

Multi-Scale Analysis and Prediction of Protein-Protein Interactions

by

Nurcan Tunçbağ

**A Thesis Submitted to the
Graduate School of Sciences and Engineering
in Partial Fulfillment of the Requirements for
the Degree of**

Doctor of Philosophy

in

Computational Sciences and Engineering

Koç University

October, 2010

Koç University
Graduate School of Sciences and Engineering

This is to certify that I have examined this copy of a doctoral dissertation by

Nurcan Tunçbağ

and have found that it is complete and satisfactory in all respects,
and that any and all revisions required by the final
examining committee have been made.

Committee Members:

Assoc. Prof. Özlem Keskin

Prof. Attila Gürsoy

Assist. Prof. Halil Kavaklı

Assist. Prof. Elif Özkırmı

Assist. Prof. Mehmet Sayar

Date: _____

ABSTRACT

Proteins act coherently in the cells and their roles span functions as diverse as being molecular machines and signaling. The mechanism behind this excellent synchronization is still uncovered. However, considerable effort has been centered on identifying of binding partners and binding regions, because the vast majority of the chores in the living cell involve protein–protein interactions. Proteins interact through their interfaces which contain *hot spots*, the residues contributing more to the binding energy. Hot spots are important for drug targeting and interaction specificity. In addition, structural modeling of protein interactions and incorporating them into the protein interaction networks are prerequisites for understanding cell function. Hence, the focus of this dissertation is directed to the question “*how do the proteins interact?*” rather than the question “*which proteins interact?*” at the top level. Towards this aim, firstly, this dissertation focuses on the prediction of hot spots in protein interfaces and their organization. Here, an efficient hot spot prediction model is developed and implemented that reaches an accuracy of 70% on the experimental data. A web server, namely HotPoint, is constructed based on this model. In another aspect, a novel graph-based method based on minimum cut trees developed to determine the organization of hot spots which reveal the cooperative relation between them. Nature presents a limited number of distinct binding site motifs and *structurally different protein pairs can use the same binding architectures*. Based on this origin, secondly, a multi-scale combinatorial strategy is illustrated to model protein complexes at proteome-level. This work shows how available structural information can help in modeling a pathway by using structural similarity. Here, the sample pathway is the tumor suppressor protein p53 pathway. Finally, the multi-partner proteins dataset is extracted from Protein Databank. Integration of time notion into protein interaction networks is demonstrated on two hub proteins, p53 and Mdm2 using both predictions and available structural data.

ÖZET

Proteinler hücre içinde birbirleriyle uyumlu şekilde hareket ederler ve rolleri moleküler duzenekten sinyal iletimine kadar geniş bir alana yayılır. Bu kusursuz senkronizasyonun ardındaki mekanizma henüz tam olarak aydınlatılamamıştır. Ancak, proteinlerin bağlandıkları eşlerinin ve bağlanma bölgelerinin belirlenmesi için kaydadeğer çabalar harcanmıştır, çünkü yaşayan hücrelerdeki işlerin büyük bir çoğunluğu proteinlerin etkileşimleriyle gerçekleşir. Proteinler arayüzey bölgelerinden etkileşirler ve bu bölgedeki *sıcak nokta* denilen bazı aminoasitler bağlanma enerjisine daha fazla katkıda bulunurlar. Sıcak noktalar ilaç hedefi olarak ve etkileşim özgünlüğü açısından önemlidir. Buna ek olarak protein etkileşimlerinin yapısal olarak modellenmesi ve bunların protein etkileşim ağlarıyla birleştirilmesi hücre fonksiyonunun anlaşılması açısından bir başka olmazsa olmazdır. Bu yüzden bu tez çalışmasının en üst seviyedeki odak noktası *hangi proteinlerin etkileştiği* sorusundan daha çok *proteinlerin nasıl etkileştiği* sorusuna yönlendirilmiştir. Bu amaçla, ilk olarak protein arayüzeylerindeki sıcak noktaların tahmin edilmesi ve organizasyonuna yoğunlaşmıştır. Burada, deneysel bilgi üzerinde %70 doğruluğa ulaşan verimli bir sıcak nokta tahmin modeli oluşturulmuş ve bu model kullanılarak HotPoint isimli bir ağ sunucusu yapılmıştır. Ayrıca, sıcak noktaların organizasyonuna karar vermek için, bunların müşterek ilişkilerini gösteren ağ-tabanlı minimum kesik ağacına dayalı bir metod geliştirilmiştir. Doğada sınırlı sayıda farklı bağlanma bölgesi motifi vardır ve yapısal olarak farklı protein çiftleri aynı bağlanma yapılarını kullanabilir. Bu esasa dayanarak, ikinci olarak protein komplekslerinin proteom seviyesinde modellenmesi için geliştirilmiş çoklu ölçekte tümleşik bir strateji gösterilmiştir. Bu çalışma, varolan yapısal bilginin bir biyolojik yolun yapısal benzerlik kullanılarak modellenmesine nasıl yardım edebileceğini göstermektedir. Buradaki örnek biyolojik yol, tümör baskılayıcı protein p53 biyolojik yoludur. En son olarak, Protein Veri Bankasındaki çoklu-eşe sahip olan proteinlerin veri kümesi çıkarılmıştır ve zaman kavramının protein etkileşim ağlarıyla birleştirilmesi iki merkezli protein, p53 ve Mdm2, üzerinde tahmin edilen ve bilinen etkileşimlerle gösterilmiştir.

ACKNOWLEDGEMENT

First and foremost, I offer my sincerest gratitude to my supervisors, **Assoc. Prof. Özlem Keskin** and **Prof. Attila Gürsoy** whose encouragement, friendship and continuous support from the preliminary to the concluding levels enabled me to complete this dissertation. I am heartily thankful them for giving me the chance of being a member of their research group. Their intuition and experiences inspire my growth from being a student to being a researcher. I could not wish for better supervisors and I am truly indebted to them more than they know.

I offer my deep appreciation to **Prof. Ruth Nussinov** for her guidance and sharing her endless experience with us. I would like to thank my thesis committee members **Asst. Prof. Halil Kavaklı**, **Asst. Prof. Elif Özkırmı** and **Asst. Prof. Mehmet Sayar** for their critical reading and useful comments. I am thankful to **Asst. Prof. Sibel Salman** for laying the basis of a chapter of this dissertation. I would like to thank also the Scientific and Technological Research Council of Turkey (TUBITAK) for their financial support during my PhD study.

I thank everyone for the good times throughout the 5 years at Koc University; every single moment was wonderful. My special thanks go to **Özge Engin** and **Besray Ünal** who are my colleques, my homemates and my friends. I would like to thank specially my past officemates **Cengiz Ulubas** and **Emre Güney** for our fruitful scientific collaborations and enjoyable times. I thank my current officemates (Gozde Kar, Billur Engin, Ece Ozbabacan, Engin Cukuroglu) and all people at our partner office 110.

Finally, I thank my family for their patience and continuous support throughout the PhD years.

TABLE OF CONTENTS

List of Tables	x
List of Figures	xii
Nomenclature	xiii
Chapter 1: Introduction	1
Chapter 2: Literature Review	4
2.1 Proteins Interact Through Their Interfaces.....	4
2.2 From Pairwise Protein Interactions to Structural Networks.....	5
2.3 Structural Aspects of Protein – Protein Interactions	8
2.3.1 Characteristics of Protein Interfaces.....	8
2.3.2 Architectures of Protein Interfaces.....	9
2.3.3 Structural Characteristics of Binding Regions on Hub Proteins	10
2.3.4 Artificial interfaces from crystal complexes	11
2.3.5 Cooperativity and Allostery in Protein Interactions.....	12
2.4 Critical Residues in Protein Binding: “Hot Spots”	13
2.4.1 Characteristics of Hot Spots	14
2.4.2 Prediction of Hot Spots in Protein Interfaces.....	14
2.4.3 Hot Regions: Modular Nature of Protein – Protein Interfaces.....	16
2.5 Structure-Based Modeling of Protein-Protein Interaction.....	17
2.5.1 Docking Strategies	17
2.5.2 Template-Based Modeling of Proteins Interactions.....	18
2.5.3 Construction of a Non-redundant Template Set for High-Quality Predictions...	22

Chapter 3: Identification of Computational Hot Spots in Protein Interfaces 25

3.1 Methodology for Identification of Computational Hot Spots	25
3.1.1 Training Set	25
3.1.2 Test Set	26
3.1.3 Features	26
3.1.4 Assessment of the Prediction Performance	28
3.1.5 Determination of Computational Hotspots.....	29
3.2 Results	30
3.2.1 Distribution of features of hot spots and non-hot spots.....	30
3.2.2 Comparison of Emprical Hot Spot Detection Formulations	33
3.2.3 Machine Learning Based Approaches.....	36
3.2.4 Comparison with Other Hot Spot Prediction Methods	37
3.2.5 Case studies	39
3.3 HotPoint: Hot Spot Prediction Server for Protein Interfaces	41
3.3.1 Input for the Server.....	42
3.3.2 Output of the Server	42
3.3.3 An independent case study: Interleukin-2 and its receptor complex.....	43
3.4 Concluding Remarks	45

Chapter 4: Analysis and Network Representation of Hot Spots in Protein Interfaces Using Minimum Cut Trees 46

4.1 Methodology.....	46
4.1.1 Dataset of Experimental Hotspots.....	46
4.1.2 Construction of Weighted Residue Contact Graph and Minimum Cut Tree of a Protein Complex.....	47
4.1.3 Algorithm I: Determining the Critical Residue Subtree.....	49
4.1.4 Algorithm II: Iterative Clustering of the Interface Residues.....	50
4.2 Results	51

4.2.1 Analyzing Mincut Trees for Other Protein Complexes.....	54
4.2.2 Organization of Hot Regions.....	58
4.3 Concluding Remarks	61
Chapter 5: Multi-Scale Combinatorial Docking of the Proteome for Functional Predictions	63
5.1 Methods	64
5.1.1 Datasets	64
5.1.2 Prediction Phases.....	65
5.1.3 Validation Procedure	67
5.2 Results	67
5.2.1 Validation of the Method	67
5.2.2 Comparison of Running Times with Docking	70
5.2.3 Structural Interaction Network of p53 Pathway.....	72
5.2.4 The Nucleotide Excision Repair (NER) Subsystem in the p53 Pathway.....	74
5.2.5 Cyclin/CDK Subsystem	78
5.2.6 Some Promising Interactions Predicted by Our Approach	80
5.3 Conclusions	83
Chapter 6: Multi-Partner Proteins	85
6.1 Structural Dataset of Multi-Partner Proteins	85
6.2 Towards Inferring Time Dimensionality in Protein – Protein Interaction Networks by Integrating Structures	91
6.2.1 p53 and its Binding Partners	91
6.2.2 Mdm2 and its Binding Partners.....	96
6.3 Concluding Remarks	100
Chapter 7: Conclusion	101

Appendix A: Appendix	104
A.1. Webservers, Softwares, Tools, Databases	104
A.1.1. NACCESS.....	104
A.1.2. MULTIPROT.....	104
A.1.3. ClustalW	105
A.1.4. CytoScape (network visualization and analysis)	105
A.1.5. VMD (molecule visualization)	105
A.1.6. FiberDock	106
Bibliography	107
Vita	129

LIST OF FIGURES

2.1	Illustration of the protein – protein interfaces.....	5
2.2	The protein interaction network of human taken from DIP.....	6
2.3	Classification of template-based approaches.....	19
3.1	Distribution of hot spot and non hot spot residues in the available data set with respect to their (A) relCompASA, (B) relDiffASA, (C) conservation, (D) pair potential.....	31
3.2	The comparison of hot spots and non-hot spots for each feature (relCompASA, conservation, pair potential, relDiffASA) plotted by ANOVA.	32
3.3	The distribution of hot spot and non-hot spot residues with the empirical formula and discriminant functions.	38
3.4	A case study of computational hot spot prediction using our empirical model.....	40
3.5	Streptococcal Protein (pdbID: 1fcc, chain C).....	41
3.6	The output page of HotPoint for the p53 DNA binding domain/53BP2 protein complex (pdb:1yca, chain A and B).....	43
3.7	IL-2 receptor complex.....	44
4.1	Illustration of construction steps of mincut tree with an example.....	49
4.2	Analysis of the Erythropoietin (EPO) receptor and EPO mimetic peptide complex with mincut tree.....	53
4.3	Analysis of the barnase/barstar complex with mincut tree.	56
4.4	Mincut tree for the 1ahwBC interface and illustration of the extracted subtree on the 3D structure.....	57
4.5	Mincut tree of the bipartite graph of the TEM1 – BLIP complex at first, 15 th and 16 th steps of the iteration.	60
4.6	The hvb2.1-TSST1 complex.....	61
5.1	The concept figure of the prediction algorithm..	65

5.2	Illustration of the extra interactions predicted by our method.	68
5.3	Some examples found in docking benchmark..	70
5.4	Comparison of running times of our template-based method with docking on large scale.....	72
5.5	The nucleotide excision repair subsystem as a sequence of reactions.....	76
5.6	Structural representation of the G2/M phase checkpoint.....	79
5.7	The modeled interaction between NF κ B and ASPP2 proteins.	81
5.8	Structural representation of predicted and known partners of the p27 protein.....	82
5.9	Some possible partners of Skp2 are predicted from the interface between Skp2 and Skp1..	83
6.1	Two binding sites of actin along its partners available in PDB are shown.....	87
6.2	Subtilisin and its inhibitor molecules.....	89
6.3	The cysteine protease Falcipain-2 and its inhibitors.....	90
6.4	Protein interaction network derived from PDB.	86
6.5	The fragments of the p53 protein and the available crystal structures	92
6.6	Predicted partners of the p53 DNA-binding domain (left panel), with representation of some in the complexed state (right panel).	93
6.7	Conserved contacts of Cdk2, Chk1 and Crk with p53 DBD predicted with MAPPIS.	94
6.8	The Mdm2–pRb complex predicted by PRISM and the E2F1–pRb complex available in the PDB.....	97
6.9	Predicted partners interacting at the pocket region of Mdm2.....	99
6.10	Predicted partners of the Swib domain of Mdm2 (left panel) and representation of some of them in the complex state (right panel).	100

LIST OF TABLES

3.1	The results of each features computed by ANOVA.	32
3.2	Performance values of various empirical prediction methods used to identify hot spots in the protein interfaces.....	33
3.3	Prediction results for the structures in test set (BID).....	36
3.4	Machine learning based models Results. These are the corresponding implementations from Weka	37
3.5	Hot spot prediction performances on test set (BID).	39
3.6	Performance values of linear and quadratic discriminant functions.....	39
4.1	The most connected node in the mincut tree for several complexes.	55
5.1	Number of predicted interactions and verification on the experimental data.....	73
5.2	The highest ranking solutions in the case studies.	77

NOMENCLATURE

<i>PDB</i>	Protein Data Bank
<i>ASA</i>	Accessible Surface Area
<i>PPI</i>	Protein – Protein Interaction
<i>RMSD</i>	Root Mean Square Deviation
<i>SCOP</i>	Structural Classification of Proteins
<i>PRISM</i>	Protein Interactions by Structural Matching
<i>MIM</i>	Molecular Interaction Map
<i>ASEdb</i>	Alanine Scanning Energetics Database
<i>BID</i>	Binding Interface Database
<i>ANOVA</i>	Analysis of Value

Chapter 1

INTRODUCTION

Proteins function in several processes varying from transcription regulation to signaling or from being molecular machines like ribosome to enzymatic reactions. Considering their unquestionable role in the cell make the researchers focus on the prediction of the protein function. The easiest but the most complicated way to do this is identifying interactions between proteins which is at the heart of functional genomics; their prediction is crucial for drug discovery. Through the network of these interactions we can map cellular pathways, their interconnectivities and their dynamic regulation. Besides the question “*which proteins are interacting with which others?*”, “*how does the interaction take place?*” is more challenging towards the elucidation of the complete mechanisms in the cell. The hints for the second question is hidden in structural biology.

Proteins interact through interfaces. Protein interfaces constitute patches of structural and energetic components. From an energetic perspective, the residues in protein interfaces do not have equal contribution in binding, rather a subset of these residues, called “hot spots”, play an exceptional role [1]. Hot spots may also be considered as drug targets and the existence of hot spots addresses the question how just a small molecule can disrupt a whole interaction having a large binding area. The attacks of the drug molecules can damage the undesired interactions by binding and blocking the hot spots. Hence, identifying hot spots in protein interfaces and analyzing their organization have a crucial importance.

On the other hand, besides the studies at molecular level, providing the detailed three-dimensional (3D) global structural network of the organism functional proteome is a challenging goal in structural biology [2]. In the passing decades, large-scale experimental methods were employed to determine protein interactions for several organisms and thousands of interacting proteins were identified [3-7] (reviewed in [8]) presenting new challenges, potential interpretation pitfalls, and rewards in translating the data into high resolution physical interactions [9]. There has further been an exponential increase in

structural information [10] obtained by NMR and crystallography. Nonetheless, the gap between the number of known interactions obtained from high-throughput experiments and structurally known protein complexes is large. Given experimental limitations, it is crucial to develop computational methods to predict the 3D structures of protein complexes. Prediction algorithms [11, 12] would provide atomistic details of interactions which are important for designing therapeutics and for mapping the cellular network [13].

This dissertation, primarily focuses on structural analysis of protein interactions both at the molecular and the proteome levels where prediction and organization of hot spots in protein interfaces and modeling of protein complexes towards construction of structural protein interactions network at large scale are studied. Each chapter is readable on its own independent from the other chapters; however, all the chapters attempt to address the question “*how do the proteins interact?*” and serve to functional genomics, drug design and pathway analysis at the top level.

The outline of this dissertation is as follows:

In Chapter 2, an extended and most recent review focusing on the structural aspects of protein interactions is presented. This chapter includes the corresponding works related to characteristics and architectures of protein interfaces, hot spots, modularity of protein interfaces and structural modeling of protein complexes.

Chapter 3 includes a new intuitive simple approach in hot spot prediction yet with its high accuracy and computational effectiveness. Here, first, the method is illustrated with the used dataset and features. Then, distributions of the features, the performance of the method and its comparison with other available hot spot prediction methods are presented. Also, the prediction power of the method is shown with some case studies. At the end of this chapter, the web server HotPoint is introduced which provides a user-friendly interface to run this method for online prediction of hot spots.

In Chapter 4, graph based algorithms are illustrated to analyze and visualize residue contact networks of protein interfaces employing minimum cut trees (mincut tree). These algorithms are demonstrated on protein complexes having experimental mutation data. This chapter shows that the proposed algorithms are useful at the molecular level for both identification of critical paths in the protein interfaces and extraction of hot regions by clustering of the interface residues.

Chapter 5 is designed to introduce multi-scale combinatorial docking of proteome for functional predictions. In this chapter, our knowledge-based method combined with flexibility and energy calculations to structurally model the protein complexes is explained. The rationale of the knowledge-based part of the method argues that if particular surface regions of any two proteins are spatially similar to the complementary partners of a known interface, in principle these two proteins can interact with each other via these regions. Validation of the method, computational time comparisons with docking strategies and application of the method on the tumor suppressor protein p53 interaction network are included in the results section of this chapter.

In Chapter 6, the multi-face nature of proteins is illustrated on a structural perspective. In the first part, a multi-partner dataset of proteins constructed from PDB is presented with some case studies. In the second part, the multi-face structure of two hub proteins – p53 and Mdm2 – is demonstrated both using experimental and predicted structural data.

This dissertation ends with a chapter discussing the results, explaining future directions, and finally with presenting major conclusions of the study.

Chapter 2

LITERATURE REVIEW

In this chapter, a comprehensive review of the studies related to protein interfaces, their characteristics, structural modeling of the protein interactions is presented. In the first section, general properties and architectures of protein – protein interfaces are explained. Then, critical residues in protein interfaces and the modular organization of the hot regions are reviewed. Finally, structure based protein interaction prediction algorithms are reviewed.

2.1 Proteins Interact Through Their Interfaces

Protein–protein interactions occur at the surface of a protein and are biophysical phenomena, governed by the shape, chemical complementarity, and flexibility of the molecules involved as well as the environmental conditions. The physical binding of protein structures occur through weak, non-covalent interactions. The particular region where two protein chains come into contact is termed a binding site or an interface. As an example, **Figure 2.1** shows the interface present between α - and β -globulin of the human hemoglobin (PDB ID: 1yhe). Protein interfaces have long been studied at both the protein level and the domain level. They have been represented as interface data sets and deposited into databases to be used in identification of general properties of them. Some of these databases are SCOPPI [14], InterPare [15], 3DID [16, 17], PIBASE [18], ProtCom [19] and PRINT [10, 20]. Towards the common goal of understanding how proteins interact, a number of studies have characterized the properties of interfaces between proteins [20-26] which is crucial to help binding site prediction algorithms. In the continuing parts of this chapter, the detailed analysis of interface characteristics is illustrated.

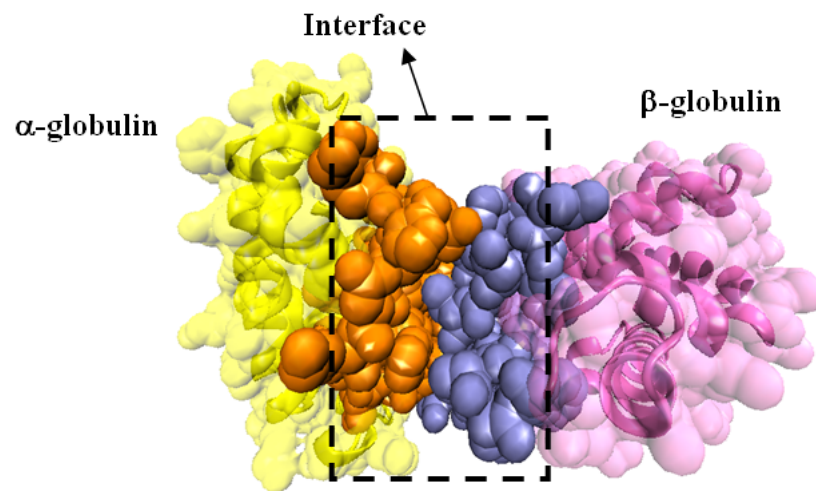


Figure 2.1 Illustration of the protein – protein interfaces. The interface region of chain α -globulin is shown in orange; the rest of the chain is colored yellow. Similarly, the interface region of β -globulin is shown in blue as the rest is colored pink. The interface region is enclosed with the dashed box.

2.2 From Pairwise Protein Interactions to Structural Networks

One of the primary objectives of the post-genomic era is the elucidation of the interactome in model cellular systems. The detailed knowledge of the full network of protein-protein interactions, *i.e.*, the distribution and the number of interactions as well as the presence of key nodes in these networks, is expected to provide new insights into the structures and properties of biological systems. The availability of interactomes for many organisms will continue to provide understanding of the global organization of cellular processes. Still, these interactomes will lack structural and chemical characteristics of each interaction. With the help of these characteristics, a deeper understanding of the physical phenomena taking place in the cells can be appreciated.

Experimentally, the pairwise interaction data of proteins on large scale can be obtained by several methods, such as yeast two-hybrid [7], phage display [27], protein arrays [28], affinity purification [29] techniques. The experimental databases – such as DIP [30], MINT [31], BIND [32] and IntAct [33]– catalog the data gained by these techniques to serve researchers dealing with protein interactions. Several external factors such as post-translational modifications and disorder generally lead to false positive and false negatives in the experimental data, mainly in yeast two-hybrid [34]. In addition, distinct experimental information from different resources may conflict with each other resulting in high false

positive rates [35]. However, there is a decreasing trend in the false-positive rate of the high-throughput experiments [9]. Other methods, such as, mass spectrometry is used to identify the components of the protein complexes and site-directed mutagenesis is used to identify which residues have critical role in binding. In line, chemical foot-printing is utilized to find the buried surface upon complexation.

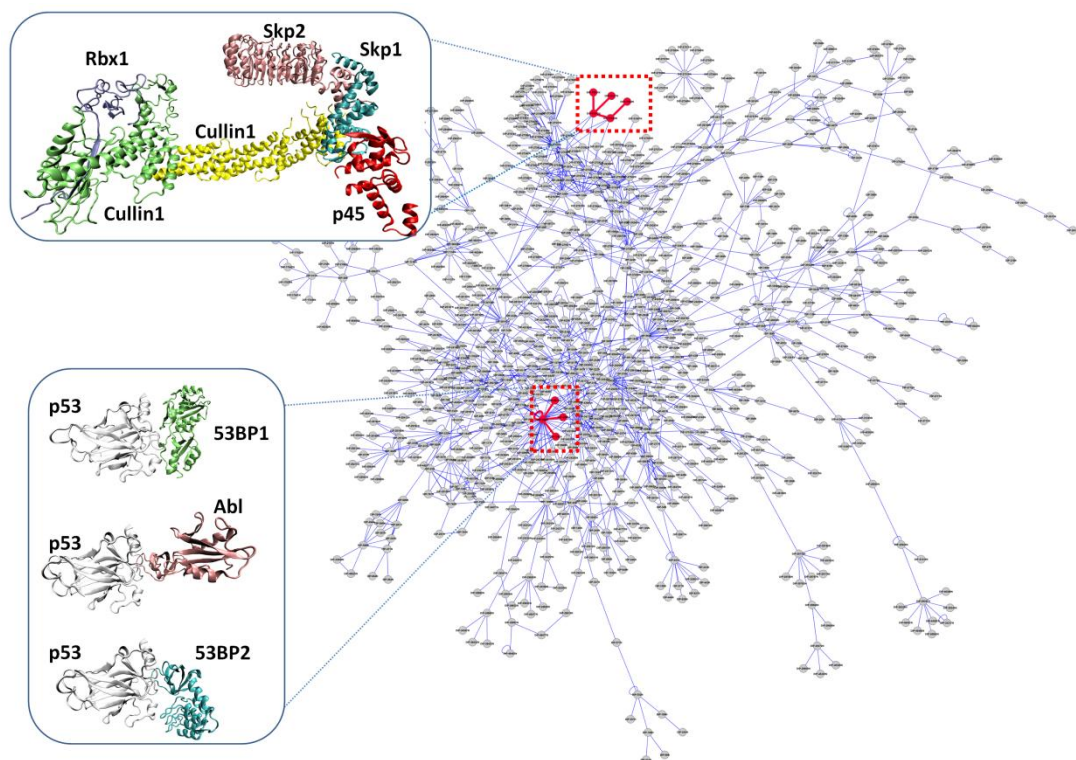


Figure 2.2 The protein interaction network of human taken from DIP. The network picture just informs about the pairwise interaction of the proteins; however, when the details are examined, structural data incorporates one more dimension into this static network. Here, the Rbx1/Cullin1/Skp1/Skp2/p45 complex and some partners of p53 is illustrated.

These techniques just provide which proteins interact with which others and gives a static picture excluding relative orientations of the interacting proteins and the residue level details. In **Figure 2.2**, a sample protein interaction map for human taken from DIP is illustrated. Incorporation of the residue level details of protein interactions – *how they take place* – can be achieved just by obtaining structural data about them. In **Figure 2.2**, two modules are highlighted and the details are shown with structural data. While the classical interaction map says that the 5 proteins in the first module are interacting with each other,

the structural information shows that these 5 proteins are actually interacting simultaneously and form a large protein complex (Rbx1/Cullin1/Skp1/Skp2/p45 complex) which is functioning in ubiquitination pathway. On the other hand, the structural details of the second module shows that the DNA binding domain of p53 uses same region to interact with Abl, 53BP1 and 53BP2 and these interactions are not simultaneously possible. In this way, the classical interaction map gains another dimension with the structural data.

The structure determination of the protein complexes is achieved by X-ray crystallography, NMR spectroscopy [36] and cryo-electron microscopy (EM) [37] at several resolutions. X-ray crystallography and NMR spectroscopy provide information at atomistic level. X-ray crystallography is the most widely used technique and gives a static information about protein structure. NMR data is obtained in solution where several structural, thermodynamic and kinetic properties can be analyzed and limited by the size of the protein complex. Transient complexes are underrepresented in structural data because of their crystallization issues. Therefore, the methods like cryo-EM is useful for visualization of transient complexes. The relative positions of the subunits in a protein complex and the interacting residues can be obtained at low resolution, but the full details are not distinguishable. The limitations of the experimental technique can be complemented by each other or by computational techniques. For example, combining the electron microscopy data for a large complex with the crystal data of the subunits may provide a complete picture for this large complex. The structural and mechanistic models of clathrin lattice is obtained by hybrid experimental methods [38]. In another example, Sali and his co-workers modeled the structure of the large molecule human-RNAPII by combining several experimental resources and computational techniques where 3D atomic details of the subunits are from ModBase, the pairwise interaction data of the subunits are from BioGrid and the generated model is put into the EM density map with an optimization stage [39].

