexes.

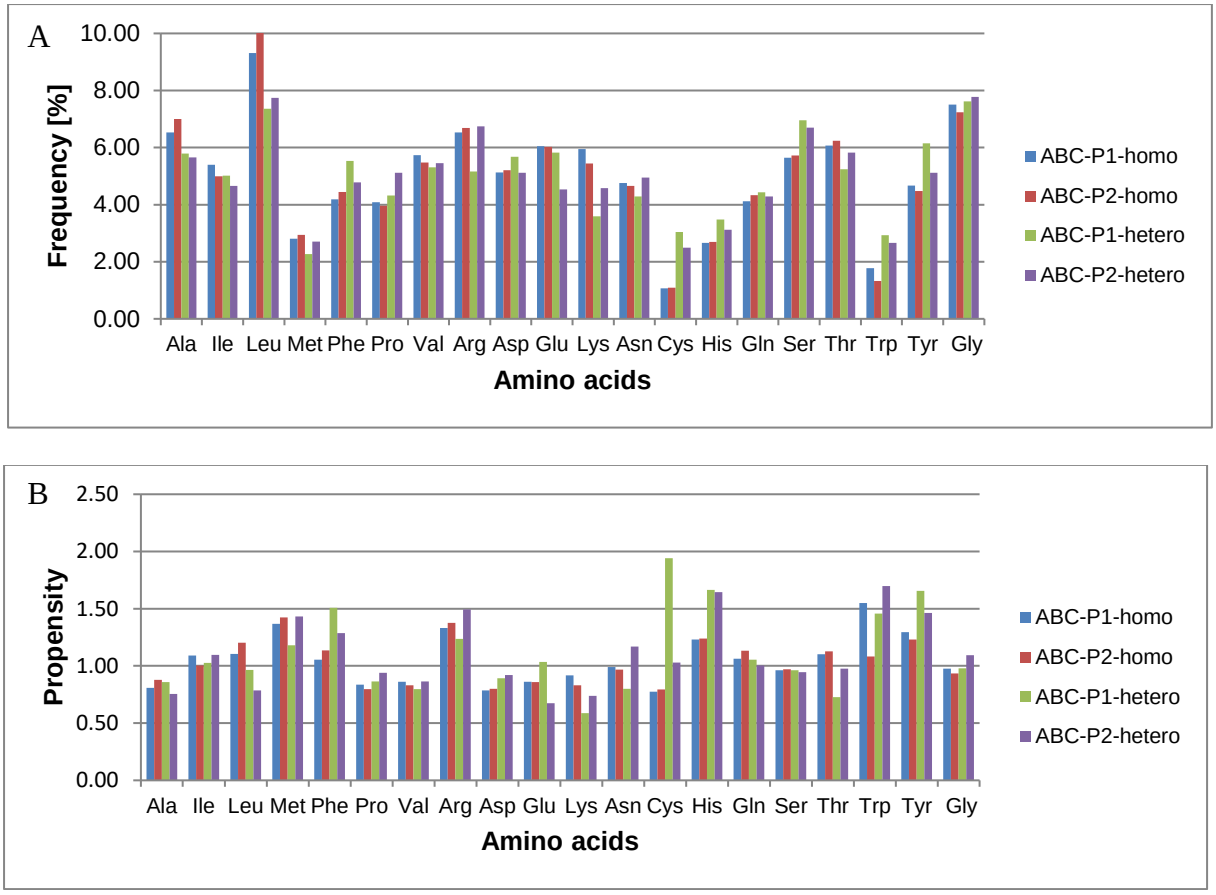

Figure 3.3 Percentage frequencies and propensities of amino acid residues at protein interfaces of PP complexes from the ABC dataset. (A) Percentage frequencies of amino acid residues at protein interfaces. (B) Propensities of amino acid residues at protein interfaces.
PP, protein-protein; ABC, ABC dataset; homo, homodimeric PP interface; hetero, heterodimeric PP interfaces; P1, protein interface of the first protein (Pi1); P2, protein interface of the second protein (Pi2).

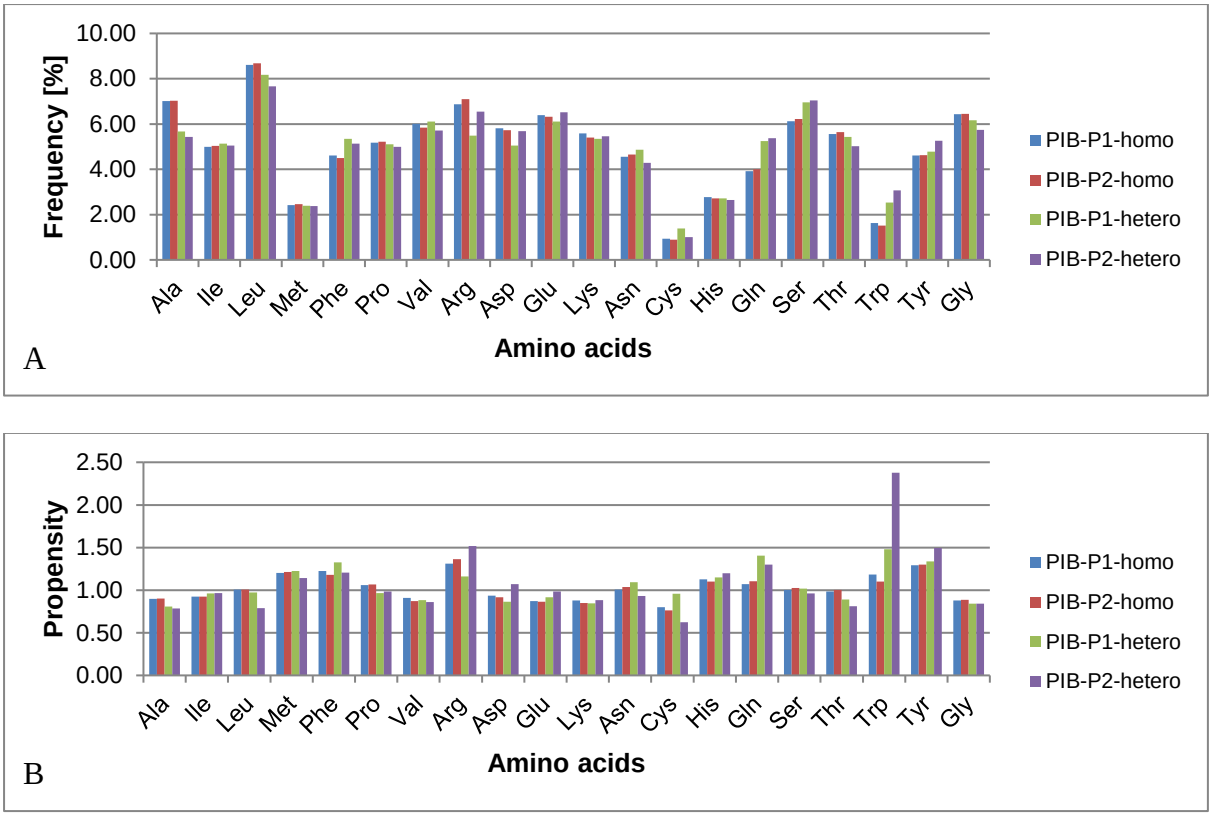

Figure 3.4 Percentage frequencies and propensities of amino acid residues at protein interfaces of PP complexes from the PIBASE dataset. (A) Percentage frequencies of amino acid residues at protein interfaces. (B) Propensities of amino acid residues at protein interfaces.
PP, protein-protein; PIB, PIBASE dataset; homo, homodimeric PP interface; hetero, heterodimeric PP interfaces; P1, protein interface of the first protein (Pi1); P2, protein interface of the second protein (Pi2).

As expected, hydrophobic and polar residues make up the largest portion of protein interfaces. In fact, this is one of the challenges for targeting PPIs with small molecules as the contact surfaces between proteins typically involve many hydrophobic and polar interactions distributed over a large interface with buried area of ~1500 – 3000 $Å^2$ [128]. According to the classification by Eisenberg, the fractions of hydrophobic, polar, charged and glycine residues are 36.95%, 33.38%, 22.11%, 7.56% for the first protein (Pi1), 37.70%, 32.48%, 22.35%, 7.46% for the second protein (Pi2) of the PP complexes from the ABC dataset, 38.60%, 30.93%, 24.09%, 6.38% for the first protein (Pi1), and 38.20%, 31.35%, 24.22%, 6.23% for the second protein (Pi2) of the PP complexes from the PIBASE dataset. Although there are minor differences between the two datasets (slightly more charged and fewer glycine residues in the PIBASE dataset), we found the composition to be overall remarkably similar.

At the interfaces of both homodimeric and heterodimeric PP complexes from the ABC and PIBASE datasets, the frequencies of methionine and tryptophan at protein interfaces are at most 3.07%. However, both amino acids have normalized interface propensities clearly larger than one, suggesting that these residues play important roles and thus occur more frequently at protein interfaces rather than elsewhere on the protein surface. Overall, tryptophan, tyrosine and arginine each have propensities above 1.0 at both protein interfaces of homodimeric and heterodimeric PP complexes from the ABC and PIBASE datasets. This reflects that aromatic amino acids and arginine play important roles in protein interfaces, which is a well-known fact. For example, Bogan and Thorn [129] reported that hotspot regions at protein interfaces are enriched in tryptophan, tyrosine and arginine. Also, Jones, Marin and Thornton [130] found that hydrophobic residues including tryptophan and tyrosine as well as arginine are moderately enriched at protein interfaces compared to the whole surface. Jones and Thornton [10] reported that with the exception of methionine, all hydrophobic residues show a greater preference for the interfaces of homodimers than for those of heterocomplexes. Based on our analysis, only leucine is clearly enriched at homodimer interfaces. Janin, Bahadur and Chakrabarti [125] wrote that relative to the accessible protein surface, the interfaces are depleted in glutamic acid, aspartic acid, and lysine, and enriched in methionine, tyrosine and tryptophan. Our findings are in good agreement with this. In our case, the enriched category also includes phenylalanine, histidine and arginine. The underrepresented category also includes alanine, proline and valine. Talavera et al. [131] provided a rather recent compilation of amino acid frequencies and propensities, separately for homomeric and heterodimeric PP complexes. A possible concern about their work is that they applied a rather generous homology threshold of 80% identity. They found tyrosine, tryptophan, methionine, cysteine, phenylalanine, leucine, valine and isoleucine to be enriched at the interfaces of homo-complexes. In the case of hetero-complexes, cysteine fell out from this list. On the other hand, lysine, asparagine, aspartic acid and glutamic acid were underrepresented in homo-complexes. The same ones plus serine and glycine were found for hetero-complexes.

The distributions of the percentage frequencies and propensities of amino acids at the protein interfaces of the PP datasets derived from ABC and PIBASE were compared with the non-parametrical Friedman test as the datasets do not have a normal distribution. As suggested by the graphical representation of Figures 3.3 and 3.4, the ABC and PIBASE datasets do not differ significantly (percentage frequencies, p-value = 0.99 and propensities, p-value = 0.97).

The fractions of hydrophobic, polar, charged and glycine residues at protein binding interfaces of PL complexes are 34.08%, 36.97%, 20.31%, 8.64% (ABC dataset), 38.25%, 35.39%, 18.12%, 8.24% (PIBASE dataset) and 42.32%, 32.61%, 18.60%, 6.47% (Timbal dataset), see Figure 3.5A. Compared to PP interfaces, the ligand-contacting protein interfaces of the Timbal dataset contain about 5% more hydrophobic residues, and about 5% fewer charged residues. In contrast, the ligand-contacting protein interfaces from the ABC and PIBASE datasets contain 3-4% more polar residues than PP interfaces and 3-4% less charged residues.

In the PL complexes of the ABC dataset, the five amino acids with the highest propensities found at protein interfaces are cysteine (2.20), tryptophan (2.18), histidine (1.75), tyrosine (1.74), and phenylalanine (1.47). In the PL complexes of the PIBASE dataset, the most enriched ones are tryptophan (2.25), tyrosine (1.93), phenylalanine (1.92), histidine (1.89), and methionine (1.66). In the PL complexes of the Timbal dataset, methione has the highest propensity of 1.85, followed by phenylalanine (1.78), tryptophan (1.78), histidine (1.54) and tyrosine (1.53), respectively. In all datasets of PL complexes, tryptophan, phenylalanine, histidine, and tyrosine are found most often at the protein interfaces (Figure 3.5B) complemented by either cysteine (ABC) or methionine (PIBASE, Timbal).

The distributions of percentage frequencies and propensities of amino acids acids at the protein interfaces in the datasets derived from ABC, PIBASE and Timbal did not differ significantly (percentage frequencies, p-value = 0.86 and propensities, p-value = 0.96, Friedman rank sum test).

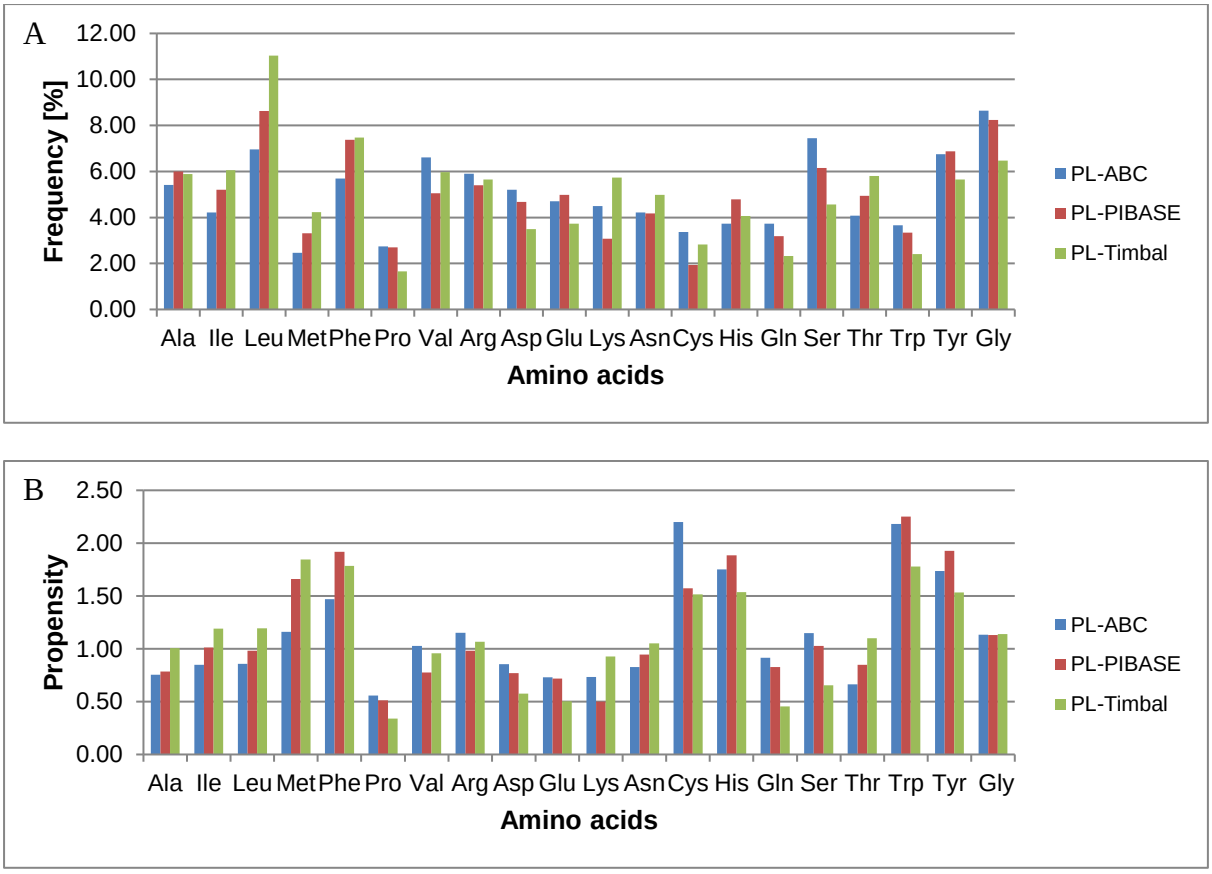

Figure 3.5 Percentage frequencies and propensities of amino acids residues at protein interfaces of PL complexes from the ABC, PIBASE and Timbal datasets. (A) Percentage frequencies of amino acids residues at protein interfaces. (B) Propensities of amino acids residues at protein interfaces. PL, protein-ligand; PL-ABC, PL complexes from the ABC dataset; PL-PIBASE, PL complexes from the PIBASE dataset; PL-Timbal, PL complexes from the Timbal dataset.

3.3.2 Amino Acid Contacts

The propensities of amino acid contacts in PP complexes between amino acids of the first protein (Pi1) and amino acids of the second protein (Pi2) were obtained by counting the absolute number of contacts and normalizing this number against the appearance probability of the two involved residues at the surface. In Figures 3.6 and 3.7, the propensity values were log2 transformed to ensure a balanced view of over- and under-representation. Contacts with high propensities were observed among residues pairs of different polarity types. In PP complexes from the ABC dataset, the five most over-represented interactions were found between the pairs of tryptophan (6.32), cysteine (4.66), phenylalanine (3.61) and histidine (3.50) as well as between tryptophan and phenylalanine (3.36), see Figure 3.6. In PP complexes from the PIBASE dataset, the five most over-represented interactions were pairs of tryptophan (7.50), methionine (4.34), phenylalanine (3.96), tyrosine (3.57), and cysteine (3.43), see Figure 3.7. These results are consistent with previous studies of protein-protein interfaces that reported

an enrichment of contacts between cysteine, hydrophobic contacts and aromatic contacts [123, 126, 132-134]. Further studies noticed that besides disulfide bonds and hydrophobic interactions, also salt-bridges contribute to stabilizing protein-protein interactions [126,132–134]. In our analysis, contacts between lysine and negatively charged amino acids (Asp, Glu) are only mildly enriched (propensity 1.23 on average), whereas those between arginine and either Asp or Glu are about two-fold enriched (2.06), see Tables A and B in S5 File, what reflects the enriched of arginine at protein interfaces. The propensities of amino acid contacts between amino acids of the first protein (Pi1) and amino acids of the second protein (Pi2) in PP complexes between datasets from the ABC and the PIBASE did not differ statistically significantly (p-value = 0.76, Wilcoxon signed rank test).

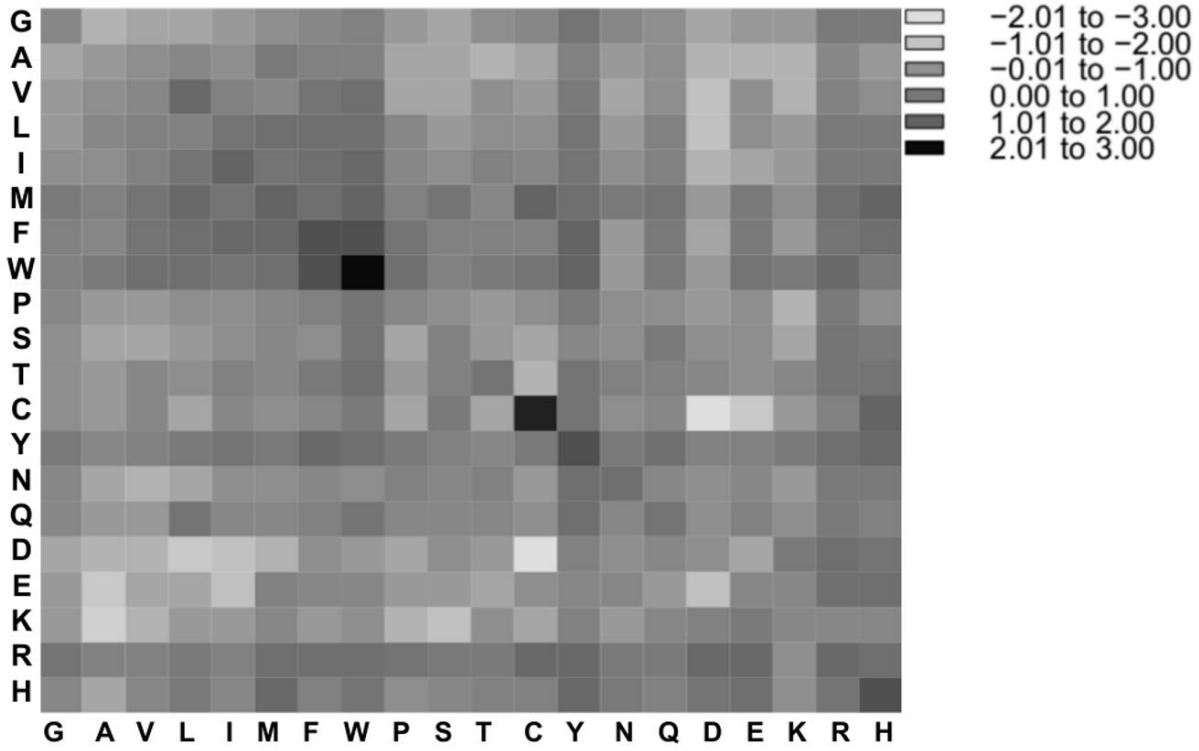

Figure 3.6 Amino acid pairing propensities (in $\log_2$-format) for interfaces of PP complexes from the ABC dataset.
PP, protein-protein.

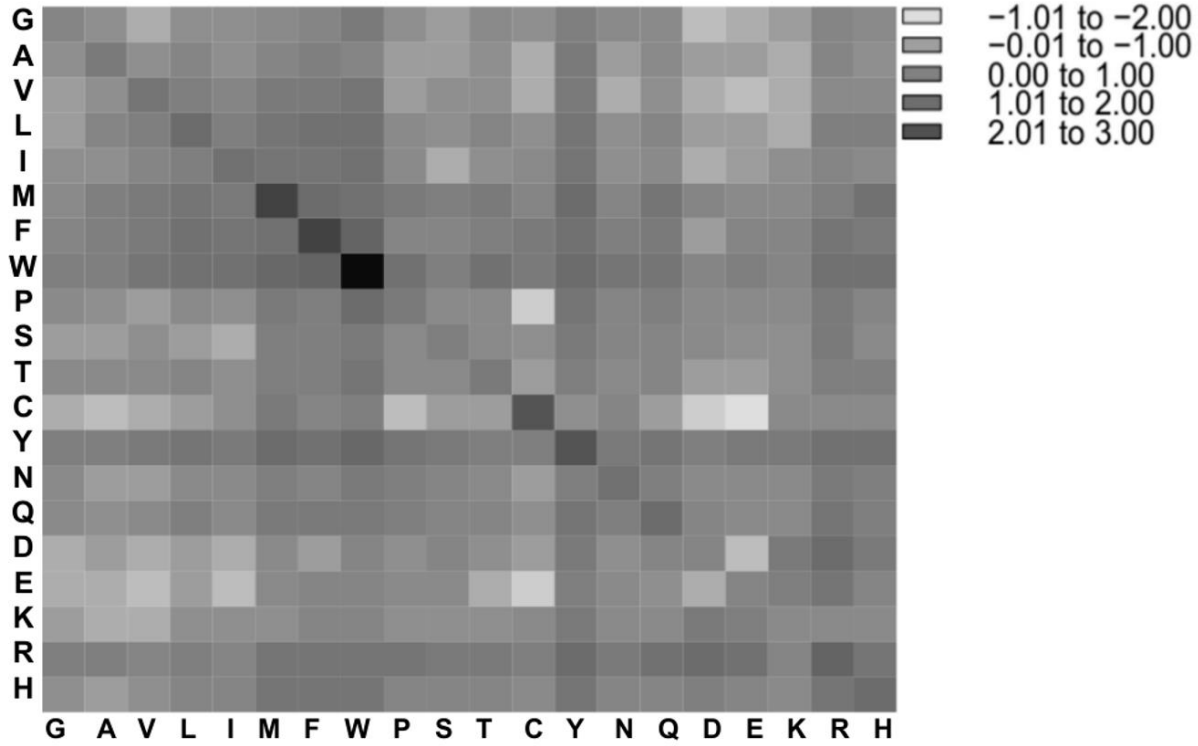

Figure 3.7 Amino acid pairing propensities (in log_2-format) for interfaces of PP complexes from the PIBASE dataset.
PP, protein-protein.

Based on the counts of amino acids, we computed the average number of amino acid residues at the interfaces of the two proteins Pi1 and Pi2 of PP complexes and the Pi3 protein of PL complexes using three different atom distances (3Å, 4Å and 5Å). At the distance threshold of 3Å, both interfaces at Pi1 and Pi2 contain less than 10 amino acids on average. For thresholds of 4Å and 5Å, the average size of the protein interfaces is 26.22 (ABC dataset) and 38.69 amino acids (PIBASE dataset) (Table 3.1).

Table 3.1 The average number with standard deviation of amino acid residues at the interfaces of PP complexes in the ABC and PIBASE datasets.

	PP complexes			
	ABC dataset		**PIBASE dataset**	
Atom distance	Pi1	Pi2	Pi1	Pi2
3 Å	7.67 ± 6.68	7.49 ± 6.85	9.61 ± 15.62	9.53 ± 15.52
4 Å	27.17 ± 19.18	26.22 ±19.7	31.13 ± 24.53	30.76 ± 24.61
5 Å	34.52 ± 23.31	32.8 ± 24.11	38.69 ± 27.97	38.09 ± 27.94

PP, protein-protein; Pi1, first protein; Pi2, second protein.

Table 3.2 shows the average number of residues at the interfaces of Pi3 in PL complexes from the ABC, PIBASE and Timbal datasets. At the distance threshold of 3 Å, the average size of the interfaces is less than 3 amino acids for all datasets. At 4 Å and 5 Å atom distances, the average sizes of the interfaces are between 6.31 amino acids (ABC dataset) and 13.54 amino acids (Timbal dataset). Although the PL interfaces from the ABC dataset are clearly smaller than those from the PIBASE and Timbal datasets, the average ligand size in the ABC dataset (20.48 atoms without hydrogen atoms) is only slightly smaller than the average ligand size in the Timbal dataset (21.53 atoms) and in the PIBASE dataset (21.42 atoms), respectively.

Table 3.2 The average number with standard deviation of amino acid residues at the interfaces of PL complexes in the ABC, PIBASE and Timbal datasets.

	PL complexes		
Atom distance	**ABC dataset**	**PIBASE dataset**	**Timbal dataset**
3 Å	1.64 ± 1.93	2.58 ± 2.08	2.54 ± 2.52
4 Å	6.31 ± 4.66	10.04 ± 4.39	9.99 ± 6.32
5 Å	8.84 ± 5.79	13.43 ± 5.63	13.54 ± 8.06

PL, protein-ligand.

3.3.3 Atomic Contacts in Protein-protein and Protein-ligand Complexes

In this section, we analyzed the atomic contacts in the datasets of PP and PL complexes. For atom pairs between the first and second proteins (Pi1–Pi2) in PP complexes and between protein and ligand (Pi3 – Lj) in PL complexes, we counted contacts of less or equal to 5 Å between six types of heavy atoms, namely carbon (C), flourine (F), nitrogen (N), oxygen (O), phosphorus (P) and sulfur (S). This resulted in 36 atomic pair contacts. Table 3.3 lists the appearance frequency of these 36 atomic contact types in PP and PL complexes from the ABC, PIBASE and Timbal datasets. In all datasets, the most frequent contacts are C…C (> 41%), O…C (> 10%), C…O (> 8%), and C…N (>7%).

Chen and Kurgan [135] previously characterized the binding interfaces of proteins with small molecules, irrespective of whether they also bind to other proteins. As expected, interactions with organic molecules are dominated by van der Waals contacts, hydrogen bonds, and covalent contacts, whereas those with charged species also involve electrostatic interactions. Hakulinen et al. [136] argued that small molecules frequently contact phenylalinine, histidine, tyrosine and tryptophan residues of proteins because their aromatic ring carbons prefer other aromatic carbons. Both findings match well with the results of this analysis. The atomic contacts in PP complexes of the ABC and PIBASE datasets did not differ

significantly (p-value = 0.76, Wilcoxon signed rank test). Also the frequencies of the atomic contacts between the PL complexes of the ABC, PIBASE and Timbal datasets did not differ significantly (p-value = 0.11, Friedman rank sum test).

Tables 3.4 and 3.5 list the percentage frequencies and normalized propensities of apolar, polar and other atomic contacts in PP complexes and PL complexes, respectively. The content of apolar contacts (45.52% for the PP complexes in the ABC dataset and 45.25% for the PIBASE dataset) and of polar contacts (13.85% vs 13.70%) is highly similar between the two PP datasets. In contrast, the PL complexes of the PIBASE dataset (46.45%) contained more apolar contacts than the Timbal dataset (44.84%) and the ABC dataset (43.04%). Concerning polar contacts in PL complexes, the Timbal dataset (14.71%) and the ABC dataset (14.48%) contain more such contacts than the PIBASE dataset (12.95%). Overall, the differences of the normalized propensities seem minor, among the PP and PL datasets, as well as between PP and PL datasets, which agrees with the findings of [135]. In all datasets, C-C contacts are slightly overrepresented (1.04 to 1.11 times the randomly expected number of contacts). N-N contacts are always more frequent (1.07 to 1.32) than O-O contacts (0.70 to 0.95).

Table 3.3 The percentage frequencies of the 36 atomic contact types in PP and PL complexes.

		PP complexes		PL complexes		
		ABC dataset	PIBASE dataset	ABC dataset	PIBASE dataset	Timbal dataset
Atom1[a]	Atom2[b]	%	%	%	%	%
C	C	44.08	44.20	41.24	44.57	43.49
C	N	10.82	10.95	8.50	9.03	7.05
C	O	10.42	10.68	12.63	11.85	8.88
C	S	0.85	0.57	0.90	0.88	1.00
C	P	0.00	0.00	0.00	0.00	0.00
C	F	0.00	0.00	0.00	0.00	0.00
N	C	8.29	8.12	5.81	5.16	5.51
N	N	2.76	2.70	1.58	1.25	1.30
N	O	2.85	2.83	2.18	1.77	1.57
N	S	0.14	0.08	0.17	0.09	0.13
N	P	0.00	0.00	0.00	0.00	0.00
N	F	0.00	0.00	0.00	0.00	0.00
O	C	10.93	11.17	13.87	12.78	15.03
O	N	4.48	4.40	4.89	4.38	6.58
O	O	3.37	3.45	4.79	4.43	4.60
O	S	0.17	0.14	0.18	0.25	0.22
O	P	0.00	0.00	0.00	0.00	0.00
O	F	0.00	0.00	0.00	0.00	0.00
S	C	0.56	0.47	0.90	1.00	0.35
S	N	0.14	0.10	0.26	0.26	0.08
S	O	0.11	0.12	0.31	0.23	0.04
S	S	0.03	0.02	0.04	0.04	0.01
S	P	0.00	0.00	0.00	0.00	0.00
S	F	0.00	0.00	0.00	0.00	0.00
P	C	0.00	0.00	0.59	0.45	1.64
P	N	0.00	0.00	0.50	0.28	1.18
P	O	0.00	0.00	0.19	0.22	0.40
P	S	0.00	0.00	0.02	0.01	0.00
P	P	0.00	0.00	0.00	0.00	0.00
P	F	0.00	0.00	0.00	0.00	0.00
F	C	0.00	0.00	0.32	0.77	0.75
F	N	0.00	0.00	0.04	0.11	0.09
F	O	0.00	0.00	0.08	0.15	0.10
F	S	0.00	0.00	0.01	0.04	0.00
F	P	0.00	0.00	0.00	0.00	0.00
F	F	0.00	0.00	0.00	0.00	0.00
Total		**100.00**	**100.00**	**100.00**	**100.00**	**100.00**

PP, protein-protein; PL, protein-ligand.

[a]For PP complexes, atom1 belongs to the first protein and for PL complexes, atom1 belongs to the protein.

[b]For PP complexes, atom2 belongs to the second protein and for PL complexes, atom2 belongs to the ligand.

Table 3.4 Percentage frequencies (with normalized propensity values in parentheses) of apolar, polar and other atomic contacts of PP complexes from the ABC and PIBASE datasets.

		PP complexes	
		ABC dataset	**PIBASE dataset**
Apolar contacts:	C…C	44.08 (1.10)	44.20 (1.10)
	C…S	0.85 (2.53)	0.57 (1.91)
	S...C	0.56 (1.82)	0.47 (1.63)
	S…S (not in Cys-Cys bridge)	0.03 (10.94)	0.01 (6.60)
	Total	**45.52**	**45.25**
Polar contacts:	N…O	2.85 (0.91)	2.83 (0.90)
	O…N	4.48 (1.40)	4.40 (1.40)
	O…O	3.37 (0.70)	3.45 (0.73)
	N…N	2.76 (1.31)	2.70 (1.29)
	O…S	0.16 (1.40)	0.14 (1.38)
	S…O	0.11 (1.05)	0.12 (1.24)
	N…S (from Cys)	0.05 (0.68)	0.02 (0.35)
	S…N (from Cys)	0.07 (0.93)	0.03 (0.50)
	Total	**13.85**	**13.7**
Other contacts:	C…N	10.83 (1.19)	10.95 (1.20)
	N…C	8.28 (0.91)	8.12 (0.88)
	C…O	10.42 (0.76)	10.68 (0.78)
	O…C	10.93 (0.78)	11.17 (0.81)
	N…S (S not from Cys)	0.09 (1.14)	0.06 (0.86)
	S…N (S not from Cys)	0.08 (1.08)	0.07 (1.05)
	S…S (in Cys-Cys bridge)	0.002 (0.84)	0.003 (1.39)
	C…P/P…C	0.00 (0.00)	0.00 (0.00)
	C…F/F…C	0.00 (0.00)	0.00 (0.00)
	N…P/P…N	0.00 (0.00)	0.00 (0.00)
	N…F/F…N	0.00 (0.00)	0.00 (0.00)
	O…P/P…O	0.00 (0.00)	0.00 (0.00)
	O…F/F…O	0.00 (0.00)	0.00 (0.00)
	S…P/P…S	0.00 (0.00)	0.00 (0.00)
	S…F/F…S	0.00 (0.00)	0.00 (0.00)
	P…P	0.00 (0.00)	0.00 (0.00)
	P…F/F…P	0.00 (0.00)	0.00 (0.00)
	F…F	0.00 (0.00)	0.00 (0.00)
	Total	**40.63**	**41.05**
	Grand Total	**100**	**100**

PP, protein-protein.

Table 3.5 Percentage frequencies (with normalized propensity values in parentheses) of apolar, polar, and other atomic contacts of PL complexes from the ABC, PIBASE, and Timbal datasets.

		PL complexes		
		ABC dataset	**PIBASE dataset**	**Timbal dataset**
Apolar contacts:	C…C	41.24 (1.04)	44.57 (1.04)	43.49 (1.11)
	C…S	0.90 (2.86)	0.88 (2.82)	1.00 (2.60)
	S…C	0.90 (1.26)	1.00 (1.14)	0.35 (0.99)
	Total	**43.04**	**46.45**	**44.84**
Polar contacts:	N...O	2.18 (1.20)	1.77 (1.02)	1.57 (0.94)
	O…N	4.89 (1.31)	4.38 (1.39)	6.58 (1.68)
	O…O	4.79 (0.85)	4.43 (0.95)	4.6 (0.80)
	N…N	1.58 (1.32)	1.25 (1.07)	1.30 (1.14)
	O…S	0.18 (1.45)	0.25 (2.50)	0.22 (1.33)
	S…O	0.31 (1.22)	0.23 (0.77)	0.04 (0.36)
	N…S	0.17 (4.07)	0.09 (1.24)	0.13 (2.75)
	S…N	0.26 (1.59)	0.26 (1.30)	0.08 (0.97)
	N…F	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
	F…N	0.04 (0.56)	0.11 (0.46)	0.09 (0.88)
	O…F	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
	F…O	0.08 (0.80)	0.15 (0.43)	0.10 (0.70)
	S…F (S from Cys)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
	F…S (S from Cys)	0.00 (0.00)	0.02 (3.21)	0.00 (0.00)
	Total	**14.48**	**12.95**	**14.71**
Other contacts:	C…N	8.50 (0.93)	9.03 (0.91)	7.05 (0.77)
	N…C	5.81 (1.12)	5.16 (1.03)	5.51 (1.13)
	C…O	12.63 (0.91)	11.85 (0.81)	8.88 (0.66)
	O…C	13.87 (0.86)	12.78 (0.94)	15.03 (0.90)
	C…P	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
	P…C	0.59 (0.57)	0.45 (0.94)	1.64 (1.07)
	C…F	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
	F…C	0.32 (1.14)	0.77 (0.77)	0.75 (1.77)
	N…P	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
	P…N	0.50 (2.08)	0.28 (2.52)	1.18 (3.30)
	O…P	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
	P…O	0.19 (0.51)	0.22 (1.33)	0.40 (0.77)
	S…S	0.04 (6.30)	0.04 (6.34)	0.01 (2.09)
	S…P	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
	P…S	0.02 (2.18)	0.01 (2.90)	0.00 (0.00)
	S…F (S not from Cys)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
	F…S (S not from Cys)	0.01 (5.40)	0.01 (1.84)	0.00 (0.00)
	P…P, P…F/F…P and F...F	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
	Total	**42.48**	**40.6**	**40.45**
	Grand Total	**100**	**100**	**100**

3.3.4 Polarity Ratio and Interface Atom Ratio

Then, we analyzed the polarity ratio (PR), namely the fraction of polar N, O, S atoms at the interface areas of both PP and PL complexes. The interface atom ratio (IR) indicates the fraction of surface atoms that are involved in protein contacts at the interface. As mention before, the interface areas were defined as those residues that are closer than 3 Å (or 4 Å and 5 Å) to at least one residue from the binding partner. Both IR and PR were computed for the datasets of PP and PL complexes from the ABC, PIBASE, and Timbal datasets.

At 3 Å distance threshold, the differences in IR and PR ratios are not representative because only the shortest-distance contacts are considered. For example, when a 3 Å cut-off is used, most carbon atoms are not considered as part of the interfaces as this short distance is shorter than twice the van der Waals radius of carbon (1.7 Å) [137]. Table 3.6 shows that, as expected, only very small differences were observed when computing PR and IR of PP complexes between the first protein (Pi1) and the second protein (Pi2), as both of them exhibit similar characteristics at binding interfaces. For the larger cut-off distances (4 Å and 5 Å), the polarity ratio (PR) decreases quickly because now all carbon atoms at the surface are included. On the other hand, the interface atom ratio (IR) of 8.0% (4 Å) and 14.0% (5 Å) shows that, expectedly, only a small fraction of the protein surface atoms are included in the interface.

Table 3.6 Interface atom ratio (IR) and polarity ratio (PR) (with standard deviations in parentheses) for interfaces of PP complexes from the ABC and PIBASE datasets.

		PP complexes from the ABC dataset			PP complexes from the PIBASE dataset		
		Atom distance			Atom distance		
		3 Å	4 Å	5 Å	3 Å	4 Å	5 Å
IR							
	Pi1	0.01 (±0.01)	0.08 (±0.05)	0.14 (±0.08)	0.01 (±0.07)	0.08 (±0.08)	0.13 (±0.09)
	Pi2	0.01 (±0.01)	0.09 (±0.08)	0.15 (±0.13)	0.01 (±0.07)	0.08 (±0.08)	0.13 (±0.10)
PR							
	Pi1	0.87 (±0.22)	0.38 (±0.07)	0.34 (±0.06)	0.72 (±0.20)	0.37 (±0.06)	0.34 (±0.05)
	Pi2	0.85 (±0.22)	0.37 (±0.06)	0.34 (±0.05)	0.71 (±0.20)	0.37 (±0.06)	0.34 (±0.05)

IR, interface atom ratio; PR, polarity ratio; PP, protein-protein; Pi1, first protein; Pi2, second protein.

Table 3.7 lists the IR and PR ratios of 161 PL complexes from the ABC dataset, 196 PL complexes from the PIBASE dataset, and 89 PL complexes from the Timbal dataset. At the distance threshold of 3 Å, almost no ligands atoms are considered as interfacial atoms whereas the opposite is the case for 5 Å where 93% (PIBASE) and 94% (Timbal) of the ligand atoms are considered as interfacial atoms compared to 78% for ABC. This is suggesting that the

PIBASE and Timbal ligands bind more flat on the protein surfaces and/or bind deeper into pockets on the protein surface than the ABC ligands. Finally, the polarity ratios of the proteins in the PL dataset are comparable to the proteins in the PP dataset.

Table 3.7 Interface atom ratio (IR) and polarity ratio (PR) (with standard deviations in parentheses) for interfaces of PL complexes from the ABC, PIBASE, and Timbal datasets.

		PL complexes from the ABC dataset			PL complexes from the PIBASE dataset			PL complexes from the Timbal dataset		
		Atom distance			Atom distance			Atom distance		
		3 Å	4 Å	5 Å	3 Å	4 Å	5 Å	3 Å	4 Å	5 Å
IR	Pi3	0.002 (±0.004)	0.01 (±0.02)	0.03 (±0.03)	0.003 (±0.003)	0.02 (±0.01)	0.03 (±0.02)	0.003 (±0.003)	0.02 (±0.02)	0.03 (±0.03)
	Lj	0.11 (±0.14)	0.55 (±0.27)	0.78 (±0.25)	0.13 (±0.17)	0.74 (±0.25)	0.93 (±0.20)	0.16 (±0.14)	0.75 (±0.21)	0.94 (±0.19)
PR	Pi3	0.83 (±0.45)	0.38 (±0.20)	0.35 (±0.14)	0.85 (±0.36)	0.38 (±0.14)	0.34 (±0.12)	0.86 (±0.38)	0.36 (±0.16)	0.32 (±0.12)
	Lj	0.76 (±0.45)	0.38 (±0.23)	0.35 (±0.18)	0.79 (±0.38)	0.33 (±0.17)	0.31 (±0.18)	0.84 (±0.41)	0.38 (±0.25)	0.36 (±0.20)

IR, interface atom ratio; PR, polarity ratio; PL, protein-ligand; Pi3, protein; Lj, ligand.

3.4 Conclusions

In this study, we characterized the residue and atom composition of overlapping protein-protein and protein-ligand interfaces from the ABC and PIBASE databases and compared these to a dataset derived from the Timbal database. According to the statistics, both interface types have, in general, a very similar composition. Among the three datasets of PL complexes, the protein interfaces of the Timbal dataset contain more hydrophobic residues and fewer polar residues than the two other datasets. The ligands in the PIBASE and Timbal datasets bind more flat on the protein surfaces or bind deeper into pockets on the protein surface than ABC ligands. Depending on the respective application in a ligand design project, researchers may consider to bias their principal dataset in one or the other direction. Selecting the appropriate set of reference data may slightly affect the physiochemical characteristics of designed ligands.

Chapter 4
STIM and ORAI Genes, Interactions with Transcription Factors, Differential Gene Expression and Co-expression Analysis on Breast Invasive Carcinoma Dataset

My contribution was to design the research project together with the co-authors Riccha Sethi, Mohamed Hamed, and Volkhard Helms. The analysis of transcription factors and STIM and ORAI genes was done by me and the differential gene expression, and co-expression analysis was done by Riccha Sethi. I and Volkhard Helms wrote the manuscript.