As the complete list of pairwise protein-protein interactions and the structures of the constituting proteins become available, it will be possible to construct reliable protein-protein interaction networks with molecular details, such as the structural knowledge of the binding sites of the constituting proteins (as illustrated in **Figure 2.2**). Such approaches and novel representations integrating different levels of details (**Figure 2.2**) will enable inferring new knowledge which would be difficult if not possible otherwise. Further, since

traditional representation of protein interaction networks signifies each interaction equally, combining structure with networks will provide better targets for drug design. These models with time series data will be able to explain the dynamic behavior of living systems.

2.3 Structural Aspects of Protein – Protein Interactions

2.3.1 Characteristics of Protein Interfaces

Several studies incorporating thermodynamic, kinetic and structural information imply that proteins interact rapidly and strongly in a specific manner to find their partners in the crowd of macromolecules in the cell. Interactions have been studied on a large number of complexes by numerous groups [1, 20-23, 25, 40-53]. Some protein interactions are tight and long-lived, called obligate interactions while some proteins continuously associate and dissociate, called transient interactions [54]. This is probably due to different strength of interactions between the proteins, the former relying more on salt bridges and hydrogen bonds, whereas the latter rely more on hydrophobic attractions [23, 50]. Also, despite their rare occurrence, disulphide bonds have large contribution on stabilization of the binding [55]. Transient complexes are underrepresented in structural data because of their crystallization issues [56]. So, the general patterns learned from structural information generally represent the obligate interactions. In general, interfaces tend to be planar or well packed depending on the type of interaction [22, 47]. The residue composition usually differs for those complexes that are transient versus those that are obligate. Weak transient interfaces are characterized as flat, small and polar contact regions [57]. Interface residues of the obligate complexes have tendency to evolve slower when compared to transient complexes [58].

Binding regions of the proteins evolve to optimize the binding affinity for a particular function and specificity to its partner. Long-range interactions are predominant in the stability and specificity of interfaces. These are electrostatic interactions, hydrogen bond interactions, van der Waals attractions and repulsions and hydrophobic forces. In order to maximize the effect of these forces, a good shape complementarity is necessary. It is known that in order to have a stable complex, both geometric and electrostatic complementarity between the two sides of interfaces is necessary [50, 51, 59, 60].

Homo-dimers are more hydrophobic and larger in size, when compared to hetero-complexes. Small interfaces prefer to interact through small pockets to exclude solvent. In spite of detecting significant differences, there is not a strict pattern to identify different types of protein-protein interactions [61]. Protein interfaces may change in size. Their sizes are based on the change in their solvent accessible surface area (Δ ASA) when going from a monomeric to a dimeric state. On average, interfaces bury 1600 \AA^2 with the same packing density as the protein interior [50]. One-third of the interfaces have a distinguishable large hydrophobic core, whereas the remaining interfaces have smaller hydrophobic patches with polar contacts and water mediated interactions. All interfaces have buried residues containing core region which is surrounded by a rim region. Core region of the interface is similar to interior of the protein in residue frequency. However, rim region is similar to protein surface [62]. Amino acid frequencies in the interface region are also an important parameter in characterization of these regions [63]. Using only amino acid composition and residue-contact preferences, an accuracy of 63-100 % is reached to predict interaction types [64]. Additionally, interface regions are rich in aromatic and hydrophobic clusters. The residue propensity of the interface regions is found to be similar to the interior of the proteins [21].

Interface residues are more conserved than the rest of the surface by a chance higher than random [65]. However, conservation is not enough alone to completely distinguish the protein binding sites, but it can be combined with other interface properties [66]. The binding site of an unbound monomer (prior to binding) is enclosed by more water molecules and has less flexibility as depicted by low temperature factors compared to the rest of the surface [55]. Similarly, the interfacial residues are less flexible than the rest of the protein surface [67].

2.3.2 Architectures of Protein Interfaces

Besides these general properties of protein interfaces and interaction types, the structural characterization of protein interfaces and analysis of interface architectures are also crucial. In a remarkable work of Keskin et al, a structurally non-redundant set of protein – protein interfaces is divided into three classes where Type I clusters contain similar interface architectures coming from similar global folds, Type II clusters include similar interface architectures coming from dissimilar global folds and Type III clusters contain one side of

the interface conserved [20]. Type I clusters are expected cases, because the interaction of similar pairs with similar architecture is usual; however, Type II clusters are interesting cases which implies that nature reuses these favorable architectures for the interaction of different proteins. One step forward is that one side of the interface architecture is structurally conserved and the partner chain is changing (Type III). This type of interfaces is important for the analysis of promiscuous interactions and for addressing the question how a given site bind to different binding sites nonspecifically. These types of clusters are important because they show the architectural conservation of protein interfaces and the specificity of the proteins in their interactions.

In the extension of this work, the growth trend of the interface dataset and structural data in PDB is analyzed [10] and we demonstrate how far we are from the complete structural information and where we are currently in structural biology. As a result, we found that the number of unique interface architectures continues to grow up, and also the population of clusters enlarges. Also, some interface architectures are more favorable and frequently used in protein–protein associations. In agreement with the broadly-accepted notion that binding and folding are similar processes, we observe that most populated folds are structurally similar to the most populated interface architectures. Nature appears to use similar preferred fold templates for single chains and for interfaces.

2.3.3 Structural Characteristics of Binding Regions on Hub Proteins

In protein-protein interaction networks, most proteins have only a few interactions, whereas, a small number of ‘hubs’ are highly connected [68]. Further they connect many crucial cellular processes [69]. Scale-free characteristics of networks make them very resilient against accidental failures: even if 80% of randomly selected proteins fail, the remaining 20% still continue to carry out the cellular functions [70]. However, protein networks with hubs are shown to be vulnerable to systematic attacks at highly connected proteins. Removal of such hub proteins is often lethal [71] , which makes these proteins essential in the cell [72]. From a structural point of view, structure can help in understanding the roles of these hubs in the affinity of the interactions. Since a single protein cannot interact with a large number of partners at the same time, they should adapt either multiple binding sites on their surfaces where multiple partners may interact through these sites simultaneously, or they reuse the same binding site for different partners. The

binding region of multi-partner proteins mainly composed of alpha helices and these interfaces are not well organized and packed when compared to other proteins [2]. Schroeder and his co-workers also similarly stated that hub proteins use different surface regions to interact with different partners. Further, they speculated that the ancient interfaces are coming from symmetric homodimers, and heterodimers are evolved from these complexes [73]. In another work, Aloy et al stated that close homologues interact in the same way; however similar folds can interact differently [74]. In an elegant study, Kim et al [75] combined structural modeling with network analysis. They used the interfaces in protein surfaces and found that for two or more proteins interacting with a common partner protein, there are two possibilities: i) they can use the same interface on the partner (single interface), then they classified the interactions as mutually exclusive. ii) or the two proteins can use different interfaces (multi-interface), then they called the interactions as simultaneously possible. They found that, most of the mutually exclusive interactions were transient, because they cannot occur at the same time. On the other hand, simultaneously possible interactions are enriched in permanent associations, connecting members of the same complex.

Our previous analysis highlights that there exist conserved interactions of a given site when interacting with multiple partners. This means that even though the partner proteins are different there are some critical residues in the binding site that make conserved interactions with multiple proteins [47]. Thus, while the patterns of the local interactions are similar in multi-partners and in single-partners, the multi-partners have been optimized by evolution to accommodate different ligand shapes, sizes and composition. For example, crystal structures of the Elongin B/Elongin C/VHL and Elongin B/Elongin C/SOCS2 [76, 77] complexes provide such clues. The concept of functional switches in transcriptional regulation [78] was emphasized by Beckett, focusing on the ability of proteins to bind alternative proteins at the same binding site.

2.3.4 Artificial interfaces from crystal complexes

If two proteins interact with each other *in vivo*, then they form a biological complex, and their interaction is formed through a biological contact. However, sometimes, the entries in the Protein Data Bank (PDB) have artifacts of crystallization meaning that some of the

complexes there would not occur in solution or in the physiological state. Determining which contacts are biological and which are not is often difficult [79].

There are some ways to differentiate biological contacts from non-biological ones. Several groups have successfully used conservation scores to predict biological protein–protein binding sites [79, 80] which conclude that conservation in combination with other factors can accurately discriminate homodimers from crystal contacts. Further, interface size is an important characteristic in distinguishing crystal and biological complexes [21, 79, 81]. A threshold of 400 Å² was used in PQS for identifying biological complexes [82]. Similarly, when the number of interface residues was less than 10, the interface was considered to result from the crystal packing rather than biological inter-subunit interactions [20, 52]. Shoemaker et al. (2006) generated a set of conserved domain–domain interactions by structural alignment to distinguish the biologically relevant interactions. According to these conserved modes, they reached an accuracy of 90% on all globin interacting pairs without false positives [83]. Another significant factor to distinguish artificial interfaces is amino acid frequency on the interaction site. If the binding site amino acid composition is similar to the rest of the surface of that protein, then this interaction is possibly crystal packing [84]. In addition, multiple features (interface area, interface area ratio, amino acid composition, correlation between surface and interface regions, gap volume index, and conservation score) can be also combined for predicting biological complexes using machine learning approaches as in NOXclass [85]. The contact frequencies of buried residues in protein interfaces used as the feature set and it performed well to discriminate artificial interfaces [86].

2.3.5 Cooperativity and Allostery in Protein Interactions

Analysis of the interaction between two proteins supplies useful information; however, the effect of a third protein to the binding process, in words, “*cooperativity*”, is also important. Here, the association of two proteins is dependent on the third protein. Experimentally, there are some evidences that binding affinity of a specific protein increases to its partner during the occurrence of a third protein although they do not interact directly. For example, the interaction between the transcriptional co-activator CREB binding protein (CBP) and its transcriptional activator partners is cooperative. The binding affinity of pKID increases two fold when it binds to KIX-MLL complex; however, the interaction between pKID and

KIX monomer is weaker [87]. Besides the cooperativity between proteins, there is also cooperation between domains in protein interactions. As an example, Klemm et al found the cooperative binding of POU-specific domain and the POU homeo domain in protein Oct-1 [88]. Different patches in protein interfaces are not completely additive, rather they significantly cooperate [89].

Allostery is defined as regulation of a protein through a change in its quaternary structure induced by a small molecule [90]. Kuriyan and Eisenberg broaden this definition as the change induced by a small molecule or another protein [91]. Nussinov and her colleagues argued that all proteins are potentially allosteric [92]. Evolutionarily some regions of proteins are very sensitive to mutations whereas some regions are robust to mutations. Ranganathan and his co-workers present a statistical approach which considers evolutionary data of the targeted protein family and calculate the pairwise energetic coupling of the residues [93]. They found sparse but connected network of residues passing through the protein core and connecting the active site with distant sites. In contrast to this work, Chi et al found that spatially close residues are coupled energetically, instead of distant residues. [94]. Nussinov and her co-workers approach allostery from graph theoretic perspective and they represent proteins as residue contact networks. They identified the centrally conserved residues which are important for long range interactions and these residues are correlated with the experimentally suggested residues which are important for allostery [95].

2.4 Critical Residues in Protein Binding: “Hot Spots”

Studies on protein interfaces have revealed that energies are not uniformly distributed. Instead, there are certain critical residues called *hot spots* comprising only a small fraction of interfaces yet accounting for the majority of the binding energy [1, 42]. Experimentally, a hot spot can be found by evaluating free energy change upon mutating it to an alanine, playing key roles on the stability of the protein association. Thorn and Bogan [96] deposited hot spots from alanine scanning mutagenesis experiments, in the Alanine Scanning Energetics Database (ASEdb). Binding Interface Database (BID) [97] presents experimentally verified hot spots at interfaces collected from literature.

2.4.1 Characteristics of Hot Spots

Analysis of amino acid composition of hot spots shows that some residues are more favorable. The most frequent ones, Tyr, Arg, and Trp, are critical due to their size and conformation in hot spots [1]. In addition, Bogan and Thorn reported that hot spots are surrounded by energetically less important residues that most likely serve to occlude bulk solvent from the hot spots (O-Ring hypothesis) [1]. Occlusion of solvent is found to be a necessary condition for highly energetic interactions [1, 98, 99]. It is observed that hot spot residues prefer to sit in complemented pockets, and are disfavored in unfilled pockets [24]. Predicted clefts using physicochemical properties and conservation of protein surfaces may correspond to binding hot spot regions [100]. Another study has illustrated that there is a correlation between energy change and decrease in the accessible surface area of individual residues as a consequence of complexation [101]. Moreira et al have supported that hotspots are protected from solvent by a rim region; however, they concluded that more computational analysis should be applied to elucidate this theory [102]. Hot spots are usually found in strong obligate complexes. In a recent work, Volkov et al found that transient complexes also have binding energy hot spots [103]. There are several studies focusing on the detection of hot spots based on conservation: Correlation between hot spot residues and structurally conserved residues were found to be remarkable [104]. However, hot spots are marginally more conserved than the rest of the interface in sequence [98, 99]. Hot spots in interfaces are found to be less flexible compared to the rest of the protein surface and interface residues [105-107]. These residues are highly packed and form clusters among themselves [44]. They are observed to form pre-organized binding motifs at protein interfaces even in the unbound cases. This organization might be to minimize the entropic cost upon complex formation.

2.4.2 Prediction of Hot Spots in Protein Interfaces

Hot spot information from experimental studies are available only for a very limited number of complexes, therefore, there is a need for computational methods to identify hot spots of protein interaction sites [108]. Several research groups developed learning based, energy based models [109, 110] learning based models [98, 99, 111-116], molecular dynamics based models [106, 117, 118] and graph based models [119, 120] to predict hot spot residues computationally.

In a pioneering work, Kortemme and Baker [121] proposed a physical model (Robetta) to detect hot spots at protein-protein interfaces accounting for energies of packing interactions, hydrogen bonds and solvation. *Computational hot spots*, the residues they identified computationally based on their model, show accordance with experimental hot spots in ASEdb. Similarly, Gao et al. used non-covalent interactions to estimate energetic contribution of interfacial residues to binding. They reported an 88 % success rate for predicting hot spots obtained from alanine scanning mutagenesis experiments [114]. Another energy based model developed by Serrano and co-workers [109] was used to predict the energetic effect of mutations on protein complexes. The calculated energy change of mutations agreed well with the experimental results. Their method is applicable to hot spot predictions as well.

Molecular dynamics (MD) simulations can provide detailed analysis of protein interfaces at the atomic level for more accurate prediction of hot spots [117]. Rajamani et al. [106] studied 11 protein complexes and found that anchoring residues in protein interfaces show restricted mobility and may act as hot spots. Kollman and co-workers [118] applied MD to find computational alanine scanning of 1:1 human growth hormone-receptor complex and reported a good agreement with the experimental data. Although these energy and MD based methods are successful to identify hot spots of individual protein complexes, they are not applicable, in practice, for large scale hot spot predictions due to their computational cost.

A neural network based approach using various features of interfaces such as sequence profiles, solvent accessibility and evolutionary conservation is employed in computational hot spot prediction [98]. The method has advantage of using only sequence; thus, it is applicable when the structure is not available and also when the binding partner is unknown. A hybrid computational model combining decision tree (using atomic contacts, physicochemical properties and shape specificity contributions) with computational alanine scanning method is proposed to predict hot spots [111]. A neural network-based approach using interface features such as sequence profiles, solvent accessibility and evolutionary conservation has been employed in computational hot spot prediction (an adaptation of ISIS) [98]. In another work, Grosdidier et al [122] predict hot spots by using docking methods without protein complex knowledge. Their performance on a subset of Kortemme's dataset reached a precision value of 0.78 and sensitivity 0.46.

Graph/network-based algorithms are also frequently utilized to study the identification, organization, and packing of hot spots. Brinda *et al.* [119] used graph representation of homodimeric protein complexes and applied spectral analysis to the residue networks to predict hot residues. del Sol and O'Meara [120] used the small-world network approach to predict hot residues in protein-protein interfaces. In their work, highly central residues are considered and they stated that 77% of the predicted residues, conserved and buried ones, are either experimental hotspot or in direct contact with an experimental hotspot. Haliloglu *et al.*[123] applied Gaussian Network Model (GNM) on several antigen-antibody and enzyme-inhibitor complexes to predict anchoring residues.

Although these methods are successful to identify hot spots of individual protein complexes, molecular dynamics based and energy based methods are not applicable, in practice, for large scale hot spot predictions due to their computational cost. Also, machine learning based methods perform well to predict binding hot spots and these methods are computationally very effective [98, 111]. Most machine learning based hot spot prediction methods learn complex relations between training data and hot spots; however, it is very difficult to translate these relations into simple, intuitive rules [98].

2.4.3 Hot Regions: Modular Nature of Protein – Protein Interfaces

As stated in the previous section, energy distribution is not uniform along the interfaces. Further, there are pockets and cavities in the interface. Chakrabarti and Janin stated that small binding sites are composed of single continuous patch; however, larger interfaces may be composed of several patches which are distant to each other [21]. Hot spots come together and form tightly packed regions, called *hot regions* [44] resulting in densely packed clusters of networked hot spots. Contribution of distinct hot regions to stability is additive whereas contribution is cooperative within clusters [44]. Schreiber and his co-workers also analyzed the modular architecture of protein interfaces. They stated that protein interfaces are composed of individual residue clusters where residues within clusters are strongly connected whereas individual clusters are weakly connected to each other. They showed the modular architecture of TEM1- β -lactamase and its inhibitor protein (BLIP) experimentally using multiple mutations and checked whether the intracluster mutations are additive or cooperative [124]. On the other hand, Moza et al (2006) declare the necessity of long distance interactions for protein – protein interactions

and showed the cooperativity between distinct hot regions which are 20Å far-away from each other. Using combinatorial mutagenesis, they showed the cooperative energetics between two hot regions in the interface of T cell receptor and a bacterial superantigen. As a result, in contrast to Schreiber's work, they stated that mutations both within and between hot regions are energetically cooperative [89]. On one hand, some works state that protein interfaces are composed of energetically independent subregions; on the other hand, some others propose that there should be cooperativity between distinct regions for long distance communication. At a large scale, structure based computational study of Carbonell et al, the binding specificity and affinity of the protein interactions are examined using modular distribution of hot spots in the binding site. As expected, they found that interaction strength of specific interactions is stronger than the promiscuous ones and hot spots in different modules interact with different partners [125]. In a recent work, the organization of hot regions is analyzed for hub proteins and hot spots are found to be more organized in date hubs when compared to party hubs [126].

2.5 Structure-Based Modeling of Protein-Protein Interaction

Combination of all knowledge about the characteristic of protein interactions helps to design high performance prediction algorithms to model the protein interactions. In this section, structure-based prediction algorithms are reviewed in two classes; docking strategies and template-based approaches, respectively.

2.5.1 Docking Strategies

From the structural point of view, modeling of the protein complexes is frequently achieved *via* 'docking'. The aim of docking is to find the best match for a given 3D coordinates of receptor and ligand proteins. Several docking algorithms are developed throughout the years [127-132]. Docking strategies are composed of a fast search algorithm to obtain the candidate conformations and a high quality scoring function for the ranking of these conformation towards finding the near native model. Geometric and chemical complementarity, electrostatic, van der Waals forces or knowledge-based potentials [133] are mainly considered in scoring functions. For the performance assessment of docking strategies, Critical Assessment of Predicted Interactions (CAPRI) is usually used where at every round several docking methods are used to predict the complex states of the given

unbound protein structures [134, 135]. CAPRI is useful to see the situation of current docking strategies and improvements throughout the years [136]. Also, the bound and unbound forms of protein structures deposited in Docking Benchmarks are another source for testing [137, 138]. From past to now, the performances of docking strategies are improved [135, 139]. However, scoring functions are still not fully optimized; hence, the correct assessment of the modeled protein complexes with these scoring functions is still an important issue [140]. While docking is very useful to find the 3D model of the protein complexes it is very challenging, especially at large scale. First of all, in the lack of the additional information about the interaction of two proteins, they give several false positive binding orientations. In addition, docking is computationally very expensive at large-scale. Recently, the first large-scale docking effort is performed and 3,000 putative protein complexes are modeled for yeast protein network [141]. Another issue in docking is flexibility. While two proteins are interacting, they are exposed to conformational change; both side-chain and backbone. To accurately find the native state of the protein complexes, flexibility should also be incorporated to the rigid body docking algorithms. Towards this aim, refinement algorithms are developed to re-assess the rigid-body docking solutions and to re-rank the modeled interactions [142-144].

In recent years, docking is directed from blind testing to combination of binding site prediction approaches to restrict the wide range of solution space. As an example, all binding modes of a targeted protein (PSD-95) and its homologs are extracted and this information is used to limit the solution space for docking purposes [145]. In another work, binding site prediction algorithms are first applied to filter the candidate conformations and then the docking procedure is applied which improves the performance of the method on docking benchmark proteins [146].

2.5.2 Template-Based Modeling of Proteins Interactions

Nature presents a limited number of protein folds [147]; hence, the number of distinct binding site motifs are also limited [13]. With the growth of the structural data for protein complexes, template-based methods take more attention where a protein complex is modeled using sequence or structural similarity to a known protein complex. We can analyze template-based methods in two classes: homology-based (sequence-homology and structure homology) and interface-based (see **Figure 2.3**).

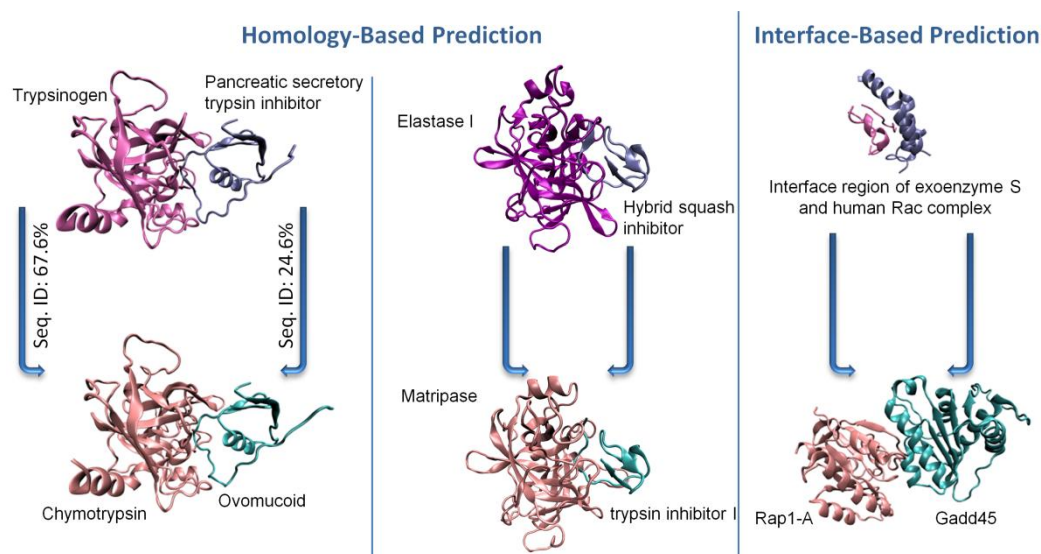


Figure 2.3 Classification of template-based approaches. In sequence homology-based prediction, protein complexes having sequence similarity with target proteins are the templates. Here, chymotrypsin/ovomuroid complex is predicted from the trypsinogen/pancreatic trypsin inhibitor. In structural-homology, independent from sequence similarity, overall structure of the template and target proteins are similar as in the interaction between matripase and trypsin inhibitor I predicted from Elastase-1/hybrid squash inhibitor complex. In interface-based prediction, just the interface region of the protein complexes are used to search similarity with target proteins. Rap1-A/Gadd45 complex is modeled using the interface between exoenzyme S and human Rac.

Homology-based methods are first appeared in the content of template-based prediction approaches. The basis depends on that proteins associate in a similar way if their sequence similarity is as high as 30% [74], while exceptional cases are available [148]. Aloy and Russell [149] searched the sequence homologues of the known protein complexes and scored the predicted interaction between homolog proteins using empirical potentials derived from known protein interactions. These potentials are usually used to assess the predicted complex, because implementation and computation of knowledge-based potentials is simple and their success is proven previously in fold recognition [150, 151]. If the score of a predicted complex is high enough, the homologous protein pair associates in a similar way with the template complex. This method is utilized to model protein complexes in yeast interaction network and fibroblast growth factor/receptor system. Later on, InterPreTS is designed as a web server to predict protein interactions using this method for a given set of protein sequences which uses Blast2 search tool to find the homologues [152]. Skolnick and his co-workers spend pioneering efforts on homology-based prediction of protein interactions. Their method, Multiprospector is based on multimeric-threading

[153]. It uses a template library composed of protein complexes besides monomers. The algorithm is composed of two phases; in the first phase, each target sequence is assigned to a protein structure in the template library. In the second phase, single chain threading is extended to multi-chain threading and each target protein pair is assigned to a group of quaternary structures. The quality of the predictions is assessed by the interfacial potentials and Z-scores. The knowledge-based statistical potentials are shown to discriminate native interactions from artificial ones with an accuracy of 90% [154]. Multiprospector exceeds the previous methods by its functionality in the lack of sequence similarity between target proteins and template complex. Multiprospector is applied on a large scale to yeast interaction network and 7321 interactions are modeled. The quality of the predictions are assessed with co-localization and molecular function [155]. The drawbacks of this algorithm are presented as that it cannot consider the conformational changes upon binding and also it cannot correctly balance the relative position of the proteins. M-Tasser is developed to solve these problems and it explicitly combines backbone flexibility with the threading to find predictions [156]. In another template-based approach, structural homology is considered to form the template set with complexes between single domain containing proteins. Also, weakly binding complexes are eliminated from this dataset. By superimposition of the target domains onto template complexes, models for protein complexes are generated and the models are ranked according to an energy function [157]. Alexov and his co-workers also used sequence homology to predict 3D structures of protein complexes [158]. Recently, Kundrotas et al constructed the GWIDD database composed of the experimental and homology-based models for several species towards the structural representation of all genome [159]. Homology-based protein modeling can just cover at most 20% of the overall protein networks. The advantage of homology-based prediction is that many proteins are unstructured in their unbound state and this method can produce prediction for these proteins.

Besides the homology, the structural similarity of overall protein structures, mainly the domain information, is also integrated to search putative protein complexes. Sali and his co-workers used matching of the overall domains to predict protein interactions. Scoring of these putative complexes are performed again using statistical potentials. The potentials derived from side chain-side chain contacts are found to distinguish the non-native contacts with an accuracy of 0.993 [160]. This method is later applied to a target dataset composed

of host and pathogen proteins to model the interactions between them [161]. Aloy et al., considered overall structural fold similarity besides sequence similarity to make more complete the structural interaction network of yeast [13].

As implied in the previous sections, the structure of the protein is evolutionarily more conserved than the sequence; further the interface of the protein is more conserved than the overall structure [66][ref]. Based on the recognition that binding and folding are similar processes with similar underlying principles, we proposed that interface structural similarity exists not only between *homologous* protein pairs; *different protein fold-pairs* can also interact *via* similar interface architectures and these architectures are similar to those observed in single chain proteins [20, 52]. Hence, the illustration of this concept inspired the idea that just using the interface region - independent from sequence similarity or global fold similarity of protein pairs – can produce promising models for protein complexes. PRISM [162, 163] is the first algorithm presenting this concept to model protein complexes where if two complementary partners of a template interface are similar to the surface of two target proteins, these two proteins are principally interact with each other using this template architecture. This method is used to generate a structural interaction map of cancer proteins [164] and to show the multi-face nature of the hub proteins [99]. This method was utilized to characterize the human cancer proteins interaction network [164] and to predict interactions for multimeric hub proteins states with shared binding sites [165]. Similar methods appeared following this idea [166-168]. One of them is ISearch which depends on the same basis with PRISM with one exception where the template interfaces in this method are domain-domain interfaces [166]. In a most recent work, Vakser and his colleagues implies the necessity of template-based docking independent from sequence homology and global fold similarity [169]. The predicted models are ranked according to a score coming from matching which changes between 0 and 1. Their results show that local structural alignments give more accurate models than global structural alignments. If the sequence similarity between target and template is very low, homology-based methods fail to find any similarity; hence, interface templates are useful to detect interaction even for the dissimilar target and template sequences.

In all template-based methods, the procedure is very smart and straightforward: select a high quality non-redundant template dataset composed of known protein complexes, extrapolate the known structural data to identify unknown interactions by using the

sequence or structural similarity, rank the predictions according to a scoring function i.e. statistical potentials or energy functions. When we look at all these template-based methods in a comparative manner, we notice that the most critical step is the choice of the template for all of them to model an interaction, because mistaken templates can produce false positive binding regions. In addition, template-based methods mainly decrease the solution space and helps docking approaches by limiting the possible orientations. In this way, these methods are computationally more efficient than the docking strategies and easily applicable at proteome scale. The most important issue in all of them is the optimized scoring of the models. The accuracy of the template-based methods are also sufficient for proteome-scale studies [170].

The limitation of these template-based methods is the availability of the similar templates in the dataset. If there is no similar template available it is not expected to find a prediction. However, if there are similar templates in the dataset, to find a prediction is very fast and reliable. With the exponential growth of the number of protein complexes in PDB, template-based prediction methods will take much more attention in the near future.

2.5.3 Construction of a Non-redundant Template Set for High-Quality Predictions

As implied in the previous section, the construction of a non-redundant template set is the most challenging part of the template-based prediction methods. Hence, each template-based prediction method has different approaches to generate a non-redundant template set. In Dockground, the representative set of the protein complexes is obtained by the elimination of sequence and structural similarities and for the remaining the highest resolution one is selected as representative. For the modeling of protein complexes, 11,932 protein complexes are considered whose resolution is less than 3 Å and the sequence similarity is less 90%. Also, the interface definition (the distance cutoff between any two atoms) is varied from 6 Å to 16 Å and 12 Å is found to be optimal cutoff. In another study, the template dataset is generated from domain-domain interfaces with a resolution of less than 3.5 Å. Further, domain-domain interfaces are clustered at superfamily-superfamily level. Interfaces in the clusters are ranked according to the interface size, resolution and domain size and the first ranking interfaces are selected as representative. However, it is stated that just clustering according to the overall domain superfamily is not enough to illustrate the structurally similar interfaces. In a homology-based method, the protein

complexes from the ProtCom database [19] is filtered according to a sequence similarity threshold to form the template set. Protein complexes having sequence similarity less than 40% is considered and crystal artifacts are not considered resulting in 463 protein complexes in the template set [158]. SCOPPI [14] and PIBASE [18] are two of the domain level interface databases used to generate template set for structural modeling of protein complexes. Clustering of the protein interfaces in SCOPPI database is achieved by the angle between two domains and structural alignment of the domains [14]. Protein interfaces in PIBASE is clustered according to the topology of the secondary structure elements [18].

Structural alignment of the protein interfaces is another approach to generate a non-redundant set. Geometric hashing is one of the very powerful methods to align interfaces and to find structural similarity between them. All available protein interfaces are clustered by geometric hashing using just Ca atoms at different time periods [10, 20, 52]. In these datasets, all interfaces are compared to each other iteratively and if they are structurally similar, they go to the same cluster. The structural similarity within clusters is very high while the similarity is very low between clusters. I2ISiteEngine [171] also uses geometric hashing for interface alignment; however, it considers the physicochemical properties of atoms (donor, acceptor etc.) besides their spatial arrangements. Another program specialized in the alignment of protein interfaces is Galinter [172]. This method represents the van der Waals interactions and hydrogen bonds within the interface as vectors and aligns these vectors with each other to find similarity between two interfaces. When compared to I2ISiteEngine, despite the different features used, their performances are similar. The most recent work about alignment of protein interfaces is iAlign [173] which uses heuristics techniques and iterative dynamic programming. The performance comparison of iAlign with I2ISiteEngine on the same test set shows that the accuracy of iAlign is higher than the accuracy of I2ISiteEngine; further, the computational time comparison points out that iAlign algorithm is faster.

When the template generation methods are examined, we see that the differences come from interface definitions, sequence similarity thresholds, clustering methods, and domain/chain level interfaces. Also, for high quality predictions the template set should as structurally diverse as possible to be able to cover most of the interactions. Hence, the structural alignment of the interfaces – independent from their sequence and global fold –

has essential role in template generation. Before the scoring of the predicted interactions, optimization of the parameters to generate a diverse and non-redundant template set is crucial for high-quality modeling of the interactions.

Chapter 3

IDENTIFICATION OF COMPUTATIONAL HOT SPOTS IN PROTEIN INTERFACES

Hot spots are residues comprising only a small fraction of interfaces yet accounting for the majority of the binding energy as reviewed in Chapter 2. In this chapter, a new intuitive efficient method is presented to determine computational hot spots based on conservation (C), solvent accessibility (ASA) and statistical pairwise residue potentials (PP) of the interface residues. Combination of these features is examined in a comprehensive way to study their effect in hot spot detection. The predicted hot spots are observed to match with the experimental hot spots with an accuracy of 70% and a precision of 64% in Alanine Scanning Energetics Database (ASEdb), and accuracy of 70% and a precision of 73% in Binding Interface Database (BID). Several machine learning methods are also applied to predict hot spots. Performance of our empirical approach exceeds learning based methods and other existing hot spot prediction methods. Residue occlusion from solvent in the complexes and pairwise potentials are found to be the main discriminative features in hot spot prediction. Our empirical method is a simple approach in hot spot prediction yet with its high accuracy and computational effectiveness.

3.1 Methodology for Identification of Computational Hot Spots

3.1.1 Training Set

Proteins that have experimental hot spot data and available crystal structures are used in developing a scoring formula. Alanine scanning data was obtained from the ASEdb, and a previously compiled data set from Robetta. The redundancy in this data set is removed using PISCES sequence culling server [174] with no sequence identity more than 35% as in the procedure of Darnell et al. [111]. In the training part, we have used only hot spots and non-hotspots to be more discriminative. The interface residues whose observed binding free energies are greater than or equal to 2.0 kcal/mol are considered as hot spots. Also, the interface residues whose binding free energy is smaller than 0.4 kcal/mol are

labeled as non-hot spots in a similar way with Gao et al [114]. Other residues having binding free energy between 0.4 and 2.0 are not included in the training to discriminate better. Actual training set used during 2-class (hot spot, non-hot spot) prediction model construction consists of 150 residues, for which both conservation and solvent accessibility information is available, of which 58 residues are hot spot and 92 residues are non-hot spot.

3.1.2 Test Set

A test set, used for assessing performance of proposed prediction models, is taken from Binding Interface Database (BID) [97]. BID contains binding free energy strengths of monomers. The test set is filtered for identical sequences in a similar fashion to the training set. The resulting set shrinks to 112 residues on 25 monomers (54 hot spots and 58 non-hot spots) when residues with known conservation scores and accessibility are considered. Hot spot residues are labeled as the ones with “strong” interaction strengths and others are tagged as non-hot spot. The data originating from training and test sets are mutually exclusive. The list of training and test sets are available as supplementary at; <http://prism.cccb.ku.edu.tr/hotpoint/supplement.doc>

3.1.3 Features

Accessibility: The accessible surface area (ASA) of each residue in monomer state and in complex state in the training and test sets are calculated by using Naccess [175]. These ASAs are then converted into relative accessibility:

$$\text{relCompASA}_i = \left(\frac{\text{ASA in Complex}_i}{\text{maxASA}_i} \right) \times 100 \quad (3.1)$$

$$\text{rel}\Delta\text{ASA}_i = \left(\frac{[\text{ASA in Monomer}_i] - [\text{ASA in Complex}_i]}{\text{maxASA}_i} \right) \times 100 \quad (3.2)$$

where “relCompASA_i” is the relative ASA in complex of ith residue and “relΔASA_i” is the relative difference ASA between complex and monomer state of ith residue; in other words, the ASA change of the residue upon complexation. “maxASA_i” is the maximum ASA of a residue in a tri-peptide state [176].

Conservation: Residue conservations are found by Rate4Site (R4S) algorithm [177]. R4S makes use of topology and branch lengths of the phylogenetic trees constructed from multiple sequence alignments (MSA) of proteins and estimates conservation rates of the amino acids based on the empirical Bayesian rule. MSAs of proteins are taken from HSSP (Homology-Derived Secondary Structure of Proteins, [178]) database. All MSAs obtained from HSSP are converted to FASTA format to be used in R4S step. Conservation scores obtained by R4S are scaled between 1 and 9. The scaled conservation score of residue i (between 1 to 9) is called Score_i .

Pair Potentials: The knowledge-based potentials have been shown to be useful in many threading, folding and binding problems [179-181]. Residue specific, non-bonded interactions taking place between sequentially distant but spatially close amino acid residues (neighbors) are recognized to play a dominant role in the stabilization of globular proteins and complexes [180, 182-187]. A practical way to obtain these potentials is to extract them from frequencies of contacts between different residues in proteins with known three-dimensional structures. Knowledge based solvent mediated inter-residue potentials [151], extracted from protein interfaces, are used in this work. Although these potentials are not very different from the potentials extracted from overall proteins, subtle changes might be important to detect interface hot spot residues. 210 distinct potentials (all possible pairs of 20 amino acids) in RT unit (R universal gas constant, T is temperature) for contacting residue pairs are supplied in the Supplementary Material. Contact potential between two residues i and j is found as;

$$\text{Pair}(i, j) = \begin{cases} \text{contact potential of type}(i, j) & \text{if } d(i, j) \leq 7.0 \text{ and } |i - j| \geq 4 \\ 0 & \text{otherwise} \end{cases} \quad (3.3)$$

where $\text{Pair}(i, j)$ is the contact potential of residues i and j and $d(i, j)$ is the distance between two residue centers [150]. We extracted the neighbors around the residues whose side chain center of mass are closer than the cutoff (7.0 Å). Overall contact potential of residue i is defined as the absolute of sum of its pair potentials:

$$\text{PP}_i = \text{abs} \left(\sum_{j=1}^n \text{Pair}(i, j) \right) \quad \text{for } |i - j| \geq 4 \quad (3.4)$$

Computational Alanine Scanning (Robetta): Robetta [110, 121] is a server which includes computational alanine scanning. Robetta server gives changes in the binding free energy ($\Delta\Delta G$) values based on an atomic energy function including Lennard Jones interactions, solvation interactions and hydrogen bonding. The calculated $\Delta\Delta G$ is named “Robetta” throughout our work. Robetta ≥ 1.0 kcal/mol is the default cutoff in the hot spot predictions in all models.

3.1.4 Assessment of the Prediction Performance

Accuracy is the ratio of number of correctly predicted residues to number of all predicted residues, formulated as;

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \quad (3.5)$$

where TP, FP, TN, FN stands for number of true positives (correctly predicted hot spot residues), number of false positives (non-hot spot residues incorrectly predicted as hot spots), number of true negatives (correctly predicted non-hot spot residues) and number of false negatives (hot spot residues incorrectly predicted as non-hot spots) respectively.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (3.6)$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (3.7)$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \quad (3.8)$$

where **Recall** is the proportion of number of correctly classified hot spot residues to the number of all hot spot residues; **Specificity** is the proportion of number of correctly predicted non-hot spot residues to the number of all non-hot spot residues; **Precision** is the ratio of number of correctly classified hot spot residues to the number of all residues classified as hot spots. Also by using F-measure we check the balance between precision and recall which is formulated as follows:

$$\text{F-Measure} = \frac{2 \times \text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}} \quad (3.9)$$

3.1.5 Determination of Computational Hotspots

Size of the experimental hot spot data is small to be used in learning based methods with large number of features to determine the hot spot characteristics. We prefer to construct our model incrementally first examining single features (base cases), and then improving our model by addition of other significant features. In the base models we use only one feature, such as relative ASA in complex, relative difference ASA, conservation, pair potentials to discriminate hot and non-hot residues. These features are selected considering following criteria: hot spots are buried [1], structurally more conserved, highly packed [163], known to be mostly of specific residue types, i.e. aromatic [1]. The performance of the base models is used as lower bounds to assess the performance of our model and several machine learning based prediction approaches.

1) Base Cases:

- a. $\text{Score}_i \geq t_{\text{score}}$
- b. $\text{rel}\Delta\text{ASA}_i \geq t_{\text{rel}\Delta\text{ASA}}$
- c. $\text{relCompASA}_i \leq t_{\text{relCompASA}}$
- d. $\text{PairPotential}_i \geq t_{\text{PairPotential}}$
- e. $\text{Robetta} \geq t_{\text{Robetta}}$

where t_{score} , $t_{\text{rel}\Delta\text{ASA}}$, $t_{\text{relCompASA}}$, $t_{\text{pairPotential}}$ and t_{Robetta} are thresholds, and currently the default values are set to 7, 30%, 20% , 18.0, and 1 respectively. The explanation and justification for these default values are given in the results section.

2) Combination of two features:

We have tested the performance of some possible two features:

$$\text{Score}_i + \text{relCompASA}_i,$$

$$\text{Score}_i + \text{PP}_i,$$

$$\text{relCompASA}_i + \text{PP}_i,$$

$$\text{relCompASA}_i + \text{Robetta}.$$

3) Addition of a third feature:

$$\text{relCompASA}_i \leq t_{\text{relCompASA}} \ \& \ (\text{Score}_i \geq t_{\text{pScore}} \ \text{or} \ \text{Robetta} \geq t_{\text{Robetta}}),$$

$$\text{relCompASA}_i \leq t_{\text{relCompASA}} \ \& \ (\text{Score}_i \geq t_{\text{pScore}} \ \text{or} \ \text{PP}_i \geq t_{\text{PairPotential}})$$

Further, we have used machine learning techniques to predict hot spots using the training set for learning. Several algorithms are employed for classification: Decision tree

(J48), decision table, SVM, BayesNet, Naïve Bayes, RBFNetwork, and Majority Voting. The features for each residue (for the learning algorithm) consist of the same ones that we have used in the formulations above, relCompASA_i , Score_i , and PP_i . The results and comparison of these formulations are discussed in the results section.

3.2 Results

3.2.1 Distribution of features of hot spots and non-hot spots

In order to decide on the threshold values, we have prepared histograms of relative complex ASA (relCompASA), relative change in ASA upon complexation ($\text{rel}\Delta\text{ASA}$), conservation score, and pair potentials for the hot spot and non-hot spot residues in ASEdb as shown in **Figure 3.1**. The mean and standard deviations of each feature are calculated for hot and non-hot residues. Further, t-tests are performed to determine if the difference between two distributions of hot and non-hot spots is statistically significant for each feature. For significant ones, we evaluate the formulas (in the Methods) by trying several threshold values between the two mean values.

Figure 3.1A shows the distribution of relCompASA . Though many of the hot spot residues have similar relCompASA values with non-hot spot residues, they have different mean values (hotspots: 11.9%, non-hotspots: 26.4%). The p-value for relCompASA is found as 4.7×10^{-7} (<0.05) which implies the significance between the means of the hot and non-hot distributions. There are significantly more non-hot spot residues which have relative complex ASA greater than 20% ($t_{\text{relCompASA}} = 20.0$). This is also consistent with previous studies indicating that hot spots are buried [1, 24, 163].

Figure 3.1B shows the distribution of change in ASA upon complexation. The means are found as 34.8 for hot spots and 26.4 for non-hot spots for $\text{rel}\Delta\text{ASA}$. This feature is also discriminative with a p-value 5×10^{-3} (<0.05). The threshold is determined as 30% for $\text{rel}\Delta\text{ASA}$ which is between the two mean values ($t_{\text{rel}\Delta\text{ASA}} = 30.0$).

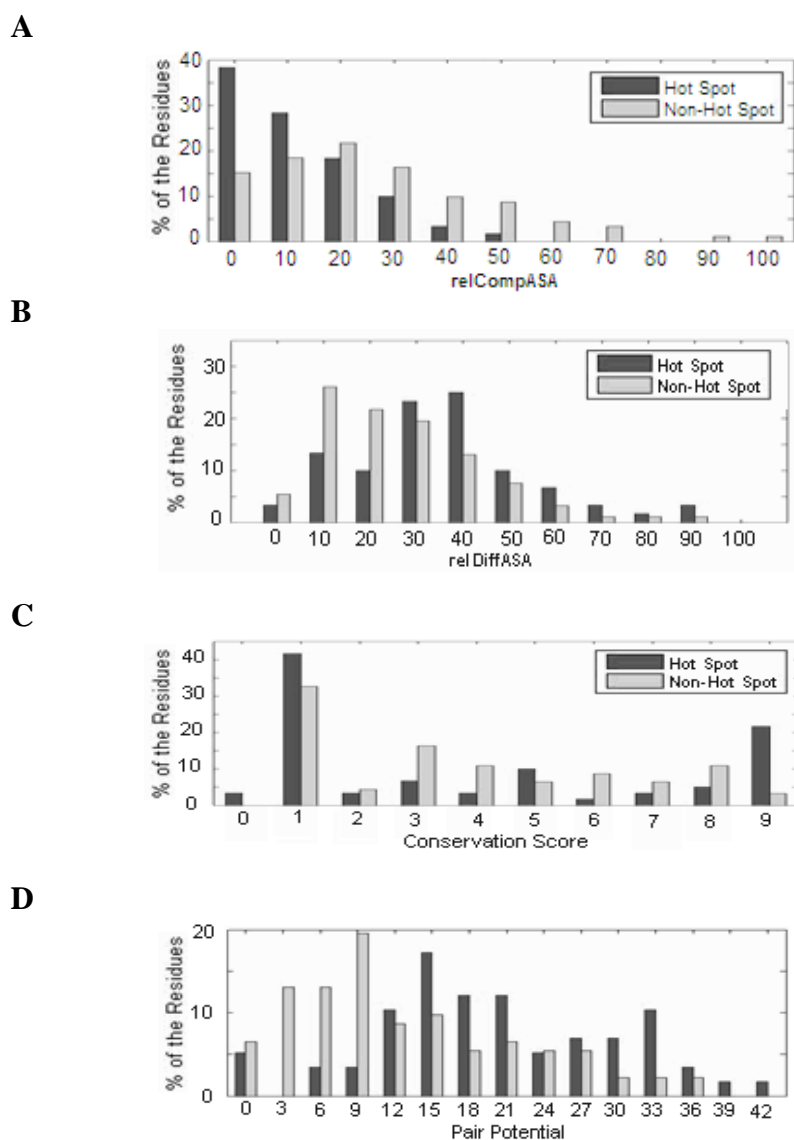


Figure 3.1 Distribution of hot spot and non hot spot residues in the available data set with respect to their (A) relCompASA, (B) relDiffASA, (C) conservation, (D) pair potential.

Figure 3.1C shows conservation score distribution which does not have a clear distinction between hot and non-hot spots. The mean value for hot residues is 4.2 and for non-hot residues 3.9. The difference between two sets is insignificant (p -value = 0.22). This indicates that conservation may not be a good discriminating factor by itself. However, to check this slight difference, we select the threshold for conservation score as 7.0 ($t_{\text{Score}}=7.0$) and test the performance of conservation in hot spot prediction.

Figure 3.1D displays the histogram for knowledge-based pair potentials of residues. The means for hot spots and non-hot spots are found as 20.3 and 12.7, respectively. This feature is statistically significant to discriminate hot spots and non-hot spots (p -value= 5.4×10^{-6}). A threshold of 18.0 ($t_{\text{pairPotential}} = 18.0$) is chosen since a residue with pair potential more than 18.0 has a higher tendency to be a hot spot.

We further performed ANOVA analysis and determined the most important features to distinguish hot spots from non-hot spots. relCompASA, rel Δ ASA and pair potentials were found to be significantly discriminative consistent with our histogram analysis (see **Figure 3.1** and **Table 3.1**).

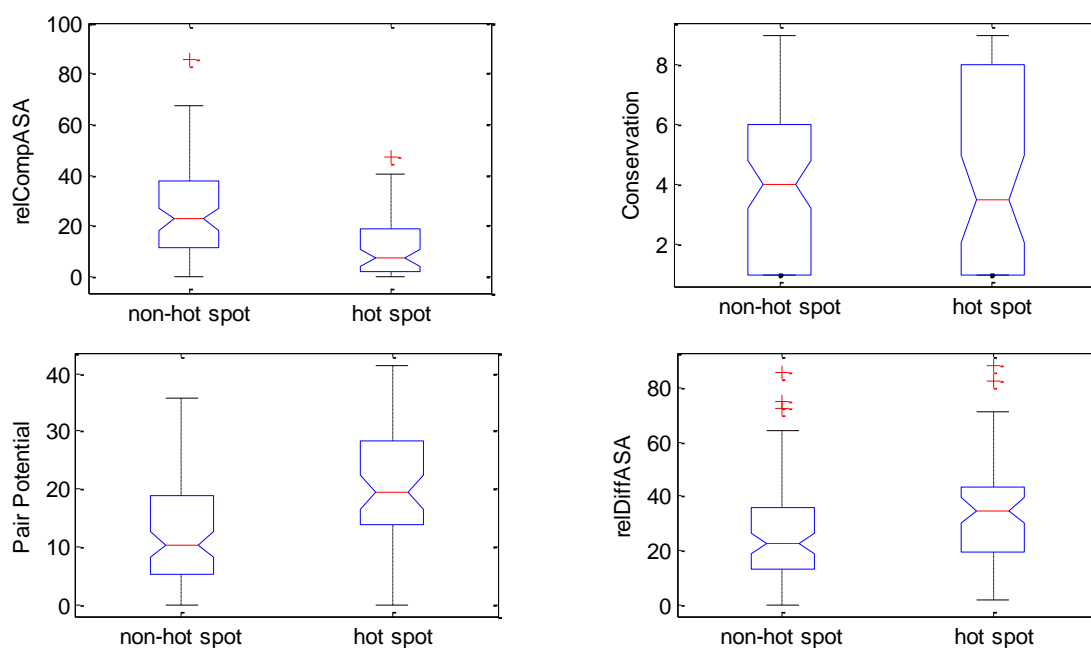


Figure 3.2 The comparison of hot spots and non-hot spots for each feature (relCompASA, conservation, pair potential, relDiffASA) plotted by ANOVA.

Table 3.1 The results of each features computed by ANOVA.

	p-Value	Mean (non-hotspot)	Mean (hot spot)	Mean (overall)
relCompASA	5.62×10^{-6}	26.64	11.96	20.96
Conservation	0.4979	3.88	4.21	4.00
Pair Potential	4.27×10^{-6}	12.70	20.30	15.64
relDiffASA	0.0088	26.39	34.76	29.62

3.2.2 Comparison of Empirical Hot Spot Detection Formulations

We have evaluated prediction performance of our models (formulations) and assessed the success of the formulations by comparing accuracy (A), recall (R), precision (P), specificity (S) and f-measure (F1). In our study, recall and specificity bear importance, since we emphasize predicting both hot spots and non-hot spots. However, precision strikes as a key determinant in quantifying how accurate the positive predictions are. The models are comprehensively tested on an independent test set (BID), and their statistical performances are presented in **Table 3.2**.

Table 3.2 Performance values of various empirical prediction methods used to identify hot spots in the protein interfaces.

	Model	Dataset	Precision	Recall	Spec.	Acc.	F1
Single Feature Performances on ASEdb and BID datasets (EMPRICAL FORMULAS)	Score ≥ 7.0	Training Set	0.50	0.33	0.79	0.61	0.40
		Test Set	0.52	0.46	0.60	0.54	0.49
	relCompASA ≤ 20.0	Training Set	0.55	0.81	0.58	0.67	0.65
		Test Set	0.60	0.67	0.59	0.63	0.63
	rel Δ ASA ≥ 30.0	Training Set	0.50	0.57	0.64	0.61	0.53
		Test Set	0.50	0.55	0.48	0.52	0.53
	Robetta ≥ 2.0	Training Set	0.82	0.47	0.93	0.75	0.59
		Test Set	0.64	0.26	0.86	0.57	0.37
	Robetta ≥ 1.0	Training Set	0.63	0.72	0.73	0.73	0.67
		Test Set	0.63	0.57	0.69	0.63	0.60
	PP ≥ 18.0	Training Set	0.56	0.55	0.73	0.66	0.56
		Test Set	0.69	0.70	0.71	0.71	0.70
Two Features Performances on ASEdb and BID datasets	relCompASA ≤ 20.0 and Score ≥ 7.0	Training Set	0.61	0.29	0.88	0.65	0.40
		Test Set	0.71	0.32	0.88	0.61	0.44
	PP ≥ 18.0 and Score _i ≥ 7.0	Training Set	0.57	0.14	0.94	0.63	0.22
		Test Set	0.75	0.33	0.90	0.63	0.46
	relCompASA ≤ 20.0 and Robetta ≥ 1.0	Training Set	0.72	0.62	0.85	0.76	0.67
		Test Set	0.75	0.50	0.85	0.68	0.60
relCompASA ≤ 20.0 and PP ≥ 18.0	Training Set	0.64	0.52	0.82	0.70	0.57	
	Test Set	0.73	0.59	0.79	0.70	0.65	
Multiple Features Performances on ASEdb and BID datasets	relCompASA ≤ 20.0 and (Score ≥ 7.0 or Robetta ≥ 1.0)	Training Set	0.64	0.69	0.75	0.73	0.66
		Test Set	0.73	0.59	0.79	0.70	0.65
	relCompASA ≤ 20.0 and (Score ≥ 7.0 or PP ≥ 18.0)	Training Set	0.63	0.67	0.75	0.72	0.65
		Test Set	0.67	0.63	0.71	0.67	0.65

The first part of the table compares single feature models. Among them, conservation is observed to have no significant effect on its own. It gives the least successful results (F1 scores and accuracies, 0.40 and 0.61 on training set, 0.49 and 0.54 on test set, respectively) compared to other features both on ASEdb and BID. This is expected according to the histogram which states hot spots are marginally conserved (**Figure 3.1C**) in line with the results of Ofran & Rost, 2007. However, conservation was found to improve predictions substantially [98] which is not the case in our results. Interface

residues are found to be more conserved than the rest of the surface residues [65, 188]; further, central interface residues were more conserved than peripheral ones [189]. Caffrey et al. analyzed interfaces using surface patches, they found that the difference between the patches and the rest was even less pronounced [66]. Here, our results suggest sequence conservation is not a discriminative characteristic of hot spots ($P=0.50$, $R=0.33$ on training set; $P=0.52$, $R=0.46$ on test set). However, we observe that the totally conserved residues (with top score 9 in our conservation scoring) are found to be substantially buried in the middle regions of the interfaces. On the other hand, not all buried residues are necessarily conserved.

Occlusion of a residue from solvent in complex state is indicated by a small relCompASA . Our results show that low relCompASA is critical for a residue to be a hot spot. Bogan and Thorn (1998) indicated that hot spots located near the center of the interface are a general property of the interfaces; and for a residue to be a hot spot, it must be largely protected from bulk solvent (corresponding to low relCompASA). Even if a residue was exposed to solvent prior to binding, it might lose a high percentage of its surface area and become protected from the solvent. This scenario is consistent with what Li et al. suggested: hot spots are either found on the complemented pockets or on the protruding surfaces [24]. Complemented pockets and their corresponding protruding residues bind to each other, eventually, to protect each other from the solvent. $\text{rel}\Delta\text{ASA}$ indicates the change in the solvent accessibility of a residue. The rationale for choosing $\text{rel}\Delta\text{ASA}_i > 30\%$ in our formulation is to be able to find the protruding residues based on this fact. However, probably due to the small number of protruding hot spot residues, this parameter ($P=0.50$, $R=0.55$, $A=0.52$, $F1=0.53$) does not perform better than relCompASA ($P=0.60$, $R=0.67$, $A=0.63$, $F1=0.63$). As a result, relCompASA seems to discriminate better hot spots from non-hot spots. Both of the energetic models (the knowledge based pair potentials and full atomistic energy terms of Robetta) seem to be quite successful to find the hot spots. Robetta's recall, precision and accuracy are higher in ASEdb (0.63, 0.72, and 0.73, respectively) but lower in BID (0.63, 0.57, and 0.63, respectively). On the other hand, pair potential performs better in BID (0.69, 0.70, and 0.71, respectively) compared to Robetta. Note that ASEdb is the training set and BID is our independent test set. As stated by Janin and his group, protein interfaces also have core and rim regions and hot spots are usually located in the cores of the interfaces [21, 62]. A residue in the core

with favorable contacts has a higher chance to be a hot spot. This could be the reason why pair potential works well. Furthermore, using full atomistic energy parameters, Robetta has computational disadvantage for large-scale predictions. In addition, the performance differences of the two models on the two distinct datasets indicate the different nature of the two datasets. The hot spots in ASEdb are defined by a single threshold of 2 kcal/mol; however, in BID, there is no single threshold but rather hot spots are divided into strong, intermediate, and neutral interactions. Thresholds change from one case to another.