Abstract
Store-operated calcium (Ca^{2+}) entry (SOCE) is ubiquitous mechanism for Ca^{2+} entry in eukaryotic cells, which regulates diverse cellular functions. SOCE is achieved primarily by the gating of the plasma membrane (PM)-localized-channel, ORAI, by the ER-localized Ca^{2+}-sensing protein, STIM. The discovery of transcription factor binding site (TFBS) motifs in specific locations on the STIM and ORAI promoters remains elusive. Moreover, the knowledge of the defects of STIM and ORAI genes expression and/or function linked to disease such as breast cancer is still obscured. Here, we used the HOCOMOCO and EPDnew databases to obtain a set of position weight matrix (PWM) and promoter sequences of STIM and ORAI, respectively, and mapped the possible binding motifs proteins using the FIMO tool. The results were then mapped with the set of transcription factors (TFs) targeting STIM and ORAI gene which were retrieved from the CheA database. We found ten predictive interactions between transcription factors bound to promoter regions of STIM and ORAI genes based on predictions using the STRING, prePPI, and mentha databases. Then, the collection of 63 TFs was used as gene of interest for co-expression and differential expression analysis on breast invasive carcinoma (BRCA) dataset. There, we found ORAI genes to be up-regulated, in contrast STIM1 and STIM2 were down-regulated. Unveiling the predicted transcription factors bound to the promoter regions of STIM and ORAI genes, the regulations of these genes and differential networks properties may suggest putative interactions for experimental studies and allows us to gain knowledge relation with breast cancer.

Keywords: STIM, ORAI, breast invasive carcinoma, promoters, transcription factors.

4.1 Introduction

Calcium (Ca^{2+}) signals control many cellular functions ranging from short-term responses such as muscle contraction, impulse transmission and secretion to longer-term regulation of transcription, growth, and cell division. Store-operated Ca^{2+} entry (SOCE), the main Ca^{2+} influx mechanism in non-excitable cells, typically is activated in response to depletion of endoplasmic reticulum (ER)-Ca^{2+} stores [138,139]. SOCE is controlled by the ER-localized Ca^{2+} sensing proteins, STIM1 and STIM2 [140–142]. Recently, SOCE and the function of Ca^{2+} released-activated Ca^{2+} (CRAC) channels were shown to involve the ORAI1 or CRACM1 [143,144] and two homologs, ORAI2 and ORAI3 [145,146].

In mammals, the stromal interaction molecule (STIM) family has two members STIM1 (Figure 4.1) and STIM2 (Figure 4.2). They share a sequence identity almost 65%, they have diverse properties what results in different functions [147]. Both STIM are single spanning transmembrane (TM) proteins containing an N-terminal EF-hand domain responsible for calcium store sensing and a COOH terminal cytoplasmic domain [141,148–151]. STIM1 is a ubiquitously expressed, protein of 77 kDa and consists of 13 exons, located at chromosome 11, in 11p15.5. The STIM2 gene contains 148 amino acids residues (aa) longer than STIM1 (105-115 kDa) and also consists of 13 exons located at chromosome 4, in 4p15.1 [148].

The human ORAI family includes three members: ORAI1, ORAI2 and ORAI3 (Figure 4.3). They share a high sequence similarity of almost 89% and are broadly expressed with different expression levels depending on the cell type. These ORAI members are localized on the plasma membrane (PM) and consists of four transmembrane domains that are flanked by cytosolic NH_2 and COOH termini [152–155]. Many studies have shown that ORAI1 functions together with STIM1 to initiate CRAC currents [156,157]. On the other hand, previous studies showed that STIM1 and STIM2, are linked to several diseases such as Alzheimers disease (AD), Parkinson's and autoimmune diseases, brain and breast cancer, ischemia and, neurodegeneration diseases [158–164]. Moreover, a few studies found that ORAI1 and ORAI3 are linked to breast cancer migration and metastasis [165–168]. However, only few studies addressed the relation of STIM and ORAI genes to breast cancer.

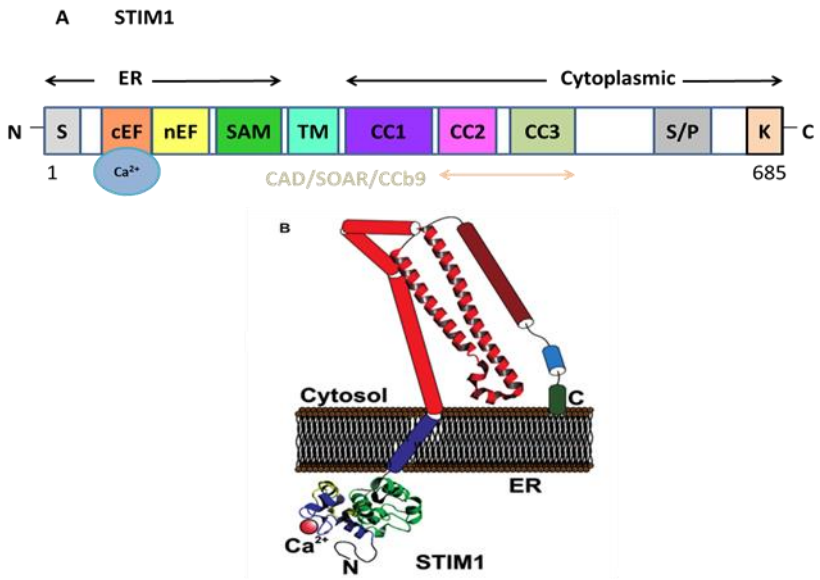

Figure 4.1 Molecular structure of STIM1 (A) STIM1 domain organization. (B) Cartoon depicting a possible model of the STIM1 monomer in the resting state. The figure 4.1(B) was taken from [147].
N, N terminus; Ca^{2+} , calcium; cEF, canonical EF-hand motif; nEF, non-canonical EF-hand motif; SAM, steril alpha motif; CC, coiled-coil domain; CAD, CRAC activation domain also called SOAR or CCb9; ER, endoplasmic reticulum; TM, transmembrane domain; S, serine-rich domain; S/P, serine- and proline-rich domain; K, poly-K and C, C terminus.

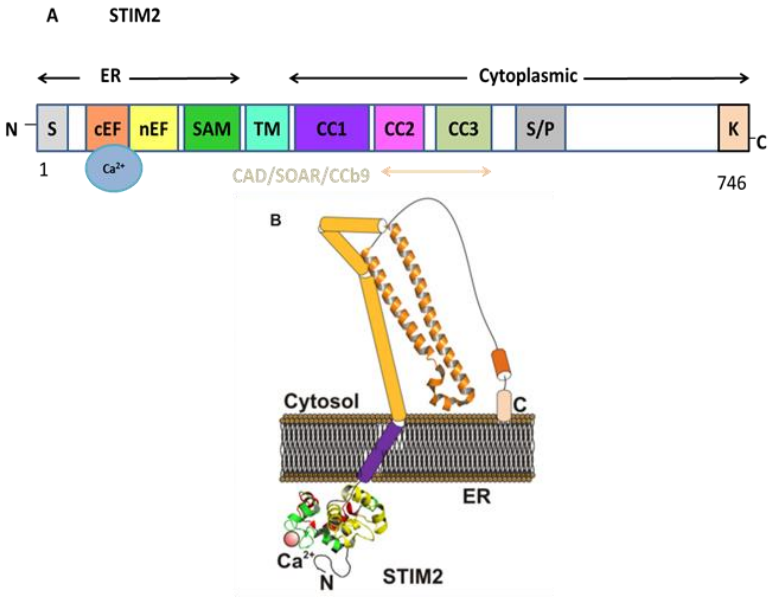

Figure 4.2 Molecular structure of STIM2 (A) STIM2 domain organization. (B) Cartoon depicting a possible model of the STIM2 monomer in the resting state. The figure 4.2(B) was taken from [147].
N, N terminus; Ca^{2+} , calcium; cEF, canonical EF-hand motif; nEF, non-canonical EF-hand motif; SAM, steril alpha motif; CC, coiled-coil domain; CAD, CRAC activation domain also called SOAR or CCb9; ER, endoplasmic reticulum; TM, transmembrane domain; S, serine-rich domain; S/P, serine- and proline-rich domain; K, poly-K and C, C terminus.

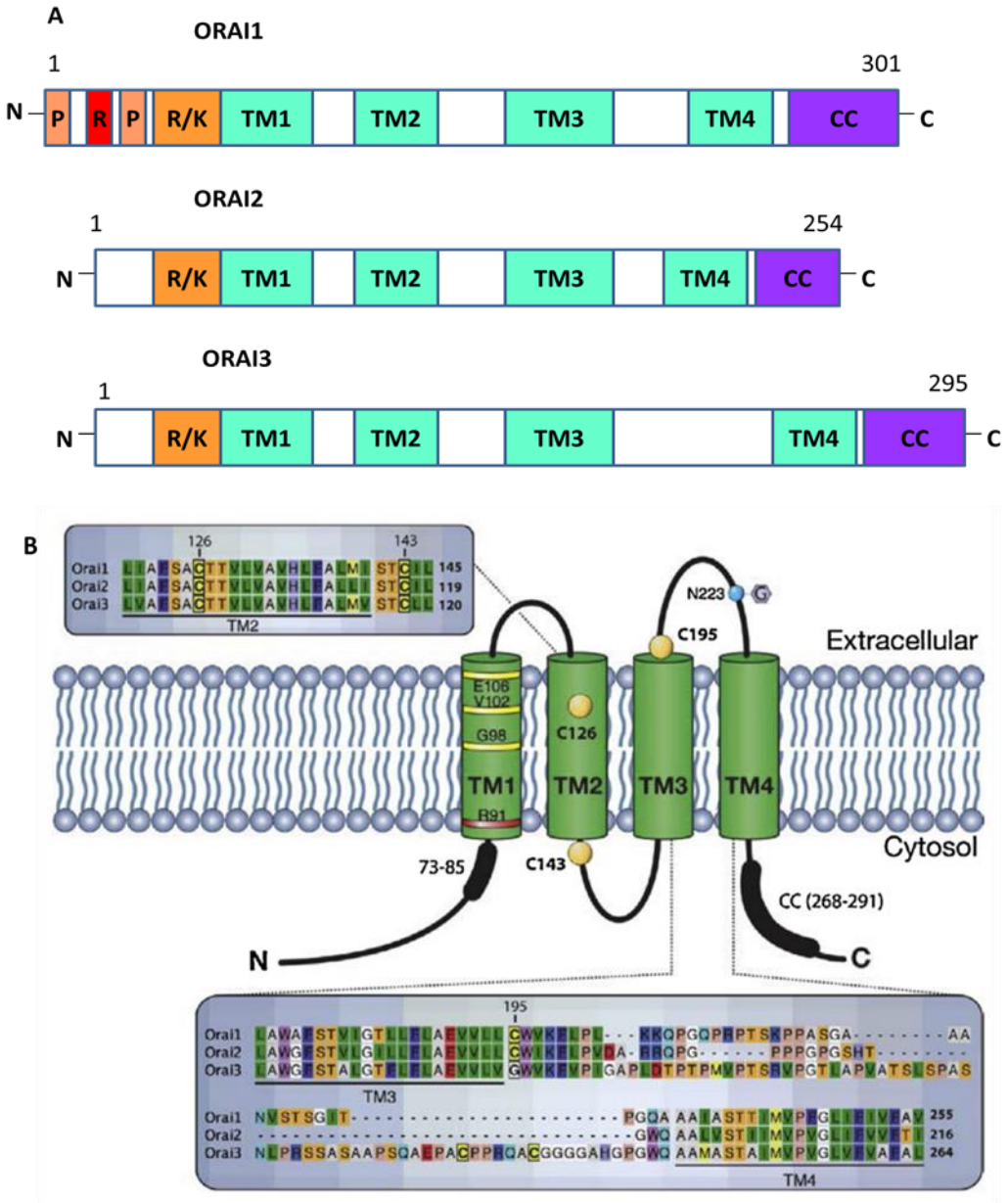

Figure 4.3 Molecular structure of ORAI1, ORAI2, and ORAI3 (A) Domain organization of ORAI genes. (B) Cartoon depicting a possible model of ORAI in the resting state and sequence alignments. The figure 4.3(B) was taken from [169].
N, N terminus; P, Proline-rich; R, Arginine-rich; R/K, Arginine/Lysine-rich; TM, transmembrane domain; CC, coiled-coil domain and C, C terminus.

The binding of transcription factors (TFs) to specific locations in the genome is important for the coordination of transcriptional regulation in cells. Hence, the identification and accurate predictions of TF binding sites (TFBS) throughout genomes is a prerequisite to understanding of gene regulation and their networks [170]. Often, TFBSs are identified by using a set of target promoter sequences to characterize the binding properties for a particular TF that is known or assumed to bind these sequences. The position weight matrix (PWM) are

computed from an alignment of these DNA sequences. Various databases are available which store TF targets such as the TRANSFAC [92] which contains experimentally verified TF target. There are also CheA [171] and Factorbook.org [172] which includes the resources of ChIP-seq and ChIP-chip data from the ENcyclopedia of DNA Element (ENCODE) project [173], HOCOMOCO [91], and JASPAR [90,174] databases. On the other hand, a number of online databases that store physical protein-protein interactions (PPIs) and networks have been developed such as the IntNetDB [175], mentha [176], OPHID [177], prePPI [178], Predictome [179], PIPs [180], and STRING [181,182]. Note that STRING also contains functional protein-protein (PP) interactions in additional to physical PPIs. These databases are the main resources of currently used for integrated research on PP interactions.

Here, by using publicly available databases and computational tools, we perform predict the transcription factors (TFs) responsible for the regulation of human STIM and ORAI genes, and perform differential gene expression and co-expression analysis for breast invasive carcinoma (BRCA) datasets. First, we created a dataset of transcription factors targeting STIM1, STIM2, ORAI1, ORAI2, and ORAI3 genes from the ChEA database and mapped them to the promoter regions of STIM and ORAI genes. Next, we searched for the predicted physical protein-protein interactions of TFs using the STRING, prePPI, and mentha databases. Then, the interacting TFs and bridge proteins were used for differential gene expression and co-expression analysis on datasets of level 3 RNASeqV2 breast invasive carcinoma (BRCA) from the TCGA database. Our main research question was to find the interacting transcription factors bound to the promoter regions of STIM1, STIM2, ORAI1, ORAI2, and ORAI3 and to assess how the regulation by these transcription factors may affect differential expression and co-expression analysis for breast invasive carcinoma datasets. The acquired knowledge should enhance our understanding of STIM and ORAI genes binding sites on promoters, their TF interactions and their regulations in differential expression gene and co-expression analysis.

4.2 Material and Methods

4.2.1 Transcription Factors Targeting STIM and ORAI Genes

We used the ChIP Enrichment Analysis (ChEA) database (amp.pharm.mssm.edu/chea) version 1 [171] to obtain the list of transcription factors (TF) targeted to STIM1, STIM2, and ORAI1-ORAI3 genes. We focused on transcription factors in human (*Homo sapiens*) only. The network of transcription factors targeting STIM1, STIM2, and ORAI1-ORAI3 genes was

visualized using the open-source platform for network analysis and visualization, Cytoscape (http://www.cytoscape.org/) [183].

4.2.2 Transcription Factors Binding Models

The dataset of transcription factor binding site (TFBS) models was obtained from the *Homo sapiens* Comprehensive Model Collection (HOCOMOCO) database version 9 (autosome.ru/HOCOMOCO/) [91]. We downloaded the probability matrices of AD curated collection in MEME text format (HOCOMOCOv9_AD_MEME.txt) which are derived against a uniform nucleotide background.

4.2.3 Sequence of Promoter Regions

A set of promoter regions of the STIM1, STIM2, ORAI1, ORAI2, and ORAI3 genes was downloaded on April 2016 from The Eukaryotic Promoter Database (EPDnew, epd.vital-it.ch/EPDnew_database.php) [89,184]. The EPDnew is a database of species-specific databases of experimentally validated promoters. The searched was filtered to *Homo sapiens*, and the default setting for the range from -499 to 100 bp relative to the transcription start site (TSS).

4.2.4 Motif Over-representation Analysis

A dataset of known binding motifs was obtained from the Find Individual Motif Occurrences (FIMO, meme-suite.org/tools/fimo) tool [185]. The FIMO tool is part of the Motif-Based Sequence Analysis Tools, the MEME Suite version 4.11.2 [186]. This tool searches a database of sequences for occurrences of known motifs and treats the motifs independently. The motif input from the HOCOMOCO database (HOCOMOCOv9_AD_MEME.txt) was uploaded in the "input the motifs" section of the FIMO tool. On the other hand, the promoter sequences of STIM1, STIM2, ORAI1, ORAI2, and ORAI3 were uploaded respectively each time in the "input the sequences" option.

4.2.5 Transcription Factors Predicted by CheA Mapped with the FIMO Results

We mapped 426 models from the HOCOMOCO database to our results of transcription factors targeted to STIM1, STIM2, and ORAI1-ORAI3 genes obtained from the ChEA database. Then the results were mapped according to their base pair (bp) location to each of the corresponding promoter regions.

4.2.6 Prediction of Physical Interactions of Protein-protein Complexes

Three different databases were used to obtain predicted physical interactions of protein-protein complexes, (i) the Search Tool for the Retrieval of Interacting Genes (STRING) database (http://string-db.org) downloaded in April 2016 [181,182], (ii) the prePPI database version 1.2.0 (https://bhapp.c2b2.columbia.edu./PrePPI) [178], and (iii) the mentha database (mentha.uniroma2.it/about.php) downloaded in April 2016 [176]. For the STRING database, the minimum required interaction score was set to the default cut-off of medium confidence (0.4). For the prePPI database, the probability ≥ 0.5 indicates true interaction. Next, we grouped the interactions according to (i) direct interactions which contain no bridge protein and (ii) indirect interactions which require bridge protein/s.

4.2.7 TCGA BRCA Dataset

Publicly accessible RNA-Seq datasets version 2, level 3 of breast invasive carcinoma (BRCA) was downloaded in May 2016 from the online data portal, The Cancer Genome Atlas (TCGA; http://tcga-data.nci.nih.gov) [8]. We downloaded two separate datasets of 113 normal samples (n= 113) and 1102 tumor samples (n=1102), respectively. Each of them contains gene expression profiles. The datasets were pre-processed using python programming language and R software (http://www.R-project.org).

4.2.8 Computing Differential Gene Expression

We applied the widely used Bioconductor package DESeq2 version 1.12.3 [187] to identify genes showing differential expression in the TCGA dataset of breast invasive carcinoma. A false discovery rate (FDR) threshold of 0.1 after Benjamini and Hochberg multiple hypothesis correction was used for statistical significance genes. We used the pheatmap package to create a heatmap of the differentially expressed genes which used the normalized counts as the input.

4.2.9 Weighted Gene Co-expression Network Analysis

The co-expression analysis network was performed using the Weighted Gene Expression Network Analysis (WGCNA) [106] package version 1.51 for R version 3.2.0. We analyzed the results separately for normal and tumor samples. The clustering and principal-components analysis (PCA) was performed to identify outliers. In normal samples, 39 outlier samples were omitted. The cut-height was set to 8 and the soft-thresholding power parameter was set to 7. In tumor samples, 455 outlier samples were removed and the cut-height was set to 12 and the soft-thresholding power parameter was set to 4.

The significant modules obtained from normal and tumor samples were used to identify the differential interactions for all selected genes. The differential interactions were computed as following:

$$Differential\ Interactions = (normal - tumor) + (tumor - normal) \quad (4.1)$$

where (normal-tumor) are the interactions that are (edges) included in modules of the normal network which are not present in the tumor network modules and (tumor-normal) is vice versa. The edge weights greater than 0.02 obtained from the WGCNA were visualized by Cytoscape [183]. Edges with higher score are represented by thicker lines.

4.2.10 Flowchart

Figure 4.4 summarizes the workflow of the analysis of this study.

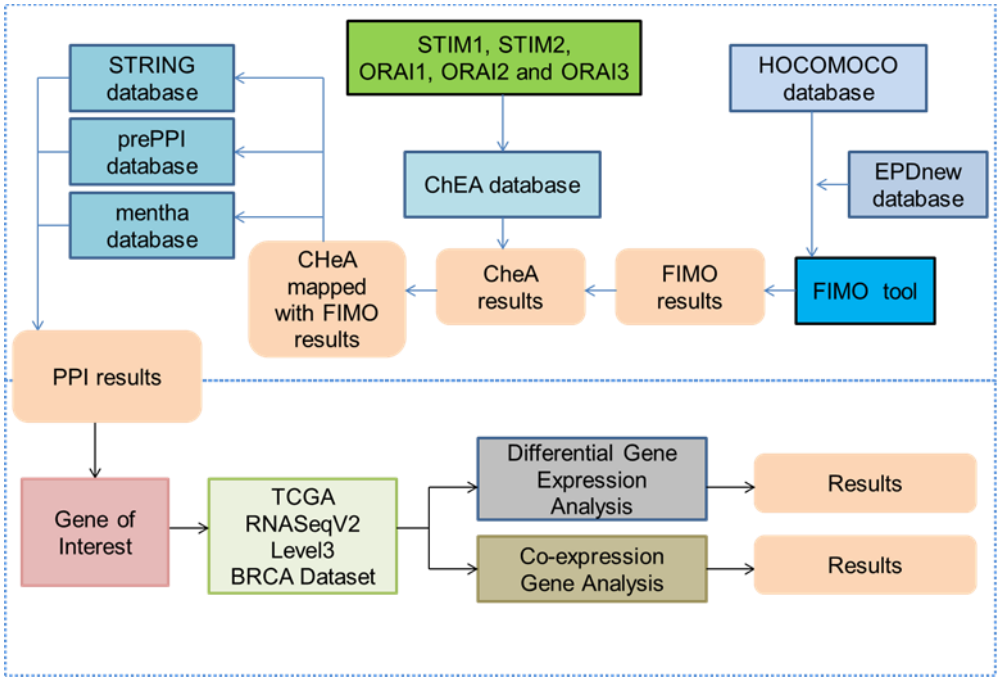

Figure 4.4 Flowchart summarizing the workflow of the analysis.

4.3 Results and Discussion

4.3.1 Transcription Factors Targeting STIM and ORAI Genes

Generally, high-throughput ChIP-seq data was used together with PWMs to identify the putative TF binding sites in the promoter regions of human STIM and ORAI genes. Thus, firstly we created a dataset of TFs targeting STIM and ORAI genes. These were downloaded from the ChEA database which contains gene lists of TFs obtained from ChIP-seq and ChIP-chip studies. We found a set of 23 TFs targeting STIM1, 29 (STIM2), 15 (ORAI1), 11 (ORAI2), and 13 (ORAI3), respectively. Overall, there are 48 non-redundant TFs targeting STIM and ORAI genes (Figure 4.5). Supplementary Information Table 4.1 lists the transcription factors targeted to STIM1, STIM2, and ORAI1-ORAI3, together with Pubmed ID (PMID), technique of experiments and cell types.

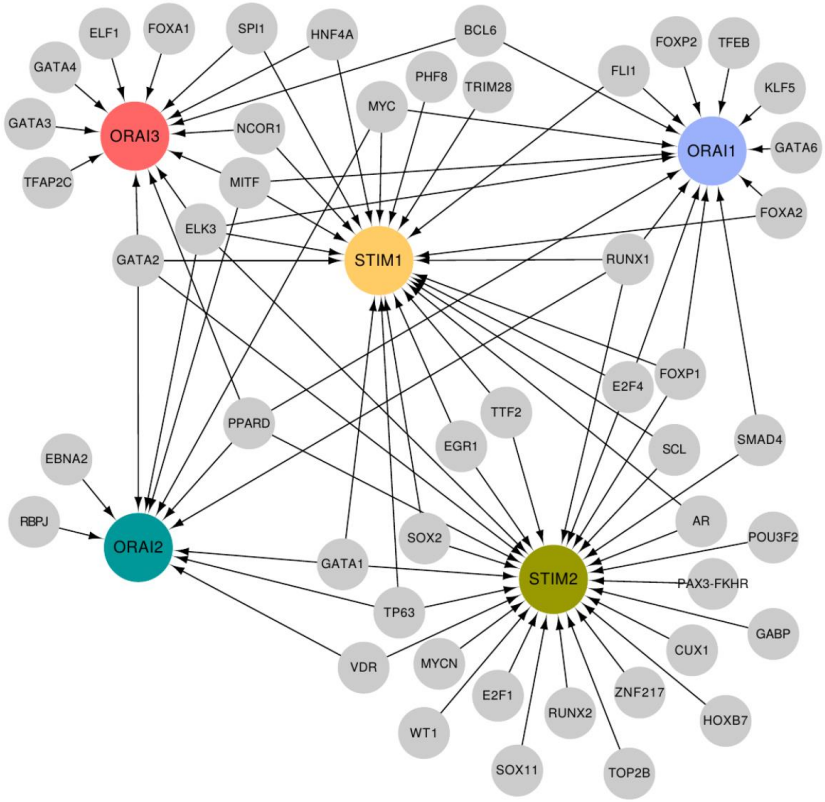

Figure 4.5 Dynamic graphical representations of 48 transcription factors targeting STIM and ORAI genes. The figure drawn using Cytoscape [183].

4.3.2 Promoter Sequences of STIM and ORAI Genes from the EPDnew Database

Two sets of promoter sequences each of STIM1 (named STIM1_1 and STIM1_2), and STIM2 (STIM2_1 and STIM2_2), and one promoter sequence for each ORAI1, ORAI2, and ORAI3 were retrieved from the EPDnew database. Supplementary Information Table 4.2 lists the promoter sequences for STIM1, STIM2, ORAI1, ORAI2, and ORAI3 genes. We used rather short promoter sequence region from -499 to 100 bp relative to the transcription start site (TSS) to focus on putative physical interactions between TFs targeted to these selected regions rather than longer promoter regions which may contain many unreliable interactions.

4.3.3 HOCOMOCO and FIMO Results

The STIM and ORAI promoter sequences were searched for occurrences of known motifs using the FIMO tool. This tool scans a sequence database for individual matches to each of the motifs provided in the HOCOMOCO dataset (HOCOMOCOv9_AD_MEME.txt) which contains 426 non-redundant curated binding models for 401 human TFs. In total, we found 411 known motifs in the promoter of STIM1_1, 79 for (STIM1_2), 276 for (STIM2_1), 83 for (STIM2_2), 304 for (ORAI1), 571 for (ORAI2), and 262 for (ORAI3), respectively.

Then, the respective motifs were mapped to the 91 transcription factors targeting STIM and ORAI genes obtained from the ChEA database. In total, there are 13 non-redundant transcription factors (E2F1, E2F4, EGR1, ELF1, ELK3, GATA1, HNF4A, MITF, MYC, PPARD, RUNX1, SOX2, and SPI1) that were found targeted STIM and ORAI genes. Table 4.1 lists the motifs found targeting STIM and ORAI genes. On the other hand, only one transcription factor was found in the promoters of STIM1_1, STIM1_2, and STIM2_2, respectively. Moreover, 12 TFs were found in the promoter of STIM2_1, 4 TFs in the promoters ORAI1 and ORAI2, and 5 TFs in the promoter ORAI3.

Table 4.1 Motifs found targeting STIM and ORAI genes based on the CheA database and FIMO tool.

Promoter Name	Motif	Strand	Start	End	p-value	q-value	Matched Sequence
STIM1_1	ELK3_f1	+	494	505	4.05e-05	0.0361	GCCTGGAAGCCG
STIM1_2	SOX2_f1	+	443	458	2.14e-05	0.0238	CTATGCATCAGAAAAG
STIM2_1	ELK3_f1	+	154	165	5.54e-05	0.0528	CTCAGGATGTGG
	E2F1_f2	-	256	269	5.86e-06	0.00441	AGGAGGCGGGGAAG
	E2F4_do	-	256	269	4.41e-05	0.0372	AGGAGGCGGGGAAG
	GATA1_si	-	290	299	7.73e-05	0.0914	GCTGATAACG
	EGR1_f2	+	339	349	9.25e-05	0.0115	GGCGGGGCTGG
	EGR1_f2	-	354	364	8.56e-06	0.00306	CGCGTGCGCGG
	E2F1_f2	+	404	417	8.14e-05	0.0204	GGGAGGCGGGGGAT
	EGR1_f2	+	489	499	5.1e-06	0.00306	GGAGGGGGCGG
	E2F1_f2	+	491	504	3.92e-05	0.0147	AGGGGGCGGGGGGA
	EGR1_f2	+	510	520	1.94e-05	0.00361	CGCGGCGGCGG
	EGR1_f2	+	513	523	1.23e-05	0.00306	GGCGGCGGCGG
	EGR1_f2	+	516	526	7.86e-05	0.0115	GGCGGCGGCGC
STIM2_2	SOX2_f1	-	304	319	5.86e-05	0.0544	TTTTACAAAATAATGA
ORAI1	E2F4_do	+	67	80	2.48E-005	0.0183	AGTGGGCGCCAAAT
	E2F4_do	+	280	291	6.29E-005	0.0231	GGTGGGCGGGGAGC
	ELK3_f1	-	387	400	3.68E-005	0.0338	TCCTGGAAGCGC
	PPARD_f1	+	396	409	3.91E-005	0.0446	CGGGGCACAGGTGG
ORAI2	RUNX1_f1	+	121	130	8.89E-005	0.101	TCTGTGGGTA
	PPARD_f1	+	519	532	1.41E-005	0.0161	TGGGCCACAGGCCA
	MYC_f1	+	528	538	2.47E-005	0.0114	GGCCACGCGGC
	MYC_f1	-	529	539	5.66E-005	0.0131	GGCCGCGTGGC
ORAI3	HNF4A_f1	+	38	50	5.71e-06	0.0063	GGACCAAAGGCCG
	MITF_f1	+	466	475	6.78e-05	0.0589	ATCATGTGGC
	SPI1_si	-	496	512	5.55e-07	0.000628	CAAAACAGGAACTGGGA
	ELF1_f1	-	499	508	8.68e-05	0.0473	ACAGGAACTG
	ELF1_f1	-	583	592	5.74e-05	0.0473	CCAGGAAGAG

We prepared two diagrams to illustrate the transcription factors targeting STIM and ORAI on their promoter regions within the range of -499 to 100 bp, where 0 bp defines as the transcription start site (TSS) (Figure 4.6). We are assuming that possible physical interactions may occur between overlapping and neighbouring or adjacent transcription factors within ≤ 50bp in the promoter regions. We identified nine such possible physical protein-protein interactions E2F1:E2F4, E2F1:EGR1, E2F1:GATA1, and E2F4:GATA1 in the promoter region of STIM2_1 and PPARD: E2F4 in the promoter region of ORAI1, PPARD: MYC (promoter region of ORAI2), MITF: SPI1, MITF: ELF1, and SPI1:ELF1 (promoter region of ORAI3) (Table 4.2). However, no putative interaction was found on promoter regions of

STIM1_1, STIM1_2 and STIM2_2 because only one transcription factor bound at each of them. In addition, an experimental study by Eylenstein et al. 2012 [188], found that NFKB1 are related to STIM1 and ORAI1. The FIMO tool predicted NFKB1 to targeted STIM1 and ORAI2, but these were not found in the CheA database. Though, we considered the interactions between NFKB1 and RUNX1 in the promoter region of ORA12.

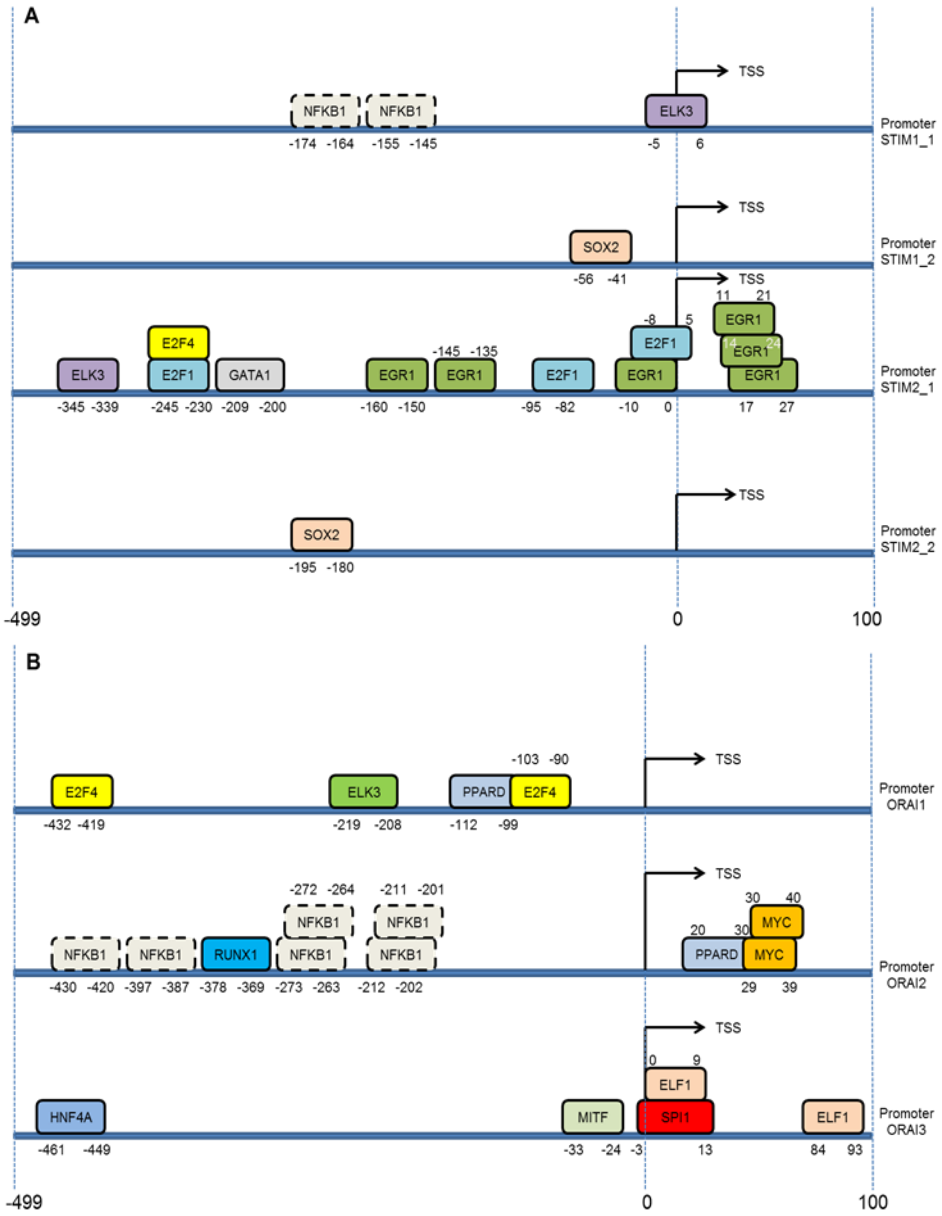

Figure 4.6 Schematic illustration of transcription factors binding motifs in the promoters of the genes STIM1, STIM2, ORAI1, ORAI2, and ORAI3. (A) Transcription factor binding motifs in the promoters of genes STIM1 and STIM2. (B) Transcription factors binding motifs in the promoters of genes ORAI1, ORAI2, and ORAI3.
TSS, transcription start site; Promoter STIM1_1, first promoter of STIM1; Promoter STIM1_2, second promoter of STIM1; Promoter STIM2_1, first promoter of STIM2; Promoter STIM2_2, second promoter of STIM2; Promoter ORAI1, promoter of ORAI1; Promoter ORAI2, promoter of ORAI2; Promoter ORAI3, promoter of ORAI3.

Table 4.2 List of ten predicted physical transcription factor interactions, location in the promoter and their overlap or gap by base pair.

			Location on promoter		
Promoter Name	**TF1**	**TF2**	**TF1**	**TF2**	**Overlap/gap by bp**
STIM2_1	E2F1	E2F4	-245 to -230		overlap by 15bp
	EGR1	E2F1	-145 to -135	-95 to -82	gap by 40bp
			-10 to 0	-8 to 5	overlap by 8bp
	*E2F1	EGR1	-8 to 5	11 to 21	gap by 6bp
			-8 to 5	14 to 24	gap by 9bp
			-8 to 5	17 to 27	gap by 12bp
	E2F1	GATA1	-245 to -230	-209 to -200	gap by 21bp
	E2F4	GATA1	-245 to -230	-209 to -200	gap by 21bp
ORAI1	PPARD	E2F4	-112 to -99	-103 to -90	overlap by 4bp
ORAI2	NFKB1	RUNX1	-430 to -420	-378 to -369	overlap by 42bp
			-397 to -387	-378 to -369	overlap by 9bp
	PPARD	MYC	20 to 30	29 to 40	overlap by 1bp
			20 to 30	30 to 40	gap by 0bp
ORAI3	MITF	SPI1	-33 to -24	-3 to 13	gap by 21bp
	MITF	ELF1	-33 to -24	0 to 9	gap by 24bp
	SPI1	ELF1	-3 to 13	0 to 9	overlap by 9bp

*The interaction between EGR1:E2F1 is assumed to be the same as interaction between E2F1:EGR1.
bp, base pair; TF1, first transcription factor and TF2, second transcription factor.

4.3.4 Physical Protein-protein Interactions

We used three well known databases STRING, prePPI, and mentha, to search for putative physical protein-protein interactions between transcription factors targeting STIM and ORAI genes on their promoter regions. Generally, these databases will search for possible physical protein-protein interactions either as direct or indirect interactions involving further bridge protein/s. Every database presented different results, scores, outputs, and networks.

Table 4.3 lists the results obtained from the STRING database. Figure 4.7 shows the network of interactions between 13 TFs targeting STIM and ORAI genes generated by the STRING database. The STRING database predicted eight possible protein-protein interactions either direct or involving bridge protein/s. The interactions of node1 and node2 which have a score of ≥0.5 suggest true interactions (Supplementary Information Table 4.3). In total, three direct interactions were observed between E2F1:E2F4, E2F4:GATA1 in the promoter of STIM2_1, and MITF:SPI1 in the promoter of ORAI3. Additionally, besides being direct interactions, E2F1:E2F4 was also connected by the bridge protein MYC and E2F4:GATA1 by MYC and EGR1.

Table 4.3 List of the predicted interactions between transcription factors obtained from the STRING database.

Promoter name	Putative Interactions	Type of interaction	Bridge Proteins	Interactions
STIM2_1	E2F1:E2F4	Direct		
	E2F1:E2F4	Bridge protein	MYC	E2F1>MYC>E2F4
STIM2_1	E2F1:EGR1	Bridge protein	SPI1	E2F1>SPI1>EGR1
	E2F1:EGR1	Bridge protein	E2F4	E2F1>E2F4>EGR1
	E2F1:EGR1	Bridge protein	MYC	E2F1>MYC>EGR1
STIM2_1	E2F1:GATA1	Bridge protein	ELF1	E2F1>ELF1>GATA1
	E2F1:GATA1	Bridge protein	SPI1	E2F1>SPI1>GATA1
	E2F1:GATA1	Bridge protein	MYC	E2F1>MYC>GATA1
STIM2_1	E2F4:GATA1	Direct		
	E2F4:GATA1	Bridge protein	MYC	E2F4>MYC>GATA1
	E2F4:GATA1	Bridge protein	EGR1	E2F4>EGR1>GATA1
ORAI1	PPARD:E2F4	Bridge protein	EGR1	PPARD>EGR1>E2F4
ORAI2	NFKB1:RUNX1	Bridge protein	EGR1	NFKB1>EGR1>RUNX1
	NFKB1:RUNX1	Bridge protein	MYC	NFKB1>MYC>RUNX1
	NFKB1:RUNX1	Bridge protein	SPI1	NFKB1>SPI1>RUNX1
ORAI2	PPARD:MYC	Bridge protein	EGR1	PPARD>EGR1>MYC
	PPARD:MYC	Bridge protein	HNF4A	PPARD>HNF4A>MYC
ORAI3	MITF:ELF1	Bridge proteins	SPI1, E2F1	MITF>SPI1>E2F1>ELF1
	MITF:ELF1	Bridge proteins	SPI1, GATA1	MITF>SPI1>GATA1>ELF1
ORAI3	MITF:SPI1	Direct		
ORAI3	SPI1:ELF1	Bridge protein	GATA1	ELF1>GATA1>SPI1
	SPI1:ELF1	Bridge protein	E2F1	ELF1>E2F1>SPI1

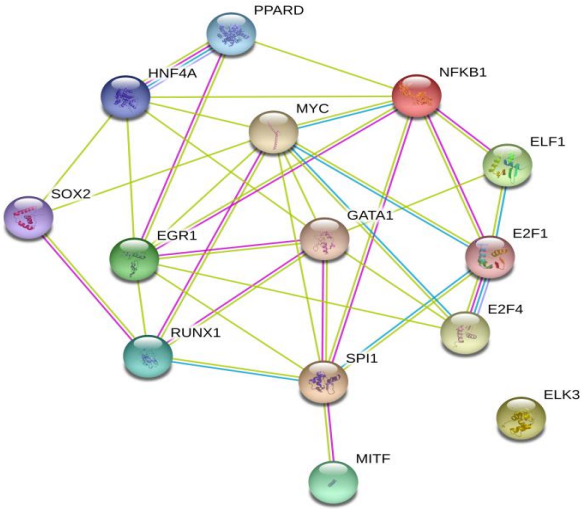

Figure 4.7 Network of 14 transcription factors targeting STIM and ORAI genes. Evidence view of the STRING database output depicting the transcription factors targeting STIM and ORAI genes obtained from http://string-db.org/.

The prePPI database results include prePPI LR, database LR, final probability and prediction code. The prediction code was labelled as S, T, G, E, M, C, and P to represent their sources of evidence used in the prediction. In our cases, the code of S was found which referred to structural modelling between the interactions of E2F1:E2F4, E2F1:GATA1, and NFKB1:RUNX1, T represents protein-peptide modelling (PPARD:E2F2 and PPARD:MYC) and G means GO term similarity (function similarity) (E2F1:EGR1 and MITF:SPI1). Generally, the prePPI database predicts only direct interactions. In total, four putative physical protein-protein interactions are predicted true by the prePPI database with a final probability ≥0.5. This includes the interactions between E2F1:E2F4 (final probability of 1.00), E2F:EGR1 (final probability of 0.68), MITF:SPI1 (final probability of 1.00), and SPI1:ELF1 (final probability of 0.99), respectively (Table 4.4).

Table 4.4 List of predicted interactions between transcription factors obtained from the prePPI database.

Promoter name	Predicted Interactions	prePPI LR	Database LR	Final Probability	Prediction Code
STIM2_1	E2F1:E2F4	109392.00	957.82	1.00	S
STIM2_1	E2F1:EGR1	1268.49	Not available	0.68	G
STIM2_1	E2F1:GATA1	31.17	Not available	0.05	S
STIM2_1	E2F4:GATA1	Not found	Not found	Not found	Not found
ORAI1	PPARD:E2F4	23.76	Not available	0.04	T
ORAI2	NFKB1:RUNX1	388.29	Not available	0.39	S
ORAI2	PPARD:MYC	61.64	Not available	0.09	T
ORAI3	MITF:ELF1	Not found	Not found	Not found	Not found
ORAI3	MITF:SPI1	47.73	4625.64	1.00	G
ORAI3	SPI1:ELF1	60755.6	Not available	0.99	S

LR, likelihood ratio; G, GO term similarity; S, structural modelling and T, protein-peptide modelling.

Generally, the mentha database can predict direct and indirect interactions. Table 4.5 lists the predicted interactions obtained from the mentha database. In total, three interactions with bridge proteins are predicted between E2F1:EGR1, PPARD:E2F4, PPARD:MYC, and one direct interaction between E2F1:E2F4. We identified one bridge protein (HDAC1) in the NFKB1:RUNX1 interaction, three bridge proteins (HDAC1, HDAC3, and NCOR2) in PPARD:E2F4 interactions, four bridge proteins (CDKN2A, CREBBP, EP300, and SP1) associate with E2F1:EGR1 interactions, and five bridge proteins (EP300, HDAC1, HDAC2, HDAC3, and KDM1A) with PPARD:MYC interactions, respectively. However, interactions

between E2F1:GATA1, E2F4:GATA1, MITF:SPI1, MITF:ELF1, and SPI1:ELF1 were not found in mentha.