We have further tested the effect of combining features. First, we combined two features: (relCompASA + Score), (PP + Score), (relCompASA + Robetta), and (relCompASA + PP). We observe that in all these cases, adding a second feature increases the precision, specificity and accuracy but decreases the recall. In other words, fewer positive hot spot predictions are made with higher percentage of true cases; in addition, non-hot predictions improve compared to single feature models. On ASEdb, (relCompASA + Robetta) model has the highest F1 score (0.67) whereas on BID, (relCompASA + PP) has the highest F1 score (0.65). Compared to the single feature performances, adding the relCompASA in the Robetta model increases the precision from 0.63 to 0.75 and in the PP model from 0.69 to 0.73 in BID. Similarly, specificity increases from 0.69 to 0.85 (for Robetta) and 0.71 to 0.79 (for PP) in BID. The model starts to pick the hot spots and non-hot spots with higher specificity and precision. Further, adding relCompASA improves the performance of pair potentials on ASEdb with respect to pair potentials only while maintaining the BID performance. Similarly, relCompASA improves performance of Robetta on BID compared to using Robetta only. Our results indicate that hot spots are mostly buried and form a network of favorable interactions with other residues as reported by Nussinov and her colleagues [163]. When conservation score is added to these two-feature models, both the precision and specificity decrease. We should note that, we tried many other features, combinations and thresholds, but listed only the high scoring ones. As a result, our prediction based on relCompASA and pair potentials demonstrated 0.70 *accuracy* and 0.73 *precision*, 0.59 *recall*, and 0.79 *specificity* on the independent test set. It performed better than the base models and the machine learning based models (discussed in the next section). We predict 32 of the hot spots correctly with 12 false positives. On the other hand, 46 of the non-hot spots are correctly classified with 22 false negatives (**Table 3.3**).

Table 3.3 Prediction results for the structures in test set (BID).

Interface ID	Monomer ID	TP	FP	TN	FN	Total	Accuracy (%)
1cdlAE	1cdlA	1	2	0	0	3	33
1cdlAE	1cdlE	3	1	2	0	6	83
1ddmAB	1ddmA	2	0	3	2	7	71
1ddmAB	1ddmB	0	0	3	5	8	38
1dfjEI	1dfjE	0	0	0	1	1	0
1dvaHX	1dvaH	2	5	6	1	14	57
1dvaHX	1dvaX	4	1	2	0	7	86
1dziAC	1dziA	1	0	5	2	8	75
1ebpAC	1ebpA	2	0	1	1	4	75
1ebpAC	1ebpC	0	0	2	1	3	67
1es7AB	1es7A	0	1	0	0	1	0
1fccAC	1fccC	3	0	4	0	7	100
1foeAB	1foeB	0	0	1	0	1	100
1gl4AB	1gl4A	3	0	2	2	7	71
1jatAB	1jatA	1	0	0	0	1	100
1jatAB	1jatB	1	0	0	0	1	100
1k4uSP	1k4uP	1	1	2	1	5	60
1lqbCD	1lqbD	0	0	2	0	2	100
1mq8AB	1mq8B	1	0	0	0	1	100
1nfiBF	1nfiF	1	0	1	0	2	100
1ub4AC	1ub4C	1	0	1	1	3	67
2hhbAD	2hhbD	0	0	1	0	1	100
2nmbAB	2nmbA	1	0	0	1	2	50
2nmbAB	2nmbB	0	1	1	0	2	50
3sakAC	3sakA	4	0	7	4	15	73
Total		32	12	46	22	112	70

3.2.3 Machine Learning Based Approaches

The machine learning (ML) methods fail to create a distinctive improvement over our proposed model. Performance of ML based models is illustrated in **Table 3.4** with the details of the classifiers on 10-fold cross-validation and on test set. In general, ML based models do not exceed our empirical formula ($A=0.70$). The main reason for this relative failure is probably deficiency of training data. Nevertheless, decision trees play an indispensable role in determination of relative importance of the features. We have applied decision tree for three features; relative compASA, pair potential and conservation score. The decision tree model determines pair potential as the most discriminating feature followed by relCompASA in accordance with our model. Testing on BID dataset gives an accuracy of 0.63 with a recall of 0.52. We have also applied other classifiers and their combination by majority voting. Best classifier among them is BayesNet based on F1-score. Its accuracy is 0.68 on 10-fold cross validation test and 0.64 on BID test set.

Table 3.4 Machine learning based models Results. These are the corresponding implementations from Weka [190]

Classifier	Testing	Pre.	Rec.	Spec.	Acc.	F1
BayesNet	10 – fold	0.58	0.64	0.71	0.68	0.61
	Test set	0.64	0.63	0.67	0.65	0.64
Naïve Bayes	10 – fold	0.57	0.66	0.69	0.67	0.61
	Test set	0.63	0.67	0.64	0.65	0.65
RBFNetwork	10 – fold	0.59	0.55	0.76	0.68	0.57
	Test set	0.67	0.48	0.78	0.63	0.56
SVM	10 – fold	0.57	0.36	0.83	0.65	0.44
	Test set	0.73	0.44	0.85	0.65	0.55
Decision Tree (J48)	10 – fold	0.47	0.59	0.59	0.59	0.52
	Test set	0.65	0.52	0.74	0.63	0.58
Decision Table	10 – fold	0.58	0.64	0.71	0.68	0.61
	Test set	0.64	0.63	0.67	0.65	0.64
Majority Voting (all except SVM)	10 – fold	0.56	0.62	0.70	0.67	0.59
	Test set	0.64	0.63	0.67	0.65	0.64

3.2.4 Comparison with Other Hot Spot Prediction Methods

Robetta is designed to find the computational alanine scanning mutagenesis and gives $\Delta\Delta G$ values for individual residues. In their work, interface residues whose experimental $\Delta\Delta G$ value is greater or equal to 1.0 kcal/mol are considered as experimental hot spots [121]. Also, if predicted $\Delta\Delta G$ values are greater or equal to 1.0 kcal/mol, the corresponding residues are labeled as computational hot spots. The predictive performance of Robetta on BID (with $\Delta\Delta G \geq 1.0$ kcal/mol) is as follows; P=0.63, R=0.57, S=0.69, A=0.63, and F1=0.60. Our empirical formula achieved a slightly better success rate compared to Robetta with its precision of 0.73, accuracy of 0.70 and F1 score of 0.65 (as shown in **Table 3.5**). A recently described method, KFC shows a precision of 0.51 and a recall of 0.36 with 0.42 F1-score on BID (performance data are taken from Darnell et al). KFCA is a hybrid method which combines KFC with Robetta. KFC is trained on a bunch of features such as residue size, atomic contacts, hydrogen bonds, chemical type etc. Our method – comprising relCompASA and pair potentials – performs better than both KFC and KFCA with its precision of 0.73 and recall of 0.59 with 0.65 F1-score compared to KFC (0.51, 0.36 and 0.42) and KFCA (0.53, 0.48 and 0.41) (**Table 3.5**). We further applied ISIS, a sequence based approach, on BID giving following performance: P=0.48, R=0.70, F1=0.57. Although the precision of ISIS is low we should note that the method is not

designed for hot spot prediction but rather finding binding site residues and it does not use structure information. Therefore, it is not fair to compare it with the structure based methods.

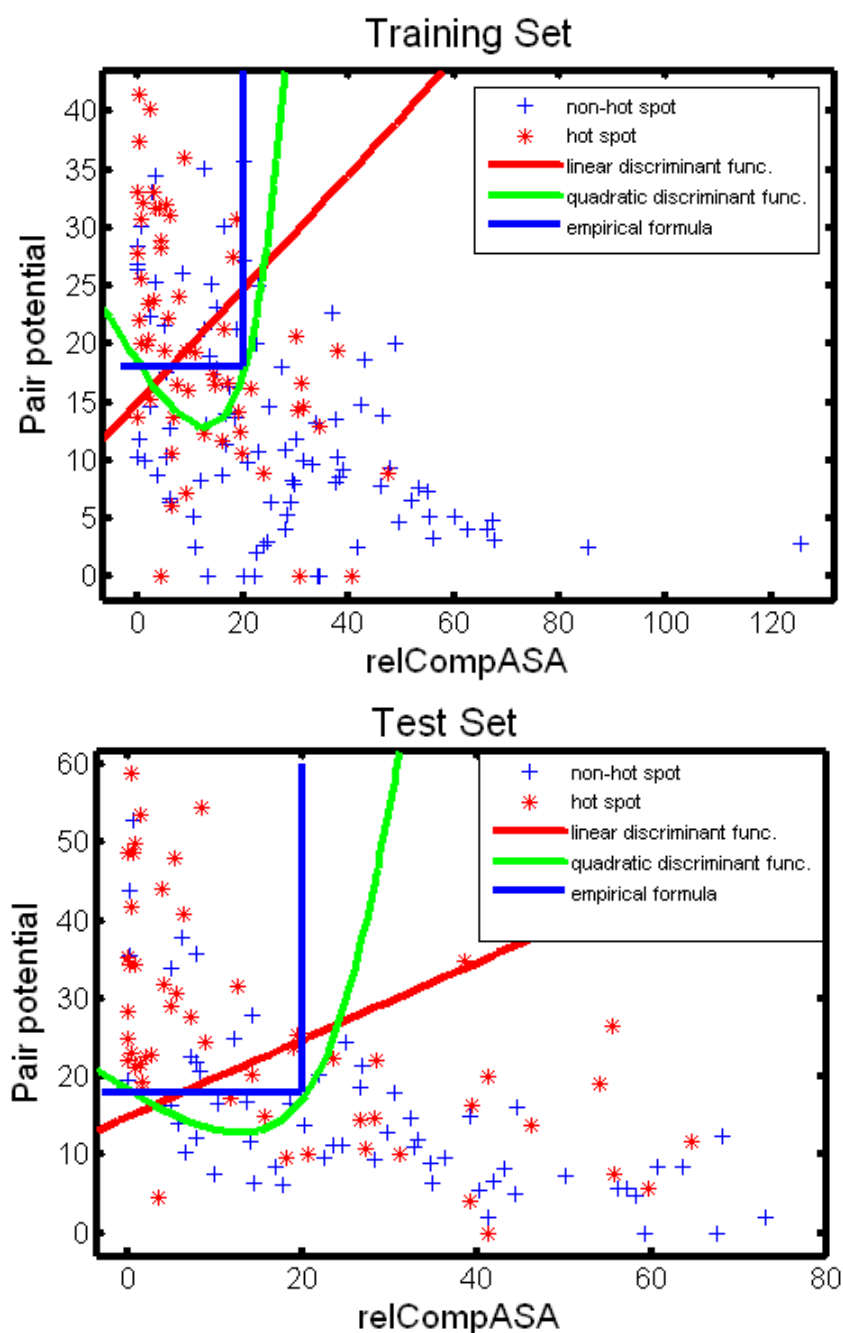


Figure 3.3 The distribution of hot spot and non-hot spot residues with the empirical formula and discriminant functions.

We also performed discriminant analysis (both linear and quadratic), trained on ASEdb and tested on BID, resulting in comparable performance.. The distribution of the data

points are illustrated in **Figure 3.3** both for training set and for test set. The performance values of linear and quadratic discriminant functions are given in **Table 3.6**. LDA results in comparable performance (see **Table 3.5**). However, our method has advantages of presenting a simple and intuitive rule relating physical properties to hot spots.

Table 3.5 Hot spot prediction performances on test set (BID).

Method	Precision	Recall	F1
Robetta	0.63	0.57	0.60
KFC	0.51	0.36	0.42
KFCA	0.53	0.48	0.51
LDA	0.72	0.57	0.64
Our Formula	0.73	0.59	0.65

When we analyze overall performances, we noticed that our results are similar to Robetta; however, it outperforms any machine learning based predictions including KFC. Besides its high prediction performance, another advantage of our method over Robetta is its computational effectiveness and applicability to the large scale datasets.

Table 3.6 Performance values of linear and quadratic discriminant functions.

	Dataset	Precision	Recall	Specificity	Accuracy	F1-Score
Linear	Training set	0.61	0.47	0.82	0.68	0.53
	Test set	0.72	0.57	0.79	0.69	0.64
Quadratic	Training set	0.58	0.60	0.73	0.68	0.59
	Test set	0.68	0.63	0.72	0.68	0.65

3.2.5 Case studies

Erythropoietic Receptor (EPOR) – EPO Mimetic Peptide

Erythropoietin (EPO) is a hormone participating in the regulation of proliferation and differentiation of immature erythroid cells. EPO mimetic peptide (EMP1) functions as a mimetic of EPO. There is a competition between EMP1 (pdbID:1ebp, chainC) and EPO to bind the EPOR (pdbID:1ebp, chainA) [191]. Despite the unrelated sequences of EMP1 and EPO, both can bind to the EPOR stimulating biological activity. Experimentally defined hot spots in 1ebpAC interface are F93_A, M150_A, F205_A and W13_C. In addition, T151_A, L11_C, T12_C are found experimentally to be non-hotspots (in BID). Our empirical method predicts two of the four hot spots correctly which are F205_A and M150_A. Despite their high contact potentials, because F93_A and W13_C are exposed to

solvent, they are predicted as non-hot spots. Also, all of the 3 non-hot spots are predicted correctly. In total, 5 of the 7 residues are correctly predicted (**Figure 3.4**). KFC predicts all seven residues as non-hot spots. Robetta identifies M150_A and W13_C as hot spots correctly and the rest as non-hot spots.

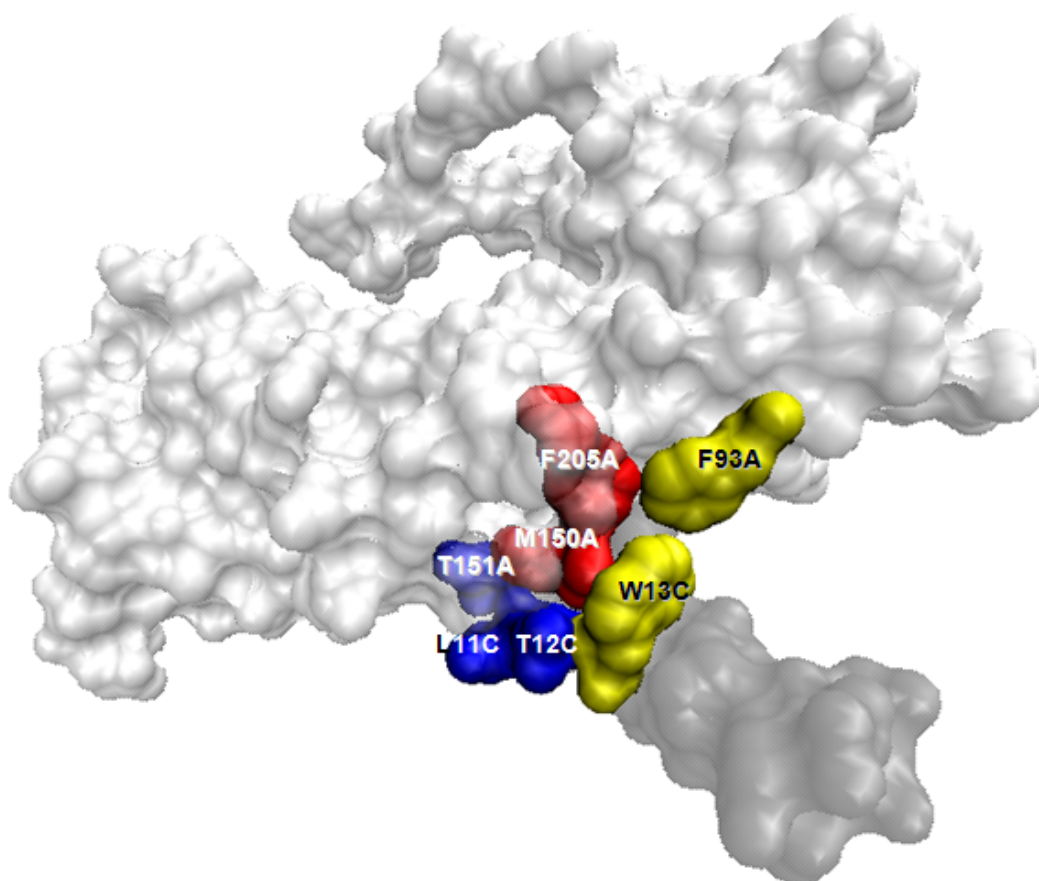


Figure 3.4 A case study of computational hot spot prediction using our empirical model. The complex formed between erythropoietin (EPO) receptor (chain A, colored white) and EPO Mimetics peptide (chain C, colored gray) (PDB ID: 1ebp, interface ID: 1ebpAC) is visualized (using VMD [192]). Red residues are experimental hot spots and correctly predicted. Yellow residue indicates the residue that the residue is an actual hot spot predicted as non-hot spot. Blue residues are non-hot spots which are also predicted as non-hot spots. In this case, 5 of the 7 residues are predicted correctly by our proposed model.

Streptococcal Protein G – Mammalian Immunoglobulin

Streptococcal protein G (pdbID: 1fcc, chain C) is a cell wall protein which binds mammalian immunoglobulin (pdbID: 1fcc, chain A) [193]. Protein G has experimentally determined 3 hot, 4 non-hot spots in its binding site to immunoglobulin. Mutations of

residues E27_C, K31_C and W43_C in protein G strongly affect its binding to immunoglobulin. These hot spots are located in the middle of the binding site of protein G to immunoglobulin and form a cluster of hot spots. Our method labels all these residues as hot spots. Also, non-hot residues are distributed the edges of binding site and more accessible to the solvent and they have less contact to other residues. All of them are predicted as non-hot by our model (**Figure 3.5**). Robetta and KFC perform similar. They identify E27_C and W43_C as hot spots correctly and the rest as non-hotspots. These two cases are selected from BID randomly; however, when other cases are examined we noticed that our predictions correlate with Robetta.

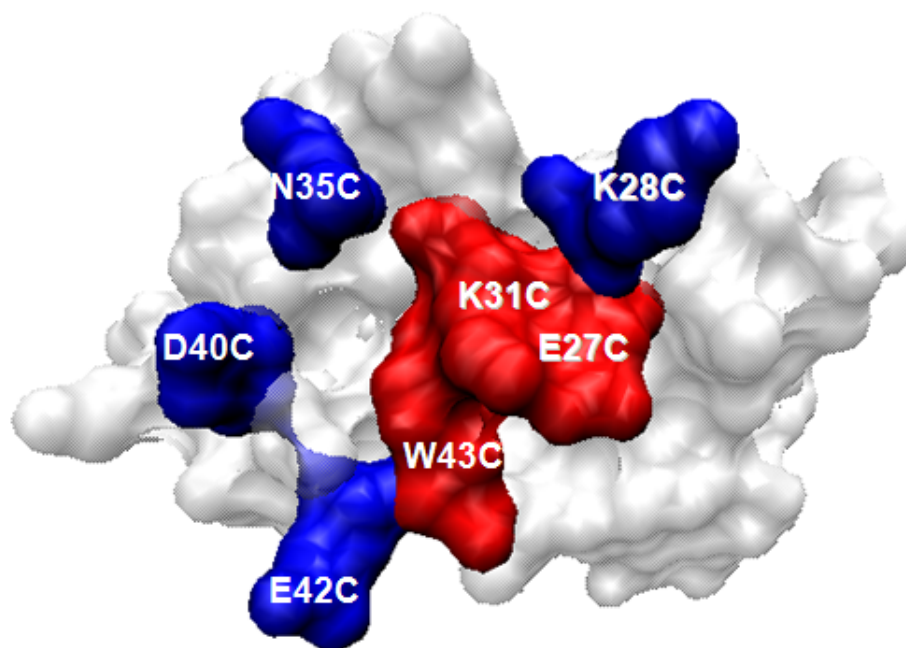


Figure 3.5 Streptococcal Protein (pdbID: 1fcc, chain C). Chain A is not shown. Red residues are actual hot spots predicted correctly; blue residues are actual non-hot spots predicted correctly. Hot residues form hot region and this region is located in the middle of the binding site of 1fccC.

3.3 HotPoint: Hot Spot Prediction Server for Protein Interfaces

Here, we present HotPoint web server, which provides a user-friendly interface to run the method explained in the previous sections [99] for online prediction of hot spots in protein interfaces. Our aim is to provide an efficient server at a single location for analysis of any protein-protein interface which can be utilized by researchers interested in protein binding

sites. The method principally considers the solvent accessibility and the total contact potential of the interface residues. The output tabulates the interface residues with the highlighted hot spots and their features. Additionally, it provides an interactive 3D visualization of the submitted protein-protein interface with the predicted hot spots for observing their localization. Distinct features of HotPoint from existing servers (Robetta and KFC server) are the improved efficiency and accuracy. The calculation of solvent accessibility and pair potentials of residues are faster than atomic level computations performed by Robetta, and the prediction accuracy is higher than both Robetta and KFC server. The HotPoint web server is available at <http://prism.cccb.ku.edu.tr/hotpoint>. Server interface is coded in PHP. The code to predict hot spots is written in Python.

3.3.1 Input for the Server

Input data is the protein structure in PDB formatted coordinate file, two chain identifiers forming the interface and the interface definition. User can either run the server with default distance thresholds to extract interface residues or can change the interface definition by submitting a distance threshold. There are two options to submit a structure file. User can enter the four letter PDB code of a protein which is directly downloaded from the ftp site of PDB. The second option is uploading a structure file that is in the PDB format. HotPoint requires two chain identifiers which confine to a protein interface. Server does not work for PDB files containing only one chain and returns an error. For NMR structures, it uses the first model in the prediction and gives results for the first model. HotPoint is specific to protein-protein interfaces; chains corresponding to DNA structures return a warning in the web server.

When there is not enough input data, the server informs the users of what is missing. The HotPoint web server is free and open to all users and there are no login requirements.

3.3.2 Output of the Server

When a protein structure with its chain identifiers is submitted, HotPoint server starts the calculation of three consecutive steps: extraction of interface residues, calculation of the features, prediction based on empirical model. During the processing, the server informs users about the steps it is performing. The output of the server is a table consisting of the interface residues with their features (**Figure 3.6**). The interface residues are tabulated with

chain names, one letter residue names, residue numbers, their relative ASA in complex, relative ASA in monomer and total pair potentials. In the last column of the table, the prediction is presented as H (hot spot) or NH (non-hot spot). Background of the predicted hot spots is highlighted with red color. The prediction results as a text file and interface residue coordinates in PDB file format are also downloadable by the user. In this way, the results can be visualized in any visualization tool. Besides the downloadable files, overall complex, the interface residues and hot spots can be visualized interactively using the Jmol [194] applet window in the HotPoint server.

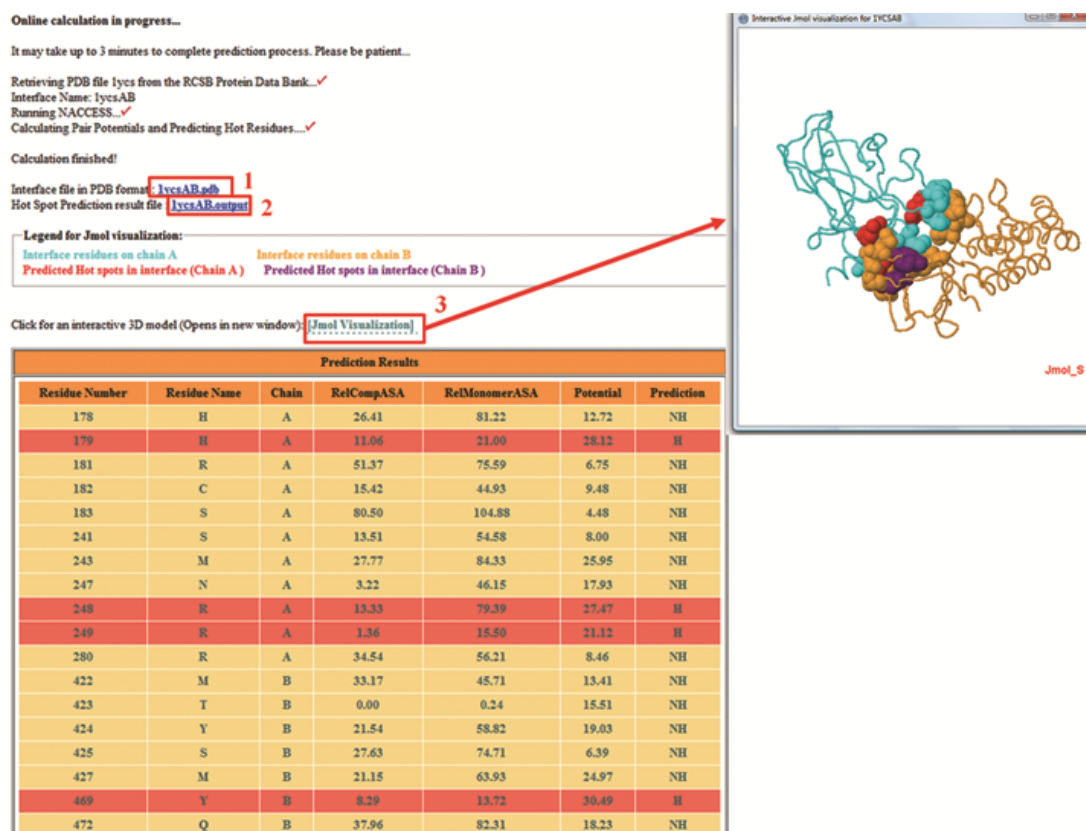


Figure 3.6 The output page of HotPoint for the p53 DNA binding domain/53BP2 protein complex (pdb:1ycs, chain A and B). Interface residues of this complex are shown in the table with hot spot predictions. 1. The coordinates of interface residues can be downloaded, 2. Hot spot prediction results are also downloadable, 3. Interface with predicted hot spots can be visualized by Jmol.

3.3.3 An independent case study: Interleukin-2 and its receptor complex

Interleukin-2 (IL-2) is a cytokine immune system signaling molecule. IL-2 gets functional when it associates with the IL-2 receptor. To find the residues necessary for binding, several residues (K35, R38, M39, T41, F42, K43, F44, Y45, E62, P65, V69 and L72) on

IL-2 are mutated to alanine. Among these residues, F42, Y45 and E62 reduce binding affinity of IL-2 to its receptor more than 100 folds. Further, small inhibitor molecule SP4206 also targets these hot spots of the receptor [195].

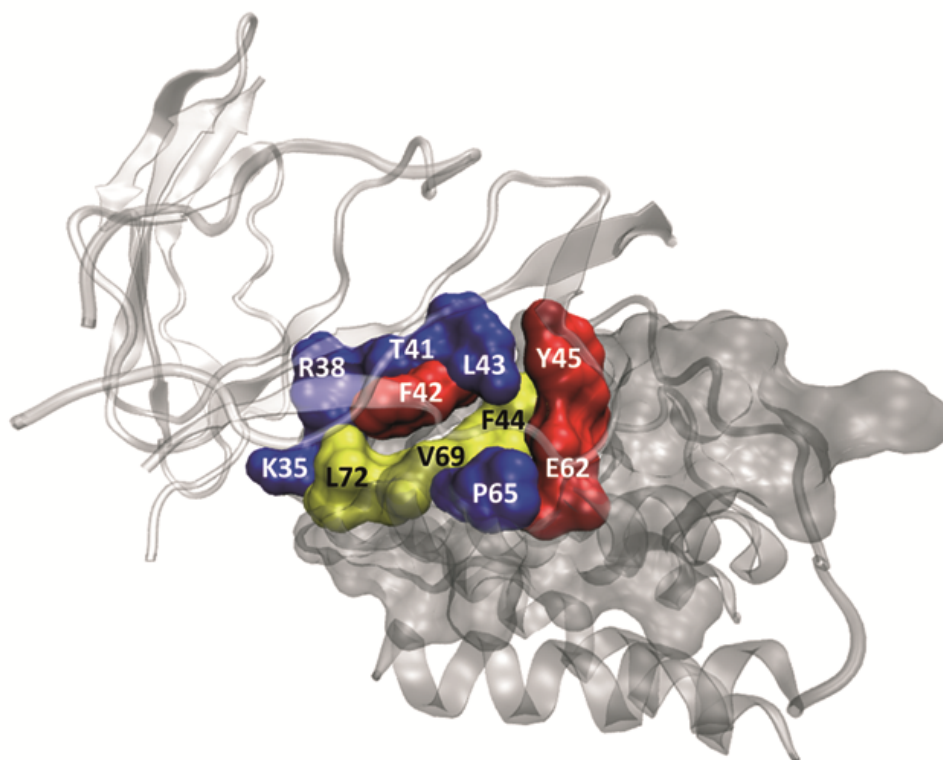


Figure 3.7 IL-2 receptor complex. The PDB code for this complex is 1z92. The red colored residues are correctly predicted hot spots. The blue colored ones are correctly predicted non-hot spots. The yellow colored residues represent non-hot spot residues which are incorrectly predicted as hot spots.