Table 4.5 List of the predicted interactions between transcription factors obtained from the mentha database.

Promoter name	Putative Interactions	Type of interaction	Bridge Proteins	TF1	TF2	Score	PMID
STIM2_1	E2F1:E2F4	Direct	NULL	E2F4	E2F1	0.126	16357170
STIM2_1	E2F1:EGR1	Bridge protein	CDKN2A	CDKN2A	E2F1	0.507	11314038
				CDKN2A	EGR1	0.623	19057511
	E2F1:EGR1	Bridge protein	SP1	E2F1	SP1	0.902	10547281, 8657141, 10409740, 8657142
				EGR1	SP1	0.523	20121949
	E2F1:EGR1	Bridge protein	CREBBP	CREBBP	E2F1	0.523	12748276, 8932363
				CREBBP	EGR1	0.623	9806899
	E2F1:EGR1	Bridge protein	EP300	EP300	E2F1	0.507	24112038, 15123636, 23001041
				EP300	EGR1	0.93	9806899, 15225550, 20089040, 20018936
STIM2_1	E2F1:GATA1						
	Not found						
STIM2_1	E2F4:GATA1						
	Not found						
ORAI1	PPARD:E2F4	Bridge protein	NCOR2	NCOR2	PPARD	0.454	11867749
				NCOR2	E2F4	0.376	22508987
	PPARD:E2F4	Bridge protein	HDAC1	HDAC1	PPARD	0.569	18037904, 11867749
				E2F4	HDAC1	0.523	9724731, 23060449, 9858615
	PPARD:E2F4	Bridge protein	HDAC3	PPARD	HDAC3	0.454	11867749,12943985
				E2F4	HDAC3	0.376	22508987
ORAI2	NFKB1:RUNX1	Bridge protein	HDAC1	NFKB1	HDAC1	0.91	25241761 16319923 24448807 12972430 17827154 11931769 17962807
				HDAC1	RUNX1	0.73	22498736, 16652147, 21059642
ORAI2	PPARD:MYC	Bridge protein	HDAC3	PPARD	HDAC3	0.454	11867749, 12943985
				MYC	HDAC3	0.692	22002311,18483244, 23079660
	PPARD:MYC	Bridge protein	KDM1A	PPARD	KDM1A	0.49	23455924
				MYC	KDM1A	0.332	23455924
	PPARD:MYC	Bridge protein	HDAC2	HDAC2	PPARD	0.309	25241761, 11867749
				HDAC2	MYC	0.472	17314511, 20195357, 22286234
	PPARD:MYC	Bridge protein	HDAC1	HDAC1	PPARD	0.569	18037904, 11867749
				HDAC1	MYC	0.911	22286234, 18003922, 26496610, 17314511, 18271930, 24951594

	PPARD:MYC	Bridge protein	EP300	EP300	PPARD	0.309	16930961
				EP300	MYC	0.911	17157259, 15616592, 16287840, 16126174
ORAI3	MITF:SPI1						
	Not found						
ORAI3	MITF:ELF1						
	Not found						
ORAI3	SPI1:ELF1						
	Not found						

PMID, Pub Med ID.

Table 4.6 summarizes the type of interactions either as direct interaction or using bridge protein/s predicted from the STRING, prePPI, and mentha databases. If one of these databases predicts any interaction; which the first TF (TF1) interact with the second TF (TF2), we labelled "YES" in "Consider as TFs pair?" column. Overall, 14 non-redundant TFs are found to act as bridge proteins predicted by the STRING and mentha databases. Next, we used these 63 non-redundant genes (including 48 TFs predicted by the ChEA database and STIM1, STIM2, ORAI1, ORAI2, and ORAI3 genes itself) as genes of interest for differential expression gene analysis and co-expression analysis of the breast invasive carcinoma dataset (Table 4.7).

Table 4.6 List of predicted interactions between transcription factors obtained from the STRING, prePPI, and mentha databases.

Promoter name	TF1	TF2	Type of Interaction	STRING	prePPI	mentha	Consider as TF pair?
STIM2_1	E2F1	E2F4	Direct	YES	YES	YES	YES
			Bridge Protein	YES			
	E2F1	EGR1	Direct		YES		YES
			Bridge Protein	YES		YES	
	E2F1	GATA1	Direct				
			Bridge Protein	YES			YES
	E2F4	GATA1	Direct	YES			YES
			Bridge Protein	YES			
ORAI1	PPARD	E2F4	Direct				
			Bridge Protein	YES		YES	YES
ORAI2	NFKB1	RUNX1	Direct				
			Bridge Protein	YES		YES	YES
	PPARD	MYC	Direct				
			Bridge Protein	YES		YES	YES
ORAI3	MITF	ELF1	Direct				
			Bridge Protein	YES			YES
	MITF	SPI1	Direct	YES	YES		YES
			Bridge Protein				
	SPI1	ELF1	Direct		YES		YES
			Bridge Protein	YES			

Table 4.7 List of 63 non-redundant genes of interest obtained from the CheA, STRING, prePPI, and mentha databases.

AR	FOXA1	HNF4A	PAX3-FKHR	STIM1
BCL6	FOXA2	HOXB7	PHF8	STIM2
CDKN2A	FOXP1	KDM1A	POU3F2	TFAP2C
CREBBP	FOXP2	KLF5	PPARD	TFEB
CUX1	GABP	MITF	RBPJ	TOP2B
E2F1	GATA1	MYC	RUNX1	TP63
E2F4	GATA2	MYCN	RUNX2	TRIM28
EBNA2	GATA3	NCOR1	SCL	TTF2
EGR1	GATA4	NCOR2	SMAD4	VDR
ELF1	GATA6	*NFKB1	SOX11	WT1
ELK3	HDAC1	ORAI1	SOX2	ZNF217
EP300	HDAC2	ORAI2	SP1	
FLI1	HDAC3	ORAI3	SPI1	

***** NFKB1 was found to be related to STIM and ORAI genes by experimental study and was predicted by the FIMO tool.

4.3.5 Differential Gene Expression Analysis

To identify specific genes that were differentially expressed in breast invasive carcinoma dataset, we used DESeq2 packages. This dataset contains 113 normal and 1102 tumor samples. Aforementioned, we used a set of 63 genes of interest (Table 4.7) for differentially expressed analysis with STIM and ORAI genes as the main focus. By setting the p-value ≤ 0.05 and FDR value of 0.1, DESeq2 analysis identified 45 out of 63 genes that were differentially expressed with 26 genes being significantly up-regulated and 19 genes are down-regulated (Table 4.8). The ten most significantly differentially up-regulated genes were E2F1, ORAI2, CDKN2A, SOXII, GATA3, TRIM28, RUNX2, FOXA1, HDAC1, and KDM1A. On the other hand, the ten most significantly down-regulated genes were EGR1, FLI1, FOXP2, BCL6, TP63, MITF, SMAD4, STIM2, RBPJ, and ELK3. The ORAI1, ORAI2, and ORAI3 genes were found up-regulated. This agrees with the results of several studies stating that the expression of ORAI1 [167,189,190] and ORAI3 [168,190] genes increased in primary human breast cancer cells or tissues. On the other hand, we found STIM1 and STIM2 to be down-regulated genes.

Following data pre-processing, a total of 59 genes were identified. Four genes EBNA2, GABP, PAX3-FKHR, and SCL were not found in both normal and tumor samples of the breast invasive carcinoma dataset. Based on principal component analysis (PCA) for the gene expression profile data, it was found that the gene expression level in tumor (breast invasive carcinoma) samples was dissimilar to that in normal samples, thus the two groups were distinguished absolutely at the gene expression level (Figure 4.8).

Table 4.8 List of differentially expressed genes in the dataset of breast invasive carcinoma (BRCA).

	Log fold change	p-value	FDR-adjusted		Log fold change	p-value	FDR-adjusted
(a) Up-regulated genes				**(b) Down-regulated genes**			
E2F1	2.81	5.29E-132	2.81E-130	EGR1	-2.68	7.75E-84	2.05E-82
ORAI2	1.05	2.78E-63	4.92E-62	FLI1	-1.22	3.04E-55	3.22E-54
CDKN2A	2.95	8.34E-63	1.10E-61	FOXP2	-2.38	5.07E-50	4.48E-49
SOX11	3.06	2.24E-43	1.19E-42	BCL6	-1.12	2.54E-49	1.92E-48
GATA3	1.87	5.02E-42	2.42E-41	TP63	-2.56	6.32E-45	4.19E-44
TRIM28	0.83	3.19E-40	1.41E-39	MITF	-1.21	1.55E-44	9.12E-44
RUNX2	1.28	6.42E-38	2.62E-37	SMAD4	-0.53	1.13E-28	4.00E-28
FOXA1	1.86	3.20E-33	1.21E-32	STIM2	-0.50	1.56E-28	5.16E-28
HDAC1	0.59	1.51E-27	4.70E-27	RBPJ	-0.49	2.32E-25	6.83E-25
KDM1A	0.52	1.80E-22	5.02E-22	ELK3	-0.69	2.40E-21	6.07E-21
MYCN	1.75	8.31E-22	2.20E-21	MYC	-0.90	7.03E-16	1.49E-15
ZNF217	0.72	1.02E-18	2.45E-18	SP1	-0.25	4.66E-12	9.14E-12
HDAC2	0.61	6.77E-17	1.56E-16	KLF5	-0.82	4.02E-08	6.45E-08
TTF2	0.57	2.97E-16	6.55E-16	NCOR1	-0.31	1.637E-06	2.55E-06
ORAI1	0.55	4.46E-13	9.10E-13	TFAP2C	-0.47	2.00E-05	2.87E-05
SPI1	0.67	7.14E-12	1.35E-11	STIM1	-0.26	3.01E-05	4.10E-05
VDR	0.47	2.44E-11	4.45E-11	EP300	-0.18	0.0009369	0.0011822
PHF8	0.36	3.40E-11	6.01E-11	GATA6	-0.33	0.006	0.007
PPARD	0.39	1.52E-10	2.59E-10	FOXP1	-0.17	0.013	0.015
ORAI3	0.46	1E-09	1.69E-09				
CUX1	0.30	3.87E-06	5.86E-06				
GATA2	0.68	5.90E-06	8.69E-06				
NCOR2	0.22	2.66E-05	3.71E-05				
HDAC3	0.17	3.45E-05	4.57E-05				
RUNX1	0.25	0.0004872	0.0006298				
TOP2B	0.11	0.023	0.027				

FDR, False Discovery Rate.

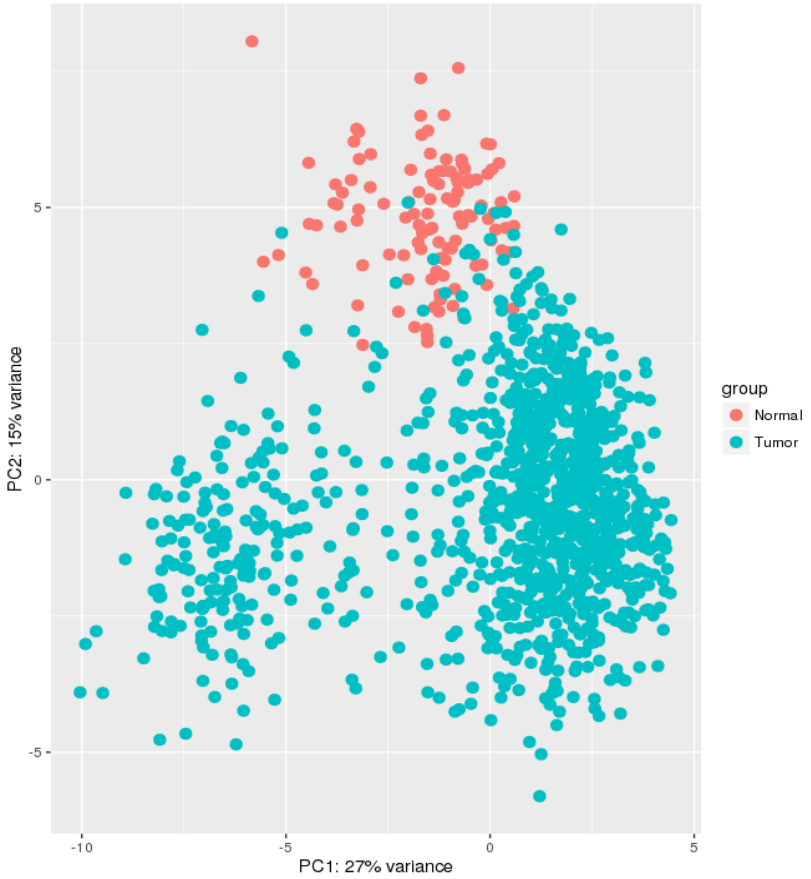

Figure 4.8 Plot of principal component analysis of 113 normal samples and 1102 tumor samples.
PC, principal component.

4.3.6 Results of Gene Co-expression Analysis

The analysis of gene co-expression of 113 normal samples and 1102 tumor samples was done separately using the WGCNA program.

4.3.6.1 Normal Samples

By observing the result obtained from the clustering dendogram and the topological overlap matrix (TOM) heatmap for all genes in normal samples (Figure 4.9), we found two significant modules. The blue module contains 15 genes (including ORAI2) and the turquoise module contains 18 genes (Table 4.9).

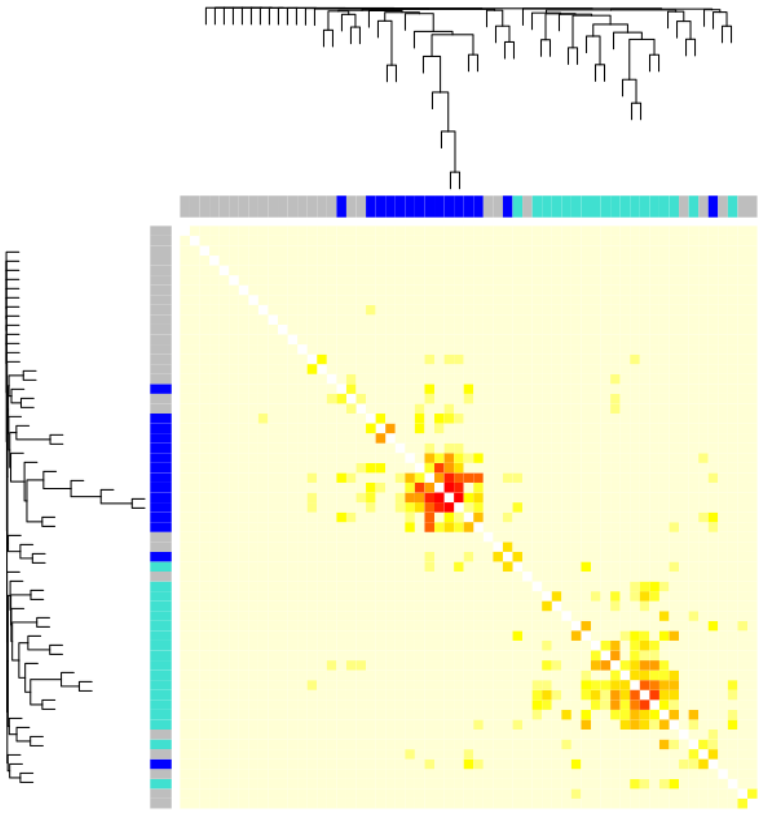

Figure 4.9 Topological overlap matrix (TOM) heatmap corresponding to the two co-expression modules in normal samples. Each row and column of the heatmap represents a single gene. Red indicates high levels of co-expression genes. The dendograms on the upper and left sides show the hierarchical clustering tree of genes.

4.3.6.2 Tumor Samples

In the tumor samples, we identified three significant modules. The brown module contains eight genes, the blue module 11 genes, and the turquoise module 13 genes, respectively (Figure 4.10). On the other hand, we found STM2 in brown module and ORAI1 in blue module. Table 4.9 summarizes all genes found in normal and tumor samples. The expression patterns of the individual modules will provide some clue about their functions.

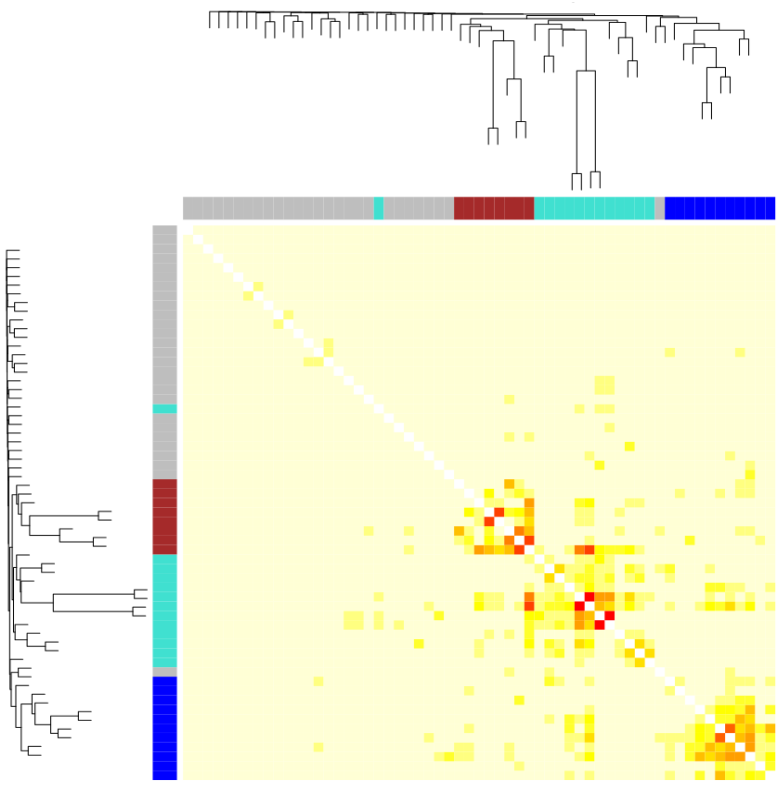

Figure 4.10 Topological overlap matrix (TOM) heatmap corresponding to the three co-expression modules in tumor samples. Each row and column of the heatmap represents a single gene. Red indicates high levels of co-expression genes. The dendograms on the upper and left sides show the hierarchical clustering tree of genes.

Table 4.9 Number of genes and gene name of the significant modules obtained from the normal and tumor samples identified by the WGCNA program.

Module	No. of genes	Gene Name			Module	No. of genes	Gene Name		
(a) Normal samples					**(b) Tumor samples**				
Blue	15	FOXP1	TOP2B	CUX1	Blue	11	SMAD4	PHF8	CREBBP
		SMAD4	ORAI2	AR			ELF1	SP1	
		FOXA1	PHF8	EP300			NCOR1	CDKN2A	
		TTF2	SP1	GATA3			TOP2B	AR	
		NCOR1	CDKN2A	CREBBP			ORAI1	EP300	
Turquoise	18	FOXP2	NCOR2	HOXB7	Turquoise	13	FOXP2	NFKB1	RUNX1
		HDAC1	TP63	VDR			HDAC1	TRIM28	
		HDAC3	RBPJ	GATA6			MITF	GATA6	
		KDM1A	KLF5	GATA2			TP63	E2F1	
		MITF	TRIM28	ELK3			EGR1	RUNX2	
		HDAC2	FLI1	TFAP2C			KLF5	ELK3	
					Brown	8	FOXA1	ZNF217	GATA3
							STIM2	FLI1	E2F4
							SPI1	TFEB	

4.3.7 Differential Interaction Networks

Following the analysis with the WGCNA package, we extracted all the genes differentially expressed from the significant modules of the TOM plot to construct networks of differential interactions using the Cytoscape [183]. The differential interactions were computed to observe which interactions are not in normal samples (labelled as normal-tumor) and vice versa for tumor samples (tumor-normal). We used the edges scores of the threshold ≥ 0.02. This resulted in 83 differential interactions in normal-tumor samples and 61 interactions in tumor-normal samples (Supplementary Information Table 4.4). The differential interactions networks suggesting that one of our focus genes, ORAI2 interacts with the SP1 and SMAD4 genes in normal-tumor, while STIM2 interacts with the ELK3, FOXP2 and FLI1 in tumor-normal (Figure 4.11). In the normal-tumor network, the three highest scored edges are the interactions between EP300:CREBPP (0.2822), NCOR1:EP300 (0.2343), and SP1:EP300 (0.2093), respectively. In the tumor-normal network, the three highest scored edges are the interactions between FOXP2:ELK3 (0.1546), MITF:RUNX2 (0.1528), and FOXA1:GATA3 (0.1112), respectively. On the other hand, genes with the highest number of edges which known as hubs play centred roles in the analysis of the networks. Thus, we identified two genes CREBBP and ELK3 as hub genes in the normal-tumor network which contains eight edges (Figure 4.11 (A)) while in the tumor-normal network, ELK3 is the hub gene which contains 12 edges (Figure 4.11 (B)).

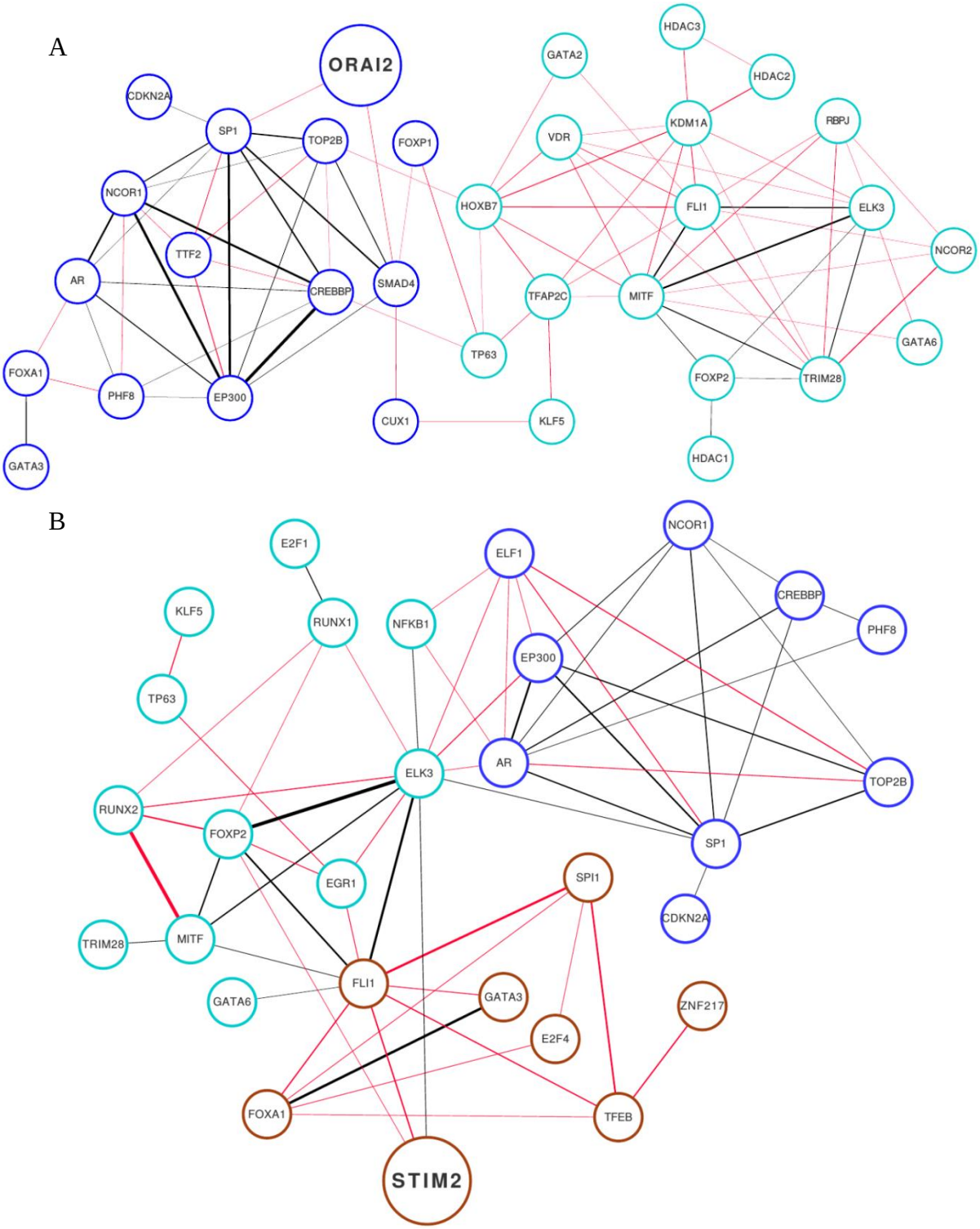

Figure 4.11 Differential Interaction Networks (A) Differential interactions network of normal-tumor. (B) Differential interactions network of tumor-normal.
Red edges indicate interactions that found in differential interactions and black edges indicate the interactions which were not found in differential interactions. Blue nodes indicate the genes found in the blue module, brown (blue module), and turquoise (turquoise module), respectively.

4.3.8 Regulation of STIM and ORAI Genes on Normal and Tumor Breast Cancer Tissue

As mentioned above, we identified ten putative interactions of transcription factors bound to promoter region of STIM and ORAI. There, five of the interactions involving E2F1 and E2F4 on promoter regions of STIM2 and ORAI1 (Table 4.6). Previous studies showed that the transcription factors of E2F family (E2F1-8) is recognized to regulate many important genes and involved in many biological processes such as apoptosis, cell proliferation, differentiation, and DNA damage response [191,192]. Furthermore, we identified E2F1 was up-regulated gene in tumor samples and interact with RUNX1 in differential interaction of tumor-normal network. Moreover, our results show that E2F1 and E2F4 are found co-expressed in tumor samples. We speculate that E2F1 and E2F4 are regulating STIM2 and ORAI1 genes which also show relation to breast cancer. In addition, we identified an interaction of NFKB1:RUNX1 on the promoter region of ORAI2 which also found co-expressed in turquoise module of tumor samples. NFKB1 is one of the family members of NFKB1 transcription factors [193]. Several studies noted that NFKB among the Ca^{2+} sensitive transcription factors which are associate with STIM1 and ORAI1 to stimulate cell proliferation and differentiation [194–196]. On the other hand, two interactions are found involving the micophthalmia-associated transcription factor (MITF) on promoter region of ORAI3 (MITF:ELF1 and MITF:SPI1). We found that MITF are co-expression in both turquoise modules of normal and tumor samples. A study by Carmit and co-workers noted that MITF functions as master regulatory of melanocytes development and melanoma oncogene [197]. Furthermore, Stanisz and co-workers stated the role of STIM and ORAI in melanocytes and melanoma which significantly correlates with the expression of MITF [163,198]. Generally, our findings suggest several roles of STIM and ORAI genes in normal and breast cancer tissues.

4.4 Conclusions

In this work, we identified 13 non-redundant transcription factors targeting STIM and ORAI genes. We then found ten putative TFs interactions bound on promoter regions of STIM and ORAI genes predicted by the STRING, prePPI, and mentha databases. According to these interactions, we found 14 non-redundant TFs act as bridge proteins. Then, we identified 63 non-redundant genes as genes of interest for differential expression gene analysis and co-expression analysis of breast invasive carcinoma dataset. In the differential expression analysis, we found 26 up-regulated genes including ORAI1, ORAI2, and ORAI3, while the 19 down-regulated genes include STIM1 and STIM2. On the other hand, in the co-expression analysis,

we found two significant modules (blue and turquoise modules) in normal samples and three significant modules (brown, blue, and turquoise modules) in tumor samples which the expression patterns of the individual modules tend to provide clues about their functions. Next, we identified 83 differential interactions in normal-tumor samples and 61 interactions in tumor-normal samples. Finally, we identified CREBBP and ELK3 as hubs genes in the normal-tumor network, and ELK3 in the tumor-normal network. Overall, our findings form an important basis for identifying TFs targeting STIM and ORAI genes and demonstrate the significant involvements of STIM and ORAI genes in breast cancer. In ongoing work, we are extending the gene enrichment analysis and applying the framework to datasets of diseases reported linked to STIM and ORAI genes such as Alzheimers disease (AD) and prostate cancer.

Chapter 5

Evaluation of the Protein Pocket Identification Tools on Protein-Ligand Complexes

My contribution was to write the manuscript, designed the research project, and analyzed the results together with the co-authors Zhao Yuan, Rahmad Akbar, and Volkhard Helms. I and Rahmad Akbar co-supervised Zhao Yuan. Zhao Yuan performed the calculations.

Abstract

Binding pockets are regions on protein surfaces where substrates of enzymatic reactions or effector molecules and co-factors may bind. Thus, identifying these cavities is often a prerequisite step for structure-based drug design. Various computational methods have been developed to identify such sites on protein surfaces. In this work, we evaluated the seven tools DEPTH, DoGSiteScorer, Fpocket, GHECOM, IsoMif, PocketPicker, and ProACT2 on a dataset of 167 non-redundant protein-ligand complexes. We analyzed how well the predicted pocket-lining residues overlap with the residues that contact the ligand. We used the residue overlap to define a score as a measure of the predictive capabilities of the tools. Even though the tools predicted pockets of various sizes and shapes we found comparable performance amongst the predictions of five tools (DEPTH, GHECOM, DoGSiteScorer, Fpocket, and IsoMif) in terms of average score. Using always the most suitable tool improved the average score by 28% over randomly selecting a tool. To support users in a pocket prediction scenario, we trained a random forest model (classifier) to output a list of suitable tools for a given protein structure. This classifier should be useful for prioritizing the tools to be used for unknown proteins or proteins that are not contained in our dataset.

Keywords: classifier, protein-ligand complexes, protein pockets, and random forest.

5.1 Introduction

Proteins play major roles in practically all cellular processes. They typically interact with small molecules (ligands), nucleic acids or other proteins to perform a certain function. These interactions often occur in a particular site on the protein surface (binding site). As binding sites are thus often related directly to protein function, it is important to advance our understanding on these sites. The large collection of experimentally determined three-dimensional structures of protein-ligand complexes stored in the protein data bank (PDB) allows us to study these binding sites. For instance, it has been shown that binding sites for small molecule ligands tend to be rather hydrophobic with few selected polar and charged residues [199–201] and tend to be found in deep pockets on the proteins surface [53,202].

Based on this data, one can develop algorithms to identify cavities that may accommodate bound ligands on protein surfaces. The current batch of such algorithms fall into five categories (i) geometric methods that can be further grouped into the three subcategories grid system scanning, probe sphere filling, and alpha shape [49,50], (ii) energy based methods, (iii) evolution based methods, (iv) blind docking and molecular dynamics and, (v) combined approaches [48]. Grid system scanning basically projects a protein structure onto a three-dimensional grid of points and examines spatial overlaps on this grid. DEPTH [57,58], DoGSiteScorer [56], GHECOM [49], and PocketPicker [55] are tools that implement a grid system scanning approach. Probe sphere filling methods generate a set of probe spheres to fill cavities on protein surfaces. Pockets are then defined as those regions containing the highest amount of spheres, e.g. by the tool IsoMif [63]. Alpha shape methods rely on the alpha-shape theory and Voronoi tessellation to identify a pocket. Fpocket [68] is a representative of alpha shape methods. Energy based methods identify pockets using energetic criteria. Cavities with the largest total interaction energies are defined as pockets. For instance, ProACT2 [69,70] is a representative of energy based method. Other tools employ further strategies. For example, Rate4Site [203] uses an evolution based approach and MolSite [77] utilizes blind docking. As a wide range of different strategies and approaches are employed by current pocket identification tools, it is of interest to compare and contrast the performance of these tools.

Defining the correct pockets on protein surfaces is not an easy task [68,204]. Schematically, Figure 5.1 sketches three possible ways to define the pocket volume of a pacman-shape surface cavity that is shown in two dimensions. It is unclear what definition is correct and most useful. Here we analyzed how well the constructed pockets overlap with the protein contacts made by small molecule ligands in their X-ray conformations. One should add,

as a word of caution, that native or synthetic ligands may either be smaller than surface pockets or exceed the volume of the pocket into the solution. Figure 5.2 shows for a case system that these tools predict pockets of various sizes and shapes. Indeed, in some cases the ligand is not fully enclosed by the detected pocket whereas other tools generated rather large pockets.

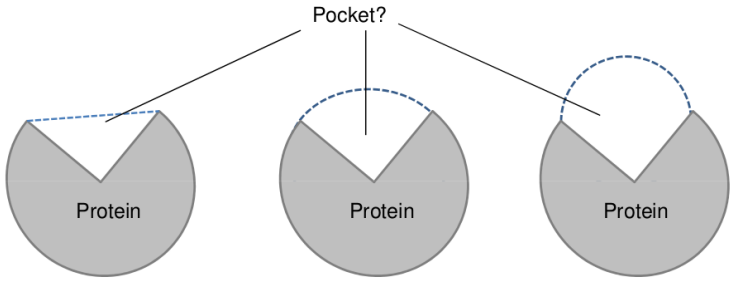

Figure 5.1 Two dimensional illustrations of three possible ways to define the pocket volume of a pacman-shape surface cavity.

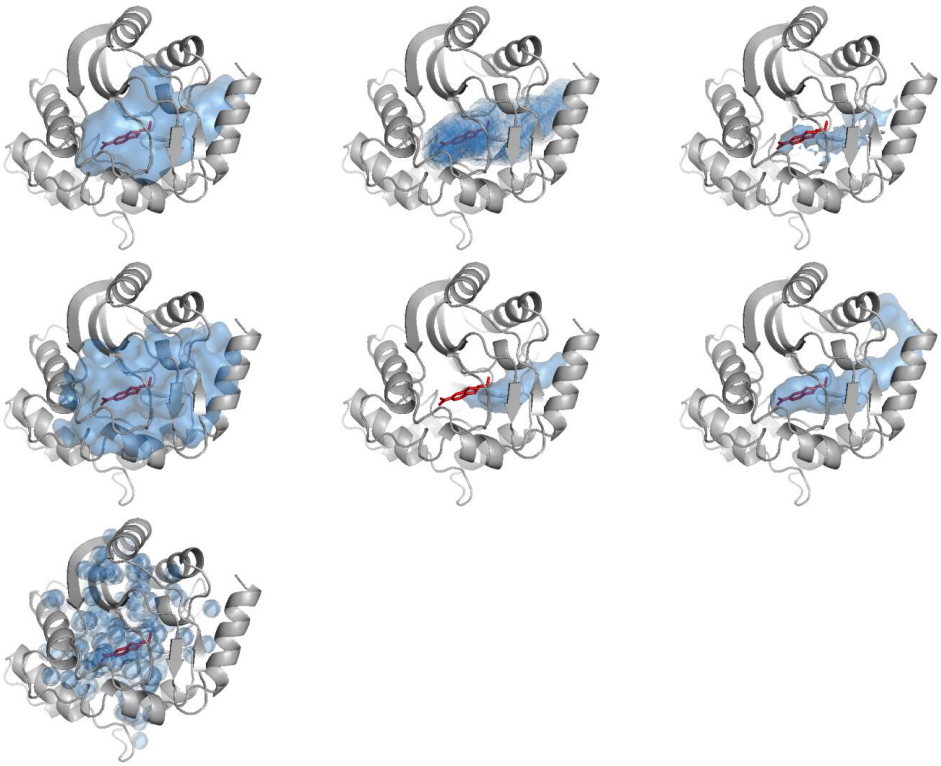

Figure 5.2 Pockets (blue) identified by DEPTH (lining residues) (top panel left), GHECOM (top panel middle), Fpocket (top panel right), DoGSiteScorer (middle panel left), PocketPicker (middle panel middle), IsoMif (middle panel right), and ProACT2 (bottom panel) that overlap with the ligand HNT (red sticks) bound to human phenylethanolamine N-methyltransferase, PNMT (PDB ID: 2G70). The figures were generated using PyMOL Molecular Graphics System [47].

Here, we compared the performance of seven tools using a set of quality metrics on a set of 167 protein-ligand complexes. We then computed a set of physico-chemical and geometric features for each ligand-bound pocket in the dataset. Correlation analysis of pocket features and the quality metrics revealed only weak correlation between pocket features and the predictive performance of the tools. In general, we found comparable performance in the predictions of five tools DEPTH [57,58], GHECOM [49], DoGSiteScorer [56], Fpocket [68], and IsoMif [63].

5.2 Material and Methods

5.2.1 Dataset

We used a non-redundant dataset of 195 protein-ligand complexes retrieved by Degac et. al. [205] from the PDBbind [37,206] database version 2014. Beside the protein, these complexes contain a single ligand with weight less than 1000 Da and the resolution of the X-ray structure must be equal to or better than 2.5 Å. Additionally, the ligand molecules must contain only common organic elements and the protein molecules include only the standard 20 amino acid residues in the area of the binding sites. Complexes with insertions and/or residues numbered with special characters (20 complexes), or that could not be processed with Fpocket [68] (five complexes) or with DoGSiteScorer [56] (three complexes) were removed. This yielded a final dataset of 167 complexes (Supplementary Information Table 5.1). Only a single chain was considered if a complex contains more than one homomer.

5.2.2 Tools

Initially, we considered a total of 25 pocket identification tools. However, among these, we were only able to automate the use of seven tools due to various limitations. Table 1 lists the tools used in this work. The complete set of tools and the respective limitations are listed in Supplementary Information Table 5.2.

Table 5.1 Names, URLs and year of creation of the seven protein pocket prediction tools.

Program	URL	Year
DEPTH	http://mspc.bii.a-star.edu.sg/tankp/help.html	2011
DoGSiteScorer	http://dogsite.zbh.uni-hamburg.de/	2012
Fpocket	http://Fpocket.sourceforge.net/	2009
GHECOM	http://strcomp.protein.osaka-u.ac.jp/GHECOM/	2010
IsoMif	http://bcb.med.usherbrooke.ca/imfi.php	2015
PocketPicker	http://gecco.org.chemie.uni-frankfurt.de/pocketpicker/index.html	2007
ProACT2	http://people.cryst.bbk.ac.uk/~ubcg66a/proact2_summary.html	2010

5.2.3 Binding Site Definition

A binding site is defined as the set of residues which are located within 5.0 Å from the ligand surface (actual positive class in the X-ray structure of the protein-ligand complexes). The remaining residues in the protein are defined as non-binding site residues (actual negative class).

5.2.4 Model Evaluations

Predicted pockets were evaluated by comparing their pocket lining residues with the residues of the corresponding binding site of the respective protein-ligand complexes using a confusion matrix. Each column of the confusion matrix represents the predicted class, whereas each row represents the actual class. We used this matrix to quantify the correct and incorrect predictions of each tool. In our case, TP (true positive) is the overlap between residues in the binding site and the lining residues of the predicted pockets, FN (false negative) are the residues in the binding site that were predicted as non-binding site. FP (false positive) are the non-binding site residues that were predicted as binding sites and TN (true negative) are the non-binding site residues that were correctly predicted as non-binding site. From the confusion matrix we computed the following quality metrics.

5.2.5 MCC

The Matthews correlation coefficient (MCC) [207] quantifies the degree of correlation between the actual and predicted classes of the residues. An MCC value of 1 indicates that all predictions are correct, -1 for completely incorrect predictions [208], and a value of zero indicates a random prediction [209]. MCC is defined in equation 5.1:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FN)(TP+FP)(TN+FP)(TN+FN)}} \tag{5.1}$$

5.2.6 Precision

Precision (P) is the proportion of correct predictions (TP) among all positive predictions (TP and FP). The precision value ranges from 0 to 1.

$$P = \frac{TP}{TP+FP} \tag{5.2}$$

5.2.7 Recall

Recall (R) is the proportion of correct predictions (TP) in condition positive (TP and FN).

The recall value ranges from 0 to 1.

$$R = \frac{TP}{TP+FN} \tag{5.3}$$

5.2.8 Overlap

Overlap measures the fraction of overlapping residues in a binding site for every tool (high probability residues, HPR). The values range from 0 to 1. An overlap value of 1 indicates that all residues in the binding site predicted by one tool are high probability residues that are found by the majority of the tools.

$$overlap\ score = \frac{HPR\ in\ a\ tool}{total\ HPR} \tag{5.4}$$

5.2.9 Correlation between Features and Quality Metrics

The correlation between features (chemical descriptors) and quality metrics was computed using the Pearson Correlation Coefficient (PCC).

$$PCC = \frac{\sum XY - \frac{(\sum X)(\sum Y)}{n}}{\sqrt{\left(\sum X^2 - \frac{(\sum X)^2}{n}\right)\left(\sum Y^2 - \frac{(\sum Y)^2}{n}\right)}} \tag{5.5}$$

where X is the value of a descriptor, Y is the value of the corresponding quality metrics and n is the total number of samples.

5.2.10 TScore

A tool score (TScore) was defined as the sum of the quality metrics (MCC, precision, recall, and overlap) normalized by the count of this metrics. Higher scores, in general, indicate better performance.

$$TScore = \frac{MCC+precision+recall+overlap}{4} \tag{5.6}$$

5.2.11 TRatio

The tool ratio (TRatio) was used to approximate the distance between a predicted TScore and the reference score. The reference score was defined separately for each protein-ligand complex as the maximum score obtained from any of the tools (maxTScore). TRatio is defined as:

$$TRatio = \frac{prediction\ score}{maxTScore} \tag{5.7}$$

5.2.12 Classifiers

Scikit-learn [210] was used to evaluate five classification methods Bayes [211,212], decision tree [213,214], random forest [215], support vector machine with radial kernel (svm_c_rbf), and support vector machine with linear kernel (svm_linear) [216,217]. These methods were trained on 75% of the dataset and tested on the remaining samples. As the training and test data were randomly split, we iterated the training and testing procedures 100 times to obtain more objective results.

5.3 Results

5.3.1 Pocket Distributions

The pocket tools output various quantities of pockets and formats for a given protein. DEPTH and ProACT2 output predicted pockets as a single file making it difficult to extract individual pockets for further analysis. GHECOM and PocketPicker output a predetermined number of five pockets per protein. In contrast, Fpocket, DoGSiteScorer, and IsoMif output all predicted pockets on the protein surface without limiting the number of predicted pockets to a fixed quantity. For the latter three tools, as they output all predicted pockets for a given protein, Figure 5.3 shows the distribution of the number of predicted pockets. Fpocket, DoGSiteScorer, and IsoMif predicted 1 to 43, 3 to 112, and 13 to 147 pockets, respectively. IsoMif predicted, on average, the largest number of pockets per protein (52.1) followed by DoGSiteScorer (18.4), and Fpocket (11.7).