HotPoint predicts all three experimental hot spots (F42, Y45 and E62) correctly for the IL-2/IL-2 receptor complex (PDB code: 1z92, chain A is IL-2 and chain B is IL-2 receptor). According to our interface definition, M39 cannot be found in the interface residues. So, for the remaining eight residues, HotPoint labels five residues (K35, R38, T41, K43 and P65) as non-hot spot, correctly. However, three residues come as false positives (F44, V69 and L72) from HotPoint prediction. As a result, 8 out of 11 alanine mutations are correctly predicted. This protein complex is independent from the training and test sets. The predictions are illustrated in **Figure 3.7** in 3D using the output files obtained from HotPoint.

3.4 Concluding Remarks

In this chapter, a new efficient method to determine computational hot spots based on pair potentials and solvent accessibility of interface residues is presented. We note that solvent occlusion is a necessary factor to define a hot spot, but not sufficient itself. Conservation has not a significant effect in hot spot prediction as a single feature. Residue occlusions from solvent and pairwise potentials are found to be the main discriminative features in hot spot prediction. The predicted hot spots are observed to match with the experimental hot spots with an accuracy of 70%. We also compared our empirical methods to several machine learning methods and other hot spot prediction methods. Our method outperforms them with its high performance. Further, the model outperforms other existing approaches. This method is used to construct the HotPoint server which allows online calculation for all protein interfaces within practical running times. It tabulates residue level features and prediction results for a given protein complex which are also downloadable.

Chapter 4

ANALYSIS AND NETWORK REPRESENTATION OF HOT SPOTS IN PROTEIN INTERFACES USING MINIMUM CUT TREES

Protein interfaces can be represented as networks, where nodes are residues and edges are the contact between residues. Typical residue contact networks have a complex structure consisting of many residues (nodes) and interactions (edges). Hence, it is difficult to identify which of these residues and interactions are critical in terms of stability and function of protein complexes. We propose a novel approach to generate a simple, yet informative representation of residue contact networks of protein interfaces. We assign knowledge-based potentials as edge weights of the network. Our approach constructs a minimum cut tree (mincut tree) from the weighted residue contact network. We propose algorithms to extract hot spots and their organization from the mincut tree. Based on the minimum cut/maximum-flow theorem, the residues identified by our approach points out the residues with maximal energetic flow. Mutating such residues would change the energy flow significantly. In this method, the residue contributing to several mincuts is the most important one in terms of energy. Computational tests on a nonredundant dataset of protein complexes, having experimental mutation data, indicate that the most connected residue in the mincut tree generally corresponds to an experimental hot spot and other critical residues are observed to form a subtree. This method does not focus only on hot spot prediction and it cannot identify all hot spots, rather it is used for the analysis of the organization of interface residues and the connection between residues. Further, we show how to use mincut trees to cluster residues corresponding to hot regions in protein interfaces.

4.1 Methodology

4.1.1 Dataset of Experimental Hotspots

Experimentally, a hot spot can be identified by evaluating the change in binding free energy upon mutating it to an alanine residue [42]. Alanine Scanning Energetics Database

(ASEdb) is an information source for the hotspots obtained via alanine scanning mutagenesis experiments [96]. Another database, namely the Binding Interface Database (BID), contains experimentally verified hot spots in interfaces collected from the literature [97]. In this work, we use the protein complexes deposited in these two databases. For the complexes in ASEdb, the residues whose change in binding free energy is at least 2.0 kcal/mol are considered as hot spots. For the complexes in BID, the residues whose interaction is “strong” are considered as hot spots. Thus, totally a non-redundant set of 38 protein complexes are examined.

4.1.2 Construction of Weighted Residue Contact Graph and Minimum Cut Tree of a Protein Complex

An undirected weighted residue contact graph $G(N,E)$ consists of a node set N and an edge set E , with positive weights w_e , for all $e \in E$. In this graph, nodes represent interface residues and edges between them represent the contacts between pairs of residues. Two residues, one from each chain, are in contact if the distance between any two atoms belonging to two residues is smaller than the sum of their van der Waals radii plus a 0.5 Å tolerance. Also, two residues, within one chain, are in contact if the distance between the C^α atoms of these residues is smaller than 6 Å.

The weights of the edges in the graph are obtained from knowledge-based solvent-mediated potentials derived by Keskin *et al.* in 1998 [151], which are in good agreement with the residue frequencies obtained in a recent work [196]. The knowledge-based potentials have been shown to be useful in many threading, folding and binding problems [150, 179, 180]. These potentials represent the interaction parameters between two residues in native proteins. A practical way to obtain these potentials is to extract them from frequencies of contacts between different residues in proteins with known three-dimensional (3D) structures [183]. We provide 210 distinct potentials (all possible pairs of 20 different aminoacids) in RT unit (R universal gas constant, T is temperature) for contacting residue pairs in the Supplementary Material. All of the entries in this matrix are negative valued. In the residue contact network, the absolute value of the corresponding entry in the pair potential matrix is used as the edge weight (w_e).

A cut in a connected graph is defined by a partition of the node set into two sets, and consists of all edges that have one endpoint in each partition. Clearly, the removal of the

cut disconnects the graph. The weight of a cut is the sum of the weights of the edges crossing the cut. For $s, t \in N$, an s - t cut is defined as a cut which puts s and t into different node sets of the partition. A *minimum weight s - t cut* (min s - t cut) is a subset of edges with minimum total weight that separates the network into at least two disconnected sets of nodes. The problem of finding a min s - t cut can be efficiently solved using a maximum flow algorithm [197]. In the residue contact graph, the minimum weight cut between two residues illustrates the minimum total contact potential to separate these two residues into two disconnected subgraphs. Furthermore, min s - t cuts for all pairs of nodes can be represented by a *minimum cut (mincut) tree* in a compact way so that both the weight of a min s - t cut in the graph and the corresponding partition is the same in the tree. Gomory and Hu showed that a mincut tree can be computed using only $n-1$ min s - t cut computations (that finds the maximum flow from s to t), where n is the number of nodes [198]. To construct a mincut tree, G should be a connected network. If this is not the case, we take the largest connected component in G and perform our calculations on this graph. The algorithm to construct a mincut tree can be found in [198], and alternatively in [199, 200]. Here, we only demonstrate it with a simple example shown in **Figure 4.1**. First, all nodes are considered as a single node. Then, two nodes of N are picked to initialize the algorithm (here, nodes 1 and 2) and the minimum weight cut that separates 1 and 2 is found to split the node set N into subsets $S_1 = \{1\}$ and $S_2 = \{2, 3, 4, 5\}$, and has total weight 5. Next, another pair of nodes (nodes 4 and 5) are picked from S_2 and a minimum weight 4-5 cut is found which has total weight 6. The algorithm continues until all subsets contain only one node. As a result, using $(n-1)$ mincut calculations; the mincut tree is constructed (shown in part B4 in **Figure 4.1**). We note that a mincut tree always exists for a connected graph but it does not need to be unique (due to the choice of nodes s, t in the s - t cut). However, we observed our procedure to be robust to different mincut trees in our computations.

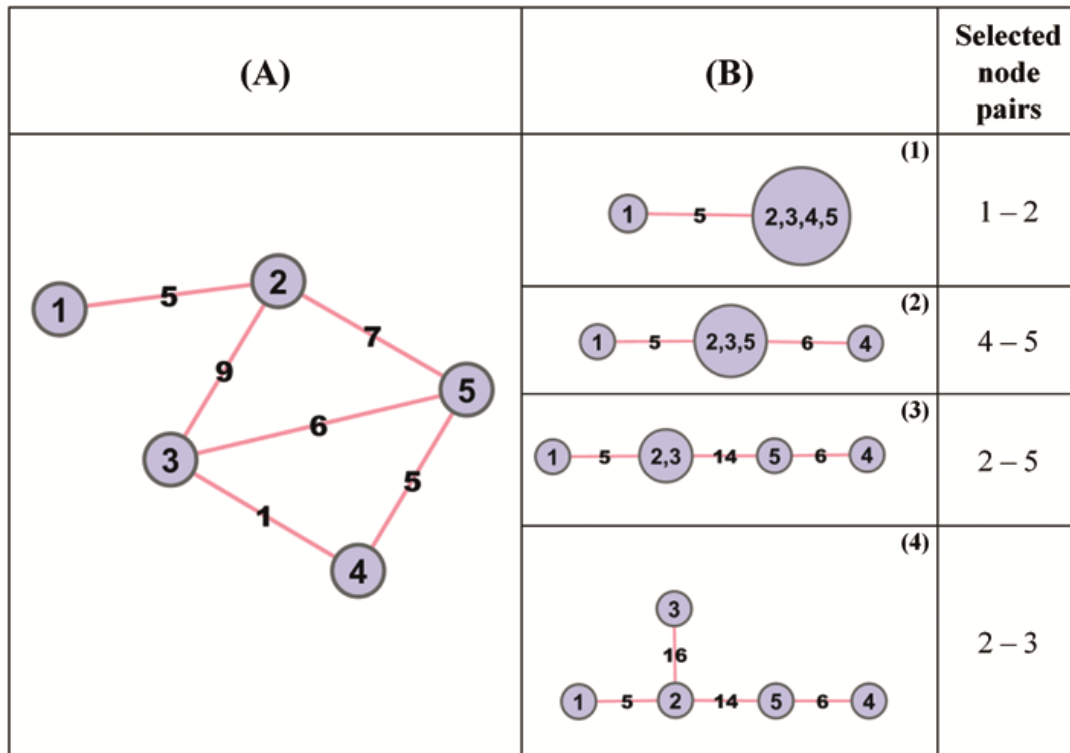


Figure 4.1 Illustration of construction steps of mincut tree with an example. Model network is shown in left panel (part A). (1) to (3) in part (B) are the intermediate steps to construct a mincut tree. The last column illustrates which node pairs are selected at each step. Here, nodes 1 and 2 are picked first, and a minimum cut 1-2 is found as 5. Then, from the remaining supernode, nodes 4 and 5 are selected and a mincut 4-5 is found as 6. (4) in part B is the resulting mincut tree of the network in part A.

4.1.3 Algorithm I: Determining the Critical Residue Subtree

To understand the interconnections among hotspots, we identify a *critical residue subtree* in the mincut tree. In the initial step of the algorithm, the seed node, S , is chosen as the node with maximum total weight on edges incident on it in the tree. Then, we look at the neighbors of this residue and we take a neighbor if it has at least a degree of δ and at least a weighted degree of W_t . We set the degree threshold as $\delta = 3$ and the weight threshold as

average weighted degree $W_t = \frac{\sum_{e \in E'} W_e}{n-1}$ in our computations, where W_e is the weight of the

edges in the mincut tree and n is the number of nodes in the graph. The weighted degree of a node i is $dW(i) = \sum_{e \in E'} W_e$, where E' is the set of edges incident on i . Then, we check an adjacent node j in T if it exists. If $dW(j) \geq W_t$ and $\text{degree}(j) \geq \delta$, then the next node is j and

it is added to the node list, L. Next, we go forward recursively by scanning the neighbors of the neighbors. At the end of the algorithm, if we cannot find any adjacent node passing the thresholds, then we output the node list, L, which corresponds to a subtree of the mincut tree. The simple steps of the algorithm are shown in **Algorithm 4.1**.

Algorithm 4.1 Critical subtree extraction algorithm.

Input: $G(N,E) \leftarrow$ a weighted undirected graph with weights w_e

Output: L, a list of nodes corresponding to a subtree in the mincut tree

$T \leftarrow$ mincut tree of G

$\delta \leftarrow 3$, degree threshold

$W_t \leftarrow \frac{\sum_{e=1}^{n-1} W_e}{n-1}$, weighted degree threshold

S \leftarrow the node with maximum weighted degree

L \leftarrow (S)

K \leftarrow (S)

while K $\neq \emptyset$

 i \leftarrow remove first node from K

 for all neighbors j of t

 if $dW(j) \geq W_t$ and $\text{degree}(j) \geq \delta$ and j is not in L

 append j to K

 append j to L

 end if

 end for

end while

return L

4.1.4 Algorithm II: Iterative Clustering of the Interface Residues

For clustering of the interface residues, we apply the iterative clustering algorithm in the work of Mitrofanova *et al.* to our problem [201]. Mitrofanova *et al.* use this algorithm to cluster unweighted network of protein-protein interactions in yeast; in this way, they aim to identify protein complexes. In our work, our purpose is to generate residue clusters in protein interfaces, to see how residues are separated from each other along the iterations and to identify the relation between hot regions. For this purpose, we construct the bipartite residue graph of the interfaces. A *bipartite graph* is defined as a graph whose nodes can be divided into two disjoint sets such that every edge is between two nodes, one from each set. In the bipartite residue graph, $G(N,E)$ where $N = U \cup V$, U represents the set of contacting residues from one chain, whereas V represents the set of contacting residues from the other chain. So, only inter-chain contacts are represented in $G(N,E)$. We assign

edge weights, w_e , as the absolute value of statistical residue contact potentials, as before. To find residue clusters, we follow an iterative procedure. After the construction of a mincut tree of the graph, we find the minimum value of the edge weights in the mincut tree (W_{\min}) and remove the edges whose weight is equal to W_{\min} from the mincut tree. Removal of an edge in the mincut tree corresponds to removal of a cut set from the residue contact network; in other words, separating the original network into at least two disjoint sets. At the i^{th} iteration, the set of the connected components is represented as $G_{\text{sub}}^i = \{G^{i,1}, G^{i,2}, \dots, G^{i,j}\}$ and the set of the subtrees is represented as $T_{\text{sub}}^i = \{T^{i,1}, T^{i,2}, \dots, T^{i,j}\}$ where j is the number of the subnetworks and subtrees. Next, we find the minimum edge weight for each tree $W_{\min}^i = \{W^{i,1}, W^{i,2}, \dots, W^{i,j}\}$ and remove all minimum weight edges from the mincut trees in T_{sub}^i . Then, we reconstruct the contact network for all of the remaining sub-trees and find mincut tree for each network. This iterative procedure continues until all subtrees have at least k connected residues where $k = 5$ in our computations. The simple steps of the algorithm are shown in **Algorithm 4.2**.

Algorithm 4.2 Clustering of the interface residues using a mincut tree.

Input: a weighted undirected bipartite graph, $G(N,E)$ with the edge weights w_e
Output: Node sets at the end of each iteration
Construct a mincut tree T of G
Find the minimum edge weight (W_{\min}) and remove the edges whose weight is equal to W_{\min} from T
while the size of each subgraph in $G_{\text{sub}}^i \geq k$ where $k = 5$
 $T_{\text{sub}}^i = \{T^{i,1}, T^{i,2}, \dots, T^{i,j}\}$ is the set of subtrees after removal of the edges
 $G_{\text{sub}}^i = \{G^{i,1}, G^{i,2}, \dots, G^{i,j}\}$ is the set of subnetworks after removal of the cut edges
 Construct the mincut tree of the subgraphs in $G_{\text{sub}}^i; T_{\text{sub}}^{i+1} = \{T^{i+1,1}, T^{i+1,2}, \dots, T^{i+1,j}\}$
 Find minimum edge weight of each tree in $T_{\text{sub}}^{i+1}; W_{\min}^{i+1} = \{W^{i+1,1}, W^{i+1,2}, \dots, W^{i+1,j}\}$
 Remove all minimum weight edges from the mincut trees in T_{sub}^{i+1}
 $i = i+1$
end while
return the set of subgraphs, G_{sub}^{i-1}

4.2 Results

Significant interactions between residues are represented by highly weighted edges using pairwise contact potentials. A mincut tree represents the weakest connections with minimum absolute contact energy in a compact structure. Thus, the complex structure of the contact network of the residues containing n nodes and m edges is simplified and

summarized by a tree with $n-1$ edges. Hence, residue contacts and closely related parts of the network can be interpreted and visualized easily.

As an example, the Erythropoietin (EPO) receptor and EPO mimetic peptide complex is analyzed. EPO is a hormone participating in the regulation of proliferation and differentiation of immature erythroid cells. EPO mimetic peptide (EMP1) functions as a mimetic of EPO. There is a competition between EMP1 (pdbID:1ebp, chainC) and EPO to bind the EPOR (pdbID:1ebp, chainA). Despite the unrelated sequences of EMP1 and EPO, both can bind to EPOR and stimulate biological activity [191]. Their interface region (1ebpAC) is shown in **Figure 4.2** both in structural representation (part A) and in graph representation (part B) (the edge weights are not shown in the figure.) The nodes in part B are colored according to the chains, and also the edges are colored according to the type of the contact (inter-chain or intra-chain). Blue colored edges are the inter-chain contacts and the red edges are intra-chain contacts. The mincut tree constructed for this residue network is illustrated in **Figure 4.2** (part C). When we compare the network in part B and the tree in part C visually, we observe two advantages of the latter; i. it is much easier to visually inspect the organization of the residues in the interface, ii. it is more informative. When we check the most connected node in the mincut tree, we notice that this node corresponds to an experimental hotspot. Other important residues are consecutive in a path in the mincut tree. Experimental data in Binding Interface Database (BID) [97] indicate that residues 93A, 150A, 205A and 13C are hot residues in the binding of EPO receptor and EPO mimetic peptide. In mincut tree of 1ebpAC interface, the most connected node is 150A. When we concentrate on the details of this analysis further, we observe that other experimentally determined key residues are sequenced in a subtree in this mincut tree.

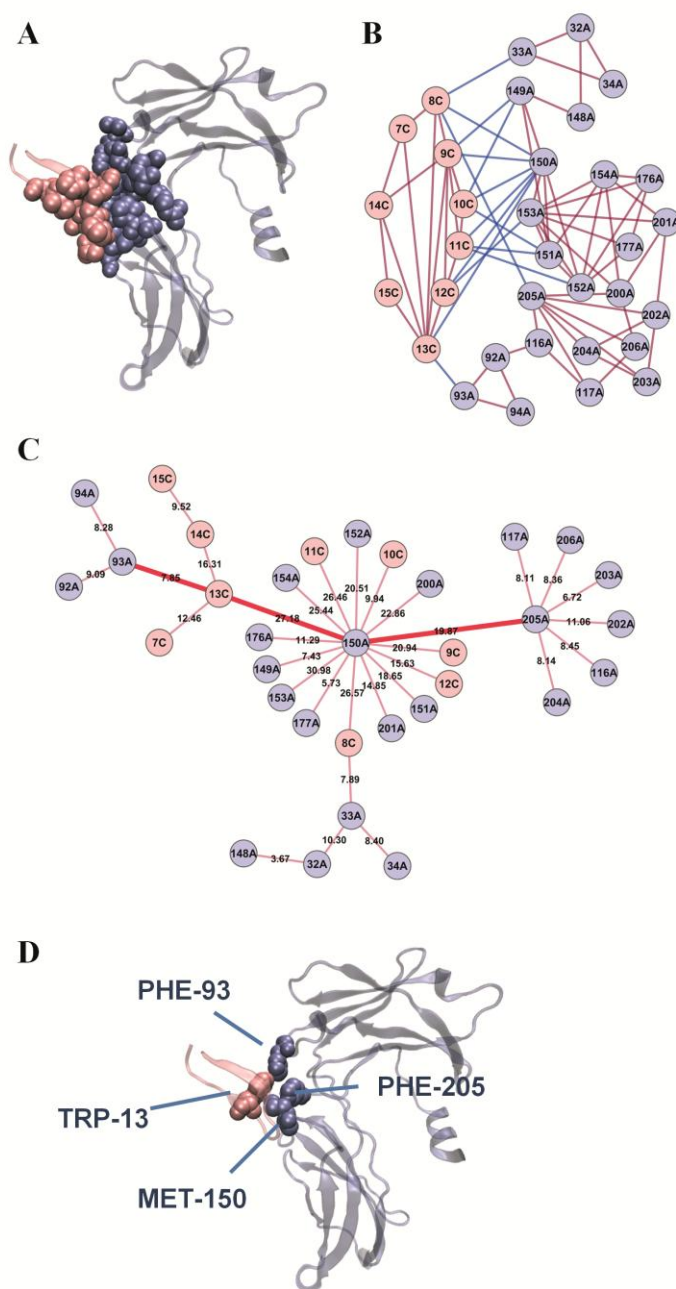


Figure 4.2 Analysis of the Erythropoietin (EPO) receptor and EPO mimetic peptide complex with mincut tree. (A) An example of protein interface and (B) its residue interaction network. (C) Constructed mincut tree of the 1ebpAC interface. Nodes are colored according to the chains. (D) Mapping of the generated subtree on the 3D structure of 1ebpAC interface.

The red arrows in **Figure 4.2** show the subtree where experimental hot spots are located. The hot residues (205A, 150A, 93A, 13C) form a path in the resulting mincut tree. This is an indication of the communication between the hotspots and also shows how

information in one chain can be transmitted to another chain. Among the hot spots, 150A forms interaction with six residues in chain C and four residues in chain A. So, it is expected that this residue should be critical in binding. However, each of 205A and 93A interacts with a single residue; 13C interacts with two residues. The rest of their interactions are formed within the chain they are involved in. Therefore, it is not clear that these residues are hot spots just by looking at the interaction network shown in **Figure 4.2B**. **Algorithm 4.1** successfully points out these residues by locating them along the same subtree of the mincut tree.

4.2.1 Analyzing Mincut Trees for Other Protein Complexes

We further construct a mincut tree for 38 complexes for which experimental hot spot information is available. We noticed a consistent trend that the most connected node in the mincut tree usually corresponds to an experimental hot spot. With this method, besides the visual compactness, critical residues can be identified as well. In **Table 4.1**, the most connected node in the mincut tree is given for 38 complexes. In this table, the most connected residue corresponds to an experimental hot spot in 20 out of 38 protein complexes. For the remaining complexes, these residues are either a close neighbor of an experimental hot spot (in 2 proteins) or they are computational hot spots predicted by other methods such as Hotpoint [99], KFC [111] (in 13 proteins). The largest weighted degree node in the residue contact network is the residue with largest energetic contribution. In 27 out of 38 cases, the residue with largest energy contribution corresponds also to the most connected node in the mincut tree. In the remaining 11 cases, the largest weighted degree node in the residue contact network does not correspond to the most connected residue in the mincut tree. In 6 out of 11 complexes, the mincut tree approach finds the hot spots; but in comparison, the largest weighted degree node in the residue contact network is a hot spot in only one complex. The advantage of the proposed approach is that besides the most connected node, a residue sub-tree is generated and several of the residues in this sub-tree either correspond to other hot spots or they are closely related to the residue with maximal flow which may act cooperatively and form a continuous path in 3D. In a classical network representation, this information is hidden and the mincut tree approach brings it out. Furthermore, the original network is too crowded to visualize, whereas the mincut tree representation provides essential information in a simpler format.

Table 4.1 The most connected node in the mincut tree for several complexes.

Protein Complex	Interface Name	The most connected node
Ribonuclease inhibitor – angiogenin complex	1a4yAB	318A – TRP
Growth hormone/receptor complex	1a22AB	179A–ILE [§]
Immunoglobulin heavy chain – tissue factor complex	1ahwBC	156C – TYR
Barnase – Barstar Complex	1brsAD	39D – ASP
E9 DNase – Im9 Complex	1bxiAB	33A – LEU
Chymotrypsin – BPTI Complex	1cbwCD	15D – LYS
Soluble tissue factor complex	1danTU	19T – PHE [§]
Immunoglobulin heavy chain – peptide complex	1dn2AE	252A – MET [§]
Fv-Fv idiotope-anti-idiotope complex	1dvfBD	98D – TYR
Cell division protein ZipA/ FtsZ fragment complex	1f47AB	85B–PHE [§]
Immunoglobulin FC/Fragment B of protein A complex	1fc2CD	136C–LEU [§]
GP120/CD4 complex	1gc1GC	28C–TRP
Interferon gamma receptor/fab fragment complex	1jrhLI	92L–TRP
TEM1- β -lactamase- inhibitor complex	1jtgAB	142B – PHE
IGG1-kappa D1.3 FV complex	1vfbAB	36A–TYR [§]
Beta trypsin/inhibitor complex	2ptcEI	14I – CYS [§]
HyHEL-10 Fab heavy chain-lysozyme complex	3hfmHY	33H – TYR
HyHEL-10 Fab light chain-lysozyme complex	3hfmLY	20Y – TYR
Human Growth Factor – Receptor Complex	3hhrAB	182A – CYS
Calmodulin – Protein Kinase Complex	1cdIAE	810E – ILE
Numb Protein Complex	1ddmAB	199A – LEU
Ribonuclease inhibitor - ribonuclease A complex	1dfjEI	259I – TRP [§]
DES-GLA factor VIIA – peptide complex	1dvaHX	34H – LEU
Integrin – collagen complex	1dziAC	220A – LEU
EPO Receptor – EPO Mimetic Peptide Complex	1ebpAC	150A – MET
Bone morphogenetic protein-2/ receptor 1A complex	1es7AD	785D – PHE [§]
Blood coagulation factor VIIA/soluble tissue factor	1fakLT	70L – CYS [§]
IGG1-Protein G complex	1fccAC	27C – GLU
Mms2/Ubc13 heterodimer	1jatAB	8B – PHE
HslUV protease/chaperone complex	1g3iAG	443A – ILE
Nidogen-1 with IG3 complex	1gl4AB	429A – HIS
Beta catenin/APC complex	1jppBD	424B – LEU
Phagocyte NADPH Oxidase complex	1k4uSP	505S – ILE [§]
alphaL I domain in complex with ICAM-1	1mq8AB	204B – LEU
IkappaBalpha/NF-kappaB complex	1nfiBF	254B – VAL [§]
MazE/MazF Complex.	1ub4AC	458C – LEU
Numb PTB domain-peptide complex	2nmbAB	199A – LEU [§]
p53 oligomerization domain complex	3sakAC	23A – PHE

Bold residues are experimental hot spots.

[§] Identified as hot spot by other prediction methods such as Hotpoint [99], KFC [111].

One of the complexes in **Table 4.1** is the barnase-barstar complex (pdb ID: 1brs). Barnase (chain A) is a ribonuclease enzyme. Barstar (chain D) inhibits barnase by blocking its active site. In this way, barstar stops barnase to damage the synthesized RNA [202]. A

mincut tree is constructed for the barnase-barstar complex (see **Figure 4.3A**) and the most connected node, 39D, found to be an experimental hotspot.

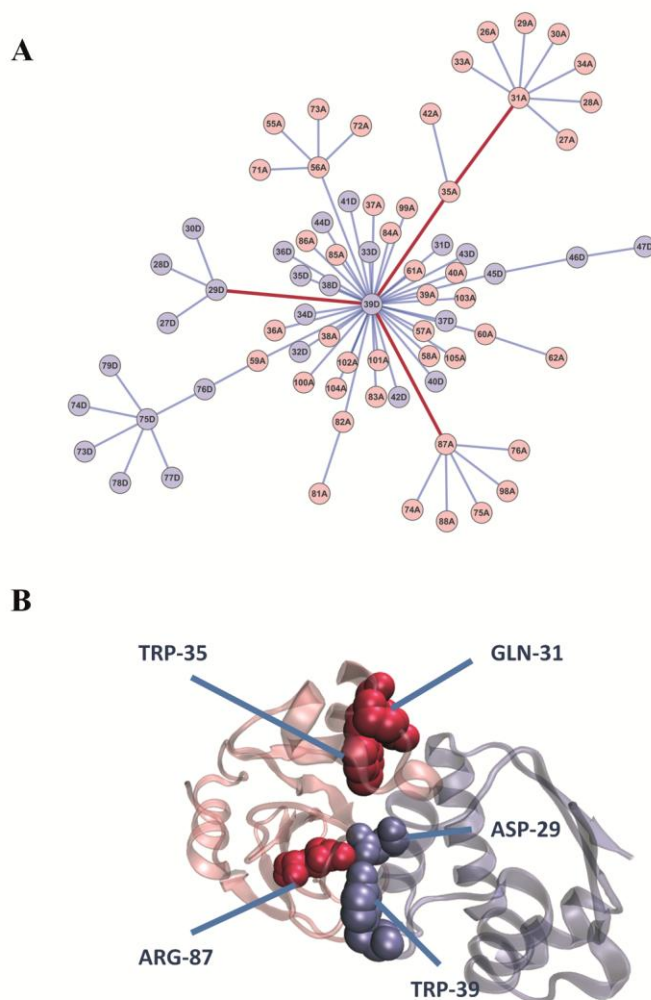


Figure 4.3 Analysis of the barnase/barstar complex with mincut tree (A) Mincut tree for the 1brsAD interface. Nodes are colored according to the chains. The red colored bold edges represent the subtree of critical residues. The critical residues in this subtree are “29D – 39D – 87A – 31A – 35A”. (B) Spatial illustration of this subtree on the protein complex.

Further, the generated subtree using **Algorithm 4.1** consists of the nodes 29D-39D-31A-35A-87A. Within this subtree, all three residues in the 29D – 39D – 87A path are experimental hot spots; they are located closely in the 3D structure and form a region (shown in **Figure 4.3B**). The mincut tree brings these three residues together. When we go from the complicated overall residue contact network to the mincut tree, we can examine and interpret the organization of the hot spots. To justify the connection between these five

residues, we take the single point and double point mutation information available for the barnase-barstar complex.[203] The observed change in binding free energy are as follows, $\Delta\Delta G_{29D} = 3.4$ kcal/mol, $\Delta\Delta G_{87A} = 5.5$ kcal/mol, $\Delta\Delta G_{39D} = 7.7$ kcal/mol, $\Delta\Delta G_{87A/39D} = 6.1$ kcal/mol, $\Delta\Delta G_{87A/29D} = 8.0$ kcal/mol. According to these energy values, the residues 87A and 39D are cooperative with each other which causes an energy difference of 7.1 kcal/mol (difference between 5.5+7.7 kcal/mol and 6.1 kcal/mol); on the other hand, simultaneous mutation of 87A and 29D causes a difference of 0.9 kcal/mol. Since mutation data related to the residues 31 and 35 in chain A are not known experimentally, we cannot comment about the relation between the distant residues 31A and 35A with the 29D, 39D and 87A. However, from literature, we found that the residues 31A and 35A have significant effect on folding of barnase.[204, 205] The effect of 31A is 1.1 kcal/mol [205]. The mutation on 35A decreases more than 70% of the fluorescence intensity of barnase [204]. So, their mutations may also have effect on 29D, 39D and 87A.

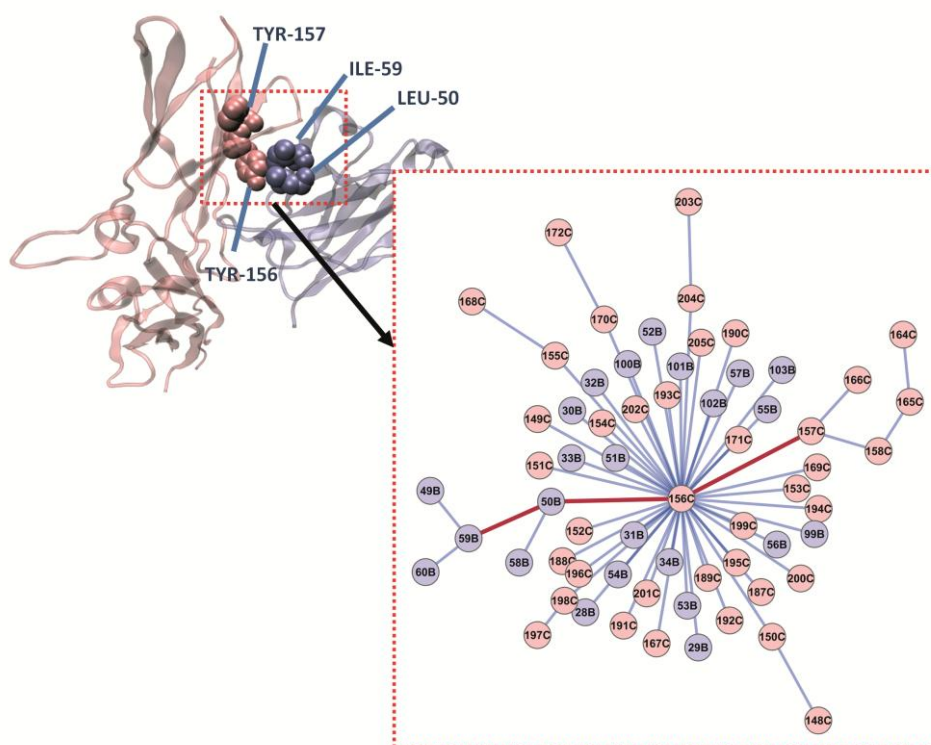


Figure 4.4 Mincut tree for the 1ahwBC interface and illustration of the extracted subtree on the 3D structure. Nodes are colored according to the chains. The red colored bold edges represent the subtree. This subtree forms a continuous residue path in 3D.

Another example is the immunoglobulin complex with tissue factor (pdbID:1ahw, between chain B and chain C) [206]. The most connected node in the mincut tree is 156C

which is also an experimental hot spot defined in Alanine Scanning Database (ASEdb) [96]. The subtree identified in the mincut tree is composed of the residues; *59B–50B–156C–157C* (see **Figure 4.4**). The hot spot (156C) is surrounded by other residues in this path as seen in the 3D picture of the immunoglobulin-tissue factor complex and they form a region in the binding site. These residues play possibly an important role in binding (For further examples, see **Table 4.1**).

4.2.2 Organization of Hot Regions

As discussed before, hotspots are clustered into hot regions,[2, 44, 47] and understanding the organization of hot spots and hot regions in protein binding sites is a major task for protein interaction prediction, as well as the design of therapeutic agents. We have two contradicting cases available from the literature. The first one is the TEM1- β -lactamase and its inhibitor protein (BLIP) complex which is analyzed by Schreiber and his co-workers [124]. They stated that distinct residue clusters are energetically additive, but the residues within the same cluster are highly cooperative. The other one is the TSST1 – hv-b2.1 complex whose distinct hot regions are energetically cooperative [89]. To analyze hot regions, we construct bipartite graphs of the residues. We apply the iterative clustering algorithm (**Algorithm 2**) to both examples, and analyze the correlation between the results. Here, our aim is to cluster the interface residues and to see which residues are important to bring clusters together along the iterations.

The clustering of the interface residues of TEM1-BLIP complex (pdb ID:1jtg, chain A and B, respectively), performed by Schreiber and his co-workers, divides the interface region into five clusters (namely C1, C2, C3, C4 and C5). Using multiple mutagenesis analysis of two clusters (C1 and C2), they stated that these two clusters are energetically independent of each other, but the intra-cluster connections are cooperative. When we construct the bipartite graph of the residue network, we see four independent subgraphs. Two of the subgraphs correspond to C3 and C5; they are not connected to the largest connected component. For mincut tree analysis we focus on the largest connected component. The largest connected component in the graph contains the residues in C1, C2 and C4. Using the mincut tree approach, we cluster the nodes and check the robustness of the clusters to deletion of edges in the minimum cut in the residue network iteratively as described in **Algorithm 4.2**. Here, we notice that 130A, 234A, 235A and 243A are

removed at the initial iteration steps (see **Figure 4.5**, part A) which corresponds to C1. All residues in C1, except 49B, are removed in the iterations which imply that the minimum cuts connecting these residues to the network are weak within this cluster. When we continue the iterations, we end up with two clusters for the largest connected component which correspond to C2 and C4, respectively. Further, we notice that two distinct clusters are connected to each other using residue 49B. The contact between 216A-49B-237A is robust to several edge cuts until the 16th iteration (see **Figure 4.5**, part A) and this part is almost the strongest part of the mincut tree. We hypothesize that the information flow from one cluster to another passes through 49B. The importance of this residue is not obvious from the original residue interaction network. However, mincut tree shows the critical role of 49B by showing that it connects two individual residue clusters. This result is in correlation with the experimental mutation data presented by Schreiber and his co-workers. The change in binding free energy upon single point mutation of 49B is 7.5 kJ/mol. On the other hand, the change is 7.1 kJ/mol upon simultaneous mutation of 49B, 130A, 235A, 243A and 234A. Both mutations have almost the same effect on binding. Further, the effect of single point mutations of 130A, 243A and 234A (1.4, 5.3 and 4.3 kJ/mol, respectively) are not as large as the effect of 49B. Probably, the effect of 49B is dominant in simultaneous mutations and 49B is the most critical residue in C1. This residue also connects two other clusters and furthermore its connection with these clusters is robust to the edge removal. Thus, the mincut tree (shown in **Figure 4.5**, part A) clearly suggests a link between C2 and C4 through the residue 49B. Here, we state that the mutation of 49B may lead to structural rearrangements in C2 and C4. When 49B is mutated to alanine, the connection between C2 and C4 might be broken according to the mincut tree. Thus, we suggest for an experimental analysis of the clusters C2 and C4 using multiple mutations. Analysis of the cooperativity between these two distinct clusters may be investigated further.

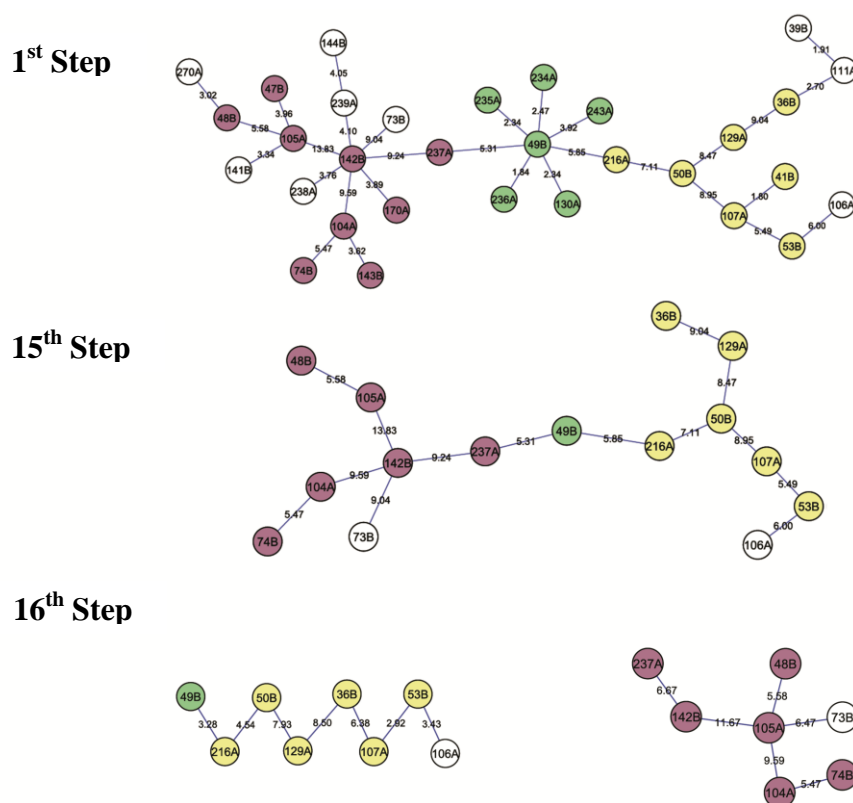


Figure 4.5 Mincut tree of the bipartite graph of the TEM1 – BLIP complex at first, 15th and 16th steps of the iteration. Each color represents a cluster and the coloring scheme is the same as in Schreiber’s work.

To check the cooperativity between distinct hot regions, Moza *et al.* [89] analyzed the interaction between hvb2.1 and TSST1. In the hvb2.1-TSST1 complex (pdb ID: 2ij0), there are two distinct hot regions on the hvb2.1 (chain E) surface which are 52E and 53E in CDR2 loop and 61E and 62E in FR3. They stated that although these two regions are distant to each other by more than 20Å, they are highly cooperative. When we apply **Algorithm 4.2** to the hvb2.1-TSST1 complex, we observe that these two hot regions are linked by the residue 17A on the surface of TSST1 (chain A) and this linkage is easily distinguished using the mincut tree (see **Figure 4.6**). Although these two hot regions are spatially far away from each other, they are located in the loop regions, at the flexible parts of the hvb2.1 and they are connected via residue 17A on the partner protein TSST1. So, their cooperativity is expected when we analyze the mincut tree. Another observation is that, when we apply iterative clustering algorithm, we obtain only one cluster and along the

iterations residues are separated one-by-one from the main mincut tree. This result shows a strong connection between two hot regions in the TSST1-hvb2.1 complex.

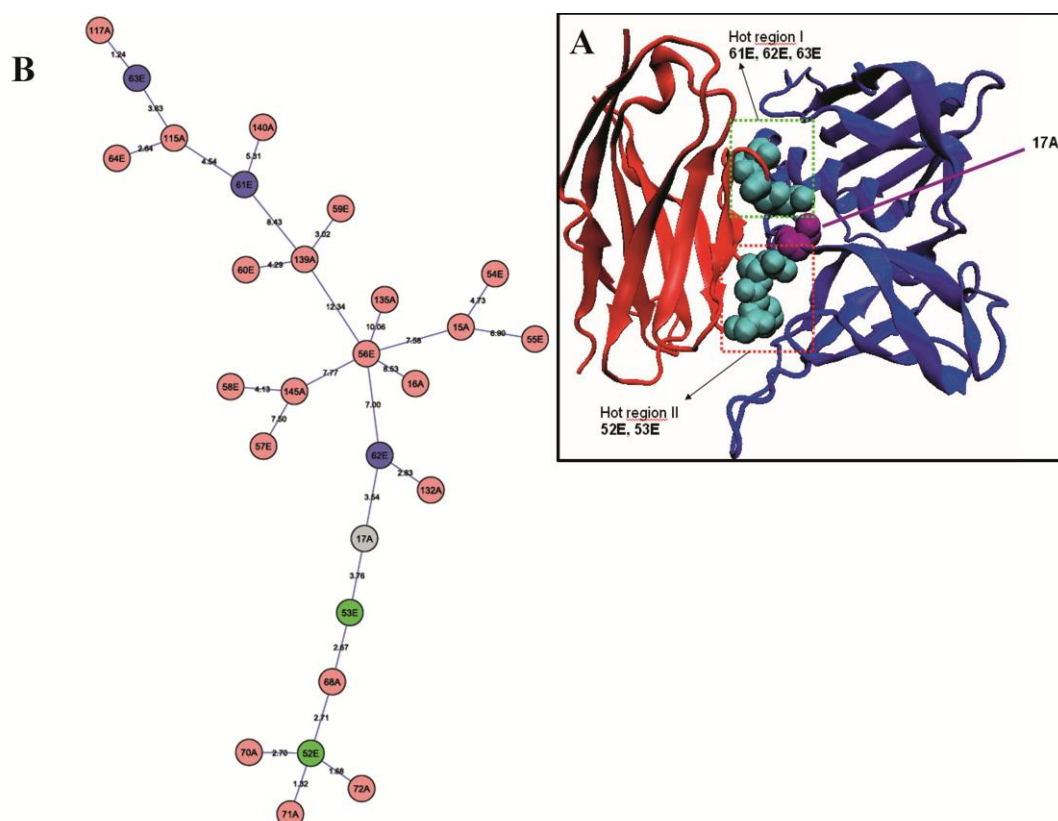


Figure 4.6 The hvb2.1-TSST1 complex (A) Three dimensional structure (B) Mincut tree of the bipartite graph.

4.3 Concluding Remarks

Proteins interact through their binding sites. Several graph based algorithms are used to characterize and analyze protein-protein interactions. In this work, we use minimum cut trees to visualize and analyze residue contact networks compactly. Edges in the contact network are weighted according to an energy function, namely knowledge-based potentials. Mincut tree representation highlights some central residues at first glance, which cannot be distinguished in classical network representation visually. This information provides us the most important node and the critical paths within the interface region. As the most connected residue in the tree usually corresponds to an experimental hot spot, and hot spots are sequenced along paths on the tree, we give an algorithm to find a subtree containing hot spot paths.

We also analyzed the dependency of the distinct residue clusters using some known protein complexes such as TEM1-BLIP and hvb2.1-TSST1. We found some critical traces to explain the cooperativity between distinct residue clusters using a clustering algorithm that runs on a mincut tree. As a future direction, one may also analyze how hotspots are communicating with each other and how information in one chain is transmitted to another chain using the proposed algorithms.

Briefly, our new approach is useful for basic biological cases at the molecular level. A mincut tree simplifies the complex residue network between two interacting proteins visually. Using this tree, residue clusters and the critical residues in binding of two proteins can be identified computationally by efficient algorithms.

Chapter 5

MULTI-SCALE COMBINATORIAL DOCKING OF THE PROTEOME FOR FUNCTIONAL PREDICTIONS

Construction of the structural protein interaction network is of crucial importance since it is a prerequisite for understanding how the proteome and thus the cell function. Yet, predicting, on the proteome scale, *which* proteins interact and *how* they interact is a daunting task. As reviewed in Chapter 2, structural predictions of protein interactions are frequently carried out via ‘docking’. However, in the absence of additional biochemical data, docking is challenging on the proteome scale because there are many favorable ways for proteins to interact. An alternative strategy is knowledge-based, using a protein-protein interface dataset. This suggests that using structural alignment of each side of known interfaces against the entire surfaces of all monomers can predict protein associations: a protein whose surface matches one side of the interface can bind a protein whose surface matches the complementary side. Yet, on their own knowledge-based methods may not be sufficient for proteome modeling because they disregard flexibility and energetics.

Here, for the first time, we combine the two approaches on a large scale. We integrate large scale rigid body structural alignments with flexible refinement and energy minimization of predicted protein-protein interactions. This leads to a powerful combinatorial strategy to predict functional associations in the proteome. Different than previous methods, the structural alignments are restricted to interface regions rather than of entire single chain folds. This strategy, which is based on our earlier observation that the number of interface motifs is restricted in nature [10, 20, 46, 52], allows broad proteome comparisons since it neglects the requirement that the folds be similar [165]. At the same time, from the technical standpoint, it necessitates a technique which is able to compare motifs consisting of discontinuous protein fragments, which can have different order in the chain [207, 208]. In our strategy, the two complementary parts of the interface of known complexes serve as templates in searches for structurally similar target protein surfaces. Predicted complexes with matching surfaces undergo flexible refinement using a new

efficient docking tool [144]. Finally, energy assessments make the prediction more physical and provide a way to score the modeled complexes. This powerful approach can be utilized to analyze any pathway, as long as structures are available. To demonstrate its usefulness and predictive power, we simulate the molecular interaction map of the p53 pathway. In particular we focus on the nucleotide excision repair (NER) and the cyclin-dependent kinase subsystems. Structural modeling of the partners of NF κ B, p27 and Skp2 proteins from known interfaces in the p53 pathway are presented as case studies.

5.1 Methods

The rationale of this method is as follows: if complementary partners of a known protein interface are similar to surface regions of *any* two monomers, these two proteins can interact with each other *via* these regions. The method utilizes structural and evolutionary similarity. No sequence similarity is used. **Figure 5.1** presents a flowchart of the algorithm.

5.1.1 Datasets

The method employs two types of datasets: template dataset which is a subset of a nonredundant set of known protein interfaces derived from the PDB [10]; and target dataset which contains the structures of protein chains in the target pathway. To generate the template sets, we utilized the structurally non-redundant interface dataset containing 49,512 interfaces clustered into 8,205 structurally distinct interface clusters. To extract interfaces, two types of residues in each chain are defined: interacting and nearby. If the distance between any two atoms belonging to two residues, one from each chain, is less than the sum of their van der Waals radii plus a 0.5 Å tolerance, these two residues are defined as 'interacting'; if the distance between the C $^{\alpha}$ of a non-interacting residue and an interacting residue in the same chain is under 6 Å, the non-interacting residue is flagged as a 'nearby' residue. Nearby residues are important for the interface architecture [10, 20, 52]. Two different template sets are used: (i) a structurally non-redundant template dataset composed of hetero-dimers (1,037 interfaces); and (ii) the p53 pathway template set, composed of all currently known and available interfaces in this pathway (59 interfaces). Interfaces hot spots, i.e. residues contributing more to the binding energy, are identified using the Hotpoint web server [209]. The molecular interaction map (MIM) [210] is used to obtain the target protein set for the p53 pathway. Several proteins in the molecular

interaction map (MIM) do not have complete structures. For example, the human DNA excision protein ERCC1 has 297 residues in full length; however, the available structures are for residues 96-227 (PDB: 2a1i, chain A) and 220-297 (1z00, chain A). Both fragments are considered in the target set. In this pathway, 77 proteins have structural information but when considering all protein fragments, the number of chains increases to 112.

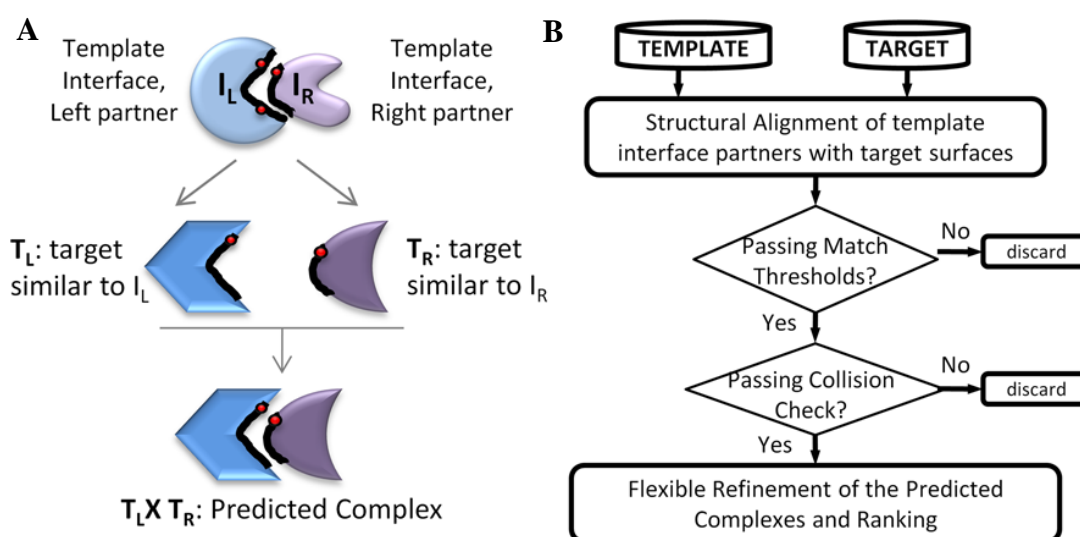


Figure 5.1 The concept figure of the prediction algorithm. (A) Schematic illustration of the concept of the prediction algorithm. If complementary partners (I_L and I_R) of a template interface are similar to surface regions of *any* two targets (T_L and T_R), these two targets can interact with each other *via* these regions. The red points are hot spots. These incorporate evolutionary information into the matching. (B) The flowchart of the algorithm. There are two datasets in the algorithm: template dataset and target dataset. Firstly, the surface of the proteins in the target dataset are extracted. Then, each partner of the template interface is aligned with the target surfaces. If the match passes the residue and hotspot matching thresholds, these targets are transformed on the template interface. If there are colliding residues between the two partner targets, the putative complexes are eliminated. Else, they pass to the flexible refinement stage where side chains are optimized, global energy of the predicted complexes are calculated and they are ranked based on the energy.

5.1.2 Prediction Phases

The prediction algorithm is composed of four consecutive phases. Proteins interact *via* surface residues; thus, in the *initial phase* the surface regions of target proteins are extracted. Protein ‘surface’ is the shell around the entire monomer surface. The surface regions in the target dataset are extracted based on the relative accessible surface area of the residues. If the relative accessibility (accessible surface area vs that of the residue in an

extended conformation) of a residue is more than 15%, it is labeled surface residue. Residues in its neighborhood are used to provide the structural scaffolds of the protein surfaces. A residue is defined as 'nearby' if the distance between its C^α atom and that of a surface residue is under 5.0 Å. Nearby residues are important in the structural alignment phase [10, 20, 52, 163].

In the *second phase*, using structural alignments, the occurrences of each side of the known interface on monomer surface regions are sought. Specifically, each interface in the template dataset is split into its constituent chains. Using the MultiProt engine [208] our method searches whether the complementary partners of a template interface are structurally similar to any region on target surfaces. MultiProt searches for spatial residue similarities disregarding the order of the residues on the chain. Because template interfaces and target surfaces may not consist of contiguous chain fragments, MultiProt is particularly appropriate for this task. Geometry and residue type (hydrophobic, hydrophilic, aromatic or glycine) are considered in the structural alignment. 40% of the residues of template chains should match the target surfaces to pass to the next step. This threshold is 60% for template chains containing less than 50 residues. At least one hot spot in each template partner should correctly match with the target surface. Hot spot filtering incorporates evolutionary similarity between target surface and template interface in addition to structural similarity.

In the *third phase*, the two chains whose surface regions are similar to the two parts of the template interface are transformed onto this template and the solution is assessed: if the two partners present many spatially-colliding residues the match is eliminated. Side chain clashes are not considered at this stage. A more rigorous refinement is performed at the last phase of the algorithm.

The *last phase* involves flexible refinement of the rigid docking solutions of MultiProt to resolve steric clashes, especially of side chains, and ranking putative complexes by the global energy. For flexible refinement, we use FiberDock [144] which considers both side chain and backbone flexibility. For side-chain flexibility it uses a rotamer library and finds optimum combination of rotamers with the lowest total energy. Here, restricted side chain optimization is performed where only clashing interface residues are assumed to be movable. The backbone flexibility is modeled by normal modes. In our computations, we consider the first 50 modes of each protein in the putative complex. 20% of the clashes

between the side-chain atoms are allowed. Finally, FiberDock calculates energies, and the predicted protein complexes are ranked according to the calculated energies. In this way, the geometric complementarity is combined with docking procedures which makes the method more physical.

5.1.3 Validation Procedure

88 rigid body test cases (from 165 protein chains) in Docking Benchmark 3.0 [137] are used for validation of the method. The benchmark contains 28 enzyme/inhibitor, 21 antibody/antigen and 39 other type of complexes. For the benchmarking, two template sets are used: i) an optimal template set which is extracted from the bound states of these proteins (88 interfaces). Here, a template is a discontinuous sequence subset of the target. Using benchmark templates, we first check whether with an optimal template set the method can find all ‘true’ solutions and no ‘false’ ones. ii) A more diverse template dataset (1,037 interfaces) which is utilized to model the interactions. With the second template set, we aim to see how many interactions the method predicts. All possible pairs of the 165 target protein chains are searched on the templates and a 165 x 165 interaction matrix is constructed to see if the method distinguishes binders from non-binders.

5.2 Results

5.2.1 Validation of the Method

The structural alignment of template interface partners with target protein surfaces is independent of global chain homology and sequence order. At the validation stage our first aim is to examine how the method performs on an optimal template set. Each of the target protein surfaces is aligned with the partner chains of those 88 interfaces. The method is applied to all possible pairs (165x165) and a matrix of interacting pairs is generated. At the matching phase, correct binding regions are found for all 88 protein complexes, except one case which is an antibody/antigen complex. Further, these correct protein complex models are ranked first by FiberDock. Besides the 87 complex models, 243 protein complexes are also modeled at the end of this run. If all 165 nodes would interact with each other, there would be 13,530 edges in the network. Our algorithm gives only 243 extra interactions; 41 of them are modeled as antibody/antigen, 55 as enzyme and inhibitor/substrate complexes, 74 as one side antibody, and the remaining are between other types of complexes (for

details, see **Figure 5.2A**). Many of these extra interactions arise from antibodies. As an example for the extra interactions, the modeled complexes of bovine trypsin are illustrated. In addition to the soybean trypsin inhibitor (1ba7), our algorithm predicts that bovine trypsin (1qqu) can interact with Bowman-Birk inhibitor (1k9b:A), pancreatic secretory trypsin inhibitor (1hpt), bovine pancreatic trypsin inhibitor (9pti), CMTI-1 squash (1lu0:B) inhibitor and TDPI from tick (2uux) which are all trypsin inhibitors. Although the overall structures and sequences of the partner proteins are dissimilar, they can bind to the bovine trypsin on the same surface, and the energy-based rankings of these interactions are high (**Figure 5.3A**).