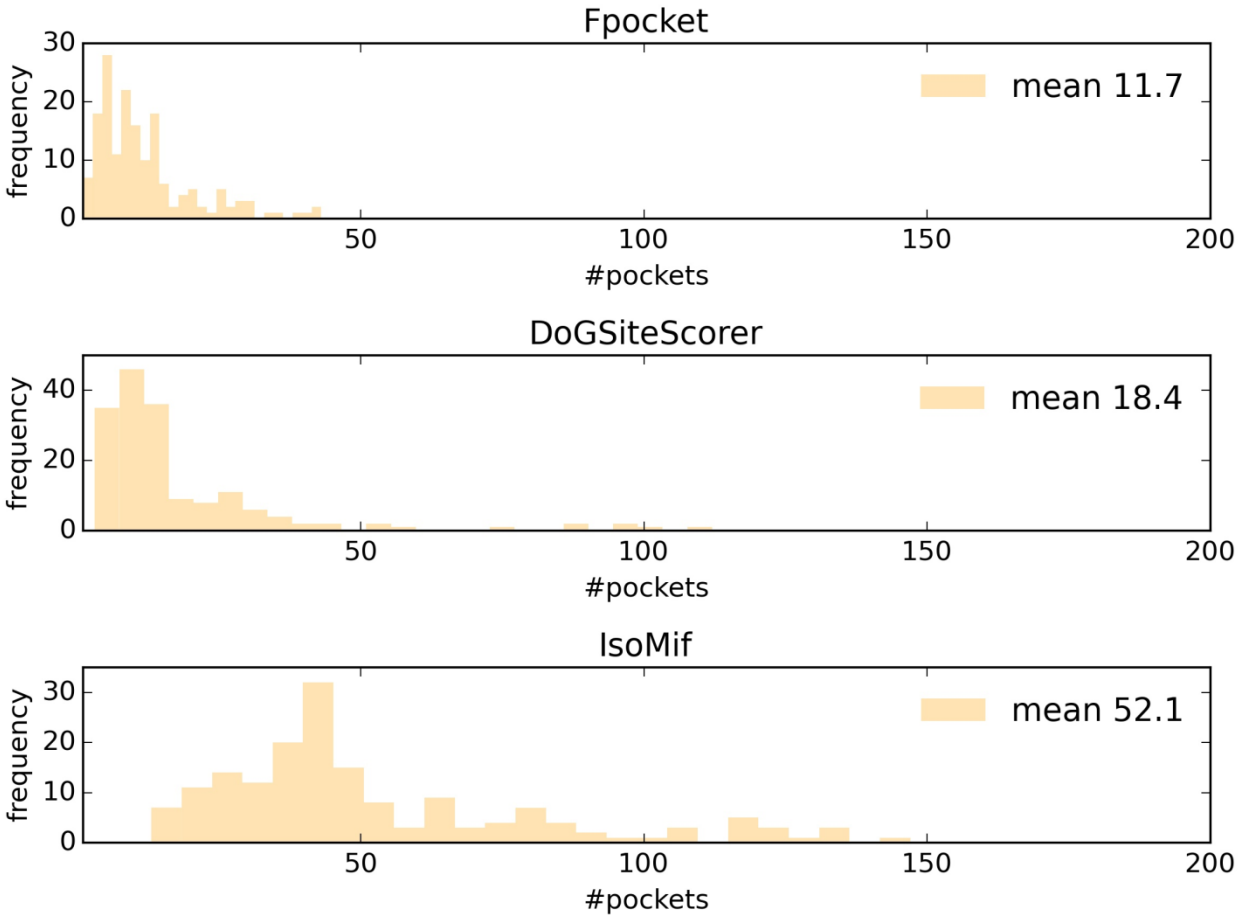

Figure 5.3 Distribution of the number of predicted pockets per protein for Fpocket, DoGSiteScorer, and IsoMif, respectively.

5.3.2 Size Distributions

Beside predicting different numbers of pockets, the tools also predict pockets of different sizes. For tools which provide individual pockets (Fpocket, DoGSiteScorer, IsoMif, GHECOM, and PocketPicker), we approximated the size of the predicted pockets by counting the number of residues that constitute a predicted pocket. As GHECOM and PocketPicker only output five pockets per protein, we always considered the top five pockets in each case. DoGSiteScorer, PocketPicker, and IsoMif predicted a similar number of residues for the top five pockets of 26.7, 26.5, and 23.6, respectively. Fpocket predicted 17.4 residues per pocket and GHECOM predicted the smallest number of residues per pocket, 9.0. Figure 5.4 shows the distribution of the number of residues per pocket for the tools.

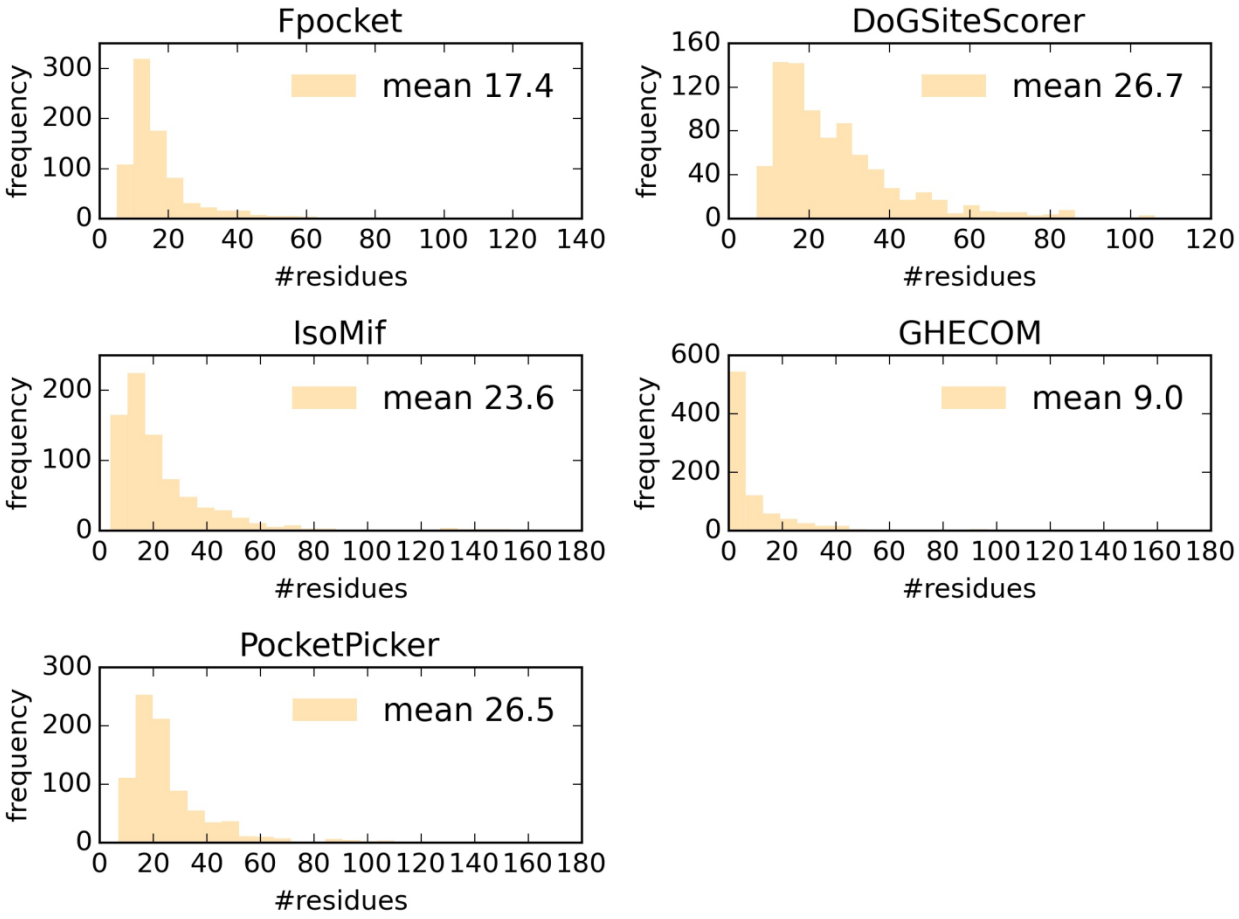

Figure 5.4 Distribution of the number of pocket lining residues per pocket (top five) for Fpocket, DoGSiteScorer, IsoMif, GHECOM, and PocketPicker, respectively.

5.3.3 Prediction versus Reality

Then, we measured how well the identified pockets match the positions of bound ligands (see also discussion section). We used four quality metrics to assess the performance of each tool against a reference binding site namely Matthews correlation coefficient (MCC), precision (pre), recall (recall), and residue overlap (overlap). The Methods section provides comprehensive definitions of these metrics. Reference binding sites encompass residues within 5.0 Å distance from any atom of the small-molecule ligand bound to the protein. Figure 5.5 shows the cumulative density of each metric for each tool. In general a tool would perform well if its corresponding quality metrics are shifted to the right of the density plots.

ProACT2 and PocketPicker consistently yielded lower performance compared to the other tools. ProACT2 is on the leftmost portion of the density plots (Figure 5.5) for MCC, precision, and overlap. Similarly, PocketPicker is near or at the leftmost portion of the density plot for MCC, precision, and recall. On the other hand, GHECOM performed best (on the rightmost of the plots) in three metrics (MCC, precision, and overlap) followed by

DoGSiteScorer in recall, and DEPTH was on the rightmost of the MCC plot. On the other hand, IsoMif was placed always in the middle range.

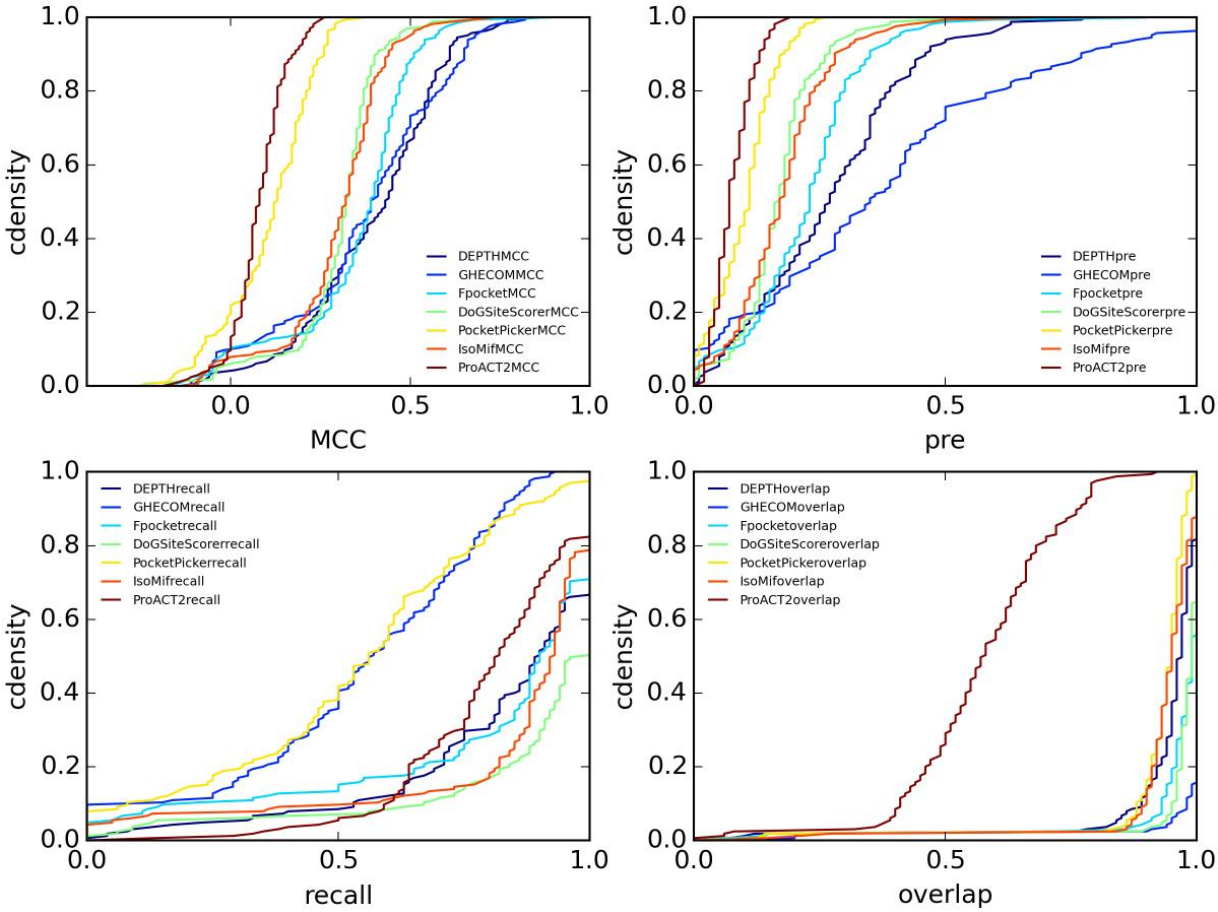

Figure 5.5 Cumulative density of Matthews correlation coefficient (MCC), precision (pre), recall, and residues overlap (overlap).

5.3.4 Features and Quality Metrics

Next we asked whether the pocket tools work better for certain types of pockets and worse for other. Several of the tools used in this work can compute features for the predicted pockets. DoGSiteScorer, Fpocket, GHECOM, PocketPicker, and ProACT2 provide 64, 19, 5, 420, and 23 descriptors for a protein, respectively. PocketPicker [55] generates the largest number of features. However, its features focus only on shapes and buriedness of a pocket neglecting physico-chemical properties entirely. In contrast, DoGSiteScorer [56] outputs both shapes and physico-chemical descriptors. For this reason we selected DoGSiteScorer to compute features for each pocket in our reference dataset. Details regarding the features used are included in Supplementary Information Table 5.3.

Similar to the previous section, we used a cumulative density plot to visualize the correlation values. The densities are centered around zero in all metrics and tools. This suggests

predominantly weak correlations between the features and the quality metrics. Furthermore, the Pearson correlation coefficients for MCC, precision, and overlap never exceeded 0.5 in all tools. For GHECOM, the feature simple score (simpleScore) correlated slightly with recall (0.56).

5.3.5 Tool Score

To rank the tools, we defined a simple tool score (TScore) by summing MCC, precision, recall, overlap and normalized the sum by four. Figure 5.6 shows the TScores per protein-ligand complex as a heatmap. We then labelled each protein with the tool that achieved the highest TScore (maxTScore). DEPTH, GHECOM, Fpocket, DoGSiteScorer, PocketPicker, IsoMif, and ProACT2 were assigned to 65, 33, 36, 15, 0, 14, and 4 proteins, respectively. The average TScores across DEPTH (0.58), GHECOM (0.54), Fpocket (0.55), DoGSiteScorer (0.55), and IsoMif (0.53) did not differ much. Compared to this group, PocketPicker (0.40) and ProACT2 (0.36) yielded considerably lower scores. Picking the tools randomly from the seven tools yielded an average TScore of 0.53. On the other hand, if one would always pick the tool with the highest TScore for each individual protein, the maximum possible TScore is 0.67. This means that an optimal choice of tools with the highest TScore for each protein may improve the average score by 28% over choosing tools randomly.

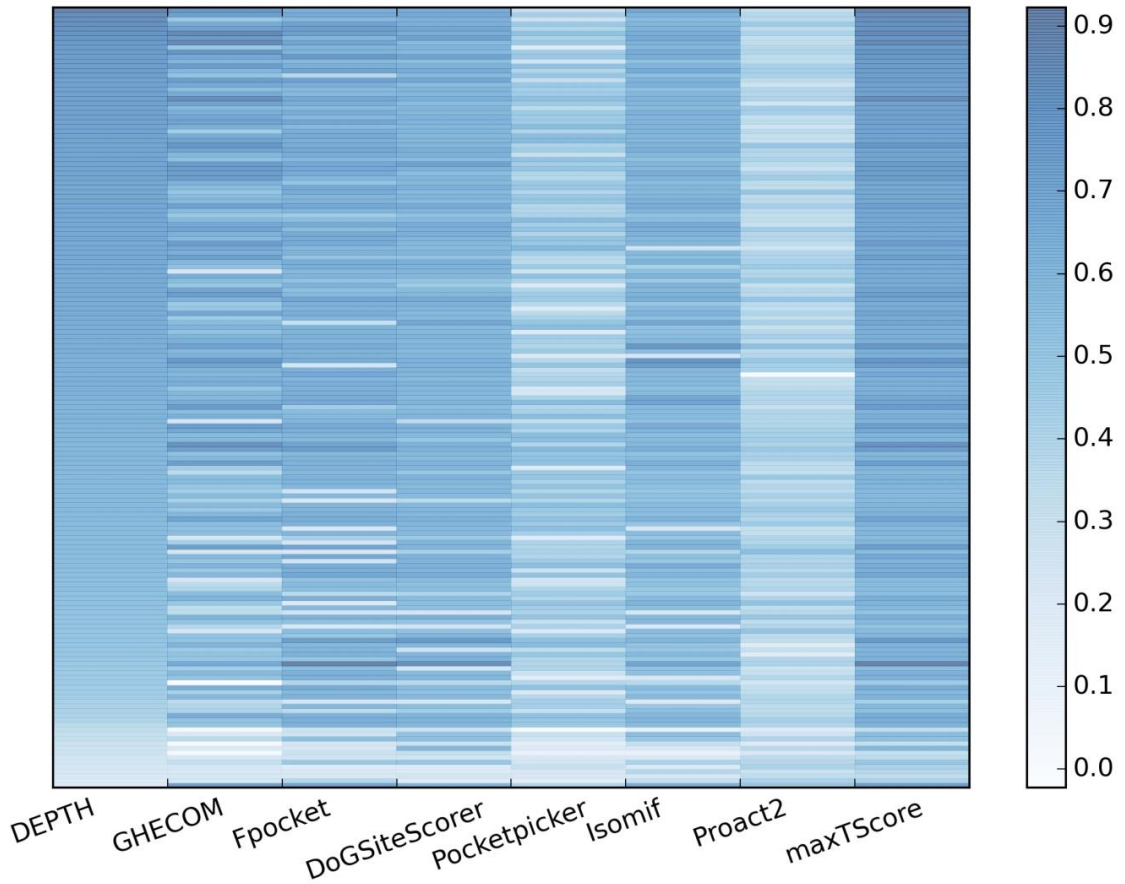

Figure 5.6 Heatmap of TScores of DEPTH, GHECOM, Fpocket, DoGSiteScorer, PocketPicker, IsoMif, ProACT2, and maxTScore, respectively. The column maxTScore contains the maximum TScores among all methods.

5.3.6 Identifying an Optimal Pocket Tool for a Protein

Since we are now in possession of a labelled dataset along with the corresponding features, we were able to train classifiers on this dataset. The idea behind is that such classifiers could be used to identify the optimal tool for an unknown protein. We trained Bayes, decision tree, random forest, support vector machine with radial and linear kernel on 75% of the total data and tested on the remaining data. As the training and testing dataset are split randomly, we needed to account for the variance on each split, hence, we iterated the training 100 times. We found that random forest yielded an accuracy of 0.60 followed by support vector machine with radial kernel (0.55), decision tree (0.50), support vector machine with linear kernel (0.38), and Bayes (0.34).

As the TScores amongst the tools are very similar, thus, we computed the ratio of the TScores the predictions in our best performing classifier (random forest) and the maximum TScores of the reference proteins (TRatio). Such a ratio approximates the distance between the TScore of the random forest predictions and the corresponding maximum TScore of reference proteins. A ratio value near one indicates that the TScore of the random forest prediction is

very similar to the maximum TScore of the reference protein. We found that TRatios are distributed around 0.90±0.03, further indicating that the best tools have similar predictive capabilities.

As TScores are very similar across the tools we decided to output a list of tools along with the corresponding weights (probability values from the random forest model) for each protein in the prediction instead of just one optimal tool. A higher weight indicates that the classifier is confident in the decision while uniform weights across the tools indicate less confidence in the decision. For example, for EPSP synthase from *Escherichia coli* (PDB ID:2QFT) DEPTH has a clearly higher weight (0.4253) than the other tools (first row, Table 2). On the other hand, in the case of 2-naphtamidine urokinase inhibitor (PDB ID:1SQA) four tools (DEPTH, Fpocket, DoGSiteScorer, and IsoMif) have similar weights (second row, Table 2). Similarly, for the PTP16-inhibitor complex (PDB ID:2QBR) DEPTH and IsoMif have similar weights (last row, Table 2). For cases like 1SQA and 2QBR it is useful to present the predictions in the form of a list of tools instead of just one tool to allow users to interfere and make a choice based on their need or objectives. Supplementary Information Table 5.4 lists the weight for each tool obtained for each protein of the training and test sets.

Table 5.2 Examples of weight for each tool for each protein complex obtained from the classifier of the random forest model.

PDB ID	DEPTH	GHECOM	Fpocket	DoGSiteScorer	IsoMif	ProACT2
2QFT	0.4253	0.1908	0.1888	0.0744	0.0817	0.0389
1SQA	0.2332	0.1385	0.2138	0.1915	0.1757	0.0472
2QBR	0.2545	0.1071	0.1713	0.1459	0.2411	0.0801

5.4 Discussion

As mentioned in the introduction section defining the correct pockets on protein surfaces is not an easy task. This is reflected in our results. Despite similar average TScores for five tools, we found clear differences in the shape and size of the predictions (Figures 5.2 and 5.4). It was previously reported that pockets vary in shapes and can be found buried deep within a protein or narrow and shallow on the protein surface [201,202]. In addition, Villar and Kauvar [218] noted that specific amino acids such as Arg, His, Trp, and Tyr are often enriched in protein binding sites compared to the entire protein. Due to these complexities, a pocket identification tool can be very successful for a set of proteins and less successful on other sets.

On the dataset studied here and using the four quality metrics defined by us, DEPTH yielded the highest aggregated TScore for 64 proteins. However, one should note that DEPTH

predicted larger pockets compared to other tools for the majority of the proteins. This is illustrated by the number of residues per predicted pocket. DEPTH, on average, predicted pockets with 77.0 residues while other tools predicted an average smaller pockets (for top five pockets), DoGSiteScorer (26.7), PocketPicker (26.5), IsoMif (23.6), Fpocket (17.4), and GHECOM (9.0). We speculate that the larger pocket sizes might have provided an advantage to DEPTH in terms of TScores since the score is computed based on these residues.

Interestingly, when TScores were correlated to maxTScore, we found that GHECOM yielded the highest correlation (0.84) followed by DoGSiteScorer (0.78), DEPTH (0.77), and IsoMif (0.75). Not surprisingly PocketPicker (0.43) and ProACT2 (0.23) yielded the lowest correlations. Even though GHECOM did not yield the highest average TScores, the tool was either in the second or third position for each protein-ligand complex in the dataset. In addition, GHECOM also predicted the smallest pockets (Figure 5.3). The high correlation to maxTScore and the relatively small pocket size indicate that GHECOM can produce quite precise predictions and could be a reasonable choice if one does not know what tool to choose for a given protein. Alternatively, the decision for a suitable tool may of course also be based on the random forest classifier that was trained here.

5.5 Conclusions

In this work we compared seven pocket identification tools. We found that these tools predict pockets of various size and shapes. For instance, DEPTH, DoGSiteScorer and PocketPicker tend to predict larger pockets, whereas GHECOM predicts smaller sized pockets. The tools also predict various numbers of pockets per protein. IsoMif identified most pockets an average (52.1), followed by DoGSiteScorer (18.4), and Fpocket (11.7). We introduced TScores as a measure of the predictive capabilities of the tools. When one applies the optimal tool for each protein, the average TScore increases by 28% over randomly labelled proteins. The tool GHECOM constructed pockets of rather compact size and its predictions were consistently among the top three tools. Finally, we trained a random forest classifier on this dataset. The classifier outputs a list of tools with a corresponding set of weights indicating the confidence of the decision boundaries. The classifier should be useful to prioritize a set of pocket tools for protein structures that are not contained in our dataset.

Chapter 6

Conclusions and Future Works

In this thesis, we presented three different projects which mainly related to the core area of protein interactions and gene regulations. We are aimed to understanding the role of protein-protein interactions and protein-ligand interactions, protein interfaces, and pockets. We also are extending our study to gain knowledge on transcription factors targeting STIM and ORAI genes, the regulation and relation with breast cancer.

First, we performed statistical analysis on the composition of overlapping protein-protein and protein-ligand interfaces. We started from the research question to find out whether small molecule ligands have similar physio-chemical features as protein binding interfaces when they bind at overlapping protein-protein or protein-ligand binding interfaces. We are using five different datasets from the ABC, PIBASE and TIMBAL databases. According to the statistics, we found that generally, both interface types have a very similar composition. Among the three datasets of PL complexes, we found that the protein interfaces of the Timbal dataset contain more hydrophobic residues and fewer polar residues than the two other datasets. In addition, we found that the ligands in the PIBASE and Timbal datasets bind more flat on the protein surfaces or bind deeper into pockets on the protein surface than ABC ligands. To further explore the findings, we will apply the framework on larger datasets of protein-protein and protein-ligand complexes.

For the second project, we addressed several angles about STIM and ORAI genes. We identified ten predictive interactions between transcription factors bound to promoter regions of STIM and ORAI genes based on predictions using the STRING, prePPI, and mentha databases. We used a set of genes of interest for co-expression and differential expression analysis on breast invasive carcinoma dataset which main focus on the regulation of STIM and ORAI genes. We found ORAI1, ORAI2, and ORAI3 genes to be up-regulated and in contrast STIM1 and STIM2 were down-regulated. We identified several roles of STIM and ORAI genes in normal and breast cancer tissues. The results presented in this study allow us to gain knowledge and unveiling the predicted transcription factors bound to the promoter regions of STIM and ORAI genes, their regulation, and relation with breast cancer. In future, we aim to extend this study by performing gene enrichment analysis and expand this workflow to new datasets of diseases which reported related to STIM and ORAI such as Alzheimers disease and prostate cancer.

In final study presented in this thesis, we evaluate performance of seven protein pocket identification tools on a dataset of protein-ligand complexes. We analyzed how well the predicted pocket-lining residues overlap with the residues that contact the ligand. We also used the residue overlap to define a score as a measure of the predictive capabilities of the tools. We found comparable performance amongst the predictions of five tools in terms of average score. We then trained a random forest model (classifier) to output a list of suitable tools for a given protein structure to assist users in a pocket prediction scenario. This classifier should be helpful for prioritizing the tools to be used for unknown proteins or proteins that are not contained in our dataset. Future work, we will present a more comprehensive evaluation of our framework on larger datasets of protein-ligand complexes.

In summary, the studies presented in this thesis led to gain knowledge of protein interactions and regulation of STIM and ORAI genes. Overall, our future works for these three projects will involve and focusing on applying the frameworks either on new or larger datasets for better understanding and comprehensive evaluation.

REFERENCES

1. Darbellay B, Arnaudeau S, König S, Jousset H, Bader C, Demaurex N, et al. STIM1- and Orai1-dependent store-operated calcium entry regulates human myoblast differentiation. J Biol Chem. 2009;284: 5370–5380. doi:10.1074/jbc.M806726200
2. Johnstone LS, Graham SJL, Dziadek MA. STIM proteins: Integrators of signalling pathways in development, differentiation and disease. J Cell Mol Med. 2010;14: 1890–1903. doi:10.1111/j.1582-4934.2010.01097.x
3. Flourakis M, Lehen'kyi V, Beck B, Raphaël M, Vandenberghe M, Abeele F V, et al. Orai1 contributes to the establishment of an apoptosis-resistant phenotype in prostate cancer cells. Cell Death Dis. 2010;1: e75. doi:10.1038/cddis.2010.52
4. Prevarskaya N, Skryma R, Shuba Y. Calcium in tumour metastasis: new roles for known actors. Nat Rev Cancer. Nature Publishing Group; 2011;11: 609–618. doi:10.1038/nrc3105
5. Gruszczynska-Biegala J, Pomorski P, Wisniewska MB, Kuznicki J. Differential roles for STIM1 and STIM2 in store-operated calcium entry in rat neurons. PLoS One. 2011;6: e19285. doi:10.1371/journal.pone.0019285
6. Moccia F, Zuccolo E, Soda T, Tanzi F, Guerra G, Lisa Mapelli, et al. Stim and Orai proteins in neuronal Ca2+ signaling and excitability. Front Cell Neuroscie. 2015;9: 1–14. doi:10.3389/fncel.2015.00153
7. Jemal A, Siegel R, Ward E, Murray T, Xu J, Smigal C, et al. Cancer Statistics, 2006. CA Cancer J Clin. 2006;56: 106–130. doi:10.3322/canjclin.56.2.106
8. Network TCGA. Comprehensive molecular portraits of human breast tumours. Nature. 2012;490: 61–70. doi:10.1038/nature11412
9. Keskin O, Tuncbag N, Gursoy A. Characterization and prediction of protein interfaces to infer protein-protein interaction networks. Curr Pharm Bi10. Jones S, Thornton JM. Principles of protein-protein interactions. Proc Natl Acad Sci USA. 1996;93: 13–20.
11. Keskin O, Nussinov R. Similar binding sites and different partners: implications to shared proteins in cellular pathways. Structure. 2007;15: 341–54. doi:10.1016/j.str.2007.01.007
12. Chakrabarti P, Janin J. Dissecting protein-protein recognition sites. Proteins Struct Funct Genet. 2002;47: 334–343. doi:10.1002/prot.10085
13. DeLano WL. Unraveling hot spots in binding interfaces: Progress and challenges. Curr

Opin Struct Biol. 2002;12: 14–20. doi:10.1016/S0959-440X(02)00283-X

14. Conte L Lo, Chothia C, Janin È. The atomic structure of protein-protein recognition sites. J Mol Biol. 1999;285: 2177–2198.
15. Moreira IS, Fernandes PA, Ramos MJ. Hot spots-A review of the protein–protein interface determinant amino-acid residues. Proteins Struct Funct Bioinforma. 2007;68: 803–812. doi:10.1002/prot
16. Keskin O, Ma B, Nussinov R. Hot regions in protein-protein interactions: the organization and contribution of structurally conserved hot spot residues. J Mol Biol. 2005;345: 1281–94. doi:10.1016/j.jmb.2004.10.077
17. Chen Y, Kortemme T, Robertson T, Baker D, Varani G. A new hydrogen-bonding potential for the design of protein-RNA interactions predicts specific contacts and discriminates decoys. Nucleic Acids Res. 2004;32: 5147–62. doi:10.1093/nar/gkh785
18. Camacho CJ, Zhang C. FastContact: rapid estimate of contact and binding free energies. Bioinformatics. 2005;21: 2534–2536. doi:10.1093/bioinformatics/bti322
19. Meireles LMC, Dömling AS, Camacho CJ. ANCHOR: a web server and database for analysis of protein-protein interaction binding pockets for drug discovery. Nucleic Acids Res. 2010;38: W407-11. doi:10.1093/nar/gkq502
20. Bromberg Y, Rost B. Comprehensive in silico mutagenesis highlights functionally important residues in proteins. Bioinformatics. 2008;24: i207–i212. doi:10.1093/bioinformatics/btn268
21. Nooren IMA, Thornton JM. Diversity of protein-protein interactions. EMBO J. 2003;22: 3486–3492.
22. Acuner Ozbabacan SE, Engin HB, Gursoy A, Keskin O. Transient protein-protein interactions. Protein Eng Des Sel. 2011;24: 635–648. doi:10.1093/protein/gzr025
23. Perkins JR, Diboun I, Dessailly BH, Lees JG, Orengo C. Transient Protein-Protein Interactions: Structural, Functional, and Network Properties. Structure. Elsevier Ltd; 2010;18: 1233–1243. doi:10.1016/j.str.2010.08.007
24. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The Protein Data Bank. Nucleic Acids Res. 2000;28: 235–242.
25. Stark C, Breitkreutz B-J, Reguly T, Boucher L, Breitkreutz A, Tyers M. BioGRID: a general repository for interaction datasets. Nucleic Acids Res. 2006;34: D535-539. doi:10.1093/nar/gkj109
26. Stark C, Breitkreutz B-J, Chatr-Aryamontri A, Boucher L, Oughtred R, Livstone MS, et al. The BioGRID Interaction Database: 2011 update. Nucleic Acids Res. 2011;39:

D698-704. doi:10.1093/nar/gkq1116

27. Chatr-aryamontri A, Ceol A, Palazzi LM, Nardelli G, Schneider MV, Castagnoli L, et al. MINT: the Molecular INTeraction database. Nucleic Acids Res. 2007;35: D572-4. doi:10.1093/nar/gkl950
28. Bader GD, D. B, V. HCW. BIND: the Biomolecular Interaction Network Database. Nucleic Acids Res. 2003;31: 248–250. doi:10.1093/nar/gkg056
29. Xenarios I, Salwínski L, Duan XJ, Higney P, Kim S-M, Eisenberg D. DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. Nucleic Acids Res. 2002;30: 303–305.
30. Hermjakob H, Montecchi-Palazzi L, Lewington C, Mudali S, Kerrien S, Orchard S, et al. IntAct: an open source molecular interaction database. Nucleic Acids Res. 2004;32: D452-5. doi:10.1093/nar/gkh052
31. Walter P, Ansari S, Helms V. The ABC (Analysing Biomolecular Contacts) -Database. J Integr Bioinform. 2007;4(1): 50.
32. Davis FP, Sali A. PIBASE: a comprehensive database of structurally defined protein interfaces. Bioinformatics. 2005;21: 1901–7. doi:10.1093/bioinformatics/bti277
33. Higueruelo AP, Jubb H, Blundell TL. TIMBAL v2: update of a database holding small molecules modulating protein-protein interactions. Database (Oxford). 2013;2013: bat039. doi:10.1093/database/bat039
34. Hu L, Benson ML, Smith RD, Lerner MG, Carlson HA. Binding MOAD (Mother Of All Databases). Proteins. 2005;60: 333–40. doi:10.1002/prot.20512
35. Benson ML, Smith RD, Khazanov NA, Dimcheff B, Beaver J, Dresslar P, et al. Binding MOAD, a high-quality protein-ligand database. Nucleic Acids Res. 2008;36: D674-8. doi:10.1093/nar/gkm911
36. Basse MJ, Betzi S, Bourgeas R, Bouzidi S, Chetrit B, Hamon V, et al. 2P2Idb: a structural database dedicated to orthosteric modulation of protein-protein interactions. Nucleic Acids Res. 2013;41: D824-7. doi:10.1093/nar/gks1002
37. Wang R, Fang X, Lu Y, Yang CY, Wang S. The PDBbind database: Methodologies and updates. J Med Chem. 2005;48: 4111–4119. doi:10.1021/jm048957q
38. Liu T, Lin Y, Wen X, Jorissen RN, Gilson MK. BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. Nucleic Acids Res. 2007;35: D198-201. doi:10.1093/nar/gkl999
39. Gilson MK, Liu T, Baitaluk M, Nicola G, Hwang L, Chong J. BindingDB in 2015: A public database for medicinal chemistry, computational chemistry and systems

pharmacology. Nucleic Acids Res. 2015;44: gkv1072. doi:10.1093/nar/gkv1072

40. Hendlich M, Bergner A, Günther J, Klebe G. Relibase: Design and Development of a Database for Comprehensive Analysis of Protein–Ligand Interactions. J Mol Biol. 2003;326: 607–620. doi:10.1016/S0022-2836(02)01408-0
41. Perot S, Sperandio O, Miteva MA, Camproux AC, Villoutreix BO. Druggable pockets and binding site centric chemical space: A paradigm shift in drug discovery. Drug Discov Today. 2010;15: 656–667. doi:10.1016/j.drudis.2010.05.015
42. Zheng X, Gan L, Wang E, Wang J. Pocket-based drug design: exploring pocket space. AAPS J. 2013;15: 228–41. doi:10.1208/s12248-012-9426-6
43. Konc J, Janezic D. Binding site comparison for function prediction and pharmaceutical discovery. Curr Opin Struct Biol. 2014;25: 34–39. doi:10.1016/j.sbi.2013.11.012
44. Fischer E. Einfluss der Configuration auf die Wirkung der Enzyme. Ber Dtsch Chem Ges. 1894;27: 2985–2993. doi:10.1002/cber.18940270364
45. Xie Z-R, Hwang M-J. Ligand-binding site prediction using ligand-interacting and binding site-enriched protein triangles. Bioinformatics. 2012;28: 1579–85. doi:10.1093/bioinformatics/bts182
46. Lazo JS, Sharlow ER. Drugging undruggable molecular cancer targets. Annu Rev Pharmacol Toxicol. 2016;56: 23–40. doi:10.1146/annurev-pharmtox-010715-103440
47. DeLano W. Pymol: An open-source molecular graphics tool. CCP4 Newsl Protein Crystallogr. 2002;700.
48. Volkamer A, Griewel A, Grombacher T, Rarey M. Analyzing the topology of active sites: On the prediction of pockets and subpockets. J Chem Inf Model. 2010;50: 2041–2052. doi:10.1021/ci100241y
49. Kawabata T. Detection of multiscale pockets on protein surfaces using mathematical morphology. Proteins Struct Funct Bioinforma. 2010;78: 1195–1211. doi:10.1002/prot.22639
50. Yu J, Zhou Y, Tanaka I, Yao M. Roll: A new algorithm for the detection of protein pockets and cavities with a rolling probe sphere. Bioinformatics. 2009;26: 46–52. doi:10.1093/bioinformatics/btp599
51. Schmidtke P, Barril X. Understanding and predicting druggability. A high-throughput method for detection of drug binding sites. J Med Chem. 2010;53: 5858–67. doi:10.1021/jm100574m
52. Kuntz ID, Blaney JM, Oatley SJ, Langridge R, Ferrin TE. A geometric approach to macromolecule-ligand interactions. J Mol Biol. 1982;161: 269–288. doi:10.1016/0022-

2836(82)90153-X

53. Laskowski RA, Luscombe NM, Swindells MB, Thornton JM. Protein clefts in molecular recognition and function. Protein Sci. 1996;5: 2438–52. doi:10.1002/pro.5560051206
54. Hendlich M, Rippmann F, Barnickel G. LIGSITE: Automatic and efficient detection of potential small molecule-binding sites in proteins. J Mol Graph Model. 1997;15: 359–363. doi:10.1016/S1093-3263(98)00002-3
55. Weisel M, Proschak E, Schneider G. PocketPicker: analysis of ligand binding-sites with shape descriptors. Chem Cent J. 2007;1: 7. doi:10.1186/1752-153X-1-7
56. Volkamer A, Kuhn D, Rippmann F, Rarey M. Dogsitescorer: A web server for automatic binding site prediction, analysis and druggability assessment. Bioinformatics. 2012;28: 2074–2075. doi:10.1093/bioinformatics/bts310
57. Tan KP, Varadarajan R, Madhusudhan MS. DEPTH: A web server to compute depth and predict small-molecule binding cavities in proteins. Nucleic Acids Res. 2011;39: 1–7. doi:10.1093/nar/gkr356
58. Tan KP, Nguyen TB, Patel S, Varadarajan R, Madhusudhan MS. Depth: a web server to compute depth, cavity sizes, detect potential small-molecule ligand-binding cavities and predict the pKa of ionizable residues in proteins. Nucleic Acids Res. 2013;41: 314–321. doi:10.1093/nar/gkt503
59. Kalidas Y, Chandra N. PocketDepth: A new depth based algorithm for identification of ligand binding sites in proteins. J Struct Biol. 2008;161: 31–42. doi:10.1016/j.jsb.2007.09.005
60. Laskowski RA. SURFNET: A program for visualizing molecular surfaces, cavities, and intermolecular interactions. J Mol Graph. 1995;13: 323–330. doi:10.1016/0263-7855(95)00073-9
61. Kawabata T, Go N. Detection of pockets on protein surfaces using small and large probe spheres to find putative ligand binding sites. Proteins. 2007;68: 516–529. doi:10.1002/prot
62. Brady GP, Stouten PF. Fast prediction and visualization of protein binding pockets with PASS. J Comput Aided Mol Des. 2000;14: 383–401.
63. Chartier M, Najmanovich R. Detection of Binding Site Molecular Interaction Field Similarities. J Chem Inf Model. 2015;55: 1600–1615. doi:10.1021/acs.jcim.5b00333
64. Edelsbrunner H, Mücke EP. Three-dimensional alpha shapes. ACM Trans Graph. 1994;13: 43–72. doi:10.1145/174462.156635
65. Edelsbrunner H, Facello M, Liang J. On the definition and the construction of pockets

in macromolecules. Computer (Long Beach Calif). 1995; 1–16.