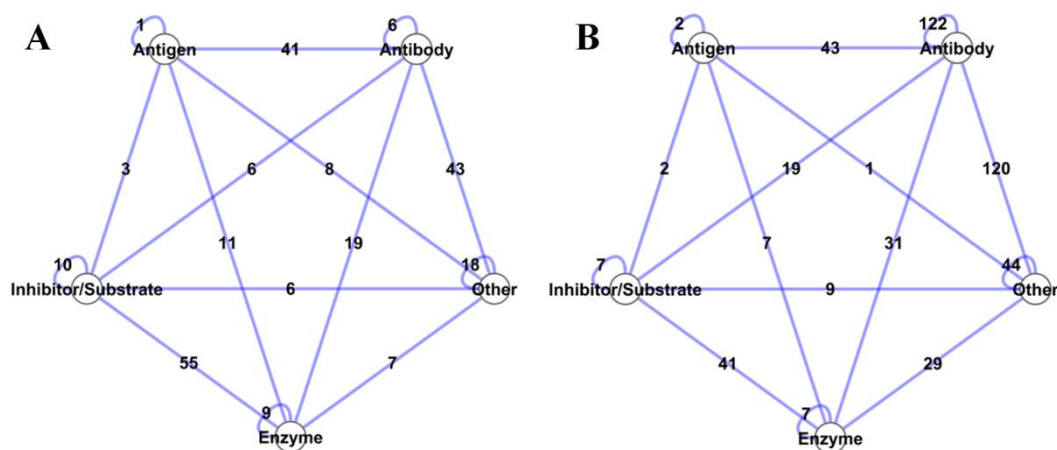


Figure 5.2 Illustration of the extra interactions predicted by our method (A) on the benchmark templates (88 interfaces) and (B) on the diverse template set (1,037 interfaces). The 165 protein chains are categorized into 5 classes in line with Docking Benchmark classification. Each node represents one class and each edge represents the number of predicted interactions between two classes.

We next compare against a more diverse template dataset containing 1,037 hetero-dimeric protein interfaces which are structurally non-redundant. This template set also contains 23 of the benchmark templates although its construction is not biased to the docking benchmark collection. The method is applied to all possible pairs (165x165) on 1,037 templates. 49 out of 88 are correctly found by the method. For the remaining ones, the algorithm can not find any similar templates. On the other hand, 484 extra interactions are found by the method of which 41 are enzyme/inhibitor complexes, 43 antibody/antigen complexes, and 292 one side antibody (for details, see **Figure 5.2B**). As an example for the

correctly predicted complexes, the interaction between bovine chymotrypsinogen (2cga) and pancreatic secretory trypsin inhibitor (1hpt) is found using the interface region in human leukocyte elastase/the turkey ovomucoid inhibitor complex (1ppfEI). The sequence similarity between elastase and chymotrypsinogen is 32%, between trypsin inhibitor and ovomucoid inhibitor is 28%. As illustrated in **Figure 5.3B**, the template interface matches well with target surfaces and the calculated global energy for this interaction is -51.10 kcal/mol. In another example, the interaction between falcipain 2 (2ghu) and cystatin (1cew) is modelled using the interface of papain/stefin B complex (1stfEI) as template. The sequence similarity between falcipain and papain is 32%, between cystatin and stefin B is 15%. This predicted complex gives a calculated global energy of -43.23 kcal/mol (**Figure 5.3C**). The subtilisin (2gkr) / ovomucoid (1scn) complex is modeled using the template interface between subtilisin/chymotrypsin inhibitor 2 (2sniEI). The sequence similarity between ovomucoid and chymotrypsin inhibitor is low (only 8%). Independent of their global fold, the structural similarity between the binding regions is very high.

To show the similarity between the binding regions and dissimilarity in the global folds, chymotrypsin inhibitor 2 is superimposed on ovomucoid and the predicted model is illustrated in **Figure 5.3D**. Our method successfully identifies the binding region on ovomucoid and correctly models the subtilisin/ovomucoid complex. Overall, the validation results show that as long as similar interfaces are available in the template set, our method efficiently finds the structurally similar regions on the target proteins, and following refinement these modeled protein complexes are ranked as first, based on energy.

Our method is knowledge-based; thus, if no similar interface exists in the template set, it cannot provide an interaction model. As in any homology, or motif-based prediction method, whether global or local, the outcome is a function of the template dataset.

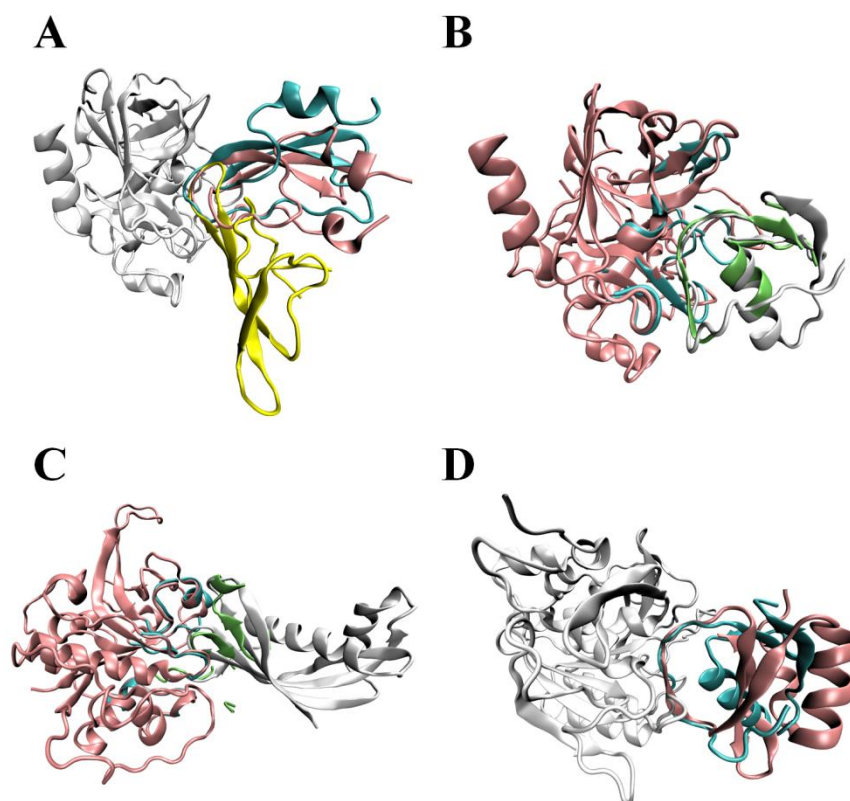


Figure 5.3 Some examples found in docking benchmark. (A) Bovine trypsin (colored white) can interact with several trypsin inhibitors using the same region and three of these partners are superimposed to show the structural similarity in their binding sites only. Although the overall structures of Bowman-Birk inhibitor (1k9b:A, yellow), bovine pancreatic trypsin inhibitor (9pti, pink), and TDPI from tick (2uux, cyan) are dissimilar, the binding region to bovine trypsin is structurally very conserved. (B) The interaction between bovine chymotrypsinogen (2cga, pink) and pancreatic secretory trypsin inhibitor (1hpt, white) is modelled on the interface region of human leukocyte elastase/the turkey ovomucoid inhibitor complex. Template interface (1ppfEI) is colored cyan and green to show the structural matching between target surfaces and template partners. (C) The falcipain 2 (2ghu, pink) and cystatin (1cew, white) interaction is modelled on the interface region of papain/stefin B complex (1stfEI). Template interface is colored cyan and green. (D) The subtilisin (2gkr, white)/ovomucoid (1scn, cyan) complex is modeled on the template subtilisin/chymotrypsin inhibitor 2 (2sniEI). Chymotrypsin inhibitor 2 (pink) is superimposed on ovomucoid to show the structural similarity in the interface region between target and template chains.

5.2.2 Comparison of Running Times with Docking

The running time of the method is completely dependent on the structural matching and rigid body refinement parts. Structural matching of one target surface with one template chain by MultiProt takes less than an alignment of two proteins because target surface and template chain are the residue subset of the overall proteins. It takes 1 sec on average

including transformation which varies depending on the template and target sizes. On the other hand, FiberDock refines a rigid body solution of Multiprot in an average time of 14 sec which varies depending on receptor size. The total running time of the refinement part linearly increases as a function of the number of rigid body solutions. For a target dataset composed of i.e. N proteins and a template set composed of $\sim 1,000$ interfaces, structural matching of all targets with all template interfaces takes $1,000 \times 2 \times N \times 1$ sec [$O(N)$]. On the other hand, rigid body docking of a protein pair takes i.e. with Zdock [211] 4 min on 16 processor, with PatchDock [212] less than 10 min on single processor. Running of all pairs of these N targets takes $(N \times (N-1) / 2) \times 10$ min [$O(N^2)$] in the best case.

In **Figure 5.4A**, the comparison of running times as a function of target dataset size is illustrated. As shown in this figure, for small number of target proteins, both methods have more or less similar running times. However, on large scale, things change and the knowledge-based method dramatically decreases the solution space and as a result the running times. As the target dataset increases, the difference between running times gets larger and the advantage of template-based method at large scale is obvious. Hence, with the fast growth of the PDB the number of distinct interface motifs will increase making such fast strategies increasingly popular and useful for the modeling of protein interactions. Also, the running time of our method is a function of template dataset size in addition to that of the target dataset. In **Figure 5.4B**, for 50 targets the running time of docking is compared to our template-based method. The figure shows that if the template dataset size would be composed of 15,000 interfaces, the running times of both methods would be same which is 15 fold larger than the current template set.

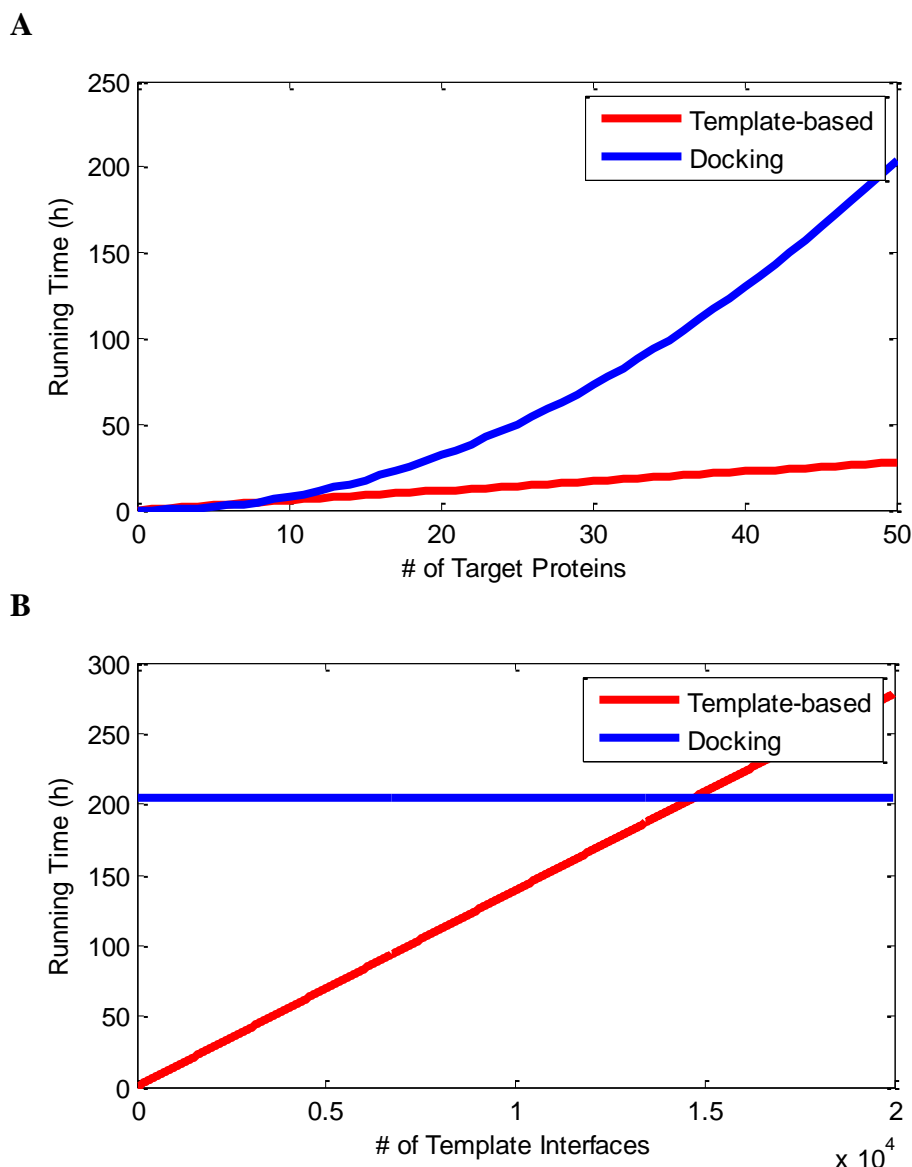


Figure 5.4 Comparison of running times of our template-based method with docking on large scale. (A) Running times are plotted as a function of the number of target proteins. The number of templates is equal to 1,000 for the template-based method (B) Running time of our template-based method is plotted as a function of the number of template interfaces for 50 target proteins. If there were 15,000 template interfaces, two methods would have same running times for 50 target proteins.

5.2.3 Structural Interaction Network of p53 Pathway

Following validation, we apply our multi scale combinatorial docking algorithm to the p53 pathway where interaction data are obtained from the human molecular interaction map (MIM) [210]. Additional interactions from other databases such as DIP [30], MINT [31], BIND [32] and IntAct [33] enrich this map. Overall, 328 interactions between 104 molecules are found from MIM and interaction databases. Among them, only for 25

interactions the structures of the complexes are available in the PDB. At this point, our method intervenes to complete the lacking network parts. By using the template interfaces in the p53 pathway with a default matching threshold, 108 interactions between 49 proteins are obtained. 46 of these 108 are known experimentally (detailed in **Table 5.1**). In this network, transcription factors such as E2F1-2-3, Max, Myc, Jun and Fos interconnect *via* multiple interactions. As expected, there is also a large number of interactions between cyclins and kinases. The template set containing 1037 interfaces gives just 53 interactions between 38 proteins with default thresholds, of which 24 interactions have experimental evidence. When the matching thresholds are relaxed by 10%, the template set in the p53 pathway gives 396 putative protein complexes between 68 protein chains of which 88 interactions are verified by interaction databases. Using the relaxed matching thresholds, 721 interactions between 71 proteins are found from the 1,037 template interfaces of which 110 interactions are verified. The results show that using strict matching thresholds give more reliable predictions, but also miss true positives. When thresholds are relaxed, true positive rate increases; however, false positives also increase.

Table 5.1 Number of predicted interactions and verification on the experimental data.

Template Dataset	# of predicted interactions ^a	# of verified interactions ^b	Extra interactions ^c	Total # of verified interactions
p53 templates (default)	108 (49)	30	16	46
1037 templates (default)	53 (38)	18	6	24
p53 templates (relaxed)	396 (68)	52	36	88
1037 templates (relaxed)	721 (71)	67	43	110

^a numbers in paranthesis represents the number of proteins; i.e. 108 interactions between 49 proteins.

^b Experimental interaction data for verification are obtained from the human molecular interaction map, DIP, MINT, BIND and IntAct.

^c For further evidence for the predicted interactions, we used the String search tool where we considered only experimental interactions and databases with medium confidence threshold (0.4).

In MIM, molecular components are clustered into putative subsystems. These subsystems are determined by mutual interactions and functional correlations. Here, apart from the overall prediction statistics, we describe simulations of two subsystems as case studies: the nucleotide excision repair (NER) and Cyclin/CDK subsystems.

5.2.4 The Nucleotide Excision Repair (NER) Subsystem in the p53 Pathway

DNA can be exposed to damaging chemical and physical agents such as UV and free radicals. The nucleotide excision repair pathway repairs DNA damage *via* sequential combination of protein complexes rather than a pre-organized protein assembly. The basic steps of NER involve DNA damage recognition, damage verification, association of repair proteins, excision of the damaged DNA and resynthesis [213, 214]. Several structures of proteins involved in NER are available in the PDB such as the ERCC1/XPA, ERCC1/XPF, and XPC/Centrin peptide complexes. However, the structural coverage of the NER pathway is still far from complete. We aim to assist in the assembly by modeling protein complexes in this pathway.

The ‘excision-resynthesis’ reaction begins with damage recognition (see **Figure 5.5**). The HR23B/XPC/Centrin complex is a sensor which identifies the damaged part. The crystal structure of the XPC/Centrin heterodimer is available (pdbID: 2a4j). Transcription factor complex TFIIH (containing Cdk7, CycH, and XPB, XPD helicases) helps in opening the strand around the damaged part. XPA confirms the damage and is essential for the following steps. The replication protein A (RPA) functions in the stabilization of the opened DNA. RPA interacts with XPA for correct positioning of endonucleases (XPG, ERCC1/XPF) [215]. RPA is a heterotrimeric protein composed of 14, 32 and 70 kDa subunits. The 70 kDa subunit of RPA (1fgu:A) associates with XPA (1xpa) to function in the assembly of the repair complex. RPA and ERCC1 associate with XPA sequentially: First RPA binds and then ERCC1 [216]. Residues 98 to 187 of XPA are required for binding to the 70 kDa subunit of RPA. XPA uses the residues 67-80 to interact with ERCC1 (pdbID: 2jnw). Thus, the binding regions on XPA do not overlap. Using our method, the structural model for RPA (70 kDa)/XPA complex is generated. With the template set of 1,037 interfaces, the algorithm can not find a similar template for the RPA/XPA complex. We then utilized all interfaces including homodimers. The template interface tRNA-ribosyltransferase homodimer (2ashCD) gives one hit for this interaction. XPA interacts with RPA at XPA's Zn-containing N-terminal (residues 102-129) [217]. The predicted region on XPA corresponds exactly to this region and is illustrated in **Figure 5.5**.

In the NER subsystem, Rad51 and Rad52 directly interact with each other [218] and the RPA/Rad52/Rad51 assembly functions in concert. The highest ranking solution for the Rad51/Rad52 complex is found from the template generated from Ras-related protein Rab-

7/Rab interacting lysosomal protein complex (1yhnAB). RPA 32 kDa/Rad52 complex is found using the template exonuclease I/exonuclease II complex (2c38SV). The putative trimer of the Rad51, Rad52 and RPA 32 kDa shows that their simultaneous interaction is possible. In addition, the binding region on Rad51 includes Phe259 which is essential for Rad52 binding [219]. RPA 32 kDa uses a region distinct from the one which interacts with 14 kDa and 70 kDa subunits.

Following the RPA/XPA complex formation, the endonucleases (ERCC1/XPF, XPG) enter the pathway and function in DNA incisions. The crystal structure of the ERCC1/XPF complex is known (1z00). Rad52 can bind to the XPF subunit of the ERCC1/XPF complex using its DNA binding domain (N-terminal region, 1h2i:A, residues 1 to 209). XPF/Rad52/ERCC1 can interact simultaneously. This assembly stimulates the endonuclease activity of XPF/ERCC1 and weakens the DNA strand annealing activity of Rad52 [220]. The first ranking solution of the method gives a binding region which does not overlap the one where ERCC1 binds to XPF. The structures of ERCC1 and Rad52 do not interpenetrate when they interact with XPF. In this way, a hypothetical trimeric complex of ERCC1/Rad52/XPF is obtained. Rad52 self-associates and forms a ring structure which functions in DNA annealing. Rad52 is also stable on its own [221]. One mechanism suggests that Rad52 forms the ring after entering the nucleus [222]. This predicted binding region on Rad52 overlaps with the self interacting region. From this predicted assembly, we suggest that when Rad52 binds to XPF, it can not form the Rad52 complex ring which binds to ssDNA, weakening its DNA annealing activity.

Following removal of the damaged part, DNA repair and resynthesis begin. PCNA, described as a 'sliding clamp', has an essential role. RF-C is required to load PCNA onto the DNA. The crystal structure of the yeast PCNA/RF-C complex is in the PDB. Another interaction partner of PCNA is Gadd45, considered as a stress sensor. The interaction between PCNA and Gadd45 stimulates DNA excision repair and inhibits the entry of cells into the S phase [223]. This indicates that Gadd45 may be considered as a link between the p53-dependent cell cycle checkpoint and DNA repair. The 94 N-terminal residues of Gadd45 contribute to strong interactions with three regions of PCNA (1-20, 61-80 and 196-215) and weak interaction with one region (121-170) [224].

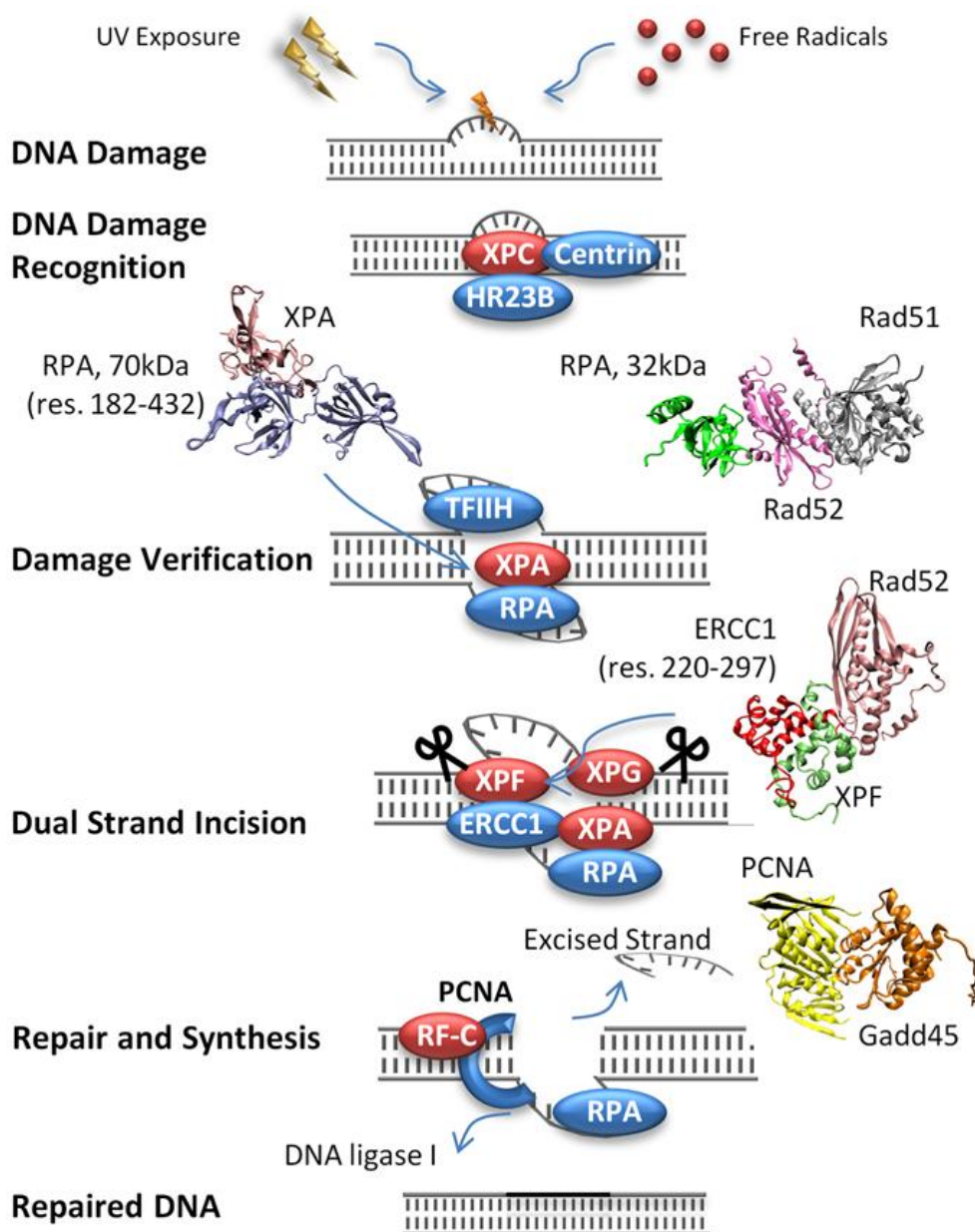


Figure 5.5 The nucleotide excision repair subsystem as a sequence of reactions. It initiates with damage recognition by HR23B/XPC/Centrin. Then, the strand around the lesion is opened by the TFIIH complex. The replication protein A (RPA) stabilizes the opened DNA and associates with XPA for the correct positioning of endonucleases. Following the RPA/XPA complex formation, the endonucleases (ERCC1/XPF, XPG) enter the pathway and function in DNA incisions. After that, repair and resynthesis process begins. At each step of the NER, structural representation of the predicted (XPA/RPA, Rad51/Rad52/RPA, Rad52/XPF/ERCC1, PCNA/Gadd45) are illustrated. The crystal structures of XPF/ERCC1, RPA complex, XPA/ERCC1, XPC/Centrin complexes are available in the PDB.

The first ranking solution, predicted from the template interface of the signaling protein complex YPD1/SLN1 (1oxbAB), finds that residues 138 to 146 of PCNA are in contact with Gadd45 which corresponds to a patch in the experimentally known weak interaction region on PCNA (residues 121-170).

Table 5.2 The highest ranking solutions in the case studies.

Target Protein 1	Target Protein 2	ΔG_{calc} (kcal/mol) ^a	Template Name
NER Subsystem			
RPA 70 kDa (1fgu:A)	XPA (1xpa)	38.80	tRNA-ribosyltransferase homodimer (2ashCD)
Rad52 (1h2i)	Rad51 (1n0w:A)	2.98	Ras-related protein Rab-7/Rab interacting lysosomal protein complex (1yhnAB)
Rad52 (1h2i)	XPF (1z00:B)	-4.55	tyrosyl-tRNA synthetase (2cycAB)
Gadd45 (2kg4)	PCNA (1axc:A)	-26.58	the signaling protein complex YPD1/SLN1 (1oxbAB)
RPA, 32 kDa (2z6k:A)	Rad52 (1h2i)	41.94	Exonuclease I/exonuclease II complex (2c38SV)
Cyclin/Cdk Subsystem			
Cdk1 (1lc9)	CycB (2b9r:A)	-38.63	Cdk2/CycA complex (1vywAB)
14-3-3 (1ywt:A)	Cdk1 (1lc9)	33.20	Cdk5/Cdk5 inhibitor p25 complex (1unlAD)
Cdk1 (1lc9)	Gadd45 (2kg4)	-92.71	pyridoxal kinase complex (1lhpAB)
Other Interactions			
Nfkb, p65 subunit (1nfi:A)	ASPP2 (1yca:B)	-45.61	Nfkb p65 subunit/Ikb complex (1nfiAF)
p27 (1jsu:C)	Cks1 (1buh:B)	-44.06	Cdk2/p27 complex (1jsuAC)
	Rad52 (1h2i)	-55.35	Cdk2/p27 complex (1jsuAC)
p27 (1jsu:C)	Mdm2	-28.87	CycA/p27 complex (1jsuBC)
p27 (1jsu:C)	TFIIH, p62 subunit (1pfj)	-37.51	CycA/p27 complex (1jsuBC)
p27 (1jsu:C)	ERCC1 (2a1i)	-29.54	CycA/p27 complex (1jsuBC)
Skp2 (2ast:B)	p19 ^{ink} (1blx:B)	-26.69	Skp1/Skp2 complex (2astAB)
Skp2 (2ast:B)	E2F4 (1cf7:A)	-9.32	Skp1/Skp2 complex (2astAB)
Skp2 (2ast:B)	p300 (1p4q:B)	-20.53	Skp1/Skp2 complex (2astAB)
Skp2 (2ast:B)	HMG (2e6o)	-20.23	Skp1/Skp2 complex (2astAB)
Skp2 (2ast:B)	APC (1m5i)	-13.01	Skp1/Skp2 complex (2astAB)
Skp2 (2ast:B)	Plk1 (1q4k:A)	-2.84	Skp1/Skp2 complex (2astAB)

^a The energy values are calculated by Fiberdock.

In the last step, PCNA anchors polymerases to the replication site. A patch is resynthesized, added to the DNA strand by DNA ligase I and the DNA is repaired. The events are summarized in **Figure 5.5**. The list of putative complexes with the calculated energies are tabulated in **Table 5.2**. In this way, we simulated functional associations in the

NER pathway using both experimental structural data and our predicted protein complexes, illustrating the potential of our approach in prediction of functional associations.

5.2.5 Cyclin/CDK Subsystem

Cyclin dependent kinases (CDKs) play an essential role in cell cycle progression. Deregulation of cell cycle usually causes cancer. Hence, CDKs are drug targets. Here, we analyze the CDK interactions and cell cycle G2/M phase from a structural perspective.