66. Peters KP, Fauck J, Frömmel C. The automatic search for ligand binding sites in proteins of known three-dimensional structure using only geometric criteria. J Mol Biol. 1996;256: 201–213. doi:10.1006/jmbi.1996.0077
67. Liang J, Edelsbrunner H, Woodward C. Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for ligand design. Protein Sci. 1998;7: 1884–1897. doi:10.1002/pro.5560070905
68. Le Guilloux V, Schmidtke P, Tuffery P. Fpocket: an open source platform for ligand pocket detection. BMC Bioinformatics. 2009;10: 168. doi:10.1186/1471-2105-10-168
69. Olsson TSG, Churchill AME, Pitt WR, Ladbury JE, Williams MA. ProACT2 - Analysis of solvent accessibility, cavities and contacts in proteins and their complexes. submitted. 2012;
70. Williams MA, Goodfellow JM, Thornton JM. Buried waters and internal cavities in monomeric proteins. Protein Sci. 1994;3: 1224–1235. doi:10.1002/pro.5560030808
71. Laurie ATR, Jackson RM. Q-SiteFinder: An energy-based method for the prediction of protein-ligand binding sites. Bioinformatics. 2005;21: 1908–1916. doi:10.1093/bioinformatics/bti315
72. Glaser F, Pupko T, Paz I, Bell RE, Bechor-Shental D, Martz E, et al. Consurf: identification of functional regions in proteins by surface-mapping of phylogenetic information. Bioinformatics. 2003;19: 163–164.
73. Berezin C, Glaser F, Rosenberg J, Paz I, Pupko T, Fariselli P, et al. ConSeq: The identification of functionally and structurally important residues in protein sequences. Bioinformatics. 2004;20: 1322–1324. doi:10.1093/bioinformatics/bth070
74. Ashkenazy H, Erez E, Martz E, Pupko T, Ben-Tal N. ConSurf 2010: Calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. Nucleic Acids Res. 2010;38: 529–533. doi:10.1093/nar/gkq399
75. de Rinaldis M, Ausiello G, Cesareni G, Helmer-Citterich M. Three-dimensional profiles: a new tool to identify protein surface similarities. J Mol Biol. 1998;284: 1211–21. doi:10.1006/jmbi.1998.2248
76. Innis CA. SiteFiNDER|3D: A web-based tool for predicting the location of functional sites in proteins. Nucleic Acids Res. 2007;35: 489–494. doi:10.1093/nar/gkm422
77. Fukunishi Y, Nakamura H. Prediction of ligand-binding sites of proteins by molecular docking calculation for a random ligand library. Protein Sci. 2011;20: 95–106. doi:10.1002/pro.540

78. Hetenyi C, Van Der Spoel D. Blind docking of drug-sized compounds to proteins with up to a thousand residues. FEBS Lett. 2006;580: 1447–1450. doi:10.1016/j.febslet.2006.01.074
79. Ben-Shimon A, Niv MY. AnchorDock: Blind and flexible anchor-driven peptide docking. Structure. Elsevier Ltd; 2015;23: 929–40. doi:10.1016/j.str.2015.03.010
80. Xie ZR, Hwang MJ. Ligand-binding site prediction using ligand-interacting and binding site-enriched protein triangles. Bioinformatics. 2012;28: 1579–1585. doi:10.1093/bioinformatics/bts182
81. Li L, Huang Y, Xiao Y. How to use not-always-reliable binding site information in protein-protein docking prediction. PLoS One. 2013;8. doi:10.1371/journal.pone.0075936
82. Del Sol Mesa A, Pazos F, Valencia A. Automatic methods for predicting functionally important residues. J Mol Biol. 2003;326: 1289–1302. doi:10.1016/S0022-2836(02)01451-1
83. Huang B. MetaPocket: a meta approach to improve protein ligand binding site prediction. OMICS. 2009;13: 325–30. doi:10.1089/omi.2009.0045
84. Huang B, Schroeder M. LIGSITEcsc: predicting ligand binding sites using the Connolly surface and degree of conservation. BMC Struct Biol. 2006;6: 19. doi:10.1186/1472-6807-6-19
85. Zhang Z, Li Y, Lin B, Schroeder M, Huang B. Identification of cavities on protein surface using multiple computational approaches for drug binding site prediction. Bioinformatics. 2011;27: 2083–2088. doi:10.1093/bioinformatics/btr331
86. Levine M, Tijan R. Transcription regulation and animal diversity. Nature. 2003;424: 147–151.
87. Nimwegen E van. Scaling laws in the functional content of genomes. Trends Genet. 2003;19: 479–484. doi:10.1016/S0168-9525(03)00202-6
88. Wasserman WW, Sandelin A. Applied bioinformatics for the identification of regulatory elements. Nat Rev Genet. 2004;5: 276–287. doi:10.1038/nrg1315
89. Dreos R, Ambrosini G, Cavin Perier R, Bucher P. EPD and EPDnew, high-quality promoter resources in the next-generation sequencing era. Nucleic Acids Res. 2013;41: D157–D164. doi:10.1093/nar/gks1233
90. Bryne JC, Valen E, Tang M-HE, Marstrand T, Winther O, da Piedade I, et al. JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. Nucleic Acids Res. 2008;36: D102-6. doi:10.1093/nar/gkm955

91. Kulakovskiy I V, Medvedeva YA, Schaefer U, Kasianov AS, Vorontsov IE, Bajic VB, et al. HOCOMOCO: a comprehensive collection of human transcription factor binding sites models. Nucleic Acids Res. 2013;41: D195-202. doi:10.1093/nar/gks1089
92. Matys V, Fricke E, Geffers R, Gößling E, Haubrock M, Hehl R, et al. TRANSFAC®: Transcriptional regulation, from patterns to profiles. Nucleic Acids Res. 2003;31: 374–378. doi:10.1093/nar/gkg108
93. Kilpatrick AM. A novel stochastic and entropy-based Expecttion-Maximisation algorithm for transcription factor binding site motif discovery. 2014. doi:10.1007/s13398-014-0173-7.2
94. D'haeseleer P. What are DNA sequence motifs? Nat Biotechnol. 2006;24: 423–425. doi:10.1038/nbt0406-423
95. Rapaport F, Khanin R, Liang Y, Pirun M, Krek A, Zumbo P, et al. Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. Genome Biol. 2013;14: R95. doi:10.1186/gb-2013-14-9-r95
96. Robinson MD, McCarthy DJ, Smyth GK. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2009;26: 139–140. doi:10.1093/bioinformatics/btp616
97. Anders S, Huber W, Nagalakshmi U, Wang Z, Waern K, Shou C, et al. Differential expression analysis for sequence count data. Genome Biol. 2010;11: R106. doi:10.1186/gb-2010-11-10-r106
98. Varet H, Coppée J-Y, Dillies M-A. SARTools: a DESeq2- and edgeR-based R pipeline for comprehensive differential analysis of RNA-Seq data. PLoS One. 2016;11. doi:10.1371/journal.pone.0157022
99. Anders S, McCarthy DJ, Chen Y, Okoniewski M, Smyth GK, Huber W, et al. Count-based differential expression analysis of RNA sequencing data using R and Bioconductor. Nat Protoc. 2013;8: 1765–1786. doi:10.1038/nprot.2013.099
100. Dillies MA, Rau A, Aubert J, Hennequet-Antier C, Jeanmougin M, Servant N, et al. A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. Brief Bioinform. 2013;14: 671–683. doi:10.1093/bib/bbs046
101. Benjamin Y, Hochberg Y. Controling the False Discovery Rate: A Practical and POwerful Approach to Multiple Testing. J R Stat Soc Ser C (Applied Stat. 1995;57: 289–300. doi:10.1038/203024b0
102. Yu H, Liu B-H, Ye Z-Q, Li C, Li Y-X, Li Y-Y. Link-based quantitative methods to identify differentially coexpressed genes and gene pairs. BMC Bioinformatics. 2011;12:

315. doi:10.1186/1471-2105-12-315

103. Amar D, Safer H, Shamir R. Dissection of regulatory networks that are altered in disease via differential co-expression. PLoS Comput Biol. 2013;9. doi:10.1371/journal.pcbi.1002955

104. Tesson BM, Breitling R, Jansen RC. DiffCoEx: a simple and sensitive method to find differentially coexpressed gene modules. BMC Bioinformatics. 2010;11: 497. doi:10.1186/1471-2105-11-497

105. Reverter A, Ingham A, Lehnert SA, Tan SH, Wang Y, Ratnakumar A, et al. Simultaneous identification of differential gene expression and connectivity in inflammation, adipogenesis and cancer. Bioinformatics. 2006;22: 2396–2404. doi:10.1093/bioinformatics/btl392

106. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics. 2008;9: 559. doi:10.1186/1471-2105-9-559

107. DiLeo M V., Strahan GD, den Bakker M, Hoekenga OA. Weighted correlation network analysis (WGCNA) applied to the tomato fruit metabolome. PLoS One. 2011;6. doi:10.1371/journal.pone.0026683

108. Mohamed R, Degac J, Helms V. Composition of overlapping protein-protein and protein-ligand interfaces. PLoS One. 2015;10: 4–6. doi:10.1371/journal.pone.0140965

109. Nooren IMA, Thornton JM. Structural characterisation and functional significance of transient protein–protein Interactions. J Mol Biol. 2003;325: 991–1018. doi:10.1016/S0022-2836(02)01281-0

110. Jones S, Daley DT, Luscombe NM, Berman HM, Thornton JM. Protein-RNA interactions: a structural analysis. Nucleic Acids Res. 2001;29: 943–54.

111. Petrescu A-J, Milac A-L, Petrescu SM, Dwek RA, Wormald MR. Statistical analysis of the protein environment of N-glycosylation sites: implications for occupancy, structure, and folding. Glycobiology. 2004;14: 103–14. doi:10.1093/glycob/cwh008

112. Lejeune D, Delsaux N, Charloteaux B, Thomas A, Brasseur R. Protein-nucleic acid recognition: statistical analysis of atomic interactions and influence of DNA structure. Proteins. 2005;61: 258–71. doi:10.1002/prot.20607

113. Davis FP, Sali A. The Overlap of Small Molecule and Protein Binding Sites within Families of Protein Structures. PLoS Comput Biol. 2010;6. doi:10.1371/journal.pcbi.1000668

114. Walter P, Metzger J, Thiel C, Helms V. Predicting where small molecules bind at protein-protein interfaces. PLoS One. 2013;8: e58583.

doi:10.1371/journal.pone.0058583

115. Koes DR, Camacho CJ. Small-molecule inhibitor starting points learned from protein-protein interaction inhibitor structure. Bioinformatics. 2012;28: 784–91. doi:10.1093/bioinformatics/btr717
116. Huang Y, Niu B, Gao Y, Fu L, Li W. CD-HIT Suite: a web server for clustering and comparing biological sequences. Bioinformatics. 2010;26: 680–2. doi:10.1093/bioinformatics/btq003
117. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics. 2006;22: 1658–9. doi:10.1093/bioinformatics/btl158
118. Holland RCG, Down TA, Pocock M, Prlić A, Huen D, James K, et al. BioJava: an open-source framework for bioinformatics. Bioinformatics. 2008;24: 2096–7. doi:10.1093/bioinformatics/btn397
119. Hubbard S, Thornton JM. Naccess. v2.1.1 [Internet]. Manchester U.K; 1996.
120. Eisenberg D, Weiss RM, Terwilliger TC, Wilcox W. Hydrophobic moments and protein structure. Faraday Symp Chem Soc. 1982;17: 109. doi:10.1039/fs9821700109
121. Higueruelo AP, Schreyer A, Bickerton GRJ, Blundell TL, Pitt WR. What can we learn from the evolution of protein-ligand interactions to aid the design of new therapeutics? PLoS One. 2012;7: e51742. doi:10.1371/journal.pone.0051742
122. Eyrisch S, Helms V. Transient pockets on protein surfaces involved in protein-protein interaction. J Med Chem. 2007;50: 3457–64. doi:10.1021/jm070095g
123. Yan C, Wu F, Jernigan RL, Dobbs D, Hanovar V. Characterization of Protein-protein Interfaces. Protein J. 2008;27: 59–70. doi:10.1007/s10930-007-9108-x.Characterization
124. Bickerton GR, Higueruelo AP, Blundell TL. Comprehensive, atomic-level characterization of structurally characterized protein-protein interactions: the PICCOLO database. BMC Bioinformatics. BioMed Central Ltd; 2011;12: 313. doi:10.1186/1471-2105-12-313
125. Janin J, Bahadur RP, Chakrabarti P. Protein-protein interaction and quaternary structure. Q Rev Biophys. 2008;41: 133–180. doi:10.1017/S0033583508004708
126. Glaser F, Steinberg DM, Vakser IA, Ben-tal N. Residue Frequencies and Pairing Preferences at Protein-protein interfaces. PROTEINS Struct Funct Genet. 2001;43: 89–102.
127. Jones S, Thornton JM. Analysis of protein-protein interaction sites using surface patches. J Mol Biol. 1997;272: 121–32. doi:10.1006/jmbi.1997.1234

128. Jin L, Wang W, Fang G. Targeting protein-protein interaction by small molecules. Annu Rev Pharmacol Toxicol. 2014;54: 435–56. doi:10.1146/annurev-pharmtox-011613-140028
129. Bogan AA, Thorn KS. Anatomy of hot spots in protein interfaces. J Mol Biol. 1998;280: 1–9. doi:10.1006/jmbi.1998.1843
130. Jones S, Marin A, Thornton JM. Protein domain interfaces: characterization and comparison with oligomeric protein interfaces. Protein Eng. 2000;13: 77–82. doi:10.1093/protein/13.2.77
131. Talavera D, Robertson DL, Lovell SC. Characterization of protein-protein interaction interfaces from a single species. PLoS One. 2011;6: e21053. doi:10.1371/journal.pone.0021053
132. McCoy AJ, Chandana Epa V, Colman PM. Electrostatic complementarity at protein/protein interfaces. J Mol Biol. 1997;268: 570–584. doi:10.1006/jmbi.1997.0987
133. Sheinerman FB, Norel R, Honig B. Electrostatic aspects of protein-protein interactions. Curr Opin Struct Biol. 2000;10: 153–9.
134. Ofran Y, Rost B. Analysing six types of protein-protein interfaces. J Mol Biol. 2003;325: 377–387. doi:10.1016/S0022-2836(02)01223-8
135. Chen K, Kurgan L. Investigation of atomic level patterns in protein--small ligand interactions. PLoS One. 2009;4: e4473. doi:10.1371/journal.pone.0004473
136. Hakulinen R, Puranen S, Lehtonen J V, Johnson MS, Corander J. Probabilistic prediction of contacts in protein-ligand complexes. PLoS One. 2012;7: e49216. doi:10.1371/journal.pone.0049216
137. Bondi A. van der Waals Volumes and Radii. J Phys Chem. 1965;68: 441–451.
138. Parekh AB, Penner R. Store depletion and calcium influx. Physiol Rev. 1997;77: 901–30.
139. Putney JW. A model for receptor-regulated Calcium entry. Cell Calcium. 1986;7: 1–12.
140. Liou J, Kim ML, Won DH, Jones JT, Myers JW, Ferrell JE, et al. STIM is a Ca2+ sensor essential for Ca2+-store- depletion-triggered Ca2+ influx. Curr Biol. 2005;15: 1235–1241. doi:10.1016/j.cub.2005.05.055
141. Roos J, DiGregorio PJ, Yeromin A V., Ohlsen K, Lioudyno M, Zhang S, et al. STIM1, an essential and conserved component of store-operated Ca 2+ channel function. J Cell Biol. 2005;169: 435–445. doi:10.1083/jcb.200502019
142. Zhang SL, Yu Y, Roos J, Kozak JA, Deerinck TJ, Ellisman MH, et al. STIM1 is a Ca2+ sensor that activates CRAC channels and migrates from the Ca2+ store to the plasma

membrane. Nature. 2005;437: 902–5. doi:10.1038/nature04147

143. Prakriya M, Feske S, Gwack Y, Srikanth S, Rao A, Hogan PG. Orai1 is an essential pore subunit of the CRAC channel. Nature. 2006;443: 230–233. doi:10.1038/nature05122

144. Yeromin A V, Zhang SL, Jiang W, Yu Y, Safrina O, Cahalan MD. Molecular identification of the CRAC channel by altered ion selectivity in a mutant of Orai. Nature. 2006;443: 226–9. doi:10.1038/nature05108

145. Lis A, Peinelt C, Beck A, Parvez S, Monteilh-Zoller M, Fleig A, et al. CRACM1, CRACM2, and CRACM3 are store-operated Ca2+ channels with distinct functional properties. Curr Biol. 2007;17: 794–800. doi:10.1016/j.cub.2007.03.065

146. Mercer JC, DeHaven WI, Smyth JT, Wedel B, Boyles RR, Bird GS, et al. Large store-operated calcium selective currents due to co-expression of Orai1 or Orai2 with the intracellular calcium sensor, Stim1. J Biol Chem. 2006;281: 24979–24990. doi:10.1074/jbc.M604589200

147. Rosado JA, Diez R, Smani T, Jardín I. STIM and Orai1 variants in store-operated calcium entry. Front Pharmacol. 2016;6: 1–9. doi:10.3389/fphar.2015.00325

148. Williams RT, Manji SSM, Parker NJ, Hancock MS, Stekelenburg L Van, Eid J, et al. Identification and characterization of the STIM (stromal interaction molecule) gene family : coding for a novel class of transmembrane proteins. Biochem J. 2001;357: 673–685.

149. Dziadek MA, Johnstone LS. Biochemical properties and cellular localisation of STIM proteins. Cell Calcium. 2007;42: 123–132. doi:10.1016/j.ceca.2007.02.006

150. Yuan JP, Zeng W, Dorwart MR, Choi Y-J, Worley PF, Muallem S. SOAR and the polybasic STIM1 domains gate and regulate Orai channels. Nat Cell Biol. 2009;11: 337–43. doi:10.1038/ncb1842

151. Thiel M, Lis A, Penner R. STIM2 drives Ca2+ oscillations through store-operated Ca2+ entry caused by mild store depletion. J Physiol. 2013;591: 1433–45. doi:10.1113/jphysiol.2012.245399

152. Feske S, Skolnik EY, Prakriya M. Ion channels and transporters in lymphocyte function and immunity. Nat Rev Immunol. Nature Publishing Group; 2012;12: 532–47. doi:10.1038/nri3233

153. Hoth M, Niemeyer BA. The neglected CRAC proteins: Orai2, Orai3, and STIM2. Curr Top Membr. 2013;71: 237–271.

154. Vig M, Peinelt C, Beck A, Koomoa DL, Rabah D, Koblan-Huberson M, et al. CRACM1 Is a Plasma Membrane Protein Essential for Store-Operated Ca2+ Entry. Science (80-).

2006;312: 1120–1223. doi:10.1007/s13398-014-0173-7.2

155. Derler I, Jardin I, Romanin C. The molecular mechanisms of STIM/Orai communication. Am J Physiol Cell Physiol. 2016;310: C643–C662. doi:10.1152/ajpcell.00007.2016

156. Muik M, Fahrner M, Schindl R, Stathopulos P, Frischauf I, Derler I, et al. STIM1 couples to ORAI1 via an intramolecular transition into an extended conformation. EMBO J. Nature Publishing Group; 2011;30: 1678–1689. doi:10.1038/emboj.2011.79

157. Muik M, Frischauf I, Derler I, Fahrner M, Bergsmann J, Eder P, et al. Dynamic coupling of the putative coiled-coil domain of ORAI1 with STIM1 mediates ORAI1 channel activation. J Biol Chem. 2008;283: 8014–8022. doi:10.1074/jbc.M708898200

158. Berna-Erro A, Woodard GE, Rosado JA. Orais and STIMs: Physiological mechanisms and disease. J Cell Mol Med. 2012;16: 407–424. doi:10.1111/j.1582-4934.2011.01395.x

159. Bojarski L, Pomorski P, Szybinska A, Drab M, Skibinska-Kijek A, Gruszczynska-Biegala J, et al. Presenilin-dependent expression of STIM proteins and dysregulation of capacitative Ca2+ entry in familial Alzheimer's disease. Biochim Biophys Acta - Mol Cell Res. 2009;1793: 1050–1057. doi:10.1016/j.bbamcr.2008.11.008

160. Schuhmann MK, Stegner D, Berna-Erro A, Bittner S, Braun A, Kleinschnitz C, et al. Stromal interaction molecules 1 and 2 are key regulators of autoreactive T cell activation in murine autoimmune central nervous system inflammation. J Immunol. 2010;184: 1536–1542. doi:10.4049/jimmunol.0902161

161. Cheng KT, Alevizos I, Liu X, Swaim WD, Yin H, Feske S, et al. STIM1 and STIM2 protein deficiency in T lymphocytes underlies development of the exocrine gland autoimmune disease, Sjogren's syndrome. Proc Natl Acad Sci. 2012;109: 14544–14549. doi:10.1073/pnas.1207354109

162. Aytes A, Mollevi DG, Martinez-Iniesta M, Nadal M, Vidal A, Morales A, et al. Stromal interaction molecule 2 (STIM2) is frequently overexpressed in colorectal tumors and confers a tumor cell growth suppressor phenotype. Mol Carcinog. 2012;51: 746–753. doi:10.1002/mc.20843

163. Stanisz H, Saul S, Müller CSL, Kappl R, Niemeyer B a., Vogt T, et al. Inverse regulation of melanoma growth and migration by Orai1/STIM2-dependent calcium entry. Pigment Cell Melanoma Res. 2014;27: 442–453. doi:10.1111/pcmr.12222

164. Sun S, Zhang H, Liu J, Popugaeva E, Xu NJ, Feske S, et al. Reduced synaptic STIM2 expression and impaired store-operated calcium entry cause destabilization of mature spines in mutant presenilin mice. Neuron. Elsevier Inc.; 2014;82: 79–93.

doi:10.1016/j.neuron.2014.02.019

165. Yang S, Zhang JJ, Huang XY. Orai1 and STIM1 are critical for breast tumor cell migration and metastasis. Cancer Cell. Elsevier Inc.; 2009;15: 124–134. doi:10.1016/j.ccr.2008.12.019

166. Motiani RK, Abdullaev IF, Trebak M. A novel native store-operated calcium channel encoded by Orai3: selective requirement of Orai3 versus Orai1 in estrogen receptor-positive versus estrogen receptor-negative breast cancer cells. J Biol Chem. 2010;285: 19173–19183. doi:10.1074/jbc.M110.102582

167. McAndrew D, Grice DM, Peters A a, Smart CE, Brown M a, Kenny P a, et al. ORAI1-Mediated Calcium Influx in Lactation and in Breast Cancer ORAI1-Mediated Calcium Influx in Lactation and in Breast Cancer. Mol Cancer Ther. 2011;10: 448–460. doi:10.1158/1535-7163.MCT-10-0923

168. Faouzi M, Hague F, Potier M, Ahidouch A, Sevestre H, Ouadid-Ahidouch H. Down-regulation of Orai3 arrests cell-cycle progression and induces apoptosis in breast cancer cells but not in normal breast epithelial cells. J Cell Physiol. 2011;226: 542–551. doi:10.1002/jcp.22363

169. Bhardwaj R, Hediger MA, Demaurex N. Redox modulation of STIM-ORAI signaling. Cell Calcium. Elsevier Ltd; 2016; doi:10.1016/j.ceca.2016.03.006

170. Hannenhalli S. Eukaryotic transcription factor binding sites - Modeling and integrative search methods. Bioinformatics. 2008;24: 1325–1331. doi:10.1093/bioinformatics/btn198

171. Lachmann A, Xu H, Krishnan J, Berger SI, Mazloom AR, Maayan A. ChEA: Transcription factor regulation inferred from integrating genome-wide ChIP-X experiments. Bioinformatics. 2010;26: 2438–2444. doi:10.1093/bioinformatics/btq466

172. Wang J, Zhuang J, Iyer S, Lin XY, Greven MC, Kim BH, et al. Factorbook.org: A Wiki-based database for transcription factor-binding data generated by the ENCODE consortium. Nucleic Acids Res. 2013;41: 171–176. doi:10.1093/nar/gks1221

173. Feingold E, Good P, Guyer M, Kamholz S, Liefer L, Wetterstrand K, et al. The ENCODE (ENCyclopedia of DNA elements) Project. 2004;306: 636–640. doi:10.1126/science.1105136

174. Sandelin A, Alkema W, Engström P, Wasserman WW, Lenhard B. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. Nucleic Acids Res. 2004;32: 91D–94. doi:10.1093/nar/gkh012

175. Xia K, Dong D, Han J-DJ. IntNetDB v1.0: an integrated protein-protein interaction

network database generated by a probabilistic model. BMC Bioinformatics. 2006;7: 508. doi:10.1186/1471-2105-7-508

176. Calderone A, Castagnoli L, Cesareni G. Mentha: a Resource for Browsing Integrated Protein-Interaction Networks. Nat Methods. Nature Publishing Group; 2013;10: 690. doi:10.1038/nmeth.2561

177. Brown KR, Jurisica I. Online predicted human interaction database. Bioinformatics. 2005;21: 2076–2082. doi:10.1093/bioinformatics/bti273

178. Zhang QC, Petrey D, Garzón JI, Deng L, Honig B. PrePPI: A structure-informed database of protein-protein interactions. Nucleic Acids Res. 2013;41: 828–833. doi:10.1093/nar/gks1231

179. Mellor JC, Yanai I, Clodfelter KH, Mintseris J, DeLisi C. Predictome: a database of putative functional links between proteins. Nucleic Acids Res. 2002;30: 306–9. doi:10.1093/nar/30.1.306

180. McDowall MD, Scott MS, Barton GJ. PIPs: human protein-protein interaction prediction database. Nucleic Acids Res. 2009;37: D651-6. doi:10.1093/nar/gkn870

181. Franceschini A, Szklarczyk D, Frankild S, Kuhn M, Simonovic M, Roth A, et al. STRING v9.1: protein-protein interaction networks, with increased coverage and integration. Nucleic Acids Res. 2013;41: D808-15. doi:10.1093/nar/gks1094

182. Szklarczyk D, Franceschini A, Kuhn M, Simonovic M, Roth A, Minguez P, et al. The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. Nucleic Acids Res. 2011;39: D561-8. doi:10.1093/nar/gkq973

183. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape : A Software Environment for Integrated Models of Biomolecular Interaction Networks. Genome Res. 2003;13: 2498–2504. doi:10.1101/gr.1239303

184. Périer RC, Junier T, Bonnard C, Bucher P. The eukaryotic promoter database (EPD). Nucleic Acids Res. 1999;27: 307–309. doi:10.1093/nar/27.1.307

185. Grant CE, Bailey TL, Noble WS. FIMO: scanning for occurrences of a given motif. Bioinformatics. 2011;27: 1017–1018. doi:10.1093/bioinformatics/btr064

186. Bailey TL, Johnson J, Grant CE, Noble WS. The MEME Suite. Nucleic Acids Res. 2015;43: W39–W49. doi:10.1093/nar/gkv416

187. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15: 550. doi:10.1186/s13059-014-0550-8

188. Eylenstein A, Schmidt S, Gu S, Yang W, Schmid E, Schmidt EM, et al. Transcription

factor NF-κB regulates expression of pore-forming Ca 2+ channel unit, Orai1, and its activator, STIM1, to control Ca 2+ entry and affect cellular functions. J Biol Chem. 2012;287: 2719–2730. doi:10.1074/jbc.M111.275925

189. Chang W-C, Fang Y-Y, Chang H-W, Chuang L-Y, Lin Y-D, Hou M-F, et al. Identifying association model for single-nucleotide polymorphisms of ORAI1 gene for breast cancer. Cancer Cell Int. Cancer Cell International; 2014;14: 29. doi:10.1186/1475-2867-14-29
190. Hoth M. CRAC channels, calcium, and cancer in light of the driver and passenger concept. Biochim Biophys Acta - Mol Cell Res. Elsevier B.V.; 2016;1863: 1408–1417. doi:10.1016/j.bbamcr.2015.12.009
191. Dimova DK, Dyson NJ. The E2F transcriptional network: old acquaintances with new faces. Oncogene. 2005;24: 2810–2826. doi:10.1038/sj.onc.1208612
192. DeGregori J, Johnson DG. Distinct and overlapping roles for E2F family members in transcription, proliferation and apoptosis. Curr Mol Med. 2006;6: 739–748. doi:10.1038/nature08544
193. Oeckinghaus A, Ghosh S. The NF-kappaB family of transcription factors and its regulation. Cold Spring Harb Perspect Biol. 2009;1: 1–14. doi:10.1101/cshperspect.a000034
194. Dragoni S, Laforenza U, Bonetti E, Lodola F, Bottino C, Berra-Romani R, et al. Vascular endothelial growth factor stimulates endothelial colony forming cells proliferation and tubulogenesis by inducing oscillations in intracellular Ca2+ concentration. Stem Cells. 2011;29: 1898–907. doi:10.1002/stem.734
195. Parekh AB. Decoding cytosolic Ca2+ oscillations. Trends Biochem Sci. Elsevier Ltd; 2011;36: 78–87. doi:10.1016/j.tibs.2010.07.013
196. Clapham DE. Calcium Signaling. Cell. 2007;131: 1047–1058. doi:10.1016/j.cell.2007.11.028
197. Levy C, Khaled M, Fisher DE. MITF: master regulator of melanocyte development and melanoma oncogene. Trends Mol Med. 2006;12: 406–414. doi:10.1016/j.molmed.2006.07.008
198. Stanisz H, Vultur A, Herlyn M, Roesch A, Bogeski I. The role of Orai / STIM calcium channels in melanocytes and melanoma. J Physiol. 2016;0: 1–11. doi:10.1113/JP271141.This
199. Sotriffer C, Klebe G. Identification and mapping of small-molecule binding sites in proteins: computational tools for structure-based drug design. II Farm. 2002;57: 243–

251. doi:10.1016/S0014-827X(02)01211-9

200. Henrich S, Salo-Ahen OMH, Huang B, Rippmann FF, Cruciani G, Wade RC. Computational approaches to identifying and characterizing protein binding sites for ligand design. J Mol Recognit. 2009;23: 209–19. doi:10.1002/jmr.984

201. Coleman RG, Sharp KA. Protein pockets: inventory, shape, and comparison. J Chem Inf Model. 2010;50: 589–603. doi:10.1021/ci900397t

202. Levitt M, Park BH. Water: now you see it, now you don't. Curr Biol Struct. 1993;1: 223–226.

203. Pupko T, Bell RE, Mayrose I, Glaser F, Ben-Tal N. Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. Bioinformatics. 2002;18 Suppl 1: S71–S77. doi:10.1093/bioinformatics/18.suppl_1.S71

204. Oliveira SHP, Ferraz FA, Honorato R V, Xavier-Neto J, Sobreira TJP, de Oliveira PSL. KVFinder: steered identification of protein cavities as a PyMOL plugin. BMC Bioinformatics. 2014;15: 197. doi:10.1186/1471-2105-15-197

205. Degac J, Winter U, Helms V. Graph-based clustering of predicted ligand-binding pockets on protein surfaces. J Chem Inf Model. 2015;55: 1944–1952. doi:10.1021/acs.jcim.5b00045

206. Wang R, Fang X, Lu Y, Wang S. The PDBbind database: collection of binding affinities for protein-ligand complexes with known three-dimensional structures. J Med Chem. 2004;47: 2977–80. doi:10.1021/jm030580l

207. Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. BBA - Protein Struct. 1975;405: 442–451. doi:10.1016/0005-2795(75)90109-9

208. Murakami Y, Mizuguchi K. Applying the Naive Bayes classifier with kernel density estimation to the prediction of protein-protein interaction sites. Bioinformatics. 2010;26: 1841–1848. doi:10.1093/bioinformatics/btq302

209. Baldi P, Brunak S, Chauvin Y, Andersen C. AF, Nielsen H. Assessing the accuracy of prediction algorithms for classification: an overview. Bioinformatics. 2000;16: 412–424. doi:10.1093/bioinformatics/16.5.412

210. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. J Mach Learn Res. 2011;12: 2825–2830. doi:10.1007/s13398-014-0173-7.2

211. Abdo A, Chen B, Mueller C, Salim N, Willett P. Ligand-based virtual screening using

Bayesian inference networks. J Chem Inf Model. 2010;50: 1012–1020.

212. Ahmed A, Abdo A, Salim N. Ligand-based virtual screening using Bayesian inference network and reweighted fragments. Sci World J. 2012;2012: 1–7. doi:10.1100/2012/410914

213. Chen X, Wang M, Zhang H. The use of classification trees for bioinformatics. Wiley Interdiscip Rev Data Min Knowl Discov. 2011;1: 55–63. doi:10.1002/widm.14

214. Zahiri J, Bozorgmehr JH, Masoudi-Nejad A. Computational prediction of protein-protein interaction networks: algorithms and resources. Curr Genomics. 2013;14: 397–414. doi:10.2174/1389202911314060004

215. Breiman L. Random forests. Mach Learn. 2001;45: 5–32. doi:10.1023/A:1010933404324

216. Cortes C, Vapnik V. Support-vector networks. Mach Learn. 1995;20: 273–297. doi:10.1007/BF00994018

217. Jorissen RN, Gilson MK. Virtual screening of molecular databases using a Support Vector Machine. J Chem Inf Model. 2005;45: 549–561.

218. Villar HO, Kauvar LM. Amino acid preferences at protein binding sites. FEBS Lett. 1994;349: 125–30. doi:10.1016/0014-5793(94)00648-2

Supplementary Information

8.1 Supplementary Information for Chapter 3

Table 3.1 Datasets of PP and PL complexes.

Dataset of PP complexes from the ABC database **(Table A)**.

PDB ID	Chain	Chain									
15C8	L	H	1HE8	B	A	1R5T	A	B	2BL0	B	A
1A09	A	B	1HFY	A	B	1R8Q	A	E	2BTO	T	A
1A6U	H	L	1HJA	B	I	1RFX	B	C	2C3N	A	B
1AB8	B	A	1HQK	E	D	1RH7	C	B	2CHP	C	D
1AHW	F	A	1HWU	B	A	1RQ7	B	A	2COG	A	B
1ASL	B	A	1I3O	B	E	1S4Y	B	A	2CUY	B	A
1AV1	B	A	1I8F	C	B	1S7Y	B	A	2D26	A	B
1AVA	A	C	1ICF	A	I	1SC1	A	B	2D4V	B	A
1AY7	A	B	1IFV	A	B	1SGF	G	B	2DSQ	I	G
1AZZ	A	C	1IIN	A	B	1SGR	E	I	2DVG	C	B
1B06	A	B	1JSU	B	C	1SND	A	B	2EV4	B	A
1B99	F	C	1JTH	D	A	1STF	E	I	2FHZ	B	A
1BJQ	C	A	1JZD	B	C	1SUV	D	F	2FR6	C	D
1BMF	F	B	1KQD	A	B	1T6G	C	A	2G2U	A	B
1BRR	A	C	1KQM	C	A	1TH7	L	K	2GD4	I	H
1BVI	A	C	1KWS	B	A	1U3R	A	B	2GMR	L	M
1BVN	P	T	1MCI	B	A	1U3W	B	A	2GPV	D	B
1BZX	E	I	1MCV	A	I	1UGH	E	I	2GQD	B	A
1C2O	D	A	1MI3	B	A	1UVC	A	B	2GVM	B	A
1CKG	B	A	1MSA	B	C	1VGQ	B	A	2H1L	E	F
1CYY	A	B	1N8O	C	E	1VLZ	B	A	2HI7	A	B
1D7F	B	A	1N9S	C	D	1W0I	B	A	2I2R	H	D
1DFJ	E	I	1NHG	B	D	1W29	C	D	2IDO	A	B
1DLE	A	B	1NMA	N	H	1X1U	A	D	2IJO	A	I
1DM5	B	D	1NW9	A	B	1X1Z	A	B	2IPJ	B	A
1DPJ	A	B	1NYS	A	C	1XBY	B	A	2IWG	D	E
1E0F	E	J	1O6S	A	B	1XMZ	B	A	2IYG	B	A
1E8U	B	A	1O81	B	A	1XX9	B	D	2J3K	A	B
1EYS	H	M	1OBQ	B	A	1Y48	E	I	2J96	B	A
1EZL	C	D	1OKT	B	A	1YGP	A	B	2NX5	A	D
1F45	A	B	1OME	A	B	1YI5	C	H	2NY7	G	H
1F46	B	A	1OO9	A	B	1YRQ	A	H	2O8A	A	I
1F88	B	A	1OPF	B	A	1ZJD	A	B	2OCC	A	B
1FAK	H	I	1P0S	H	E	1ZRS	B	A	2OL1	A	C
1FI8	A	C	1P4I	L	H	1ZUX	D	B	2P1L	A	B
1FX9	B	A	1P7L	A	B	2AAX	A	B	2PRG	A	C
1G0A	D	B	1PA0	B	A	2AGY	B	D	2PVO	D	A
1G21	H	G	1PMO	D	C	2AMT	C	B	2TNF	B	A
1G3N	A	B	1PPF	E	I	2AQ3	G	A	3PCD	A	M
1G6V	A	K	1PST	M	H	2AY5	A	B	4AKE	A	B
1GIF	B	A	1PYG	B	A	2AYQ	B	A	6PFK	C	D
1GQ3	C	B	1QGE	D	E	2B42	B	A	7TIM	B	A
1GXS	D	C	1QPX	A	B	2BE6	A	D			
1H6A	B	A	1QRN	D	E	2BEX	C	A			

Dataset of PL complexes from the ABC database **(Table B)**.

PDB ID	Chain	Ligand									
1A05	B	IPM	1IT6	A	CYU	1TI1	A	D12	2DQV	A	GAL
1A0J	A	BEN	1IZ2	A	SUM	1TR5	A	THP	2FMH	A	TRS
1A8J	L	PME	1JI5	B	MPD	1TXC	A	2AN	2FNW	A	REP
1ANK	A	AMP	1JTK	A	THU	1U0H	B	ONM	2FYD	A	PG4
1AR1	A	LDA	1KJ1	D	MAN	1U3T	B	CCB	2G2Z	A	COZ
1BG9	A	BGC	1KMH	B	TTX	1UBH	S	MPD	2G7Y	A	MO9
1BIW	B	S80	1KYN	A	KTP	1USR	A	SIA	2GJ6	D	3IB
1BLC	A	CEM	1L7Z	A	MYR	1UTM	A	PEA	2GOO	C	NDG
1BMQ	A	MNO	1L9B	L	HTO	1UX0	A	THU	2H0T	A	EPE
1BQI	A	SBA	1L9H	A	HTO	1V3V	B	5OP	2H6Y	A	MPD
1BWO	A	LPC	1LBC	A	CYZ	1V84	A	NAG	2HA3	A	P6G
1C50	A	CHI	1LIN	A	TFP	1W5F	B	G2P	2HG8	A	MLE
1CGY	A	MAL	1LOJ	A	MPD	1WB8	A	PMS	2HQU	C	DUP
1CLS	D	DEC	1LVW	D	TYD	1WV0	A	BN4	2HXM	A	302
1CPC	A	CYC	1M2Z	A	BOG	1WV7	T	FUC	2I17	A	CIT
1CY2	A	TMP	1MBQ	A	BEN	1X29	B	PMG	2IPF	A	TRS
1DBN	B	NAG	1MFI	B	FHC	1XEY	A	GUA	2IW6	A	QQ2
1DHK	B	NAG	1MPF	A	C8E	1XJI	A	D10	2IWZ	A	6NA
1EKX	C	PAL	1NGP	L	NPA	1XKD	A	NAP	2J6E	B	MPD
1EST	A	TSU	1NIP	B	ADP	1XR8	A	PG4	2J7L	A	XC2
1EWY	C	FAD	1O4H	A	772	1XXS	B	STE	2J8C	M	GGD
1F42	A	MNB	1O5D	H	CR9	1Y11	A	1PE	2J9C	C	ATP
1FLJ	A	GSH	1O6T	A	MES	1Y2F	A	WAI	2JH0	D	701
1FQ6	A	0QF	1O9T	B	ATP	1YRX	B	D9G	2NY0	A	HEZ
1G4I	A	MPD	1OAU	J	DNF	1ZL0	A	TLA	2OIZ	A	TSR
1G5N	A	SGN	1PFK	A	ADP	1ZOM	A	339	2OL4	B	JPN
1G8I	A	P6G	1Q4J	A	GTX	1ZRK	A	367	2OM9	A	AJA
1GG6	C	APF	1Q6O	B	LG6	2A01	A	AC9	2OPY	A	CO9
1GKA	A	D12	1Q6Y	A	MPD	2APX	A	MLA	2P95	A	ME5
1GMR	A	2GP	1QIW	A	DPD	2AY9	B	5PV	2PL7	B	HTG
1GOY	A	3GP	1RE2	A	NAG	2AZ5	B	307	2UUE	B	GVC
1GZR	B	C15	1RH7	C	P6G	2B0U	B	MPD	2YXJ	A	N3C
1H1B	A	151	1RHM	B	NA4	2B45	X	EPE	35C8	L	NOX
1H48	C	CDI	1RTK	A	GBS	2BUQ	A	CAQ	3LJR	A	GGC
1HJ1	A	PMB	1RYD	A	GLC	2C01	X	ATP	4LIP	D	CCP
1HUR	A	GDP	1RZH	H	CDL	2C4L	A	SIA	4VGC	B	SRD
1HVV	A	TAR	1S57	B	EPE	2C97	B	MPD	6RNT	A	2AM
1HX0	A	AC1	1S9Q	B	CHD	2CL0	X	TRS	9RSA	B	ADU
1I5G	A	TS5	1SUP	A	PMS	2CZ5	B	CIT			
1I9B	D	EPE	1SVL	C	ADP	2DCY	A	TAR			
1ICR	A	NIO	1TB6	I	MPD	2DJH	A	UM3			

Dataset of PP complexes from the PIBASE database **(Table C)**.

PDB ID	Chain	Chain												
1YM4	A	C	1CA7	A	C	1EZR	A	B	1HAK	A	B	1M43	A	B
1A22	B	A	1CBK	A	B	1EZV	B	A	1HG4	A	D	1M48	A	B
1A2K	C	B	1CD1	A	B	1EZV	A	B	1HKV	A	B	1M56	A	B
1A2V	C	D	1CD9	B	A	1EZV	C	D	1HP0	A	B	1M7N	A	B
1A5A	A	B	1CH4	A	B	1EZV	D	C	1HR6	D	C	1M7W	A	B
1A5A	B	A	1CKI	A	B	1F1J	A	B	1HR6	A	E	1MCZ	A	B
1A7K	C	D	1CMV	A	B	1F2D	A	B	1HXM	A	B	1MF8	B	A
1A8G	A	B	1CN4	A	C	1F3T	C	D	1HXM	B	A	1MG2	O	N
1A8T	A	B	1CWQ	A	B	1F5P	A	B	1HYR	A	C	1MPY	A	C
1A99	B	A	1CYD	A	B	1F60	A	B	1I0Z	A	B	1MRU	A	B
1AB8	A	B	1D0E	A	B	1F6B	A	B	1I1R	A	B	1MZJ	A	B
1ACB	E	I	1D0I	A	C	1F6R	A	C	1I3R	D	C	1MZN	A	C
1AE1	A	B	1D0O	A	B	1F8U	A	B	1I41	F	G	1N0H	A	B
1AFS	A	B	1D1Z	A	B	1FJG	E	H	1I4U	A	B	1N0S	A	B
1AG1	O	T	1D4A	A	C	1FR1	A	B	1I7I	A	B	1N1A	A	B
1AGR	A	D	1D6S	A	B	1FRT	A	B	1I85	D	B	1N4O	A	B
1AIG	P	O	1D6U	A	B	1FSK	J	L	1I85	A	B	1N7G	A	B
1AIG	L	H	1D8U	A	B	1FTM	A	C	1II6	A	B	1N9E	A	B
1AIP	A	C	1DBQ	A	B	1FUJ	A	D	1IJL	A	B	1NBU	B	C
1AJS	A	B	1DCL	A	B	1FVU	B	A	1IY8	B	C	1NCC	N	H
1AR1	B	A	1DF8	A	B	1FX0	B	A	1IYK	A	B	1NF3	B	D
1AUW	C	D	1DIR	A	B	1G1A	A	B	1IZ1	A	P	1NFD	A	D
1AZS	A	B	1DJU	A	B	1G2O	A	C	1J7E	A	B	1NFQ	A	B
1AZS	C	B	1DN0	B	A	1G4A	B	A	1J90	A	B	1NKS	A	B
1AZZ	C	A	1DO5	A	C	1G5Q	A	D	1JDS	A	C	1NTO	C	D
1B3R	B	D	1DOF	B	C	1G6O	A	B	1JEB	A	D	1NW4	A	B
1B6C	D	C	1DOH	A	B	1G6Y	A	B	1JFF	A	B	1NX9	B	D
1B7G	O	Q	1DPJ	A	B	1G73	C	A	1JG8	A	B	1O4S	A	B
1B7T	Z	A	1DTY	A	B	1G85	A	B	1JKF	A	B	1O5D	L	T
1B7T	Y	A	1DUG	A	B	1G8Y	C	D	1JSW	A	D	1O5I	A	B
1B8G	A	B	1DVR	A	B	1G9M	C	G	1JT0	A	C	1O5O	A	B
1B9C	A	B	1DZB	A	B	1GEG	B	C	1JWH	A	D	1O61	A	B
1BAI	A	B	1E3I	A	B	1GH7	A	B	1JWI	B	A	1O63	A	B
1BB3	A	B	1E3U	A	C	1GMY	A	B	1K2O	A	B	1O6E	B	A
1BCC	E	D	1E5E	A	B	1GNX	A	B	1K3F	A	D	1O9J	A	B
1BCC	G	C	1E7W	A	B	1GP7	A	B	1KBU	A	B	1OAT	B	C
1BD3	A	C	1E8T	A	B	1GPM	A	B	1KCZ	A	B	1OD2	A	B
1BEB	A	B	1E9S	E	F	1GPW	A	B	1KFY	B	C	1ODL	A	B
1BGY	R	O	1ECE	A	B	1GR7	A	D	1KFY	D	B	1OE7	A	B
1BI7	A	B	1ECS	A	B	1GRI	A	B	1KFY	C	B	1OGA	D	E
1BJF	A	B	1EE0	A	B	1GTV	A	B	1KI9	A	B	1OGA	E	D
1BKJ	A	B	1EI1	A	B	1GWN	A	C	1KOB	A	B	1OJ6	A	B
1BKO	A	B	1EK6	A	B	1GYL	A	B	1KRU	A	C	1OME	A	B
1BMF	C	D	1EKX	B	C	1GZ6	A	B	1KSG	A	B	1OPL	A	B
1BML	A	C	1EM6	A	B	1GZM	A	B	1KXJ	A	B	1ORR	A	B
1BQH	G	H	1EOC	B	A	1H48	A	B	1L8X	A	B	1ORT	C	D
1BUQ	A	B	1EP1	A	B	1H4G	A	B	1LBH	A	B	1ORW	B	D
1BVR	A	B	1EQU	A	B	1H5B	A	B	1LLU	E	F	1OVL	A	B
1BYE	C	D	1EWK	A	B	1H5Q	A	B	1LRT	A	B	1OVM	B	D
1C0T	A	B	1EYS	M	C	1H8V	D	E	1LW6	E	I	1OYJ	C	D

Continued

1OZF	A	B
1P0K	A	B
1P0S	H	E
1P4B	L	H
1P4E	A	D
1P60	A	B
1P7C	A	B
1P93	A	C
1PKG	A	B
1PKQ	A	B
1PV4	C	D
1PWE	C	D
1PWX	B	C
1PY4	A	C
1PYT	B	A
1Q0C	A	C
1Q3D	A	B
1Q3G	A	B
1Q57	A	C
1Q5Q	C	B
1Q6T	A	B
1Q8M	A	B
1Q90	B	D
1QA9	A	B
1QPB	A	B
1QZF	C	D
1R4P	E	A
1R5K	A	C
1R8Q	E	A
1RD5	A	B
1RD7	A	B
1RE5	B	C
1RJN	A	B
1RKX	A	D
1RPY	A	B
1RQR	A	C
1S2G	A	B
1SB2	A	B
1SF2	A	B
1SGF	G	Z
1SL6	D	A
1SN0	A	C
1SP8	A	B
1SPG	B	A
1SPI	A	B
1SQB	B	I
1SQI	A	B
1SQL	D	E
1SQP	A	I
1SU2	A	B
1SVX	B	A
1SXG	A	D
1T2O	B	A
1T8S	A	B
1T91	A	D
1T97	A	B
1TAH	A	B
1TC0	A	B
1TEX	A	B
1TF7	B	C
1TFC	A	B
1TPZ	A	B
1TW2	A	B
1TXT	A	B
1TZ3	A	B
1U0L	A	C
1U1I	A	B
1U46	A	B
1U5Q	A	B
1UB7	A	B
1UFH	B	A
1UI5	A	B
1UIM	A	B
1UIU	A	B
1UKM	A	B
1UKV	Y	G
1UMO	A	B
1UPA	B	D
1URZ	B	C
1USI	A	C
1UTR	A	B
1UU0	A	B
1UVQ	A	B
1UZM	A	B
1V2I	A	B
1V4J	A	B
1VC8	A	B
1VDM	B	H
1VDW	A	B
1VEA	A	B
1VF5	B	A
1VHW	A	D
1VIQ	B	C
1VIY	A	B
1VKG	A	B
1VM6	B	D
1VPX	C	B
1W0M	B	D
1W2Z	A	B
1W59	A	B
1W6T	A	B
1W6U	A	B
1W91	F	G
1WC1	B	C
1WUE	A	B
1WYE	A	B
1WZ8	A	B
1X27	A	E
1X9F	J	K
1XCA	A	B
1XDK	B	A
1XED	A	B
1XG5	A	B
1XGM	A	B
1XI9	A	B
1XKQ	A	B
1XNX	A	B
1XSE	A	B
1XTT	B	C
1XXD	B	C
1XYG	B	C
1Y1M	A	B
1Y1P	A	B
1YDE	B	P
1YO6	A	B
1YP2	C	D
1YQ2	A	B
1YTA	A	B
1YTZ	C	I
1YVB	A	I
1YXM	A	B
1Z08	B	D
1Z1B	A	B
1Z5A	A	B
1Z7M	E	A
1Z7X	X	W
1ZEM	E	G
1ZPD	B	E
2A2G	A	C
2A7K	G	H
2A9K	A	B
2A9K	B	A
2AAW	A	C
2AG5	A	B
2ANC	C	E
2B4G	A	B
2BCK	A	D
2BD0	A	B
2BEX	C	A
2BTW	A	B
2BVC	D	E
2C7F	A	D
2CLT	A	B
2D1Y	B	C
2DFT	A	C
2DNS	A	B
2DW6	A	B
2DY4	D	B
2EGH	B	A
2EVO	A	B
2EW8	B	D
2F1G	A	B
2F73	A	B
2FG6	C	D
2FM6	A	B
2FYI	A	B
2G2U	A	B
2GF0	A	B
2GIC	A	B
2GJ7	E	B
2I6A	A	D
2IFA	A	B
2IG2	H	L
2IGO	C	D
2IJ2	B	A
2J12	B	A
2NUU	B	C
2O23	A	B
2O2Y	C	D
2OKR	A	D
2ONL	C	A
2PD3	B	C
2PMT	C	D
2TAA	A	B
3TAT	C	D
7MDH	D	C

Dataset of PL complexes from the PIBASE database (**Table D**).

PDB ID	Chain	Ligand								
1A07	A	PTR	1EE2	A	CHD	1LW5	C	PLG	1U2R	A	DDE
1A2D	A	PYX	1EEI	D	GAA	1M2Z	A	BOG	1U4S	A	BIH
1A50	A	FIP	1EFA	A	NPF	1MVN	A	PCO	1U7G	A	BOG
1A54	A	MDC	1EHK	A	BNG	1N0S	A	FLU	1U9V	A	IHE
1A69	A	FMB	1EHK	B	BNG	1N3I	B	DIH	1UAY	A	ADN
1A8G	A	2ZN	1ELU	A	PDA	1N3Z	A	ADN	1UTR	A	PCB
1A9X	B	CYG	1EUP	A	ASD	1NBP	A	MHC	1UYS	A	H1L
1AB8	A	FOK	1EYN	A	2AN	1NLP	C	MN8	1VAF	A	ARR
1ACM	A	PAL	1EYS	M	BGL	1NX9	A	AIC	1VEA	A	HBN
1AHI	A	CHO	1EYS	H	BGL	1OKE	A	BOG	1VKG	B	CRI
1AOE	A	GW3	1F40	A	GPI	1OSV	A	CHC	1W9D	M	SEH
1AQU	B	EST	1FL6	L	AAH	1PKG	A	PTR	1X8I	A	BMH
1AWB	B	IPD	1FP1	D	HCC	1PP9	A	BHG	1X8V	A	ESL
1B4W	A	BOG	1FUP	A	PMA	1PP9	B	BHG	1XK9	A	P34
1B59	A	OVA	1G9V	A	RQ3	1PP9	D	BHG	1YIK	A	MM1
1B9C	A	CRO	1GG5	A	E09	1PP9	F	BHG	1YK3	D	BOG
1B9I	A	PXG	1GSE	A	EAA	1PPJ	G	BHG	1YP2	A	PMB
1B9V	A	RA2	1H2S	A	BOG	1Q0B	A	NAT	1Z7Y	A	AA5
1BAV	A	BIP	1H47	A	GPP	1Q3A	C	NGH	2A8T	A	ADN
1BCC	E	BOG	1H4G	A	FXP	1QW9	A	KHP	2AEO	A	CPT
1BDA	A	2Z0	1H61	A	PDN	1QZR	A	CDX	2AJ8	B	SC3
1BL6	A	SB6	1HAK	A	K21	1R4Z	A	RIL	2BM2	C	PM2
1BLS	A	IPP	1HJ1	A	PMB	1R8Q	A	AFB	2BT4	B	CA2
1BM7	A	FLF	1HNU	A	REO	1R8Q	E	AFB	2BU3	A	3GC
1BQM	A	HBY	1HRK	A	CHD	1REJ	A	B1L	2BVC	F	P3S
1BT5	A	IM2	1HZX	A	BNG	1RKW	A	PNT	2BXE	A	1FL
1BUQ	A	NTH	1I53	A	RTC	1RQI	A	DST	2CLE	B	F6F
1BX4	A	ADN	1I9H	A	BNI	1RQR	B	5FD	2CVD	A	HQL
1BZJ	A	PIC	1IDT	A	CB1	1RRI	A	A45	2D97	A	IYR
1C1P	A	BAI	1IEC	A	SEB	1S2C	A	FLF	2EVC	A	FC3
1C50	A	CHI	1IYL	A	R64	1S2D	A	AR4	2F11	A	IEM
1C9C	A	PP3	1J0B	B	5PA	1S36	A	CEI	2FAK	H	SA1
1C9M	A	SEB	1J12	A	EBG	1SA0	A	CN2	2GFD	A	RDA
1CGK	A	NAR	1JNH	H	ECO	1SEZ	A	OMN	2GIC	A	IUM
1COW	B	AUR	1JR1	A	MOA	1SHJ	A	NXN	2IPK	A	4DP
1COW	F	AUR	1JUT	A	LYD	1SIH	A	MBQ	2IPK	B	4DP
1CQE	A	BOG	1JVN	A	143	1SJD	A	NPG	2IYY	A	S3P
1CRX	B	PO4	1JYV	A	145	1SQI	A	869	2OOW	A	OX4
1CZI	E	PHI	1K3T	C	BRZ	1SUX	A	BTS	2OQW	A	CS3
1D3H	A	A26	1KDT	A	DOC	1SZR	A	PLG	2ORK	B	IPD
1D4F	A	ADN	1KFY	C	BRS	1T1R	B	IMB	2OT1	A	N3P
1DFG	A	NDT	1KFY	D	BRS	1T3T	A	CYG	2VLE	A	DZN
1DHT	A	DHT	1KIC	A	NOS	1T6H	A	PHI	3PCI	A	IHB
1DX6	A	GNT	1KLL	A	MC	1T83	A	HG2	4AC9	A	CMH
1DY3	A	87Y	1KRP	A	PST	1T9B	A	1CS	6PRC	L	CEB
1E2N	A	RCA	1KRV	B	147	1T9H	A	IUM	11GS	A	EAA
1E7P	B	F3S	1KYN	A	KTP	1TAU	A	BGL			
1E81	E	M91	1L0L	C	FMX	1TC1	A	FMB			
1EBA	B	DBY	1LBC	B	CYZ	1TFQ	A	998			
1ECZ	B	BOG	1LKD	A	BP6	1TXC	B	2AN			

Dataset of PL complexes from the TIMBAL database (**Table E**).

PDB ID	Chain	Ligand						
1A6A	A	NAG	2LZG	A	13Q	3M1I	A	GTP
1A7X	B	FKA	2OU7	A	ANP	3MDY	A	LDN
1AB8	A	FOK	2P1M	B	IHP	3NC0	A	IPH
1AGR	A	GDP	2P26	A	NAG	3O47	A	GDP
1AZS	C	GSP	2TNF	A	TRS	3OSK	A	NAG
1CS4	A	MES	2VJE	D	FLC	3QAZ	B	NAG
1D8D	B	FII	2VOH	A	CIT	3QBR	X	NHE
1H2K	A	OGA	2W5Y	A	SAH	3QD6	A	NAG
1I8L	A	NAG	2WBE	C	ANP	3QTK	A	TFA
1IA0	K	ACP	2WKP	A	FMN	3SL9	A	IMD
1ICF	I	NAG	2XRP	B	GTP	3TX7	B	P6L
1L2I	A	ETC	2YDS	A	NAG	3U88	A	GGB
1O9C	A	FLC	2YEM	A	WSH	3UP0	A	D7S
1P93	A	DEX	2ZKW	A	CU1	3V4P	C	TRS
1PZN	B	IMD	3A9E	B	REA	3VNG	A	FUU
1QAB	E	RTL	3BEJ	A	MUF	4A9E	A	3PF
1R6N	A	434	3CLX	D	X22	4AY6	A	12V
1S1J	A	IQZ	3DZU	D	PLB	4DBB	A	ACY
1TCO	B	MYR	3DZU	A	REA	4DEW	A	LU2
1TW6	B	BTB	3F7Q	A	1PE	4DSN	A	GCP
1U27	A	4IP	3FCS	A	IMD	4E2T	A	EPE
1V7P	B	NAG	3GBG	A	PAM	4G1E	A	NAG
1XLS	E	TCD	3HOF	B	DHC	4G1M	A	NAG
1YJD	C	NAG	3HQR	A	OGA	4GMX	C	K85
1YSG	A	4FC	3HVL	A	SRL	4GS6	A	1FM
2BRQ	A	GSH	3ICI	A	MES	4GZ9	A	NAG
2ERJ	B	NAG	3IPQ	A	965	4H71	A	PXE
2GXA	A	ADP	3IT8	D	NAG	4JWL	A	HRC
2H61	E	PG4	3K6S	A	MAN	4LOO	A	SB4
2HBH	A	XE4	3L0L	A	HC3			

Table 3.2 Dataset of PP complexes from the ABC dataset
List of PP complexes from the ABC dataset comprised of 94 homodimer complexes **(Table A).**

No	PDB ID	Resolution (Å)	Name of homodimer	Scientific source	Chain one	Length	Chain two	Length
1	1GIF	1.90	Human Glycosylation-Inhibiting Factor	*Homo sapiens*	B	115	A	15
2	1A09	2.00	Peptide Ligands Of PP60(C-SRC) SH2 Domains	*Homo sapiens*	A	107	B	107
3	1AB8	2.20	Adenylyl Cyclase	*Homo sapiens*	B	220	A	220
4	1ASL	2.60	Aspartate Aminotransferase	*Escherichia coli*	B	396	A	396
5	1AV1	4.00	Apolipoprotein A-I	*Homo sapiens*	B	201	A	201
6	1B06	2.20	Protein (Superoxide Dismutase)	*Sulfolobus acidocaldarius DSM 639*	A	210	B	210
7	1B99	2.70	Protein (Nucleoside Diphosphate Kinase)	*Dictyostelium discoideum*	F	155	C	155
8	1BJQ	2.65	Lectin	*Vigna unguiculata subsp. cylindrica (sow-pea)*	C	253	A	253
9	1BRR	2.90	Protein (Bacteriorhodopsin)	*Halobacterium salinarum*	A	247	C	247
10	1BVI	1.90	Protein (Ribonuclease T1)	*Aspergillus oryzae*	A	104	C	104
11	1C2O	4.20	Acetylcholinesterase	*Electrophorus electricus*	D	539	A	539
12	1CKG	2.20	Protein (Lysozyme)	*Homo sapiens*	B	130	A	130
13	1CYY	2.15	DNA Topoisomerase I	*Escherichia coli K-12*	A	264	B	264
14	1D7F	1.90	Cyclodextrin Glucanotransferase	*Bacillus sp. 1011*	B	686	A	686
15	1DLE	2.10	Complement Factor B	*Homo sapiens*	A	298	B	298
16	1DM5	1.93	Annexin XII E105k Mutant Homohexamer	*Hydra vulgaris*	B	315	D	315
17	1E8U	2.00	Hemagglutinin-Neuraminidase	*Newcastle disease virus*	B	454	A	454
18	1EZL	2.00	Azurin	*Pseudomonas aeruginosa*	C	128	D	128
19	1F46	1.50	Cell Division Protein Zipa	*Escherichia coli*	B	140	A	140
20	1F88	2.80	Rhodopsin	*Bos taurus*	B	348	A	348
21	1FX9	2.00	Phospholipase A2, Major Isoenzyme	*Sus scrofa*	B	124	A	124
22	1G0A	2.04	Hemoglobin Beta Chain	*Bos taurus*	D	145	B	145
23	1G21	3.00	Nitrogenase Iron Protein	*Azotobacter vinelandii*	H	289	G	289

Continued

24	1GQ3	2.01	Aspartate Carbamoyltransferase	*Escherichia coli K-12*	C	310	B	310
25	1H6A	2.50	Precursor Form Of Glucose-Fructose Oxidoreductase	*Zymomonas mobilis*	B	433	A	433
26	1HFY	2.30	Alpha-Lactalbumin	*Capra hircus*	A	123	B	123
27	1HQK	1.60	6,7-Dimethyl-8-Ribityllumazine Synthase	*Aquifex aeolicus*	E	154	D	154
28	1HWU	2.10	PII Protein	*Herbaspirillum seropedicae*	B	112	A	112
29	1I8F	1.75	Putative Snrnp SM-Like Protein	*Pyrobaculum aerophilum*	C	81	B	81
30	1IFV	2.25	Protein LLR18B	*Lupinus luteus*	A	155	B	155
31	1IIN	2.10	Glucose-1-Phosphate Thymidylyltransferase	*Salmonella enterica*	A	292	B	292
32	1KQD	1.90	Oxygen-Insensitive NAD(P)H Nitroreductase	*Enterobacter cloacae*	A	217	B	217
33	1KWS	2.10	Beta-1,3-Glucuronyltransferase 3	*Homo sapiens*	B	261	A	261
34	1MCI	2.70	Immunoglobulin Lambda Dimer MCG (Light Chain)	*Homo sapiens*	B	216	A	216
35	1MI3	1.80	Xylose Reductase	*Candida tenuis*	B	319	A	319
36	1MSA	2.29	Agglutinin	*Galanthus nivalis*	B	109	C	109
37	1N9S	3.50	Small Nuclear Ribonucleoprotein F	*Saccharomyces cerevisiae*	C	93	D	93
38	1NYS	3.05	Activin Receptor	*Rattus norvegicus*	A	105	C	105
39	1O81	1.50	Tryparedoxin II	*Crithidia fasciculata*	B	152	A	152
40	1OBQ	1.85	Crustacyanin C1 Subunit	*Homarus gammarus*	B	181	A	181
41	1OKT	1.90	Glutathione S-Transferase	*Plasmodium falciparum*	B	211	A	211
42	1OME	2.30	Beta-Lactamase	*Staphylococcus aureus*	A	258	B	258
43	1OPF	3.20	Matrix Porin Outer Membrane Protein F	*Escherichia coli*	B	340	A	340
44	1P7L	2.50	S-Adenosylmethionine Synthetase	*EScherichia coli*	A	383	B	383
45	1PA0	2.20	Myotoxic Phospholipase A2-Like	*Bothrops pauloensis*	B	121	A	121
46	1PMO	2.30	Glutamate Decarboxylase Beta	*Escherichia coli A*	D	466	C	466

Continued

47	1PYG	2.87	Glycogen Phosphorylase B	*Oryctolagus cuniculus*	B	842	A	842
48	1QPX	2.40	Papd Chaperone	*Escherichia coli*	A	218	B	218
49	1R5T	2.00	Cytidine Deaminase	*Saccharomyces cerevisiae*	A	142	B	142
50	1RFX	2.00	Resistin	*Mus musculus*	B	94	C	94
51	1RH7	3.11	Resistin-Like Beta	*Mus musculus*	C	81	B	81
52	1RQ7	2.60	Cell Division Protein FTSZ	*Mycobacterium tuberculosis*	B	382	A	382
53	1S7Y	1.75	Glutamate Receptor, Ionotropic Kainate 2 Precursor	*Rattus norvegicus*	B	259	A	259
54	1SND	1.84	Staphylococcal Nuclease Dimer	*Staphylococcus aureus*	A	143	B	143
55	1TH7	1.68	Small Nuclear Riboprotein Protein	*Sulfolobus solfataricus*	L	81	K	81
56	1U3R	2.21	Estrogen Receptor Beta	*Homo sapiens*	A	241	B	241
57	1U3W	1.45	Alcohol Dehydrogenase Gamma Chain	*Homo sapiens*	B	374	A	374
58	1UVC	2.00	Nonspecific Lipid Transfer Protein	*Oryza sativa*	A	91	B	91
59	1VGQ	2.13	Formyl-Coenzyme A Transferase	*Oxalobacter formigenes*	B	427	A	427
60	1VLZ	2.05	Chey	*Escherichia coli*	B	128	A	128
61	1W0I	2.10	3-Oxoacyl Carrier Protein Synthase	*Arabidopsis thaliana*	B	431	A	431
62	1W29	2.30	6,7-Dimethyl-8-Ribityllumazine Synthase	*Mycobacterium tuberculosis*	C	160	D	160
63	1X1Z	1.45	Orotidine 5'-Phosphate Decarboxylase	*Methanothermobacter thermautotrophicus*	A	252	B	252
64	1XBY	1.58	3-Keto-L-Gulonate 6-Phosphate Decarboxylase	*Escherichia coli*	B	216	A	216
65	1XMZ	1.38	GFP-Like Chromoprotein FP595	*Anemonia sulcata*	B	243	A	243
66	1YGP	2.80	Yeast Glycogen Phosphorylase	*Saccharomyces cerevisiae*	A	879	B	879
67	1ZRS	1.50	Hypothetical Protein	*Pseudomonas aeruginosa*	B	317	A	317
68	1ZUX	1.85	Green To Red Photoconvertible GPF-Like Protein EOSFP	*Lobophyllia hemprichii*	D	226	B	226
69	2AAX	1.75	Mineralocorticoid Receptor	*Homo sapiens*	A	275	B	275
70	2AMT	2.30	2-C-Methyl-D-Erythritol 2,4-Cyclodiphosphate Synthase	*EScherichia coli*	C	159	B	159

Continued

71	2AQ3	2.30	T-Cell Receptor Beta Chain V	*Mus musculus*	G	112	A	112
72	2AY5	2.40	Aromatic Amino Acid Aminotransferase	*Paracoccus denitrificans*	A	394	B	394
73	2AYQ	3.00	3-Isopropylmalate Dehydrogenase	*Bacillus coagulans*	B	366	A	366
74	2C3N	1.50	Glutathione S-Transferase Theta 1	*Homo sapiens*	A	247	B	247
75	2CHP	2.00	Metalloregulation DNA-Binding Stress Protein	*Bacillus subtilis subsp. subtilis str. 168*	C	153	D	153
76	2COG	2.10	branched chain aminotransferase 1, cytosolic	*Homo sapiens*	A	386	B	386
77	2CUY	2.10	Malonyl CoA-[acyl carrier protein] transacylase	*Thermus thermophilus HB8*	B	305	A	305
78	2D4V	1.90	Isocitrate Dehydrogenase	*isocitrate dehydrogenase*	B	429	A	429
79	2DVG	2.78	Galactose-Binding Lectin	*Arachis hypogaea*	C	236	B	236
80	2EV4	2.28	Hypothetical Protein Rv1264/MT1302	*Mycobacterium tuberculosis*	B	222	A	222
81	2FR6	2.07	Cytidine Deaminase	*Mus musculus*	C	146	D	146
82	2GPV	2.85	Estrogen-Related Receptor Gamma	*Homo sapiens*	D	230	B	230
83	2GQD	2.30	3-Oxoacyl-[Acyl-Carrier-Protein] Synthase 2	*Staphylococcus aureus*	B	437	A	437
84	2GVM	2.30	Hydrophobin-1	*Trichoderma reesei*	B	75	A	75
85	2H1L	3.16	Large T Antigen	*Simian virus 40*	E	370	F	370
86	2IPJ	1.80	Aldo-Keto Reductase Family 1 Member C2	*Homo sapiens*	B	321	A	321
87	2IYG	2.30	Appa, Antirepressor Of PPSR, Sensor Of Blue Light	*Rhodobacter sphaeroides*	B	124	A	124
88	2J3K	2.80	NADP-Dependent Oxidoreductase P1	*Arabidopsis thaliana*	A	345	B	345
89	2J96	2.25	Phycoerythrocyanin Alpha Chain	*Mastigocladus laminosus*	B	162	A	162
90	2OL1	1.80	Deoxyuridine 5'-Triphosphate Nucleotidohydrolase	*Vaccinia virus*	A	147	C	147
91	2TNF	1.40	Protein (Tumor Necrosis Factor Alpha)	*Mus musculus*	B	156	A	156
92	4AKE	2.20	Adenylate Kinase	*Escherichia coli*	A	214	B	214
93	6PFK	2.60	Phosphofructokinase	*Geobacillus stearothermophilus*	C	319	D	319
94	7TIM	1.90	Triosephosphate Isomerase	*Saccharomyces cerevisiae*	B	247	A	247

List of PP complexes from the ABC dataset comprised of heterodimer complexes **(Table B)**.

No	PDB ID	Resolution (Å)	Chain one	Name of chain one	Length	Chain two	Name of chain two	Length
1	15C8	2.50	L	IGG 5C8 FAB (Light Chain)	213	H	IGG 5C8 FAB (Heavy Chain)	217
2	1A6U	2.10	H	B1-8 Fv (Heavy Chain)	120	L	B1-8 FV (Light Chain)	108
3	1AHW	3.00	F	Tissue Factor	219	A	Immunoglobulin FAB 5G9 (Light Chain)	214
4	1AVA	1.90	A	Barley Alpha-Amylase 2	403	C	Barley Alpha-Amylase/Subtilisin Inhibitor	181
5	1AY7	1.70	A	Guanyl-Specific Ribonuclease SA	96	B	Barstar	89
6	1AZZ	2.30	A	Collagenase	226	C	Ecotin	142
7	1BMF	2.85	F	Bovine Mitochondrial F1-Atpase	482	B	Bovine Mitochondrial F1-Atpase	510
8	1BVN	2.50	P	Alpha-amylase	496	T	Tendamistat	71
9	1BZX	2.10	E	Protein (Trypsin)	222	I	Protein (Bovine Pancreatic Trypsin Inhibitor)	58
10	1DFJ	2.50	E	Ribonuclease A	124	I	Ribonuclease Inhibitor	456
11	1DPJ	1.80	A	Proteinase A	329	B	Proteinase Inhibitor IA3 Peptide	33
12	1E0F	3.10	E	Thrombin	259	J	Haemadin	57
13	1EYS	2.20	H	Photosynthetic Reaction Center	259	M	Photosynthetic Reaction Center	324
14	1F45	2.80	A	Interleukin-12 Beta Chain	306	B	Interleukin-12 Alpha Chain	197
15	1FAK	2.10	H	Protein (Blood Coagulation Factor Viia)	254	I	Protein (5l15)	55
16	1FI8	2.20	A	Natural Killer Cell Protease 1	228	C	Ecotin	84
17	1G3N	2.90	A	Cyclin-Dependent Kinase 6	326	B	Cyclin-Dependent Kinase 6 Inhibitor	168
18	1G6V	3.50	A	Carbonic Anhydrase	260	K	Antibody Heavy Chain	126
19	1GXS	2.30	D	P-(S)-Hydroxymandelonitrile Lyase Chain B	158	C	P-(S)-Hydroxymandelonitrile Lyase Chain A	270
20	1HE8	3.00	B	Transforming Protein P21/H-Ras-1	166	A	Phosphatidylinositol 3-Kinase Catalytic Subunit, Gamma Isoform	965
21	1HJA	2.30	B	Alpha-Chymotrypsin	131	I	Ovomucoid Inhibitor	51
22	1I3O	2.70	B	Caspase 3	110	E	Baculoviral Iap Repeat-Containing Protein 4	121
23	1ICF	2.00	A	Protein (Cathepsin L: Heavy Chain)	175	I	Protein (Invariant Chain)	65

Continued

24	1JSU	2.30	B	Cyclin A	260	C	P27	84
25	1JTH	2.00	D	Syntaxin 1a	77	A	SNAP25	82
26	1JZD	2.30	B	Thiol:Disulfide Interchange Protein DSBC	220	C	Thiol:Disulfide Interchange Protein dsbd	132
27	1KQM	3.00	A	Myosin Heavy Chain	835	C	Myosin Essential Light Chain	156
28	1MCV	1.80	A	Elastase 1	240	I	HEI-TOE I	28
29	1N8O	2.00	C	Chymotrypsin A, C Chain	97	E	Ecotin	142
30	1NHG	2.43	B	Enoyl-Acyl Carrier Reductase	229	D	Enoyl-Acyl Carrier Reductase	60
31	1NMA	3.00	N	N9 Neuraminidase	388	H	Fab Nc10	122
32	1NW9	2.40	B	Catalytic Domain Of Caspase-9	238	A	A Inhibitor Of Apoptosis Protein 3	91
33	1O6S	1.80	A	Internalin A	461	B	E-Cadherin	105
34	1OO9	X	A	Stromelysin-1	168	B	Metalloproteinase Inhibitor 1	128
35	1P0S	2.80	H	Coagulation Factor X Precursor	254	E	Ecotin Precursor	142
36	1P4I	2.80	L	Antibody Variable Light Chain	135	H	Antibody Variable Light Chain	124
37	1PPF	1.80	E	Leukocyte Elastase	218	I	Ovomucoid Inhibitor	56
38	1PST	3.00	M	Photosynthetic Reaction Center	296	H	Photosynthetic Reaction Center	237
39	1QGE	1.70	D	Protein (Triacylglycerol Hydrolase)	222	E	Protein (Triacylglycerol Hydrolase)	97
40	1QRN	2.80	D	T-Cell Receptor, Alpha Chain	200	E	T-Cell Receptor, Beta Chain	243
41	1R8Q	1.86	A	ADP-Ribosylation Factor 1	181	E	Arno	203
42	1S4Y	2.30	B	Inhibin Beta A Chain	116	A	Activin Receptor Type IIB Precursor	98
43	1SC1	2.60	A	Interleukin-1 Beta Convertase	178	B	Interleukin-1 Beta Convertase	88
44	1SGF	3.15	G	Nerve Growth Factor	237	B	Nerve Growth Factor	118
45	1SGR	1.80	E	Streptomyces Griseus Proteinase B	185	I	Turkey Ovomucoid Inhibitor	51
46	1STF	2.40	E	Papain	212	I	Stefin B	98

Continued

47	1SUV	1.75	D	Serotransferrin, N-Lobe	329	F	Serotransferrin, C-Lobe	345
48	1T6G	1.80	C	Endo-1,4-Beta-Xylanase I	184	A	Xylanase Inhibitor	381
49	1UGH	1.90	E	Protein (Uracil-Dna Glycosylase)	223	I	Protein (Uracil-Dna Glycosylase Inhibitor)	82
50	1X1U	2.30	A	Ribonuclease	110	D	Barstar	89
51	1XX9	2.20	B	Coagulation Factor XI	238	D	Ecotin	142
52	1Y48	1.84	E	Subtilisin BPN'	281	I	Chymotrypsin Inhibitor 2	64
53	1YI5	4.20	C	Acetylcholine-Binding Protein	210	H	Long Neurotoxin 1	71
54	1YRQ	2.10	A	Periplasmic [NiFe] Hydrogenase Small Subunit	264	H	Periplasmic [Nife] Hydrogenase Large Subunit	549
55	1ZJD	2.60	A	Catalytic Domain of Coagulation Factor XI	237	B	Kunitz Protease Inhibitory Domain of Protease Nexin II	57
56	2AGY	1.10	B	Aromatic Amine Dehydrogenase	361	D	Aromatic Amine Dehydrogenase	135
57	2B42	2.50	B	Endo-1,4-beta-xylanase A	185	A	Xylanase Inhibitor-I	381
58	2BE6	2.50	A	Calmodulin 2	150	D	Voltage-Dependent L-Type Calcium Channel Alpha-1C Subunit	37
59	2BEX	1.99	C	Nonsecretory Ribonuclease	135	A	Ribonuclease Inhibitor	460
60	2BL0	1.75	B	Myosin Regulatory Light Chain	145	A	Major Plasmodial Myosin Heavy Chain	63
61	2BTO	2.50	T	Thioredoxin 1	108	A	Tubulin Btuba	473
62	2D26	3.20	A	Alpha-1-Antitrypsin	358	B	Alpha-1-Antitrypsin	36
63	2DSQ	2.80	I	Insulin-Like Growth Factor IB	70	G	Insulin-Like Growth Factor-Binding Protein 1	94
64	2FHZ	1.15	A	Colicin-E5 Immunity Protein	109	B	Colicin-E5	108
65	2G2U	1.60	A	Beta-lactamase SHV-1	265	B	Beta-lactamase inhibitory protein	165
66	2GD4	3.30	I	Antithrombin-III	443	H	Coagulation Factor, Stuart Factor, Stuart-Prower Factor, Contains: Factor X Light Chain; Factor X Heavy Chain; Activated Factor Xa Heavy Chain	241
67	2GMR	2.50	L	Photosynthetic Reaction Center Protein L Chain	281	M	Photosynthetic Reaction Center Protein M Chain	307
68	2HI7	3.70	A	Thiol:Disulfide Interchange Protein dsbA	189	H	Disulfide Bond Formation Protein B	176
69	2I2R	3.35	H	Kv Channel-Interacting Protein 1	180	D	Potassium Voltage-Gated Channel Subfamily D Member 3	144
70	2IDO	2.10	A	DNA Polymerase III Epsilon Subunit	186	C	Hot Protein	83

Continued

71	2IJO	2.30	A	Polyprotein	58	I	Pancreatic Trypsin Inhibitor	58
72	2IWG	2.35	D	IG Gamma-1 Chain C	207	E	52 Kda Ro Protein	181
73	2NX5	2.70	A	HLA-B35	276	D	ELS4 TCR Alpha Chain	188
74	2NY7	2.30	G	Envelope Glycoprotein Gp120	317	H	Antibody B12, Heavy Chain	230
75	2O8A	2.61	A	Serine/Threonine-Protein Phosphatase Pp1-Gamma Catalytic Subunit	329	I	Protein phosphatase inhibitor 2	206
76	2OCC	2.30	A	Cytochrome C Oxidase	514	B	Cytochrome C Oxidase	227
77	2P1L	2.50	A	Apoptosis Regulator Bcl-X	153	B	Beclin 1	31
78	2PRG	2.30	A	Peroxisome Proliferator Activated Receptor Gamma	271	C	Nuclear Receptor Coactivator SRC-1	88
79	2PVO	3.40	D	Ferredoxin-1	96	A	Ferredoxin-thioredoxin reductase, catalytic chain	110
80	3PCD	2.10	A	Protocatechuate 3,4-Dioxygenase	200	M	Protocatechuate 3,4-Dioxygenase	238

Table 3.3 Dataset of PP complexes from the PIBASE dataset
List of PP complexes from the PIBASE dataset were grouped into 335 homodimer complexes **(Table A)**.

No	PDB ID	Resolution (Å)	Name of homodimer	Scientific source	Chain one	Length	Chain two	Length
1	1YM4	2.25	Beta-Secretase 1	*Homo sapiens*	A	408	C	408
2	1A2V	2.40	Methylamine Oxidase	*Ogataea angusta*	C	655	D	655
3	1A7K	2.80	Glyceraldehyde-3-Phosphate Dehydrogenase	*Leishmania mexicana*	C	360	D	360
4	1A8G	2.50	HIV-1 Protease	*Human immunodefiency virus*	A	99	B	99
5	1A8T	2.55	Metallo-Beta-Lactamase	*Bacteroides fragilis*	A	232	B	232
6	1A99	2.20	Putrescine-Binding Protein	*Escherichia coli*	B	344	A	344
7	1AB8	2.20	Adenylyl Cyclase	*Rattus norvegia*	A	220	B	220
8	1AE1	2.40	Tropinone Reductase-1	*Datura stramonium*	A	273	B	273
9	1AFS	2.50	3-Alpha-Hydroxysteroid Dehydrogenase	*Rattus norvegia*	A	323	B	323
10	1AG1	2.36	Triosephosphate Isomerase	*Trypanosoma brucei*	O	250	T	250
11	1AGR	2.80	Guanine Nucleotide-Binding Protein G(I)	*Rattus norvegia*	A	353	D	353
12	1AJS	1.60	Aspartate Aminotransferase	*Sus scrofa*	A	412	B	412
13	1AUW	2.50	Delta 2 Crytallin	*Anas platyrhynchos*	C	468	D	468
14	1B3R	2.80	Protein (S-Adenosylhomocysteine Hydrolase)	*Rattus norvegia*	B	431	D	431
15	1B7G	2.05	Protein (Glyceraldehyde-3-phosphate dehydrogenase)	*Sulfolobus solfatarius*	O	340	Q	340
16	1B8G	2.37	Protein (1-Aminocyclopropane-1 carboxylate synthase)	*Malus domestica*	A	429	B	429
17	1B9C	2.40	Protein (Green Fluorescent Protein)	*Auquorea victoria*	A	238	B	238
18	1BAI	2.40	Protease	*Rous sarcoma virus*	A	124	B	124
19	1BB3	1.80	Lysozyme	*Homo sapiens*	A	130	B	130
20	1BD3	1.93	Uracil Phosphoribosyltransferase	*Toxoplasma gondii*	A	243	C	243
21	1BEB	1.80	Beta-Lactoglobulin	*Bos taurus*	A	162	B	162
22	1BJF	2.40	Neurocalcin Delta	*Bos taurus*	A	193	B	193
23	1BKJ	1.80	NADPH-Flavin Oxidoreductase	*Vibrio harveyi*	A	240	B	240

Continued

24	1BKO	2.75	Thymidylate Synthase A	*Bacillus subtilis*	A	278	B	278
25	1BQH	2.80	Protein (VSV8)	*Mus musculus*	G	129	H	129
26	1BUQ	X	Protein (3-Ketosteroid isomerase-19-nortestosterone-hemisuccinate)	*Comamonas testosteroni*	A	125	B	125
27	1BVR	2.80	Protein (Enoyl-Acyl Carrier Protein (ACP) Reductase)	*Mycobacterium tuberculosis*	A	268	B	268
28	1BYE	2.80	Protein (Glutathione S-transferase)	*Zea mays*	C	213	D	213
29	1C0T	2.70	HIV-1 Reverse Transcriptase	*Human immunodeficiency virus 1*	A	560	B	440
30	1CA7	2.50	Protein (Macrophage Migration Inhibitory Factor)	*Homo sapiens*	A	114	C	114
31	1CBK	2.00	Pyrophosphokinase	*Haemophilus influenzae*	A	160	B	160
32	1CH4	2.50	Module-Substituted Chimera Hemoglobin Beta-Alpha	*Homo sapiens*	A	146	B	146
33	1CKI	2.30	Casein Kinase I Delta	*Rattus norvegia*	A	317	B	317
34	1CMV	2.27	Human Cytomegalovirus Protease	*Human herpesvirus 5*	A	256	B	256
35	1CWQ	2.25	Bacteriorhodopsin ("M" State Intermediate In Combination With Ground State)	*Halobacterium salinarum*	A	248	B	248
36	1CYD	1.80	Carbonyl Reductase	*Mus musculus*	A	244	B	244
37	1D0E	3.00	Reverse Transcriptase	*Moloney murine leukemia virus*	A	259	B	259
38	1D0I	1.80	Type II 3-Dehydroquinate Hydratase	*Streptomyces coelicolor*	A	156	C	156
39	1D0O	1.95	Bovine Endothelial Nitric Oxide Synthase Heme	*Bos taurus*	A	444	B	444
40	1D1Z	1.40	SAP SH2 Domain	*Homo sapiens*	A	104	B	104
41	1D4A	1.70	Quinone Reductase	*Homo sapiens*	A	273	C	273
42	1D6S	2.30	O-Acetylserine Sulfhydrylase	*Salmonella enterica subsp. enterica serovar Typhimurium*	A	322	B	322
43	1D6U	2.40	Copper Amine Oxidase	*Escherichia coli*	A	727	B	727
44	1D8U	2.35	Non-Symbiotic Hemoglobin	*Oryza sativa*	A	166	B	166
45	1DBQ	2.20	Purine Repressor	*Escherichia coli*	A	289	B	289
46	1DCL	2.30	MCG	*Homo sapiens*	A	216	B	216

Continued

47	1DF8	1.51	Protein (Streptavidin)	*Streptomyces avidinii*	A	127	B	127
48	1DIR	2.60	Dihydropteridine Reductase	*Rattus norvegia*	A	241	B	241
49	1DO5	2.75	Human Copper Chaperone For Superoxide Dismutase Domain II	*Homo sapiens*	A	154	C	154
50	1DOF	2.10	Adenylosuccinate Lyase	*Pyrobaculum aerophilum*	B	403	C	403
51	1DOH	2.10	Trihydroxynaphthalene Reductase	*Magnaporthe grisea*	A	283	B	283
52	1DTY	2.14	Adenosylmethionine-8-Amino-7-Oxononanoate Aminotransferase	*Escherichia coli*	A	429	B	429
53	1DUG	1.80	Chimera Of Glutathione S-Transferase-Synthetic Linker-C-Terminal Fibrinogen Gamma Chain	*Schistosoma japonicum*	A	234	B	234
54	1DVR	2.36	Adenylate Kinase	*Saccharomyces cerevisiae*	A	220	B	220
55	1DZB	2.00	SCFV Fragment 1F9	*Mus musculus*	A	253	B	253
56	1E3I	2.08	Alcohol Dehydrogenase, Class II	*Mus musculus*	A	376	B	376
57	1E3U	1.66	Beta-Lactamase Oxa-10	*Pseudomonas aeruginosa*	A	246	C	246
58	1E5E	2.18	Methionine Gamma-Lyase	*Trichomonas vaginalis G3*	A	404	B	404
59	1E7W	1.75	Pteridine Reductase	*Leishmania major*	A	291	B	291
60	1E8T	2.50	Hemagglutinin-Neuraminidase	*Newcastle disease virus (strain Kansas)*	A	454	B	454
61	1E9S	2.50	Conjugal Transfer Protein TRWB	*Escherichia coli*	E	437	F	437
62	1ECE	2.40	Endocellulase E1	*Acidothermus cellulolyticus*	A	358	B	358
63	1ECS	1.70	Bleomycin Resistance Protein	*Klebsiella pneumoniae*	A	126	B	126
64	1EE0	2.05	2-Pyrone Synthase	*Gerbera hybrid cultivar*	A	402	B	402
65	1EI1	2.30	DNA Gyrase B	*Escherichia coli*	A	391	B	391
66	1EK6	1.50	UDP-Galactose 4-Epimerase	*Homo sapiens*	A	348	B	348
67	1EKX	1.95	Aspartate Transcarbamoylase	*Escherichia coli*	B	311	C	311
68	1EM6	2.20	Liver Glycogen Phosphorylase	*Homo sapiens*	A	847	B	847
69	1EQU	3.00	Protein (Estradiol 17 Beta-Dehydrogenase 1)	*Homo sapiens*	A	327	B	327
70	1EWK	2.20	Metabotropic Glutamate Receptor Subtype 1	*Rattus norvegia*	A	490	B	490

Continued

71	1EZR	2.50	Nucleoside Hydrolase	*Leishmania major*	A	314	B	314
72	1F1J	2.35	Caspase-7 Protease	*Homo sapiens*	A	305	B	305
73	1F2D	2.00	1-Aminocyclopropane-1-Carboxylate Deaminase	*Cyberlindnera saturnus*	A	341	B	341
74	1F3T	2.00	Ornithine Decarboxylase	*Trypanosoma brucei*	C	425	D	425
75	1F5P	2.90	Hemoglobin V	*Petromyzon marinus*	A	149	B	149
76	1F6B	1.70	SAR1	*Cricetulus griseus*	A	198	B	198
77	1F6R	2.20	Alpha-Lactalbumin	*Bos taurus*	A	123	C	123
78	1FR1	2.00	Beta-Lactamase	*Citrobacter freundii*	A	361	B	361
79	1FTM	1.70	Glutamate Receptor Subunit 2	*Rattus norvegia*	A	263	C	263
80	1FUJ	2.20	PR3	*Homo sapiens*	A	221	D	221
81	1G1A	2.50	DTDP-D-glucose 4,6-Dehydratase Salmonella Enterica	*Salmonella enterica*	A	352	B	352
82	1G2O	1.75	Purine Nucleoside Phosphorylase	*Mycobacterium tuberculosis*	A	268	C	268
83	1G4A	3.00	Atp-Dependent Protease Hslv	*Escherichia coli*	B	175	A	175
84	1G5Q	2.57	Epidermin Modifying Enzyme Epid	*Staphylococcus epidermidis*	A	181	D	181
85	1G6O	2.50	CAG-Alpha	*Helicobacter pylori*	A	330	B	330
86	1G6Y	2.80	URE2 Protein	*Saccharomyces cerevisiae*	A	261	B	261
87	1G85	1.80	Odorant-Binding Protein	*Bos taurus*	A	159	B	159
88	1G8Y	2.40	Regulatory Protein REPA	*Escherichia coli*	C	279	D	279
89	1GEG	1.70	Acetoin Reductase	*Klebsiella pneumoniae*	B	256	C	256
90	1GH7	3.00	Cytokine Receptor Common Beta Chain	*Homo sapiens*	A	419	B	419
91	1GMY	1.90	Cathepsin B	*Homo sapiens*	A	261	B	261
92	1GNX	1.68	Beta-Glucosidase	*Streptomyces sp.*	A	479	B	479
93	1GP7	2.60	Phospholipase A2	*Ophiophagus hannah*	A	151	B	151
94	1GPM	2.20	GMP Synthetase	*Escherichia coli K12*	A	525	B	525

Continued

95	1GR7	1.80	Azurin	*Pseudomonas aeruginosa*	A	128	D	128
96	1GRI	3.10	Growth Factor Bound Protein 2	*Homo sapiens*	A	217	B	217
97	1GTV	1.55	Thymidylate Kinase	*Mycobacterium tuberculosis*	A	214	B	214
98	1GWN	2.10	RHO-Related GTP-Binding Protein Rhoe	*Mus musculus*	A	205	C	205
99	1GYL	3.00	Glycolate Oxidase	*Spinacia oleracea*	A	369	B	369
100	1GZ6	2.38	Estradiol 17 Beta-Dehydrogenase 4	*Rattus norvegia*	A	319	B	319
101	1GZM	2.65	Rhodopsin	*Bos taurus*	A	349	B	349
102	1H48	2.30	2C-Methyl-D-Erythritol-2,4-Cyclodiphosphate Synthase	*Escherichia coli BL21*	A	161	B	161
103	1H4G	1.10	Xylanase	*(Bacillus) agaradhaerens*	A	207	B	207
104	1H5B	1.85	Murine T Cell Receptor (TCR) Valpha Domain	*Mus musculus*	A	113	B	113
105	1H5Q	1.50	NADP-Dependent Mannitol Dehydrogenase	*Agaricus bisporus*	A	265	B	265
106	1H8V	1.90	Endo-Beta-1,4-Glucanase	*Trichoderma reesei*	D	218	E	218
107	1HAK	3.00	Annexin V	*Homo sapiens*	A	320	B	320
108	1HG4	2.40	Ultraspiracle	*Drosophila melanogaster*	A	279	D	279
109	1HKV	2.60	Diaminopimelate Decarboxylase	*Mycobacterium tuberculosis H37RV*	A	453	B	453
110	1HP0	2.10	Inosine-Adenosine-Guanosine-Preferring Nucleoside Hydrolase	*Trypanosoma vivax*	A	339	B	339
111	1HR6	2.50	Mitochondrial Processing Peptidase Alpha Subunit	*Saccharomyces cerevisiae*	A	475	E	475
112	1I0Z	2.10	L-Lactate Dehydrogenase H Chain	*Homo sapiens*	A	333	B	333
113	1I41	3.20	Cystathionine Gamma-Synthase	*Nicotiana tabacum*	F	445	G	445
114	1I4U	1.15	Crustacyanin	*Homarus gammanus*	A	181	B	181
115	1I7I	2.35	Peroxisome Proliferator Activated Receptor Gamma	*Homo sapiens*	A	292	B	292
116	1I85	3.20	T Lymphocyte Activation Antigen CD86	*Homo sapiens*	A	110	B	110
117	1II6	2.10	Kinesin-Related Motor Protein EG5	*Homo sapiens*	A	368	B	368

Continued

118	1IJL	2.60	Phospholipase A2	*Deinagkistrodon acutus*	A	123	B	123
119	1IY8	1.60	Levodione Reductase	*Leifsonia aquatica*	B	267	C	267
120	1IYK	2.30	Myristoyl-COA:Protein N-Myristoyltransferase	*Candida albicans*	A	392	B	392
121	1IZ1	2.50	LYSR-Type Regulatory Protein	*Cupriavidus necator*	A	294	P	294
122	1J7E	2.55	Vitamin D Binding Protein	*Homo sapiens*	A	458	B	458
123	1J90	2.56	Deoxyribonucleoside Kinase	*Drosophila melanogaster*	A	230	B	230
124	1JDS	1.80	5'-Methylthioadenosine Phosphorylase	*Sulfolobus solfatarius*	A	236	C	236
125	1JG8	1.80	L-Allo-Threonine Aldolase	*Thermotoga maritima*	A	347	B	347
126	1JKF	2.40	Myo-Inositol-1-Phosphate Synthase	*Saccharomyces cerevisiae*	A	533	B	533
127	1JSW	2.70	L-Aspartate Ammonia-Lyase	*Escherichia coli*	A	478	D	478
128	1JT0	2.90	Hypothetical Transcriptional Regulator In Qaca 5'Region	*Staphylococcus aureus*	A	194	C	194
129	1K2O	1.65	Cytochrome P450CAM	*Pseudomonas putida*	A	414	B	414
130	1K3F	2.50	Uridine Phosphorylase	*Escherichia coli*	A	253	D	253
131	1KBU	2.20	Cre Recombinase	*Enterobacteria phage P1*	A	349	B	349
132	1KCZ	1.90	Beta-Methylaspartase	*Clostridium tetanomorphum*	A	419	B	419
133	1KI9	2.76	Adenylate Kinase	*Methanothermococcus thermolithotrophicus*	A	192	B	192
134	1KOB	2.30	Twitchin	*Aplysia californica*	A	387	B	387
135	1KRU	2.80	Galactoside O-Acetyltransferase	*EScherichia coli*	A	203	C	203
136	1KXJ	2.80	Amidotransferase HISH	*Thermotoga maritima*	A	205	B	205
137	1L8X	2.70	Ferrochelatase	*Saccharomyces cerevisiae*	A	362	B	362
138	1LBH	3.20	Intact Lactose Operon Repressor With Gratuitous Inducer IPTG	*EScherichia coli*	A	360	B	360
139	1LLU	2.30	Alcohol Dehydrogenase	*Pseudomonas aeruginosa*	E	342	F	342
140	1LRT	2.20	Inosine-5'-Monophosphate Dehydrogenase	*Tritrichomonas suis*	A	376	B	376

Continued

141	1M43	2.40	Plasmepsin II	*Plasmodium falciparum*	A	331	B	331
142	1M48	1.95	Interleukin-2	*Homo sapiens*	A	133	B	133
143	1M7N	2.70	Insulin-Like Growth Factor I Receptor	*Homo sapiens*	A	322	B	322
144	1M7W	2.80	Hepatocyte Nuclear Factor 4-Alpha	*Rattus rattus*	A	250	B	250
145	1MCZ	2.80	Benzoylformate Decarboxylase	*Pseudomonas putida*	A	582	B	582
146	1MPY	2.80	Catechol 2,3-Dioxygenase	*Pseudomonas putida*	A	307	C	307
147	1MRU	3.00	Probable Serine/Threonine-Protein Kinase Pknb	*Mycobacterium tuberculosis*	A	311	B	311
148	1MZJ	2.10	Beta-Ketoacylsynthase III	*Streptomyces sp. R1128*	A	339	B	339
149	1MZN	1.90	RXR Retinoid X Receptor	*Homo sapiens*	A	240	C	240
150	1N0H	2.80	Acetolactate Synthase	*Saccharomyces cerevisiae*	A	677	B	677
151	1N0S	2.00	Bilin-Binding Protein	*Pieris brassicae*	A	184	B	184
152	1N1A	2.40	FKBP52	*Homo sapiens*	A	140	B	140
153	1N4O	1.85	L2 beta-lactamase	*Stenotrophomonas maltophilia*	A	276	B	276
154	1N7G	2.20	GDP-D-Mannose-4,6-Dehydratase	*Homo sapiens*	A	381	B	381
155	1N9E	1.65	Lysyl Oxidase	*Komagataella pastoris*	A	787	B	787
156	1NBU	1.60	Probable Dihydroneopterin Aldolase	*Mycobacterium tuberculosis*	B	119	C	119
157	1NFQ	2.40	Putative Oxidoreductase Rv2002	*Mycobacterium tuberculosis*	A	260	B	260
158	1NKS	2.57	Adenylate Kinase	*Sulfolobus acidocaldarius*	A	194	B	194
159	1NTO	1.94	NAD-Dependent Alcohol Dehydrogenase	*Sulfolobus solfataricus*	C	347	D	347
160	1NW4	2.20	Uridine Phosphorylase, Putative	*Plasmodium falciparum 3D7*	A	276	B	276
161	1NX9	2.20	Alpha-Amino Acid Ester Hydrolase	*Acetobacter pasteurianus*	B	652	D	652
162	1O4S	1.90	Aspartate Aminotransferase	*Thermotoga maritima*	A	389	B	389
163	1O5I	2.50	3-Oxoacyl-(Acyl Carrier Protein) Reductase	*Thermotoga maritima*	A	249	D	249

Continued

164	1O5O	2.30	Uracil Phosphoribosyltransferase	*Thermotoga maritima*	A	221	B	221
165	1O61	1.90	Aminotransferase	*Campylobacter jejuni*	A	394	B	394
166	1O63	2.00	ATP Phosphoribosyltransferase	*THermotoga maritima*	A	219	B	219
167	1O6E	2.30	Capsid Protein P40	*Human herpesvirus 4*	A	235	B	235
168	1O9J	2.40	Aldehyde Dehydrogenase, Cytosolic 1	*Elephantulus edwardii*	A	501	B	501
169	1OAT	2.50	Ornithine Aminotransferase	*Homo sapiens*	B	439	C	439
170	1OD2	2.70	Acetyl-Coenzyme A Carboxylase	*Saccharomyces cerevisiae*	A	805	B	805
171	1ODL	2.10	Purine Nucleoside Phosphorylase	*Thermus thermophilus HB8*	A	235	B	235
172	1OE7	1.80	Glutathione S-Transferase	*Schistosoma haematobium*	A	211	B	211
173	1OJ6	1.95	Neuroglobin	*Homo sapiens*	A	151	B	151
174	1OME	2.30	Beta-Lactamase	*Staphylococcus aureus*	A	258	B	258
175	1OPL	3.42	Proto-Oncogene Tyrosine-Protein Kinase	*Homo sapiens*	A	537	B	537
176	1ORR	1.50	CDP-Tyvelose-2-Epimerase	*Salmonella enterica subsp. enterica serovar Typhi*	A	347	B	347
177	1ORT	3.00	Ornithine Transcarbamoylase	*Pseudomonas aeruginosa*	C	335	D	335
178	1ORW	2.84	Dipeptidyl Peptidase IV	*Sus scrofa*	B	728	D	728
179	1OVL	2.20	Orphan Nuclear Receptor NURR1 (MSE 414, 496, 511)	*Homo sapiens*	A	271	B	271
180	1OVM	2.65	Indole-3-Pyruvate Decarboxylase	*Enterobacter cloacae*	B	552	D	552
181	1OYJ	1.95	Glutathione S-Transferase	*Oryza sativa*	C	231	D	231
182	1OZF	2.30	Acetolactate Synthase, Catabolic	*Klebsiella pneumoniae*	A	566	B	566
183	1P0K	1.90	Isopentenyl-Diphosphate Delta-Isomerase	*Bacillus subtilis*	A	349	B	349
184	1P4E	2.70	Recombinase FLP Protein	*Saccharomyces cerevisiae*	A	429	D	429
185	1P60	1.96	Deoxycytidine Kinase	*Homo sapiens*	A	263	B	263
186	1P7C	2.10	Thymidine Kinase	*Herpes simplex virus (type 1 / strain 17)*	A	343	B	343

Continued

187	1P93	2.70	Glucocorticoid Receptor	*Homo sapiens*	A	280	C	280
188	1PKG	2.90	C-Kit Protein	*Homo sapiens*	A	329	B	329
189	1PV4	3.00	Transcription Termination Factor Rho	*EScherichia coli*	C	419	D	419
190	1PWE	2.80	L-Serine Dehydratase	*Rattus norvegia*	C	327	D	327
191	1PWX	1.80	Halohydrin Dehalogenase	*Agrobacterium tumefaciens*	B	254	C	254
192	1PY4	2.90	Beta-2-Microglobulin Precursor	*Homo sapiens*	A	100	C	100
193	1Q0C	2.10	Homoprotocatechuate 2,3-Dioxygenase	*Brevibacterium fuscum*	A	365	C	365
194	1Q3D	2.20	Glycogen Synthase Kinase-3 Beta	*Homo sapiens*	A	424	B	424
195	1Q3G	2.65	UDP-N-Acetylglucosamine 1-Carboxyvinyltransferase	*Enterobacter cloacae*	A	419	B	419
196	1Q57	3.45	DNA Primase/Helicase	*Enterobacteria phage T7*	A	503	C	503
197	1Q5Q	2.60	Proteasome Alpha-Type Subunit 1	*Rhodococcus erythropolis*	C	259	B	259
198	1Q6T	2.30	Protein-Tyrosine Phosphatase, Non-Receptor Type 1	*Homo sapiens*	A	310	B	310
199	1Q8M	2.60	Triggering Receptor Expressed On Myeloid Cells 1	*Homo sapiens*	A	127	B	127
200	1QPB	2.40	Pyruvate Decarboxylase (Form B)	*Saccharomyces pastorianus Weihenstephan 34/70*	A	563	B	563
201	1QZF	2.80	Bifunctional Dihydrofolate Reductase-Thymidylate Synthase	*Cryptosporidium hominis*	C	521	D	521
202	1R5K	2.70	Estrogen Receptor	*Homo sapiens*	A	261	C	261
203	1RD5	2.02	Tryptophan Synthase Alpha Chain, Chloroplast	*Zea mays*	A	262	B	262
204	1RD7	2.60	Dihydrofolate Reductase	*EScherichia coli*	A	159	B	159
205	1RE5	2.60	3-Carboxy-Cis,Cis-Muconate Cycloisomerase	*Pseudomonas putida KT2440*	B	450	C	450
206	1RJN	2.30	menB	*Mycobacterium tuberculosis*	A	339	B	339
207	1RKX	1.80	CDP-Glucose-4,6-Dehydratase	*Yersinia pseudotuberculosis*	A	357	D	357
208	1RPY	2.30	Adaptor Protein APS	*Rattus norvegia*	A	114	B	114
209	1RQR	2.67	5'-Fluoro-5'-Deoxyadenosine Synthase	*Streptomyces cattleya*	A	299	C	299

Continued

210	1S2G	2.10	Purine Trans Deoxyribosylase	*Lactobacillus helveticus*	A	167	B	167
211	1SF2	2.40	4-Aminobutyrate Aminotransferase	*EScherichia coli*	A	426	B	426
212	1SGF	3.15	Nerve Growth Factor	*Mus musculus*	G	237	Z	237
213	1SL6	2.25	C-Type Lectin DC-Signr	*Homo sapiens*	D	184	A	184
214	1SN0	1.90	Transthyretin	*Sparus aurata*	A	130	C	130
215	1SP8	2.00	4-Hydroxyphenylpyruvate Dioxygenase	*Zea mays*	A	418	B	418
216	1SPI	2.80	Fructose 1,6-Bisphosphatase	*Spinacia oleracea*	A	358	B	358
217	1SQI	2.15	4-Hydroxyphenylpyruvic Acid Dioxygenase	*Rattus norvegia*	A	393	B	393
218	1SQL	2.20	Dihydroneopterin Aldolase	*Arabidopsis thaliana*	D	146	E	146
219	1SU2	2.70	Mutt/Nudix Family Protein	*Deinococcus radiodurans*	A	159	B	159
220	1SXG	2.75	Glucose-Resistance Amylase Regulator	*Bacillus megaterium*	A	280	D	280
221	1T2O	2.30	Sortase	*Staphylococcus aureus*	B	146	A	146
222	1T8S	2.60	Amp Nucleosidase	*EScherichia coli*	A	484	B	484
223	1T91	1.90	Ras-Related Protein Rab-7	*Homo sapiens*	A	207	D	207
224	1T97	2.70	Lysozyme	*Enterobacteria phage T4*	A	175	B	175
225	1TAH	3.00	Lipase	*Burkholderia glumae*	A	318	B	318
226	1TC0	2.20	Endoplasmin	*Canis lupus familiaris*	A	236	B	236
227	1TEX	2.60	Stf0 Sulfotransferase	*Mycobacterium smegmatis*	A	287	B	287
228	1TF7	2.80	KaiC	*Synechococcus sp.*	B	525	C	525
229	1TFC	2.40	Estrogen-Related Receptor Gamma	*Homo sapiens*	A	251	B	251
230	1TPZ	2.00	Interferon-Inducible GTPase	*Mus musculus*	A	422	B	422
231	1TW2	2.50	Carminomycin 4-O-Methyltransferase	*Streptomyces peucetius*	A	360	B	360
232	1TXT	2.50	3-Hydroxy-3-Methylglutaryl-CoA Synthase	*Staphylococcus aureus*	A	388	B	388

Continued

233	1TZ3	2.90	Putative Sugar Kinase	*Salmonella enterica subsp. enterica serovar Typhimurium str. LT2*	A	339	B	339
234	1U0L	2.80	Probable GTPase engC	*Thermotoga maritima*	A	301	C	301
235	1U1I	1.90	Myo-Inositol-1-Phosphate Synthase	*Archaeoglobus fulgidus DSM 4304*	A	392	B	392
236	1U46	2.00	Activated CDC42 Kinase 1	*Homo sapiens*	A	291	B	291
237	1U5Q	2.10	Serine/Threonine Protein Kinase TAO2	*Rattus norvegia*	A	348	B	348
238	1UB7	2.30	3-Oxoacyl-[Acyl-Carrier Protein] Synthase	*Thermus thermophilus HB8*	A	322	B	322
239	1UFH	2.20	YYCN Protein	*Bacillus subtilis subsp. subtilis str. 168*	B	180	A	180
240	1UI5	2.40	A-Factor Receptor Homolog	*Streptomyces coelicolor A3(2)*	A	215	B	215
241	1UIM	2.15	Threonine Synthase	*Thermus thermophilus*	A	351	B	351
242	1UIU	1.85	Nickel-Binding Periplasmic Protein	*EScherichia coli*	A	502	B	502
243	1UMO	2.59	Cytoglobin	*Homo sapiens*	A	190	B	190
244	1UPA	2.35	Carboxyethylarginine Synthase	*Streptomyces clavuligerus*	B	573	D	573
245	1URZ	2.70	Envelope Protein	*Tick-borne encephalitis virus (WESTERN SUBTYPE)*	B	401	C	401
246	1USI	1.80	Leucine-Specific Binding Protein	*Escherichia coli*	A	346	C	346
247	1UTR	X	Uteroglobin	*Rattus norvegia*	A	96	B	96
248	1UU0	2.85	Histidinol-Phosphate Aminotransferase	*Thermotoga maritima*	A	335	B	335
249	1UZM	1.49	3-Oxoacyl-[Acyl-Carrier Protein] Reductase	*Mycobacterium tuberculosis H37Rv*	A	247	B	247
250	1V2I	2.20	Hemagglutinin-Neuraminidase Glycoprotein	*Human parainfluenza virus 3*	A	431	B	431
251	1V4J	2.85	Octoprenyl-Diphosphate Synthase	*Thermotoga maritima*	A	299	B	299
252	1VC8	2.00	NDX1	*Thermus thermophilus HB8*	A	126	B	126
253	1VDM	2.50	Purine Phosphoribosyltransferase	*Pyrococcus horikoshii*	B	153	H	153
254	1VDW	1.30	Hypothetical Protein PH1897	*Pyrococcus horikoshii*	A	254	B	254
255	1VEA	2.80	Hut Operon Positive Regulatory Protein	*Bacillus subtilis*	A	148	B	148

Continued

256	1VHW	1.54	Purine Nucleoside Phosphorylase	*Vibrio cholerae*	A	253	D	253
257	1VIQ	2.40	ADP-Ribose Pyrophosphatase	*Escherichia coli*	B	220	C	220
258	1VIY	1.89	Dephospho-CoA Kinase	*Escherichia coli*	A	218	B	218
259	1VKG	2.20	Histone Deacetylase 8	*Homo sapiens*	A	377	B	377
260	1VM6	2.27	Dihydrodipicolinate Reductase	*THermotoga maritima*	B	228	D	228
261	1VPX	2.40	Protein (Transaldolase (EC 2.2.1.2))	*THermotoga maritima*	B	230	C	230
262	1W0M	2.50	Triosephosphate Isomerase	*Thermoproteus tenax*	B	226	D	226
263	1W2Z	2.24	Amine Oxidase, Copper Containing	*Pisum sativum*	A	649	B	649
264	1W59	2.70	Cell Division Protein FTSZ Homolog 1	*Methanocaldococcus jannaschii*	A	364	B	364
265	1W6T	2.10	Enolase	*Streptococcus pneumoniae TIGR4*	A	444	B	444
266	1W6U	1.75	2,4-Dienoyl-Coa Reductase, Mitochondrial Precursor	*Homo sapiens*	A	302	B	302
267	1W91	2.20	Beta-Xylosidase	*Geobacillus stearothermophilus*	F	503	G	503
268	1WC1	1.93	Adenylate Cyclase	*Arthrospira platensis*	B	226	C	226
269	1WUE	2.10	Mandelate Racemase/Muconate Lactonizing Enzyme Family Protein	*Enterococcus faecalis V583*	A	386	B	386
270	1WYE	2.80	2-Keto-3-Deoxygluconate Kinase	*Sulfolobus tokodaii str. 7*	A	311	B	311
271	1WZ8	1.80	Enoyl-CoA Hydratase	*Thermus thermophilus HB8*	A	264	B	264
272	1X27	2.70	Proto-Oncogene Tyrosine-Protein Kinase LCK	*Homo sapiens*	A	167	E	167
273	1XCA	2.30	Cellular Retinoic Acid Binding Protein Type II	*Homo sapiens*	A	137	B	137
274	1XED	1.90	Polymeric-Immunoglobulin Receptor	*Homo sapiens*	A	117	B	117
275	1XG5	1.53	ARPG836	*Homo sapiens*	A	279	B	279
276	1XGM	2.80	Methionine Aminopeptidase	*Pyrococcus furiosus*	A	295	B	295
277	1XI9	2.33	Putative Transaminase	*Pyrococcus furiosus DSM 3638*	A	406	B	406
278	1XKQ	2.10	Short-Chain Reductase Family Member (5D234)	*Caenorhabditis elegans*	A	280	B	280

Continued

279	1XNX	2.90	Constitutive Androstane Receptor	*Mus musculus*	A	256	B	256
280	1XSE	2.90	11beta-Hydroxysteroid Dehydrogenase Type 1	*Cavia porcellus*	A	295	B	295
281	1XTT	1.80	Probable Uracil Phosphoribosyltransferase	*Sulfolobus solfataricus*	B	216	C	216
282	1XYG	2.19	Putative N-Acetyl-Gamma-Glutamyl-Phosphate Reductase	*Arabidopsis thaliana*	B	359	C	359
283	1Y1M	1.80	Glutamate [NMDA] Receptor Subunit Zeta 1	*Rattus norvegia*	A	292	B	292
284	1Y1P	1.60	Aldehyde Reductase II	*Sporidiobolus salmonicolor*	A	342	B	342
285	1YDE	2.40	Retinal Dehydrogenase/Reductase 3	*Homo sapiens*	B	270	P	270
286	1YO6	2.60	Putative Carbonyl Reductase Sniffer	*Caenorhabditis elegans*	A	250	B	250
287	1YP2	2.11	Glucose-1-Phosphate Adenylyltransferase Small Subunit	*Solanum tuberosum*	C	451	D	451
288	1YQ2	1.90	Beta-Galactosidase	*Arthrobacter sp. C2-2*	A	1024	B	1024
289	1YTA	2.20	Oligoribonuclease	*Escherichia coli*	A	180	B	180
290	1YXM	1.90	Peroxisomal Trans 2-Enoyl Coa Reductase	*Homo sapiens*	A	303	B	303
291	1Z08	1.80	Ras-Related Protein Rab-21	*Homo sapiens*	B	170	D	170
292	1Z1B	3.80	Integrase	*Enterobacteria phage lambda*	A	356	B	356
293	1Z5A	2.20	Type II DNA Topoisomerase VI Subunit B	*Sulfolobus shibatae*	A	469	B	469
294	1ZEM	1.90	Xylitol Dehydrogenase	*Gluconobacter oxydans*	E	262	G	262
295	1ZPD	1.86	Pyruvate Decarboxylase	*Zymomonas mobilis*	B	568	E	568
296	2A2G	2.90	Protein (Alpha-2u-Globulin)	*Rattus norvegia*	A	181	C	181
297	2A7K	2.24	CarB	*Pectobacterium carotovorum*	G	250	H	250
298	2AAW	2.40	Glutathione S-Transferase	*Plasmodium falciparum*	A	222	C	222
299	2AG5	1.84	Dehydrogenase/Reductase (SDR Family) Member 6	*Homo sapiens*	A	246	B	246
300	2ANC	3.20	Guanylate Kinase	*Escherichia coli*	C	207	E	207
301	2B4G	1.95	Dihydroorotate Dehydrogenase	*Trypanosoma brucei*	A	317	B	317

Continued

302	2BCK	2.80	HLA Class I Histocompatibility Antigen, A-24 Alpha Chain	*Homo sapiens*	A	294	D	294
303	2BD0	1.70	Sepiapterin Reductase	*Chlorobium tepidum TLS*	A	244	B	244
304	2BTW	2.00	ALR0975 Protein	*Nostoc sp. PCC 7120*	A	254	B	254
305	2BVC	2.10	Glutamine Synthetase 1	*Mycobacterium tuberculosis H37Rv*	D	486	E	486
306	2C7F	2.70	Alpha-L-Arabinofuranosidase	*Ruminiclostridium thermocellum*	A	513	D	513
307	2CLT	2.67	Interstitial Collagenase	*Homo sapiens*	A	367	B	367
308	2D1Y	1.65	Hypothetical Protein TT0321	*Thermus thermophilus*	B	256	C	256
309	2DFT	2.80	Shikimate Kinase	*Mycobacterium tuberculosis*	A	176	C	176
310	2DNS	2.40	D-Amino Acid Amidase	*Ochrobactrum anthropi*	A	363	B	363
311	2DW6	2.30	Bll6730 Protein	*Bradyrhizobium japonicum*	A	389	B	389
312	2DY4	2.65	DNA Polymerase	*Enterobacteria phage RB69*	D	903	B	903
313	2EGH	2.20	1-Deoxy-D-Xylulose 5-Phosphate Reductoisomerase	*Escherichia coli str. K-12 substr. W3110*	B	424	A	424
314	2EVO	1.70	Methionine Aminopeptidase	*Escherichia coli*	A	264	B	264
315	2EW8	2.10	(S)-1-Phenylethanol Dehydrogenase	*Aromatoleum aromaticum EbN1*	B	249	D	249
316	2F1G	1.90	Cathepsin S	*Homo sapiens*	A	220	B	220
317	2F73	2.50	Fatty Acid-Binding Protein, Liver	*Homo sapiens*	A	149	B	149
318	2FG6	2.80	Putative Ornithine Carbamoyltransferase	*Bacteroides fragilis NCTC 9343*	C	338	D	338
319	2FM6	1.75	Metallo-Beta-Lactamase L1	*Stenotrophomonas maltophilia*	A	269	B	269
320	2FYI	2.80	Hth-Type Transcriptional Regulator CBL	*Escherichia coli K-12*	A	228	B	228
321	2GF0	1.90	GTP-Binding Protein DI-RAS1	*Homo sapiens*	A	199	B	199
322	2GIC	2.92	Nucleocapsid Protein	*Vesicular stomatitis Indiana virus*	A	422	B	422
323	2I6A	2.20	Adenosine Kinase	*Homo sapiens*	A	345	D	345
324	2IFA	2.30	Hypothetical Protein SMU.260	*Streptococcus mutans UA159*	A	208	B	208

Continued

325	2IGO	1.95	Pyranose Oxidase	*Trametes ochracea*	C	623	D	623
326	2IJ2	1.20	Cytochrome P450 BM3	*Bacillus megaterium*	B	470	A	470
327	2NUU	2.50	Ammonia Channel	*Escherichia coli*	B	415	C	415
328	2O23	1.20	HADH2 Protein	*Homo sapiens*	A	265	B	265
329	2O2Y	2.20	Enoyl-Acyl Carrier Reductase	*Plasmodium falciparum 3D7*	C	349	D	349
330	2OKR	2.00	Mitogen-Activated Protein Kinase 14	*Homo sapiens*	A	366	D	366
331	2PD3	2.50	Enoyl-[Acyl-Carrier-Protein] Reductase [NADH]	*Helicobacter pylori*	B	275	C	275
332	2PMT	2.70	Glutathione Transferase	*Proteus mirabilis*	C	203	D	203
333	2TAA	3.00	TAKA-Amylase A	*Aspergillus oryzae*	A	478	B	478
334	3TAT	3.50	Tyrosine Aminotransferase	*Escherichia coli*	C	397	D	397
335	7MDH	2.40	Protein (Malate Dehydrogenase)	*Sorghum bicolor*	D	375	C	375

List of PP complexes from the PIBASE dataset were grouped into 101 heterodimer complexes **(Table B)**.

No	PDB ID	Resolution (Å)	Chain one	Name of chain one	Length	Chain two	Name of chain two	Length
1	1A22	2.60	B	Growth Hormone Receptor	238	A	Growth Hormone	191
2	1A2K	2.50	C	RAN	216	B	Nuclear Transport Factor 2	127
3	1A5A	1.90	A	Tryptophan Synthase Alpha chain	268	B	Tryptophan Synthase Beta Chain	397
4	1A5A	1.90	B	Tryptophan Synthase Beta chain	397	A	Tryptophan Synthase Alpha Chain	268
5	1ACB	2.00	E	Alpha-Chymotrypsin	241	I	Eglin C	63
6	1AIG	2.60	P	Photosynthetic Reaction Center (H Subunit)	260	O	Photosynthetic Reaction Center (M Subunit)	307
7	1AIG	2.60	L	Photosynthetic Reaction Center (L Subunit)	281	H	Photosynthetic Reaction Center (H Subunit)	260
8	1AIP	3.00	A	Elongation Factor TU	405	C	Elongation Factor TS	196
9	1AR1	2.70	B	Cytochrome Oxidase	298	A	Cytochrome Oxidase	558
10	1AZS	2.30	A	VC1	220	B	IIC2	212
11	1AZS	2.30	C	GS-ALPHA	402	B	IIC2	212
12	1AZZ	2.30	C	Ecotin	142	A	Collogenase	226
13	1B6C	2.60	D	TGF-B Superfamily Receptor Type 1	342	C	FK506-Binding Protein	107
14	1B7T	2.50	Y	Myosin Regulatory Light Chain	156	A	Myosin Heavy Chain	835
15	1B7T	2.50	Z	Myosin Essential Light Chain	156	A	Myosin Heavy Chain	835
16	1BCC	3.16	E	Ubiquinol Cytochrome C Oxidoreductase	196	D	Ubiquinol Cytochrome C Oxidoreductase	241
17	1BCC	3.16	G	Ubiquinol Cytochrome C Oxidoreductase	81	C	Ubiquinol Cytochrome C Oxidoreductase	380
18	1BGY	3.00	R	Cytochrome BC1 Complex	110	O	Cytochrome BC1 Complex	379
19	1BI7	3.40	A	Cyclin-Dependent Kinase 6	326	B	Multiple Tumor Supressor	156
20	1BMF	2.85	C	Bovine Mitochondrial F1-Atpase	510	D	Bovine Mitochondrial F1-Atpase	482
21	1BML	2.90	A	Plasmin	250	C	Streptokinase	362
22	1CD1	2.67	A	CD1	315	B	CD1	99
23	1CD9	2.80	B	Protein (G-CSF Receptor)	215	A	Protein (Granulocyte Colony-Stimulating Factor)	175

Continued

24	1CN4	2.80	A	Protein (Erythroprotein Receptor)	228	C	Protein (Erythroprotein)	166
25	1DN0	2.28	B	IGM-Kappa Cold Agglutinin (Heavy Chain)	232	A	IGM-Kappa Cold Agglutinin (Light Chain)	215
26	1DN0	2.10	B	IGM-Kappa Cold Agglutinin (Heavy Chain)	232	A	IGM-Kappa Cold Agglutinin (Light Chain)	215
27	1DPJ	1.80	A	Proteinase A	329	B	Proteinase Inhibitor IA3 Peptide	33
28	1EOC	2.25	B	Protocatechuate 3,4-Dioxygenase Beta Chain	241	A	Protocatechuate 3,4-Dioxygenase Alpha Chain	209
29	1EP1	2.20	A	Dihydroorotate Dehydrogenase B (Pyrd Subunit)	311	B	Dihydroorotate Dehydrogenase B (Pyrk Subunit)	261
30	1EYS	2.20	M	Photosynthetic Reaction Center	324	C	Photosynthetic Reaction Center	382
31	1EZV	2.30	A	Ubiquinol-Cytochrome C Reductase Complex Core Protein I	430	B	Ubiquinol-Cytochrome C Reductase Complex Core Protein 2	352
32	1EZV	2.30	B	Ubiquinol-Cytochrome C Reductase Complex Core Protein 2	352	A	Ubiquinol-Cytochrome C Reductase Complex Core Protein I	430
33	1EZV	2.30	C	Cytochrome B	385	D	Cytochrome C1	245
34	1EZV	2.30	D	Cytochrome C1	245	C	Cytochrome B	385
35	1F60	1.67	A	Elongation Factor EEF1A	458	B	Elongation Factor EEF1BA	94
36	1F8U	2.90	A	Acetylcholinesterase	583	B	Fasciculin II	61
37	1FJG	3.00	E	30s Ribosomal Protein S5	162	H	30s Ribosomal Protein S8	138
38	1FRT	4.50	A	Neonatal FC Receptor	269	B	Beta 2-Microglobulin	99
39	1FSK	2.90	J	Major Pollen Allergen Bet V 1-A	159	L	Immunoglobulin Kappa Light Chain	214
40	1FVU	1.80	B	Botrocetin Beta Chain	125	A	Botrocetin Alpha Chain	133
41	1FX0	3.20	B	ATP Synthase Beta Chain	498	A	Atp Synthase Alpha Chain	507
42	1G73	2.00	C	Inhibitors Of Apoptosis-Like Protein Ilp	121	A	Second Mitochondria-Derived Activator Of Caspases	162
43	1G9M	2.20	C	T-Cell Surface Glycoprotein Cd4	185	G	Envelope Glycoprotein GP120	321
44	1GPW	2.40	A	Hisf Protein	253	B	Amidotransferase HISH	201
45	1HR6	2.50	D	Mitochondrial Processing Peptidase Beta Subunit	443	C	Mitochondrial Processing Peptidase Alpha Subunit	475

Continued

46	1HXM	3.12	A	Gamma-Delta T-Cell Receptor	229	B	Gamma-Delta T-Cell Receptor	242
47	1HXM	3.12	B	Gamma-Delta T-Cell Receptor	242	A	Gamma-Delta T-Cell Receptor	229
48	1HYR	2.70	A	NKG2-D Type Ii Integral Membrane Protein	137	C	MHC Class I Chain-Related Protein A	275
49	1I1R	2.40	A	Interleukin-6 Receptor Beta Chain	301	B	Viral IL-6	167
50	1I3R	2.40	D	Fusion Protein Consisting Of Mhc E-Beta-K Precursor, Glycine Rich Linker, And Hemoglobin Beta-2 Chain	228	C	H-2 Class II Histocompatibility Antigen, E-K Alpha Chain	192
51	1I85	3.20	D	Cytotoxic T-Lymphocyte-Associated Protein 4	110	B	T Lymphocyte Activation Antigen CD86	126
52	1JEB	2.10	A	Hemoglobin Zeta Chain	142	D	Hemoglobin Beta-Single Chain	146
53	1JFF	X	A	Tubulin Alpha Chain	451	B	Tubulin Beta Chain	445
54	1JWH	3.10	A	Casein Kinase II, Alpha Chain	337	D	Casein Kinase II Beta Chain	215
55	1JWI	2.00	B	Platelet Aggregation Inducer	125	A	Bitiscetin	131
56	1KFY	3.60	B	Fumarate Reductase Iron-Sulfur Protein	243	C	Fumarate Reductase 15 Kda Hydrophobic Protein	130
57	1KFY	3.60	C	Fumarate Reductase 15 Kda Hydrophobic Protein	130	B	Fumarate Reductase Iron-Sulfur Protein	243
58	1KFY	3.60	D	Fumarate Reductase 13 Kda Hydrophobic Protein	119	B	Fumarate Reductase Iron-Sulfur Protein	243
59	1KSG	2.30	A	Arf-Like Protein 2	186	B	Retinal Rod Rhodopsin-Sensitive Cgmp 3',5'-Cyclic Phosphodiesterase Delta-Subunit	152
60	1LW6	1.50	E	Subtilisin BPN	281	I	Ubtilisin-Chymotrypsin Inhibitor-2A	63
61	1M56	2.30	A	Cytochrome C Oxidase	566	B	Cytochrome C Oxidase	264
62	1MF8	3.10	B	Calcineurin B Subunit Isoform 1	170	A	Calmodulin-Dependent Calcineurin A Subunit, Alpha Isoform	373
63	1MG2	2.25	O	Amicyanin	105	N	Methylamine Dehydrogenase, Light Chain	131
64	1NCC	2.50	N	Influenza A Subtype N9 Neuraminidase	389	H	IGG2A-Kappa NC41 FAB (Heavy Chain)	221
65	1NF3	2.10	B	G25K GTP-Binding Protein, Placental Isoform	195	D	PAR-6B	128

Continued

66	1NFD	2.80	A	N15 Alpha-Beta T-Cell Receptor	203	D	N15 Alpha-Beta T-Cell Receptor	239
67	1O5D	2.05	L	Coagulation Factor VII	152	T	Tissue factor	218
68	1OGA	1.40	D	T-Cell Receptor Alpha Chain V Region	215	E	T-Cell Receptor Beta Chain C Region	252
69	1OGA	1.40	E	T-Cell Receptor Beta Chain C Region	252	D	T-Cell Receptor Alpha Chain V Region	215
70	1P0S	2.80	E	Ecotin Precursor	142	H	Coagulation Factor X Precursor	254
71	1P4B	2.35	L	Antibody Variable Light Chain	135	H	Antibody Variable Heavy Chain	124
72	1PKQ	3.00	A	(8-18C5) Chimeric Fab, Light Chain	241	B	(8-18C5) Chimeric Fab, Heavy Chain	252
73	1PYT	2.35	B	Procarboxypeptidase A	309	A	Procarboxypeptidase A	94
74	1Q90	3.10	B	Cytochrome B6	215	D	Cytochrome B6-F Complex Subunit 4	159
75	1QA9	3.20	A	Human Cd2 Protein	102	D	Human CD58 Protein	95
76	1R4P	1.77	E	Shiga-Like Toxin Type II B Subunit	70	A	Shiga-Like Toxin Type II A Subunit	297
77	1R8Q	1.86	E	Arno	203	A	ADP-Ribosylation Factor 1	181
78	1SB2	1.90	A	Rhodocetin Alpha Subunit	133	B	Rhodocetin Beta Subunit	129
79	1SPG	1.95	B	Hemoglobin	147	A	Hemoglobin	144
80	1SQB	2.69	B	Ubiquinol-Cytochrome C Reductase Complex Core Protein 2, Mitochondrial	453	I	Ubiquinol-Cytochrome C Reductase 8 Kda Protein	78
81	1SQP	2.70	A	Ubiquinol-Cytochrome-C Reductase Complex Core Protein I, Mitochondrial Precursor	480	I	Ubiquinol-Cytochrome C Reductase Iron-Sulfur Subunit, Mitochondrial Precursor (EC 1.10.2.2) (Rieske Iron-Sulfur Protein) (RISP) [Contains: Ubiquinol-Cytochrome C Reductase 8 Kda Protein (Complex III Subunit IX)]	78
82	1SVX	2.24	B	Maltose-Binding Periplasmic Protein	395	A	Ankyrin Repeat Protein Off7	169
83	1UKM	1.90	A	EMS16 A Chain	134	B	EMS16 B Chain	128

Continued

84	1UKV	1.50	G	Secretory Pathway GDP Dissociation Inhibitor	453	Y	GTP-Binding Protein YPT1	206
85	1UVQ	1.80	A	HLA Class II Histocompatibility Antigen	197	B	Hla Class II Histocompatibility Antigen	198
86	1VF5	3.00	A	Cytochrome B6	215	B	Subunit IV	160
87	1X9F	2.60	J	Globin II, Extracellular	145	K	Globin III, Extracellular	153
88	1XDK	2.90	B	Retinoic Acid Receptor, Beta	303	A	Retinoic Acid Receptor RXR-Alpha	238
89	1XXD	2.91	B	Coagulation Factor XI	238	C	Ecotin	142
90	1YTZ	3.00	C	Troponin C	162	I	Troponin I	182
91	1YVB	2.70	A	Falcipain 2	241	I	Cystatin	111
92	1Z7M	2.90	E	ATP Phosphoribosyltransferase	208	A	ATP Phosphoribosyltransferase Regulatory Subunit	344
93	1Z7X	1.95	X	Ribonuclease I	129	W	Ribonuclease Inhibitor	461
94	2A9K	1.73	A	Ras-Related Protein Ral-A	187	B	Mono-ADP-Ribosyltransferase C3	223
95	2A9K	1.73	B	Mono-ADP-Ribosyltransferase C3	223	A	Ras-Related Protein Ral-A	187
96	2BEX	1.99	C	Nonsecretory Ribonuclease	135	A	Ribonuclease Inhibitor	460
97	2G2U	1.60	A	Beta-lactamase SHV-1	265	B	Beta-Lactamase Inhibitory Protein	165
98	2GJ7	5.00	E	Glycoprotein E	401	B	IG Gamma-1 Chain C Region	227
99	2IG2	3.00	H	IGG1-Lambda Kol FAB (Heavy Chain)	455	L	IGG1-Lambda KOL FAB (Light Chain)	216
100	2J12	1.50	B	Coxsackievirus And Adenovirus Receptor	128	A	Fiber Protein	194
101	2ONL	4.00	C	MAP Kinase-Activated Protein Kinase 2	406	A	Mitogen-Activated Protein Kinase 14	366

8.2 Supplementary Information for Chapter 4

Table 4.1 List of the transcription factors targeted to STIM and ORAI genes with Pubmed ID (PMID), technique of experiments and cell types obtained from the CheA database.

Gene	TF	PMID (Pubmed)	Technique	Cell Type
STIM1	AR	20517297	CHIP-SEQ	VCAP
	E2F4	21247883	CHIP-SEQ	LYMPHOBLASTOID
	EGR1	20690147	CHIP-SEQ	ERYTHROLEUKEMIA
	ELK3	25401928	CHIP-SEQ	HUVEC
	FLI1	21571218	CHIP-SEQ	MEGAKARYOCYTES
	FOXA2	19822575	CHIP-SEQ	HepG2
	FOXP1	21924763	CHIP-SEQ	HESC
	GATA1	21571218	CHIP-SEQ	MEGAKARYOCYTES
	GATA1	19941826	CHIP-SEQ	K562
	GATA2	19941826	CHIP-SEQ	K562
	GATA2	21666600	CHIP-SEQ	HMVEC
	HNF4A	19822575	CHIP-SEQ	HepG2
	MITF	21258399	CHIP-SEQ	MELANOMA
	MYC	19915707	CHIP-SEQ	AK7
	NCOR1	26117541	CHIP-SEQ	K562
	PHF8	20622854	CHIP-SEQ	HELA
	RUNX1	17652178	CHIP-SEQ	JURKAT
	SCL	21571218	CHIP-SEQ	MEGAKARYOCYTES
	SOX2	21211035	CHIP-SEQ	LN229_GBM
	SPI1	20517297	CHIP-SEQ	HL60
	TP63	22573176	CHIP-SEQ	HFKS
	TRIM28	17542650	CHIP-SEQ	NTERA2
	TTF2	22483619	CHIP-SEQ	HELA
STIM2	AR	22383394	CHIP-SEQ	PROSTATE_CANCER
	AR	19668381	CHIP-SEQ	PC3
	AR	25329375	CHIP-SEQ	VCAP
	CUX1	19635798	CHIP-SEQ	MULTIPLE HUMAN CANCER CELL TYPES
	E2F1	21310950	CHIP-SEQ	MCF7
	E2F4	17652178	CHIP-SEQ	JURKAT
	EGR1	20690147	CHIP-SEQ	ERYTHROLEUKEMIA
	ELK3	25401928	CHIP-SEQ	HUVEC
	FOXP1	21924763	CHIP-SEQ	HESC
	GABP	19822575	CHIP-SEQ	HepG2
	GATA1	21571218	CHIP-SEQ	MEGAKARYOCYTES
	GATA2	19941826	CHIP-SEQ	K562
	HOXB7	26014856	CHIP-SEQ	BT474
	MYCN	21190229	CHIP-SEQ	SHEP-21N
	PAX3-FKHR	20663909	CHIP-SEQ	RHABDOMYOSARCOMA

Continued

	POU3F2	20337985	CHIP-SEQ	501MEL
	PPARD	21283829	CHIP-SEQ	MYOFIBROBLAST
	RUNX1	17652178	CHIP-SEQ	JURKAT
	RUNX2	22187159	CHIP-SEQ	PCA
	SCL	21571218	CHIP-SEQ	MEGAKARYOCYTES
	SMAD4	21799915	CHIP-SEQ	A2780
	SOX11	23321250	CHIP-SEQ	Z138-A519-JVM2
	SOX2	21211035	CHIP-SEQ	LN229_GBM
	TOP2B	26459242	CHIP-SEQ	MCF7
	TP63	23658742	CHIP-SEQ	EP156T
	TTF2	22483619	CHIP-SEQ	HELA
	VDR	23849224	CHIP-SEQ	CD4+
	WT1	25993318	CHIP-SEQ	PODOCYTE
	ZNF217	24962896	CHIP-SEQ	MCF7
ORAI1	BCL6	25482012	CHIP-SEQ	CML/JURL-MK1
	E2F4	21247883	CHIP-SEQ	LYMPHOBLASTOID
	ELK3	25401928	CHIP-SEQ	HUVEC
	FLI1	21571218	CHIP-SEQ	MEGAKARYOCYTES
	FOXA2	19822575	CHIP-SEQ	HepG2
	FOXP1	21924763	CHIP-SEQ	HESC
	FOXP2	23625967	CHIP-SEQ	PFSK-1 AND SK-N-MC
	GATA6	25053715	CHIP-SEQ	YYC3
	KLF5	25053715	CHIP-SEQ	YYC3
	MITF	21258399	CHIP-SEQ	MELANOMA
	MYC	22102868	CHIP-SEQ	BL
	PPARD	21283829	CHIP-SEQ	MYOFIBROBLAST
	RUNX1	21571218	CHIP-SEQ	MEGAKARYOCYTES
	SMAD4	21741376	CHIP-SEQ	HESC
	TFEB	21752829	CHIP-SEQ	HELA
ORAI2	EBNA2	21746931	CHIP-SEQ	IB4-LCL
	ELK3	25401928	CHIP-SEQ	HUVEC
	GATA1	19941826	CHIP-SEQ	K562
	GATA2	19941826	CHIP-SEQ	K562
	MITF	21258399	CHIP-SEQ	MELANOMA
	MYC	22102868	CHIP-SEQ	BL
	PPARD	21283829	CHIP-SEQ	MYOFIBROBLAST
	RBPJ	21746931	CHIP-SEQ	IB4-LCL
	RUNX1	21571218	CHIP-SEQ	MEGAKARYOCYTES
	TP63	22573176	CHIP-SEQ	HFKS
	VDR	24763502	CHIP-SEQ	THP-1
ORAI3	BCL6	25482012	CHIP-SEQ	CML/JURL-MK1
	ELF1	20517297	CHIP-SEQ	JURKAT
	ELK3	25401928	CHIP-SEQ	HUVEC
	FOXA1	25552417	CHIP-SEQ	VCAP
	GATA2	19941826	CHIP-SEQ	K562

Continued

GATA2	21666600	CHIP-SEQ	HMVEC
GATA3	24758297	CHIP-SEQ	MCF7
GATA4	25053715	CHIP-SEQ	YYC3
HNF4A	19822575	CHIP-SEQ	HepG2
MITF	21258399	CHIP-SEQ	MELANOMA
NCOR1	26117541	CHIP-SEQ	K562
PPARD	21283829	CHIP-SEQ	MYOFIBROBLAST
SPI1	23547873	CHIP-SEQ	NB4
SPI1	23127762	CHIP-SEQ	K562
TFAP2C	20629094	CHIP-SEQ	MCF7

TF, transcription factor.

Table 4.2 List of the promoter sequences of STIM and ORAI genes obtained from the EPDnew database.

Promoter Name	Promoter Sequence
STIM1_1	TTGGGGGCTGGGAGCTCGCCCCCGGGCCGAGCCGGGTCAGGCTGTTGTCGCCTCAGGCAG CTCCTGGGAGGCTAACGTCGTGTCCTGGGCCTCTGTTTAGACAGCTCTAGAACTGAGGCG AGTGGAGCAGCACCAAGGCCCGGAGATCGGGGCAGGGCAGCTGCTGTCGCCGCCGCCGCA GGCCTGAGTTACCTGAGTAACTGCGGGTCAGGGACCCGCCCGACGGCCCGCGGTTGGCGC TGGAGACTCTCGGTGGGGAAAGGGAAGCTGGGACTTGATCCTTTGCGCGGGATCCTGGCA AAGACTAGCGCGGGCCGGGGGTCCGGGAGAGCCCGCTAGGGGCGGGGATTCCGGGGAGCC GTCTTCACCGGTTATTCCGGGATCCAGCTGGGCGCTGGGGCTGGCCCGGGCTTCGCTGGG GACCGGGCGGCGCGGGGCGGGCGCGGAGACGCACGCCCCCGCCCGCCCCGGGCCCGCCCC GCGCCGCCCGCCCGCCTGGAAGCCGCTGTCCTGGGCCTGGCCGGTGTGCGTCCGCCTGCT GGACCTGGGCACCGCCAGCCGCCTGGGCACGGGACTGGGCGGGGGCGCTGACCTCGGCCT
STIM1_2	CTCATTTCTTTCATTGCTTACAAGTAGATAGCATTCCAGTTCATAGGTTTCTTTGAGAAA TAGACTGTAGAAAAGACGACAATGTTTATTCTCATTAGTCAGACGAACTGCAGCACAAGT GTTTCAAAGAGGAACCCCAAGAACTCCCAGGCTTGTAGGAAGCATCTGATTTTACATAAG TTTGCACAGTAGGAAACTGAGGTCCTCGGGGAGGAACAAAGCAACTTATGATCAGACAAA TGAGTCACTAGTAGAGCTGAAATGAGAAGCCAGATCTCCTAACCACTCCCACACCCCATC ACAGTGCCTCACTGCGTCTCTTACTGGTGGGCTTGATTGCTTTCCAAGGCCAGAGAAGGA AGTAGCTGTTCCTGTATGCTCAGACAGGAGTATAAATCACACTGTGATGTCAGAAGCTTC TTTTCTAGCTGGAGAAATAAAGCTATGCATCAGAAAAGAGCACCAATCTTATAGAGTAGT ATAGAATTAAGTGCTTACTTGTGGGACCTAGACCAGTGGTTTTCAACCTTTCATATCTTT TACCCCCCTCCTCTTATATCCACTGTGTTCCAGCAGAAGAGAAACAAAGTTCCATTTTTC
STIM2_1	AAAACGCTAAGCTATGCTAACCGCGTCTAAACAGCCAGCCACTTGTAAGCCCCACCTATA TTGAAATTACGGGGACGCTGCTGTTTTCCGATTGCAAGATTTTCCTGATCCATGCAATTA CTTTCGCTGCCCTCTACGAGGCTGAAACTCGCCCTCAGGATGTGGGACGTCTGGACTCTT CTCTCCGTCCCCTTGTAGCCCCCCACTCCCCCTCGCGGTGGTACCGTGAATGAGGGAGAG GTACACGTCCCCCTTCTTCCCCGCCTCCTATCTTCGCGGCTCGCTAAAGCGTTATCAGCC GCCCCACGGTACTACCGTCCGTCTAGGAACGCCTCCGGGGCGGGGCTGGGATGCCGCGCA CGCGCAGTACAGCAGCGCCGCGCCTGCGCCGTGGAGAGCCTGAGGGAGGCGGGGGATTGG TATGCGAGCGAATGTGCGAGGGGAGGGAGGCGTCCCGGCGGAGCGTGGTACTACGACCAG CGCGGGCCGGAGGGGGCGGGGGGATGCGCCGCGGCGGCGGCGGCGCGGGAGCTGGGGTTG GTGTTTGGCGGCGCCAGAGCAGCGGATCCCGGTCTCGCCGCAGCAGCAGCGCGGGTGTCG
STIM2_2	TACAGATTTATTATAAATGTGTATACTTGAGAATAAATGAGTAGAAGAACAGTGGAGGTA AGTCAAATATAGTGGAATTAGATGTGTAGTTAAATTTTATTTTTAACTTGATTTAAATAA TTAGGAATTTTTGAAAAGCTTTTTGCGGAAGAGTATTTCCCTGCTTTCCCTGTCATTTGA ACCCAGGATAACCAAAATAGCTGTATAGTAAGTTGCCTGATATTTGTATTAACCAAACTT AAGGCTAATGAAAAATGCTATGATTTCTGATTGAAATATGTATTTAATCGCTTGACCCAG TATTCATTATTTTGTAAAAAAAAATAAAACTGGAAATTTTTGTGAGGAATTTTTATTTTT ATTGTTCTATAAGGCTTGAAAAGGCACAGGAAGAAAACAGAAATGTTGCTGTAGAAAAGC AAAATTTAGAGCGCAAAATGATGGATGAAATCAATTATGCAAAGGAGGAGGCTTGTCGGC TGAGAGAGCTAAGGGAGGGAGCTGAATGTGAATTGAGTAGACGTCAGTATGCAGAACAGG AATTGGAACAGGTATTTACATTAAAAAAAAAATCACTTGTAAAGATGTTAACATTGCCAC

Continued

ORAI1	GACAGGTTTTGTCCATCTCGTTCACCACCCTACCCAGCCTCAGCACCTAGACCAGTGTTG GCACCCAGTGGGCGCCAAATAAACACTGCTTGAACTCCAGACGTCAGCCGCTCTTTTTCC TACAGACCTTGAGCCACCTTGTTCCAAAGGGGATATGGGCCTCAGGAGGCGCCCAGAGGT GACCTCAGGCGGCCCGACCCAGGAGTCCAAGCTCCAGGAGCAGGGCCACGGGAGCAGCTG CGGAGAGGGGCGGCGCCAGGAGCCGGAGCGGGCAGCCGGGCGCTTCCAGGAAAAGTGGCG GGCGGCGGCGCGCCAGGGACCGTGGGCGGTGCCGTCGGAGCGGGCGGGTCACGTGACGCC CACAACAACGCCCACTTCTTGGTGGGCGGGGCACAGGTGGGCGGGGAGCATGCAAAACAG CCCAGGGCGGCGGCCAATCGCGGCGCGCGCCGGGGGTCCAGGCCCCGGGGATCCGAGGCG CCGCCCGCGCGCAGTCTCTGGTCACTGCCGCCCGGGGGCTTTTGCCAGCGGCGCCGCGGG CCTGCGTGCTGGGGCAGCGGGCACTTCTTCGACCTCGTCCTCCTCGTCCTGTGCGGCCGG
ORAI2	ACTCCCCGCCCCTCTACGCAGCCAGCGTCCAATGCTGGGCACCCCCCGAGGCTCACCCTG CCAAGCCTGGGGCTCCCCTTTTGCGCCCGGACCCAGGGGCAGGGAAAGCCCAGCTCGTGG TCTGTGGGTAGCCGGACCCCCGATGGGGCGGTGGGGGGCCTCGCCTTGACTCCCAGAGCT GGGGCCGGGGACAGGAGCTGGGGCAGGAGGGATGCGCGCGGGTCGGGGTCTTCCCACCTC CCCTGCTCCTCTCCCTCGCGCGATCCCGGGGTGGTTCCAGGTGAGGCGGGGACCCCCACC CCCCCACTCTCCGAGGAGGCGCCGCCAGCCCCGCCCCTCCCGGCCCGGCGGGTGACGTGG CCGCGGCTCTCCCGCGGGTGGGTCACGTGTTGGCGGCGCCTGGTTGCCTTGGCAGCGGCT GCGGCGGCCGCGGGGGCGGGGTGGAGGCGGGGCCGGGGGACCCCGCGCGACCGGCGGAAG GAGGGAGGGGGCCGCGCTCGGCGCCCCGGCCGGGCCACTGGGCCACAGGCCACGCGGCCA CGCAGTCCGAGCGGGAGCCGAGCCGGGCGGGGCGAGGGCAGCTCCGGTGAGTGTGGCGGC
ORAI3	GTAACAGGGAGGTGCGCGGGTGGGGGGAGGGCTGGGCGGACCAAAGGCCGGAGGGGTGGG GCCTGGGGATAGCGAGAGGCTTGAGAATGGGGCCGCTTGGGGGAGGGAAGAGGCAGCCCG GCGAGGGGCAAGCGGGGGACCCAGCCGGGCTGGGCCCCTGGGCCCCGGGTCTGTACAATA CGGTTTGCTATAAAACTCAAAATCTTCCAGCCGGGGCTGCGGAGTTCGTGTGTGTATCTG CGGGGTCCCTACCTACAGATGAGTGGGCTCACCTCTCCTGGACTCATTTTGGGAGGGATT TGGAAGTGTGGACACCTGGGGTGTCCAGCTGTACCTTGGAGGGGGCTGGGGTTGGCGTGC ACCTCGGTGGGGTCCGGGCGCTTGGATAACGTTCTTGGTGGGTAGGGGTCGCGGGGAATC TCTGCGGGCCCGGGACTGCGGGGACTTGGTCCCCGGCTCCACCCCATCATGTGGCTAGCC CCGGCTCCGCCTCTGTCCCAGTTCCTGTTTTGGCCTCCGCTGTCCCGCTCCGGCTCCTGG GGCTCCCCGCAGACGCTGCTTTTCTTGCTCCACTGGGGGTGCCTCTTCCTGGGCGCCCGC

Table 4.3 List of the predicted interactions between transcription factors, name of bridge proteins and score obtained from the STRING database.

Promoter name	Potential Interactions	Type of interaction	Bridge Proteins	Interactions	node1	node2	Score	node1	node2	Score	node1	node2	Score
STIM2_1	E2F1:E2F4	Direct			E2F1	E2F4	0.966						
	E2F1:E2F4	Bridge protein	MYC	E2F1>MYC>E2F4	E2F1	MYC	0.982	MYC	E2F4	0.982			
STIM2_1	E2F1:EGR1	Bridge protein	SPI1	E2F1>SPI1>EGR1	E2F1	SPI1	0.905	SPI1	EGR1	0.501			
	E2F1:EGR1	Bridge protein	E2F4	E2F1>E2F4>EGR1	E2F1	E2F4	0.966	E2F4	EGR1	0.846			
	E2F1:EGR1	Bridge protein	MYC	E2F1>MYC>EGR1	E2F1	MYC	0.982	MYC	EGR1	0.928			
STIM2_1	E2F1:GATA1	Bridge protein	ELF1	E2F1>ELF1>GATA1	E2F1	ELF1	0.902	ELF1	GATA1	0.893			
	E2F1:GATA1	Bridge protein	SPI1	E2F1>SPI1>GATA1	E2F1	SPI1	0.905	SPI1	GATA1	0.995			
	E2F1:GATA1	Bridge protein	MYC	E2F1>MYC>GATA1	E2F1	MYC	0.982	MYC	GATA1	0.961			
STIM2_1	E2F4:GATA1	Direct			E2F4	GATA1	0.727						
	E2F4:GATA1	Bridge protein	MYC	E2F4>MYC>GATA1	E2F4	MYC	0.96	MYC	GATA1	0.961			
	E2F4:GATA1	Bridge protein	EGR1	E2F4>EGR1>GATA1	E2F4	EGR1	0.846	EGR1	GATA1	0.584			
ORAI1	PPARD:E2F4	Bridge protein	EGR1	PPARD>EGR1>E2F4	PPARD	EGR1	0.691	EGR1	E2F4	0.846			
ORAI2	NFKB1:RUNX1	Bridge protein	EGR1	NFKB1>EGR1>RUNX1	NFKB1	EGR1	0.671	EGR1	RUNX1	0.676			
	NFKB1:RUNX1	Bridge protein	MYC	NFKB1>MYC>RUNX1	NFKB1	MYC	0.993	MYC	RUNX1	0.61			
	NFKB1:RUNX1	Bridge protein	SPI1	NFKB1>SPI1>RUNX1	NFKB1	SPI1	0.781	SPI1	RUNX1	0.991			
	PPARD:MYC	Bridge protein	EGR1	PPARD>EGR1>MYC	PPARD	EGR1	0.691	EGR1	MYC	0.928			
	PPARD:MYC	Bridge protein	HNF4A	PPARD>HNF4A>MYC	PPARD	HNF4A	0.916	HNF4A	MYC	0.942			
ORAI3	MITF:ELF1	Bridge proteins	SPI1, E2F1	MITF>SPI1>E2F1>ELF1	MITF	SPI1	0.951	SPI1	E2F1	0.905	E2F1	ELF1	0.902
	MITF:ELF1	Bridge proteins	SPI1, GATA1	MITF>SPI1>GATA1>ELF1	MITF	SPI1	0.951	SPI1	GATA1	0.995	GATA1	ELF1	0.893
	MITF:SPI1	Direct			MITF	SPI1	0.951						
	SPI1:ELF1	Bridge protein	GATA1	SPI1>GATA1>ELF11	SPI1	GATA1	0.995	GATA1	ELF1	0.893			
	SPI1:ELF1	Bridge protein	E2F1	SPI1>E2F1>ELF1	SPI1	E2F1	0.905	E2F1	ELF1	0.902			

Table 4.4 List of differential interactions in (normal-tumor) and (tumor-normal) samples.

Node 1	Node 2	Weight	Node 1	Node 2	Weight
Normal-tumor			**Tumor-normal**		
EP300	CREBBP	0.2822	FOXP2	ELK3	0.1546
NCOR1	EP300	0.2343	MITF	RUNX2	0.1528
SP1	EP300	0.2093	FOXA1	GATA3	0.1112
NCOR1	CREBBP	0.1903	FLI1	ELK3	0.1095
MITF	ELK3	0.1636	SPI1	FLI1	0.1049
NCOR1	AR	0.1624	AR	EP300	0.0869
SMAD4	SP1	0.1448	SPI1	TFEB	0.0749
MITF	FLI1	0.1331	FOXP2	FLI1	0.0737
SP1	CREBBP	0.1330	SP1	EP300	0.0673
TOP2B	SP1	0.1270	FOXP2	RUNX2	0.0625
FLI1	ELK3	0.1135	FOXP2	MITF	0.0623
FOXA1	GATA3	0.0985	TOP2B	SP1	0.0623
NCOR2	TRIM28	0.0976	STIM2	FLI1	0.0604
MITF	TRIM28	0.0963	ZNF217	TFEB	0.0584
NCOR1	SP1	0.0929	MITF	ELK3	0.0579
AR	EP300	0.0915	SP1	AR	0.0552
TTF2	EP300	0.0881	FOXA1	FLI1	0.0552
TRIM28	ELK3	0.0829	FLI1	TFEB	0.0536
FLI1	HOXB7	0.0820	TOP2B	EP300	0.0532
KDM1A	HOXB7	0.0807	NCOR1	SP1	0.0487
SMAD4	TOP2B	0.0805	AR	CREBBP	0.0484
KDM1A	HDAC2	0.0803	ELF1	TOP2B	0.0474
KDM1A	FLI1	0.0743	TP63	KLF5	0.0470
TTF2	SP1	0.0698	EP300	ELK3	0.0440
RBPJ	TRIM28	0.0688	RUNX2	ELK3	0.0431
KLF5	TFAP2C	0.0686	E2F1	RUNX1	0.0424
HOXB7	TFAP2C	0.0670	ELF1	SP1	0.0410
HDAC3	KDM1A	0.0647	FOXP2	EGR1	0.0387
KDM1A	MITF	0.0624	TOP2B	AR	0.0378
TOP2B	EP300	0.0603	FLI1	GATA3	0.0378
FOXP1	TP63	0.0587	TP63	EGR1	0.0369
FOXP2	HDAC1	0.0549	EGR1	ELK3	0.0363
AR	CREBBP	0.0548	EGR1	FLI1	0.0360
MITF	VDR	0.0544	NCOR1	EP300	0.0349
TRIM28	FLI1	0.0539	SP1	CREBBP	0.0324
MITF	RBPJ	0.0539	STIM2	ELK3	0.0313
FOXP2	MITF	0.0536	FOXA1	E2F4	0.0310
HOXB7	VDR	0.0525	NFKB1	ELK3	0.0307
MITF	HOXB7	0.0522	NCOR1	AR	0.0306
FLI1	VDR	0.0516	ELF1	ELK3	0.0300
NCOR1	PHF8	0.0496	FOXA1	SPI1	0.0295

Continued

SMAD4	ORAI2	0.0491	MITF	TRIM28	0.0294
KDM1A	TFAP2C	0.0486	SP1	ELK3	0.0283
TP63	TFAP2C	0.0478	MITF	FLI1	0.0277
SMAD4	CUX1	0.0457	PHF8	CREBBP	0.0262
TTF2	TOP2B	0.0424	SPI1	E2F4	0.0258
SMAD4	EP300	0.0417	FOXA1	TFEB	0.0257
FOXA1	PHF8	0.0404	AR	ELK3	0.0248
KLF5	CUX1	0.0398	ELF1	AR	0.0247
FOXA1	AR	0.0392	ELF1	EP300	0.0244
ORAI2	SP1	0.0383	SP1	CDKN2A	0.0240
PHF8	AR	0.0379	PHF8	AR	0.0239
FLI1	TFAP2C	0.0378	FOXP2	STIM2	0.0235
HOXB7	GATA2	0.0370	RUNX2	RUNX1	0.0231
KDM1A	ELK3	0.0363	NCOR1	TOP2B	0.0225
GATA6	ELK3	0.0360	NCOR1	CREBBP	0.0216
RBPJ	FLI1	0.0351	NFKB1	AR	0.0212
FOXP2	TRIM28	0.0331	FOXP2	RUNX1	0.0211
TTF2	CREBBP	0.0328	ELK3	RUNX1	0.0205
FOXP2	ELK3	0.0306	ELF1	NFKB1	0.0203
TTF2	NCOR1	0.0302	FLI1	GATA6	0.0201
KDM1A	VDR	0.0293			
NCOR2	RBPJ	0.0286			
RBPJ	ELK3	0.0282			
NCOR1	TOP2B	0.0278			
TRIM28	VDR	0.0277			
MITF	TFAP2C	0.0264			
FLI1	GATA2	0.0257			
TOP2B	CREBBP	0.0257			
SP1	AR	0.0251			
MITF	GATA6	0.0248			
TP63	CREBBP	0.0247			
KDM1A	TRIM28	0.0245			
PHF8	EP300	0.0238			
VDR	ELK3	0.0237			
NCOR2	FLI1	0.0236			
MITF	NCOR2	0.0225			
TP63	HOXB7	0.0222			
FOXP1	SMAD4	0.0217			
HDAC3	HDAC2	0.0211			
SP1	CDKN2A	0.0207			
PHF8	CREBBP	0.0201			
TOP2B	HOXB7	0.0201			

8.3 Supplementary Information for Chapter 5

Table 5.1 Dataset of 167 protein-ligand complexes.

1Q8T	1Q8U	1XD0	2YKI	3JVS	2J62	2W66	4DEW
1VSO	2WBG	3B3W	1W3K	2CBJ	3OZT	3GCS	2X97
3OE5	1LOR	2QBP	3F3E	3I3B	3UO4	3ZSX	2V7A
3L4W	2QBR	3PXF	2YFE	3OV1	3HUC	2WTV	3F3C
2ZXD	2VVN	3F3A	2D3U	1H23	2ZWZ	3SU2	3BFU
2P4Y	1YC1	2J78	2X8Z	3GBB	2VW5	3GE7	1R5Y
4DE2	3UDH	3G2Z	3EBP	2GSS	2XHM	3MSS	1U33
1SLN	3F17	2X00	2BRB	3AO4	4DE1	4GID	3D4Z
2ZCR	3IMC	10GS	1GPK	1PS3	3PE2	3ACW	3FCQ
3CYX	2CET	3K5V	2V00	1OS0	1Z95	3BKK	3MFV
2ZJW	3DD0	3N86	3EHY	3E93	1LOQ	3F80	1U1B
3NW9	3IVG	3N7A	1NVQ	2OBF	3G0W	1N2V	3VH9
1LBK	3AG9	4G8M	2G70	3SU5	2VO5	3S8O	3B68
3PWW	2XY9	3ZSO	2FVD	1KEL	3UEX	2ZCQ	2YGE
3SU3	1SQA	2D1O	2VOT	2WEG	2XNB	3B3S	3CJ2
3FK1	4TMN	4DJR	1P1Q	2XDL	4GQQ	3GNW	3DXG
2X0Y	3MUZ	1HNN	3EJR	2QFT	2YMD	2QMJ	2XYS
3VD4	2HB1	1HFS	3L3N	3OWJ	3L7B	1W3L	3COY
3NQ3	3NOX	2ZX6	3LKA	3FV1	3MYG	2PQ9	3KV2
2OLE	1LOL	1JYQ	2IWX	3L4U	2VL4	3KWA	4DJV
3U9Q	3UEU	2WCA	1E66	2XB8	3G2N	1QI0	

Table 5.2 Names, URLs, year of creation and respective limitations of the 18 pocket identification tools.

Program	URL	Year	Limitation
APROPOS	http://www.csb.yale.edu/poststructure/apropos/apropos.html	1996	Need CSB core
BetaCavityWeb	http://voronoi.hanyang.ac.kr/betacavityweb/	2015	Available only as webserver
CASTp	http://sts.bioe.uic.edu/castp/calculation.php	2003/2006	Available only as webserver
CAVER Analyst	http://www.caver.cz	2014	No output file, not run automatically
ConCavity	http://compbio.cs.princeton.edu/concavity/	2009	Not compatible with Debian 16.0-4-amd64
Epock	http://epock.bitbucket.org	2014	Need VMD
HotPatch	http://hotpatch.mbi.ucla.edu/	2007	Incomplete documentation
LIGSITEcs	http://projects.biotec.tu-dresden.de/pocket/	1997	Need BALL algorithm
McVol	http://www.bisb.uni-bayreuth.de/index.php?page=data/mcvol/mcvol	2010	Does not take PDB as input
Metapocket2.0	http://projects.biotec.tu-dresden.de/metapocket/	2011	Available only as webserver
MolSite	http://presto.protein.osaka-u.ac.jp/myPresto4/index.php?lang=en	2011	Incomplete documentation
PASS	http://www.ccl.net/cca/software/UNIX/pass/overview.shtml	2000	Not compatible with Debian 16.0-4-amd64
POCASA	http://altair.sci.hokudai.ac.jp/g6/service/pocasa/	2010	Available only as webserver
PocketAnalyzer	http://sourceforge.net/projects/papca/ or http://cpclab.uni-duesseldorf.de/downloads	2011	Not compatible with Debian 16.0-4-amd64
QSiteFinder	http://www.bioinformatics.leeds.ac.uk/qsitefinder	2005	Not compatible with Debian 16.0-4-amd64
Rate4Site	bioinfo.tau.ac.il/ConSurf	2002	Only available in Windows OS
SITEHOUND	http://scbx.mssm.edu/sitehound/sitehound-web/Input.html	2009	Not compatible with Debian 16.0-4-amd64
SURFNET	http://www.ebi.ac.uk/thornton-srv/software/SURFNET/	1995	Not compatible with Debian 16.0-4-amd64

VMD,Visual Molecular Dynamics; PDB, protein data bank and OS, operating system.

Table 5.3 List of the descriptors obtained from the DoGSiteScorer tool.

Group	Label	Description
Ligand descriptors	lig_cov	Percentage of ligand covered by the predicted pocket
	poc_cov	Percentage of the pocket covered by the co-crystallized ligand
Size and shape descriptors	volume	Pocket volume in A^3 calculated via grid points
	surface	Pocket surface in A^2 calculated via grid points
	lipo_surf	solvent accessible lipophilic surface;
	depth	Depth of the pocket in A
	ellips c/a or b/a	Ellipsoid main axes ratios, with a > b > c
	enclosure	Rratio of number of surface to hull grid points
Functional group descriptors	H-don	Number of hydrogen bond donors
	H-acc	Number of hydrogen bond acceptors
	Met	Number of metals
	Hphob	Number of hydrophobic contacts
	siac ratio	Relative number of hydrophobic site interaction centers (SIACs, from flex)
Element descriptors	nof_dif_atms	Number of surface atoms lining the pocket
	elem_x	Number of elements of specific type in active site; types: C, N, O, S or other (X)
Amino acid composition	aa_apol, aa_pol and aa_neg	Relative number of amino acids apolar, polar, positive, and negative)
Amino acid descriptors	ALA, ARG, ASN, ASP, CYS, GLN, GLU, GLY, HIS, ILE, LEU, LYS, MET, PHE, PRO, SER, THR, TRP, TYR and VAL	Number of amino acids in pocket, 3-letter code of 20 amino acid types

Table 5.4 List of the weight of each tool for each protein.

PDB ID	DEPTH	GHECOM	Fpocket	DoGSiteScorer	IsoMif	ProACT2
1Q8T	**0.4133**	0.1937	0.2018	0.0677	0.0871	0.0363
1Q8U	**0.4149**	0.2008	0.1893	0.0711	0.0875	0.0363
1XD0	**0.3998**	0.1663	0.1428	0.0929	0.1599	0.0383
2YKI	**0.3981**	0.2065	0.1720	0.0887	0.1021	0.0327
3JVS	0.2649	0.0929	0.1151	0.1324	**0.3245**	0.0701
2J62	**0.2634**	0.1428	0.2079	0.1629	0.1664	0.0567
2W66	**0.2709**	0.1183	0.1988	0.1527	0.1975	0.0618
4DEW	**0.2799**	0.1049	0.1460	0.1810	0.2037	0.0845
1VSO	**0.2675**	0.1421	0.2116	0.1451	0.1718	0.0619
2WBG	**0.3796**	0.2107	0.1432	0.0787	0.1425	0.0453
3B3W	**0.3621**	0.1279	0.1229	0.1004	0.2305	0.0563
1W3K	**0.2837**	0.1284	0.2061	0.1617	0.1569	0.0632
2CBJ	**0.3433**	0.1332	0.1343	0.1027	0.2234	0.0631
3OZT	**0.2524**	0.1665	0.2263	0.1449	0.1583	0.0515
3GCS	**0.3593**	0.2084	0.1833	0.0800	0.1359	0.0331
2X97	**0.3794**	0.1841	0.2701	0.0657	0.0743	0.0263
3OE5	**0.3443**	0.2091	0.1697	0.0897	0.1499	0.0373
1LOR	**0.2656**	0.1598	0.1891	0.1647	0.1675	0.0533
2QBP	**0.2676**	0.1080	0.1657	0.1399	0.2425	0.0763
3F3E	**0.3170**	0.0885	0.1060	0.1394	0.2675	0.0816
3I3B	**0.2605**	0.1602	0.2600	0.1412	0.1333	0.0448
3UO4	**0.3595**	0.1491	0.1715	0.0998	0.1794	0.0407
3ZSX	**0.3797**	0.1957	0.1857	0.0835	0.1116	0.0439
2V7A	**0.3591**	0.1501	0.1335	0.0890	0.2123	0.0561
3L4W	**0.2539**	0.1482	0.2218	0.1698	0.1647	0.0416
2QBR	**0.2545**	0.1071	0.1713	0.1459	0.2411	0.0801
3PXF	**0.3847**	0.1979	0.1989	0.0812	0.1083	0.0290
2YFE	**0.4139**	0.2012	0.1885	0.0714	0.0963	0.0287
3OV1	**0.2798**	0.1208	0.1715	0.1646	0.2176	0.0457
3HUC	**0.3762**	0.2115	0.1956	0.0845	0.0991	0.0331
2WTV	**0.3581**	0.1694	0.1975	0.0967	0.1398	0.0385
3F3C	**0.2674**	0.1753	0.2301	0.1415	0.1372	0.0485
2ZXD	**0.3690**	0.2211	0.1714	0.0840	0.1248	0.0297
2VVN	**0.2577**	0.1213	0.2214	0.1546	0.1911	0.0539
3F3A	**0.3711**	0.1865	0.1614	0.0977	0.1388	0.0445
2D3U	**0.3539**	0.1333	0.1462	0.0964	0.2297	0.0405
1H23	**0.3562**	0.2298	0.1538	0.0793	0.1429	0.0381
2ZWZ	**0.3785**	0.1522	0.1206	0.0819	0.2225	0.0442
3SU2	**0.3553**	0.1255	0.1467	0.1062	0.2203	0.0460
3BFU	**0.2563**	0.1304	0.2397	0.1555	0.1566	0.0615
2P4Y	**0.4201**	0.1800	0.2041	0.0750	0.0897	0.0311
1YC1	**0.3806**	0.1821	0.1683	0.0905	0.1289	0.0496

Continued

2J78	**0.3884**	0.2140	0.1579	0.0844	0.1135	0.0418
2X8Z	**0.3873**	0.1739	0.2635	0.0687	0.0769	0.0297
3GBB	**0.4069**	0.1970	0.1688	0.0785	0.1123	0.0365
2VW5	**0.3799**	0.1929	0.1595	0.0866	0.1300	0.0511
3GE7	**0.3484**	0.1782	0.2081	0.0867	0.1327	0.0458
1R5Y	0.2391	0.1301	**0.2643**	0.1558	0.1491	0.0617
4DE2	**0.2651**	0.1292	0.2076	0.1613	0.1668	0.0701
3UDH	**0.4061**	0.1951	0.2027	0.0788	0.0873	0.0299
3G2Z	**0.2581**	0.1122	0.2129	0.1620	0.1860	0.0688
3EBP	**0.3483**	0.1386	0.1527	0.1007	0.2143	0.0453
2GSS	**0.2590**	0.1345	0.2499	0.1361	0.1651	0.0553
2XHM	**0.3878**	0.1977	0.2380	0.0734	0.0809	0.0222
3MSS	**0.3987**	0.2165	0.1761	0.0735	0.1015	0.0337
1U33	**0.3343**	0.1734	0.1813	0.1213	0.1471	0.0427
1SLN	**0.3597**	0.1971	0.1919	0.0851	0.1304	0.0359
3F17	**0.3821**	0.1596	0.1650	0.1002	0.1493	0.0439
2X00	0.2487	0.1189	0.1581	0.1612	**0.2579**	0.0551
2BRB	**0.3511**	0.2136	0.1689	0.0864	0.1398	0.0402
3AO4	0.2528	0.1397	**0.3007**	0.1416	0.1215	0.0437
4DE1	**0.2556**	0.1245	0.2134	0.1675	0.1739	0.0652
4GID	**0.4089**	0.1918	0.2181	0.0707	0.0828	0.0277
3D4Z	**0.2929**	0.0859	0.1403	0.1359	0.2788	0.0662
2ZCR	**0.4160**	0.1872	0.1901	0.0751	0.0962	0.0354
3IMC	**0.3658**	0.1946	0.1752	0.0920	0.1290	0.0434
10GS	0.2425	0.1485	**0.2463**	0.1468	0.1689	0.0470
1GPK	**0.3622**	0.2275	0.1535	0.0772	0.1415	0.0381
1PS3	**0.3089**	0.0839	0.1357	0.1309	0.2764	0.0643
3PE2	**0.3659**	0.1701	0.1457	0.0823	0.1911	0.0449
3ACW	**0.4033**	0.1907	0.2083	0.0743	0.0889	0.0345
3FCQ	**0.2641**	0.1451	0.2276	0.1387	0.1711	0.0535
3CYX	**0.3889**	0.1417	0.1471	0.0891	0.1770	0.0563
2CET	**0.3820**	0.2205	0.1569	0.0859	0.1138	0.0409
3K5V	**0.3689**	0.2124	0.1660	0.0861	0.1275	0.0392
2V00	**0.3695**	0.1891	0.1833	0.0811	0.1287	0.0483
1OS0	**0.3317**	0.1728	0.1963	0.0973	0.1623	0.0397
1Z95	**0.4029**	0.1514	0.1545	0.0835	0.1675	0.0403
3BKK	**0.3863**	0.1892	0.2387	0.0697	0.0835	0.0327
3MFV	**0.3560**	0.1737	0.1672	0.0955	0.1601	0.0474
2ZJW	**0.3655**	0.2112	0.1799	0.0777	0.1320	0.0337
3DD0	**0.2764**	0.1277	0.1889	0.1473	0.2013	0.0585
3N86	**0.2669**	0.1487	0.2331	0.1452	0.1492	0.0569
3EHY	**0.3535**	0.1721	0.1775	0.1085	0.1447	0.0436
3E93	**0.3951**	0.2141	0.1760	0.0691	0.1096	0.0361

Continued

1LOQ	**0.3578**	0.2029	0.1504	0.1009	0.1505	0.0375
3F80	**0.3698**	0.1705	0.1628	0.0973	0.1491	0.0506
1U1B	**0.2616**	0.1033	0.2079	0.1703	0.2027	0.0541
3NW9	**0.2511**	0.1711	0.2099	0.1463	0.1625	0.0591
3IVG	**0.4155**	0.1983	0.1835	0.0734	0.0945	0.0347
3N7A	**0.2713**	0.1502	0.2387	0.1393	0.1464	0.0541
1NVQ	**0.3624**	0.1580	0.1569	0.0853	0.1903	0.0472
2OBF	**0.3819**	0.1906	0.1879	0.0873	0.1146	0.0377
3G0W	**0.3903**	0.1891	0.1840	0.0842	0.1185	0.0339
1N2V	**0.3244**	0.1948	0.1935	0.0965	0.1465	0.0443
3VH9	**0.2863**	0.1119	0.1313	0.1361	0.2704	0.0639
1LBK	**0.3877**	0.2125	0.1763	0.0713	0.1189	0.0332
3AG9	**0.3723**	0.1294	0.1280	0.1019	0.2185	0.0499
4G8M	**0.3749**	0.1991	0.1951	0.0865	0.1081	0.0361
2G70	**0.3613**	0.1873	0.1955	0.0948	0.1221	0.0390
3SU5	**0.3655**	0.1380	0.1343	0.1049	0.2136	0.0437
2VO5	**0.2533**	0.1543	0.2135	0.1504	0.1642	0.0643
3S8O	**0.2515**	0.1237	0.1910	0.1645	0.2228	0.0465
3B68	**0.3807**	0.1694	0.1623	0.0819	0.1648	0.0410
3PWW	**0.3497**	0.1793	0.1529	0.1021	0.1746	0.0414
2XY9	**0.3882**	0.1905	0.2272	0.0776	0.0839	0.0325
3ZSO	**0.3119**	0.1780	0.2135	0.1119	0.1237	0.0609
2FVD	**0.3752**	0.2055	0.2073	0.0765	0.1058	0.0297
1KEL	**0.3009**	0.1216	0.2132	0.1412	0.1685	0.0545
3UEX	**0.3623**	0.1668	0.1639	0.0943	0.1785	0.0342
2ZCQ	**0.4172**	0.1855	0.1895	0.0747	0.1019	0.0311
2YGE	**0.3661**	0.1964	0.1641	0.0809	0.1414	0.0511
3SU3	**0.3632**	0.1373	0.1341	0.1067	0.2134	0.0453
1SQA	**0.2332**	0.1385	0.2138	0.1915	0.1757	0.0472
2D1O	**0.3658**	0.1863	0.1786	0.0900	0.1384	0.0409
2VOT	**0.4032**	0.1359	0.1175	0.0817	0.2115	0.0503
2WEG	**0.4151**	0.1397	0.1293	0.0887	0.1809	0.0463
2XNB	**0.3553**	0.1969	0.1881	0.0851	0.1378	0.0368
3B3S	0.2495	0.0982	0.1517	0.1665	**0.2637**	0.0704
3CJ2	0.2575	0.0997	0.1736	0.1493	**0.2597**	0.0602
3FK1	0.2670	0.1479	**0.2725**	0.1439	0.1109	0.0578
4TMN	0.2402	0.1521	**0.2409**	0.1543	0.1514	0.0610
4DJR	**0.3558**	0.1428	0.1231	0.1099	0.2175	0.0509
1P1Q	**0.4020**	0.1787	0.2029	0.0812	0.0987	0.0365
2XDL	**0.3686**	0.1537	0.1479	0.0972	0.1790	0.0537
4GQQ	0.2517	0.0979	0.1211	0.1647	**0.2953**	0.0695
3GNW	**0.3769**	0.2130	0.1803	0.0793	0.1181	0.0324
3DXG	**0.2593**	0.1136	0.2285	0.1737	0.1713	0.0535

Continued

2X0Y	**0.2764**	0.1535	0.2337	0.1458	0.1389	0.0516
3MUZ	**0.2579**	0.1597	0.2509	0.1475	0.1367	0.0473
1HNN	**0.3899**	0.1849	0.1960	0.0812	0.1089	0.0391
3EJR	0.2715	0.0863	0.1446	0.1441	**0.2895**	0.0640
2QFT	**0.4253**	0.1908	0.1888	0.0744	0.0817	0.0389
2YMD	**0.2671**	0.1069	0.1763	0.1738	0.2201	0.0557
2QMJ	**0.3813**	0.1194	0.1027	0.0948	0.2598	0.0420
2XYS	**0.3611**	0.1461	0.1326	0.0991	0.2165	0.0446
3VD4	**0.2621**	0.1286	0.1520	0.1450	0.2527	0.0596
2HB1	0.2612	0.1244	**0.2735**	0.1507	0.1431	0.0471
1HFS	**0.2732**	0.1491	0.2359	0.1328	0.1661	0.0430
3L3N	**0.3594**	0.2314	0.1736	0.0763	0.1234	0.0359
3OWJ	**0.3834**	0.1963	0.1951	0.0727	0.1128	0.0397
3L7B	**0.3913**	0.1925	0.2401	0.0686	0.0771	0.0304
1W3L	**0.2803**	0.1223	0.1995	0.1619	0.1733	0.0627
3COY	**0.4210**	0.1946	0.1849	0.0743	0.0918	0.0333
3NQ3	**0.3653**	0.1797	0.1717	0.0968	0.1520	0.0345
3NOX	0.2513	0.0961	0.1297	0.1398	**0.3114**	0.0717
2ZX6	**0.3739**	0.1797	0.1630	0.0853	0.1582	0.0399
3LKA	**0.3635**	0.1766	0.1665	0.1055	0.1439	0.0441
3FV1	**0.3933**	0.2019	0.1709	0.0787	0.1163	0.0388
3MYG	**0.3527**	0.2009	0.1989	0.0861	0.1263	0.0351
2PQ9	**0.4107**	0.1839	0.2146	0.0721	0.0837	0.0351
3KV2	**0.3541**	0.1728	0.1762	0.0965	0.1539	0.0464
2OLE	**0.2768**	0.1041	0.1787	0.1509	0.2360	0.0535
1LOL	**0.3349**	0.1709	0.1585	0.1129	0.1865	0.0363
1JYQ	**0.2774**	0.0991	0.1475	0.1526	0.2646	0.0589
2IWX	**0.4143**	0.2165	0.1744	0.0714	0.0933	0.0301
3L4U	0.2695	0.0859	0.1381	0.1432	**0.3048**	0.0585
2VL4	**0.2695**	0.1279	0.1900	0.1509	0.1963	0.0653
3KWA	**0.2726**	0.1229	0.1911	0.1413	0.2076	0.0645
4DJV	**0.2601**	0.1085	0.1834	0.1655	0.2207	0.0617
3U9Q	**0.4046**	0.2014	0.2074	0.0717	0.0879	0.0270
3UEU	**0.3611**	0.1628	0.1580	0.0961	0.1789	0.0430
2WCA	**0.3901**	0.1985	0.1615	0.0863	0.1305	0.0331
1E66	**0.3604**	0.2369	0.1500	0.0743	0.1390	0.0394
2XB8	**0.2739**	0.1603	0.2293	0.1365	0.1443	0.0557
3G2N	**0.3911**	0.1987	0.2305	0.0737	0.0744	0.0317
1QI0	**0.2703**	0.1301	0.2058	0.1737	0.1650	0.0551