Cell cycle: G2/M Checkpoint prevents cells from entering mitosis (M phase) if the genome is damaged. Under normal conditions, Cdc25 is activated by Plk1; Cdk1 (Cdc2) is activated by Cdc25. Cdk1 bound to CycB drives the cells from the G2 phase to the M phase.

The structure of the Cdk1/CycB complex is unavailable. To simulate this pathway using structures, the Cdk1/CycB complex should first be modelled. Cdk1 (PDB: 1lc9) is structurally very similar to other cyclin-dependent kinases. Structurally, CycB (2b9r:A) is also similar to other cyclins. The first ranking putative Cdk1/CycB interaction (-38.63 kcal/mol) is found using the interface between Cdk2 and CycA (1vywAB). Finding the binding region of the Cdk1/CycB complex is trivial because the overall structures of the target proteins are very similar to the partner chains of the template interface. Using sequence order independent alignment of interface partners (which are discontinuous segments) with the surface region of the target proteins, we obtain a favorable energy value which indicates that our method can successfully model this interaction.

When DNA is damaged, two cascades get activated. The first halts the G2 to M phase process by attacking Cdc25. The second targets Cdk1, keeping it inactive. **In the first cascade**, Chk kinases phosphorylate and inactivate Cdc25. The 14-3-3 protein plays a role in the regulation of the signaling pathways and functions in nuclear export. The phosphorylated Cdc25 is exported from the nucleus by the 14-3-3 before it interacts with Plk1. **In the second cascade**, p53 dissociates from Mdm2 when it is phosphorylated, binds DNA and promotes production of all three proteins, 14-3-3, Gadd45 and p21 which inhibit Cdk1. 14-3-3 functions in nuclear export as in the first case, but here the target is Cdk1 to be moved out of the nucleus. The first ranked solution of 14-3-3/Cdk1 complex is predicted from the template interface between Cdk5/Cdk5 inhibitor p25 complex (1unlAD). The predicted region on Cdk1 overlaps with its CycB binding region.

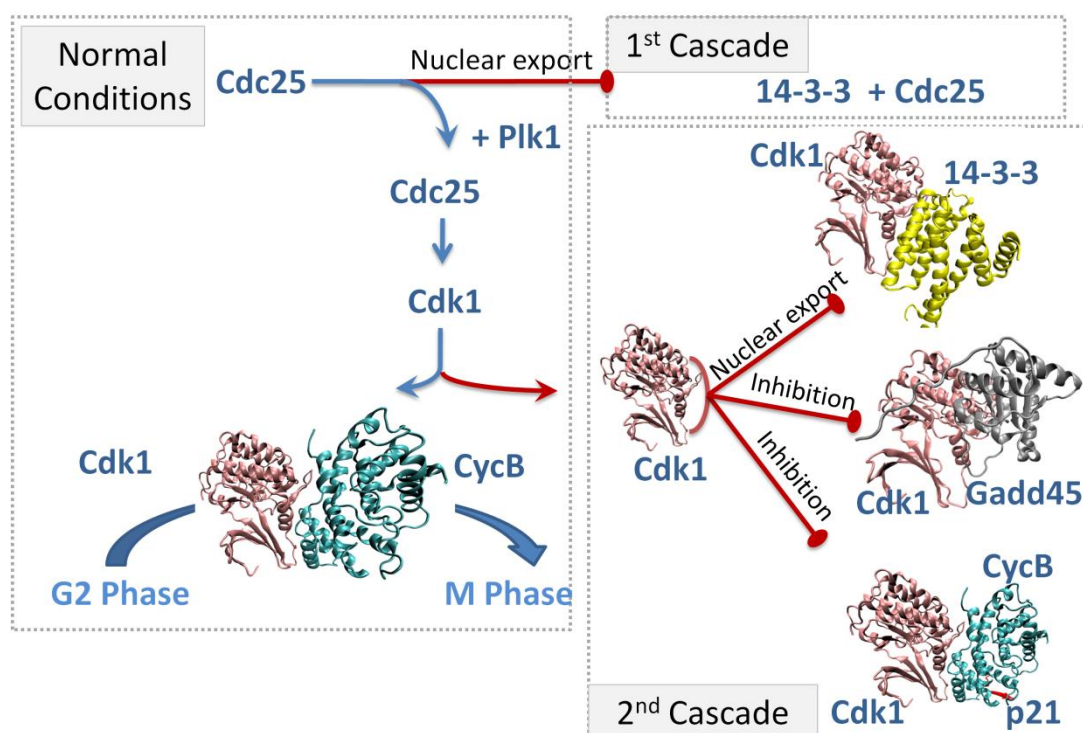


Figure 5.6 Structural representation of the G2/M phase checkpoint. In this figure, all interactions are modelled using our prediction method, except the p21/Cdk1/CycB complex. Under normal conditions, Plk1 activates Cdc25; then, Cdc25 activates Cdk1 which drives the cell into mitosis (shown in left panel). DNA damage initiates two parallel cascades to inactivate the Cdk1/CycB complex (shown in the right panels). Activations are shown with blue arrows; inhibitions and nuclear exports are shown with red oval arrows.

Experimental observations indicate that Gadd45 inhibits the activity of Cdk1/CycB as strongly as p21 which is also a Cdk/Cyc inhibitor. In addition, Gadd45 does not associate with CycB which indicates two possible mechanisms: either Gadd45 associates with free Cdk1 or it displaces CycB from Cdk1/CycB [225]. The inhibitory domain of Gadd45 includes residues between 65 and 84 [226]. We predict two possible binding orientations for the Cdk1/Gadd45 protein pair. The first ranking putative interaction is predicted from the template interface in the pyridoxal kinase complex (1lhpAB). The second ranked solution comes from the acetylglutamate kinase complex (2ap9AB). The experimentally determined region of Gadd45 (residues 65 to 84) matches exactly the predicted region on the Gadd45 surface in this solution. In the first solution, the predicted region overlaps with the region where Cdk1 interacts with CycB. Thus, this prediction supports the first mechanism where Gadd45 associates with free Cdk1 and prevents it from interacting with

CycB. In the second solution, the predicted region does not overlap with the CycB interacting region of Cdk1 which supports the second mechanism.

Figure 5.6 illustrates the simulation of this pathway with the predicted interactions where all are first ranking solutions. The modelled quaternary structure of the Cdk1/CycB/p21/PCNA complex is taken from Ref. [227]. Since only 22 residues of p21 have available coordinates, p21 cannot serve as a target protein in our algorithm. **Table 5.2** gives the energies and template interfaces for the putative complexes.

5.2.6 Some Promising Interactions Predicted by Our Approach

NFκB–ASPP2 Interaction

NFκB is a transcription factor playing a role in the regulation of apoptosis. In the cytoplasm it is present in an inactive state. When NFκB gets activated, it passes into the nucleus and binds specific DNA fragments [228]. NFκB is composed of two subunits, p65 and p50 (1nfi, p65: chain A, p50: chain B). Ikb (1nfi, chain F), an ankyrin repeats-containing protein, keeps NFκB in a resting state in the cytoplasm. Using the interface between Ikb and NFκB (1nfiAF) as a template, we predict possible partners of the NFκB protein among all target proteins in the p53 pathway. We find a possible interaction between NFκB (1nfi:A) and ASPP2 (1ycs:B). ASPP2 is a proapoptotic protein which associates with several proteins including the DNA binding domain of p53 (1ycs:A). It contains ankyrin repeats like Ikb.

Figure 5.7 illustrates the structure of the predicted complex. The ASPP2 structure matches well the Ikb side of the template. Hence, ASPP2 may be considered an inhibitor of NFκB and it may block the DNA binding activity of NFκB in the nucleus. The predicted region corresponds to a previous model [229], where two regions were proposed for binding of NFκB to ASPP2. The predicted NFκB/ASPP2 complex shows that the simultaneous interaction of NFκB subunits and ASPP2 is possible. However, the ASPP2/p53 heterodimer cannot interact with the p65 subunit of NFκB in the presence of the p50 subunit, because of the geometric clash between them. Further, Ikb and ASPP2 use the same region on NFκB; thus, their interaction with NFκB is also mutually exclusive.

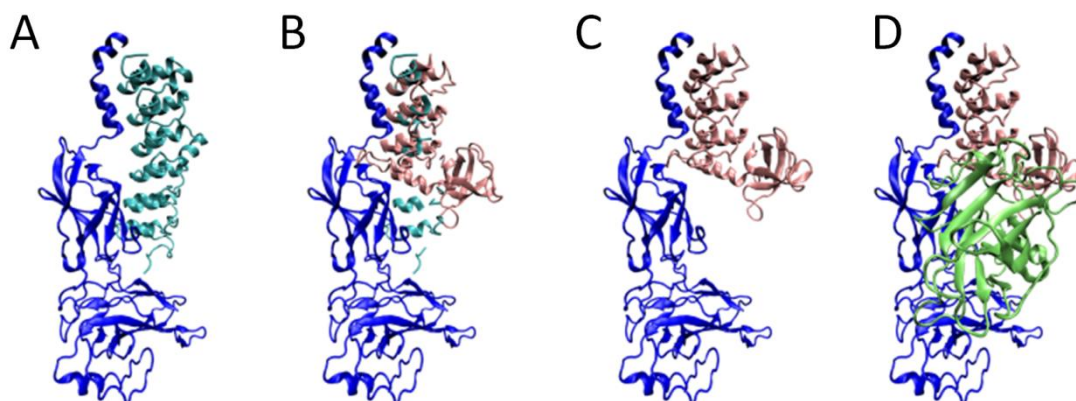


Figure 5.7 The modeled interaction between NF κ B and ASPP2 proteins. (A) The NF κ B/I κ b complex (1nfiAF). (B) The predicted NF κ B/ASPP2 complex, NF κ B is in blue, the right partner of the template interface is in cyan and ASPP2 is in pink. (C) The NF κ B/ASPP2 complex following flexible refinement. (D) The possible NF κ B/ASPP2/p53 trimer.

The p27 Protein and its Possible Partners

p27 is a cyclin dependent kinase inhibitor. The trimeric complex of p27/Cdk2/CycA (1jsu) generates two interfaces with one side being p27. From these two template interfaces, several interaction partners for p27 are found by our method. One of them is Cks1 which is similar to the Cdk2 part of the interface between p27 and Cdk2. Four β -sheets match well with the β -sheets of Cdk2 (**Figure 5.8**) which is the highest ranking solution (-44.06 kcal/mol). Three new interaction partners for p27 are predicted from the interface between p27 and CycA: Mdm2, TFIIH (p62 subunit) and ERCC1 (-28.87, -37.51, -29.54 kcal/mol, respectively). String [230] provides experimental evidence for the Mdm2/p27 and ERCC1/p27 predicted complexes.

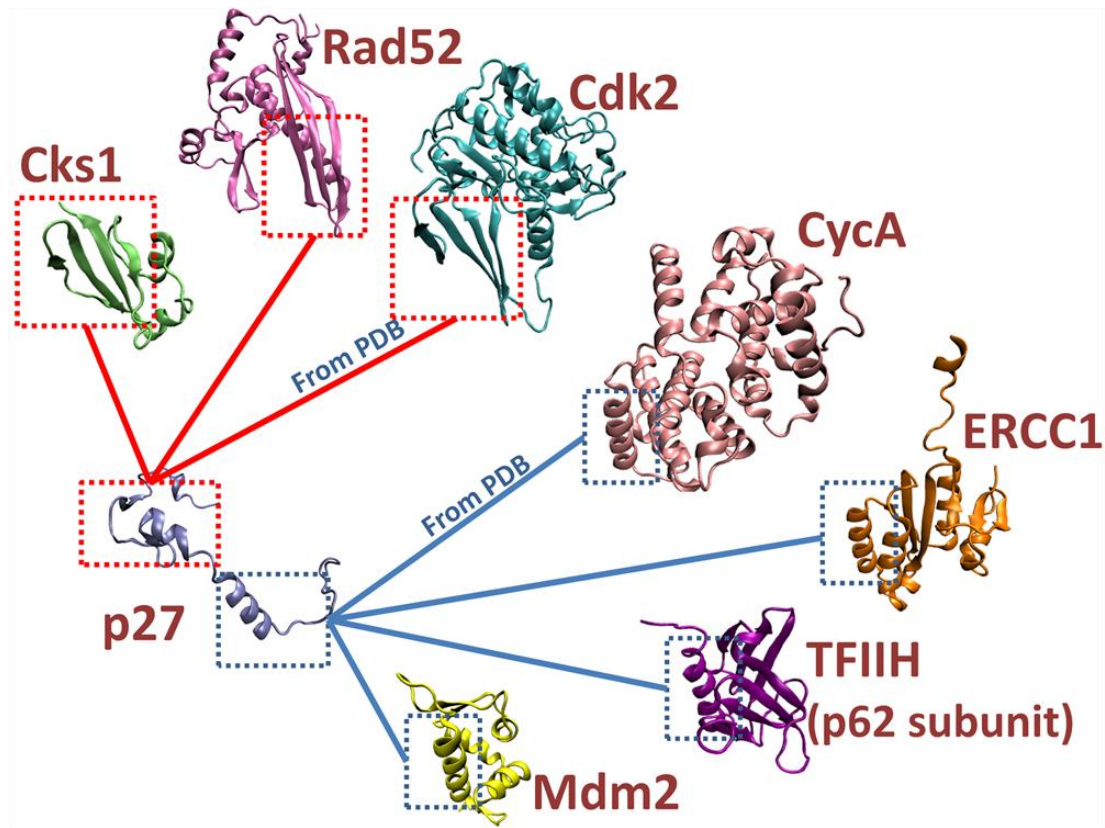


Figure 5.8 Structural representation of predicted and known partners of the p27 protein. p27 has two binding sites, one is shown in red colored dashed box and the other in blue. The edges between p27 and its partners are colored in the binding site colors.

The Skp2 Protein and its Possible Partners

The S-phase kinase-associated protein (Skp2) is an F-box protein. Association of p27 with Skp2 is the rate limiting step in ubiquitin-mediated degradation of p27 [231]. When we compare all target proteins in the template set of the p53 pathway interfaces, we notice that Skp2 uses the same region to interact with other proteins. One of these partners is another kinase inhibitor, p19^{ink}, whose ankyrin repeats match well the Skp1 side of the interface (-31.11 kcal/mol). We speculate that in addition to p27, p19^{ink} may be a degradation target of Skp2. As shown in **Figure 5.9**, APC, E2F4, HMG, p300 and Plk1 also have structurally similar segments on their surfaces and are predicted as interacting partners of Skp2.

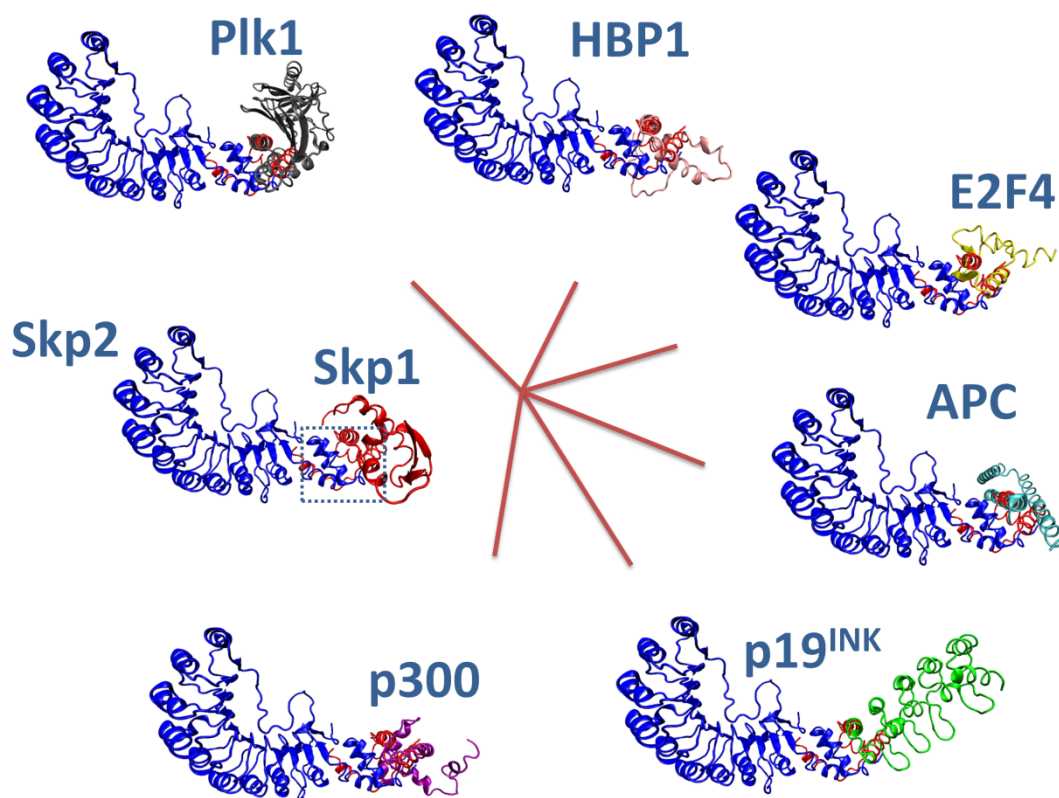


Figure 5.9 Some possible partners of Skp2 are predicted from the interface between Skp2 and Skp1. The template interface is shown in dashed box on the Skp1/Skp2 complex. In the continuing part of this figure red colored segments represent the interface part of Skp1. These segments match well the surface of other target proteins, E2F4, HBP1, p300, APC, Plk1 and p19^{ink}.

5.3 Conclusions

Here we presented a combinatorial approach to effectively predict the functional associations of proteins on a large scale. This approach relies on the expectation that the number of protein-protein interface architectures in nature is limited; thus extrapolation of the known architecture space on target protein surfaces may help to identify protein interactions. This knowledge-based approach is made more physical by combining with docking strategies, flexible refinement of the solutions and energy calculations to rank them. We show how available structural information can help in modeling a pathway by using structural similarity *independent* from global fold homology and how such a knowledge-based approach can be enhanced by filters, flexibility and energy calculations. To show its functionality, we modeled the molecular interaction map of the p53 pathway. The nucleotide excision repair (NER) subsystem, cyclin-dependent kinase subsystem and some promising putative functional associations are illustrated in the p53 pathway.

Our strategy models the proteome based on local, sequence-order-independent homology to each side of an interface whose structure has been determined experimentally. As in any strategy which is motif-based there *must* be a similar motif in the template set. If there is no such motif, we cannot expect the method to find it. As in homology modeling, or threading of protein chains, a motif-based strategy is an advantage and a disadvantage: the advantage is that if such motif is available, the method is fast and reliable (which is why single chain homology modeling is so popular). At the same time it is a disadvantage, since the outcome depends on the presence of the motif in the template set. The Structural Genomics initiative is widely expected to assist in protein folding via the generation of all motifs which could be used for modeling. Similar considerations apply to modeling of protein-protein interaction. We expect that with the fast growth of the PDB the number of distinct interface motifs will grow making such fast strategies increasingly popular and useful for the modeling of protein interactions.

Chapter 6

MULTI-PARTNER PROTEINS

In this chapter, the multi-partner proteins are analyzed along their distinct binding sites. Adapting of multiple binding sites or reutilizing of a single site by several partners is crucial for interaction with many different proteins. Multi-partner proteins can interact with their partners at different time periods through the same region (i.e., *mutually exclusive interactions*); or at the same time through different regions (i.e., *simultaneous interactions*), or both. In the first part, the dataset of multi-partner proteins available in PDB are presented along some case studies; then, two hub proteins – p53 and Mdm2 – are illustrated with PRISM predicted and experimental interactions.

6.1 Structural Dataset of Multi-Partner Proteins

Datasets of protein-protein interfaces are constructed by considering pairwise contacts of protein chains within a single protein complex in PDB. For proteins interacting with many other partners, there are different crystal structures containing these multiple partner proteins. In the current work, we will follow a new approach to reveal the multi-faced structure of proteins in PDB. For this aim, we take all proteins in PDB clustered according to their sequence similarity (70% homologous protein chains in each cluster). On the other hand, our pairwise interface dataset [10] is updated with new structures. The interfaces of the proteins in each cluster are searched in this chain level interface dataset. One protein is fixed and its partners available in the interface dataset transformed on this protein using Multiprot [208]. For two interactions to be simultaneously possible, first of all their binding sites should not overlap; further, the rest of the structures of two partners of a protein should not interpenetrate into each other. For this purpose, we set two thresholds. If less than five residues are overlapping of two binding regions and less than 5 residues are interpenetrating, these two partners can interact with the center protein simultaneously. All possible combinations of these simultaneous interactions give the possible multimeric states of the proteins.

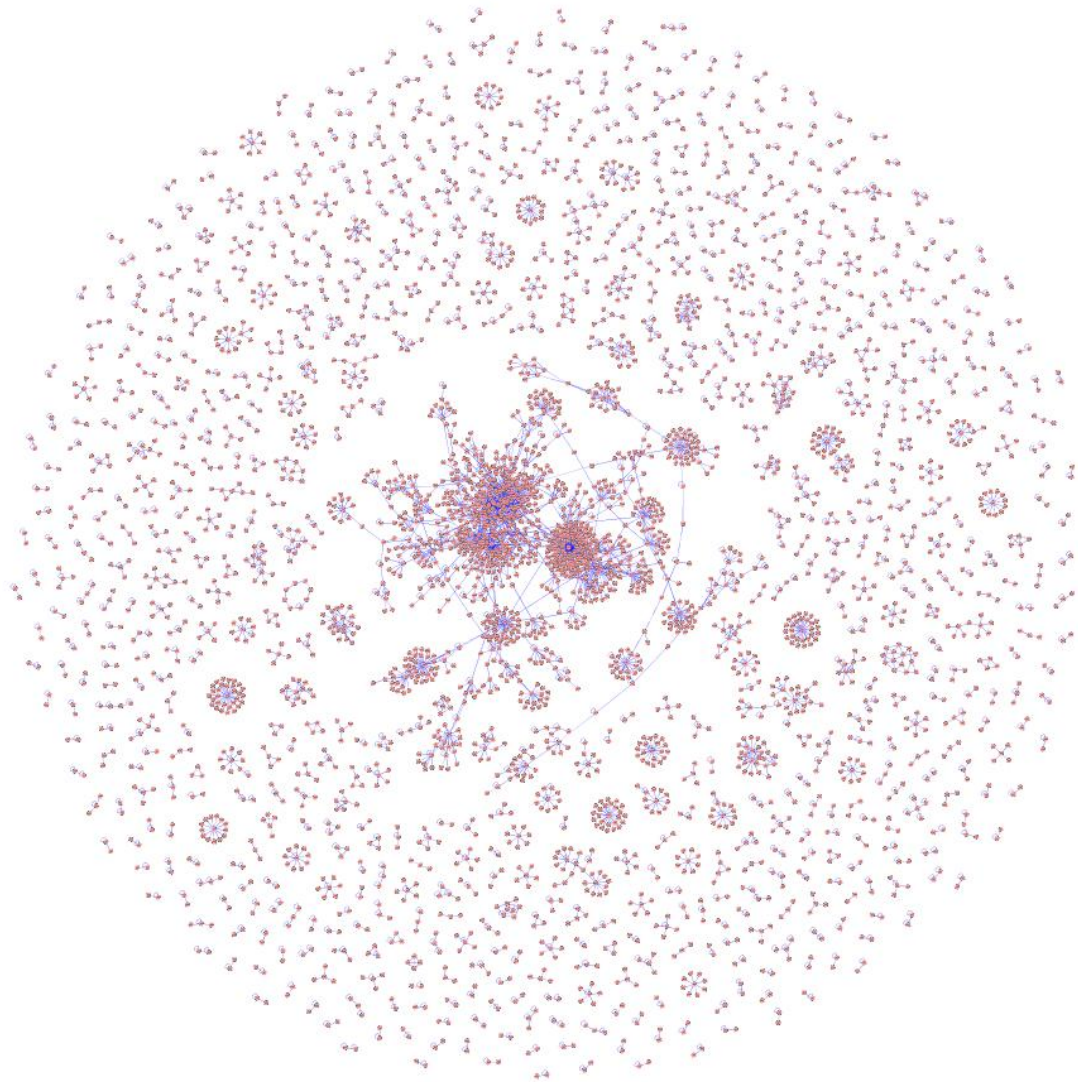


Figure 6.1 Interaction network of multi-partner proteins derived from PDB.

As a result, compact multimeric states of the proteins are obtained. To construct the multi-partner proteins dataset, 64,352 protein structures are downloaded from PDB (as of April 2010). These protein structures lead to 127,510 two-chain protein interfaces. After application of the procedure, we obtain 1,491 multi-partner proteins in the dataset of which 389 have a binding region re-used several times by their partners and 1,401 use more than one region to interact with their partners. This procedure is useful because it collects the disorganized data for a specific protein. In this way, all known partners of a targeted protein can be easily picked. Also, this dataset shows the overlapping and distinct binding

sites on the protein surfaces. These multi-partner proteins can be represented as a network to visualize the connectivity of the proteins in PDB. The network contains 3920 protein nodes 4915 interaction edges. There is a large connected component and several isolated entities which are not connected to the largest one (see **Figure 6.1**). With the increase in the number of protein structures in PDB, this network will get more connected in the future. However, still this is a rich source to figure out which interactions can occur simultaneously and which are mutually excluded. It assigns implicitly the time dimensionality in protein networks and transforms node-and-edge maps into cellular processes, and their regulation.

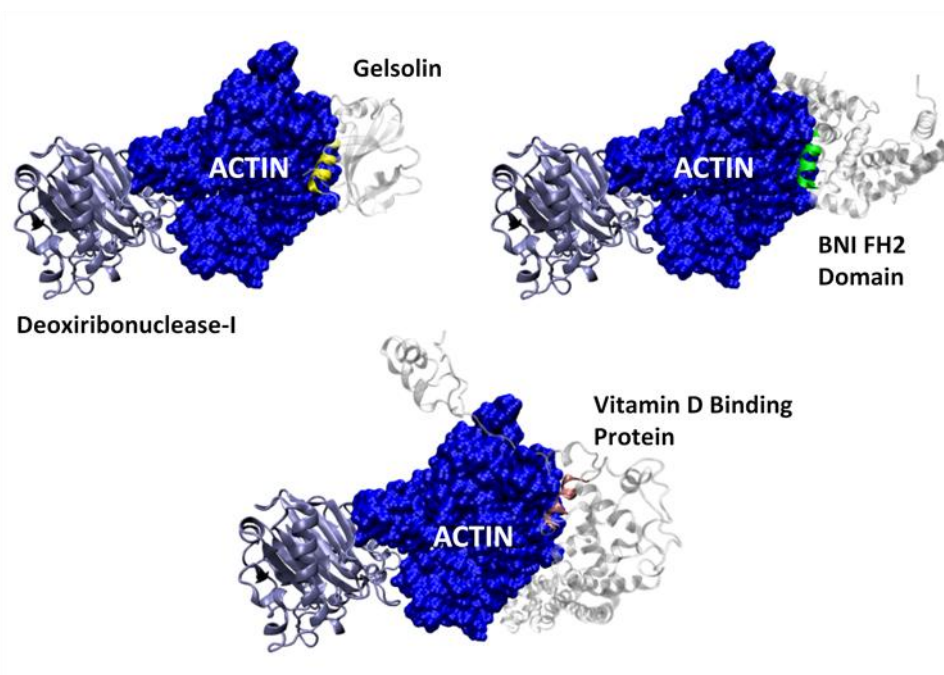


Figure 6.2 Two binding sites of actin along its partners available in PDB are shown. The conserved helices are colored yellow, green and pink for gelsolin, BNI FH2 domain and vitamin D binding protein, respectively.

As an example for multi-partner proteins in PDB, actin is illustrated in **Figure 6.2**. Here, two binding sites of actin are shown. It reuses the same binding site to interact with Vitamin D binding protein, gelsolin and BNI1 FH2 domain. On the other hand, deoxyribonuclease-1 interacts with actin through the second region. When we examine the details of the binding sites, we noticed that an 11-aa long helix is available in the binding region of all three partner proteins. The two computational hot spots 143 and 345 are

conserved in all three interactions contacting with the mentioned helices and the complementary hot spots in the partner proteins form a continuous region in the interfaces. The size of the interface between actin and gelsolin (1yvnAG) is 1021 \AA^2 , between BNI1 FH2 domain (1y64AB) is 1343 \AA^2 , between vitamin D binding protein (1ma9AB) is 1711 \AA^2 . Although the overall interface architectures are not completely similar and sizes of the interfaces are different, a local similarity mediates the interaction between these proteins.

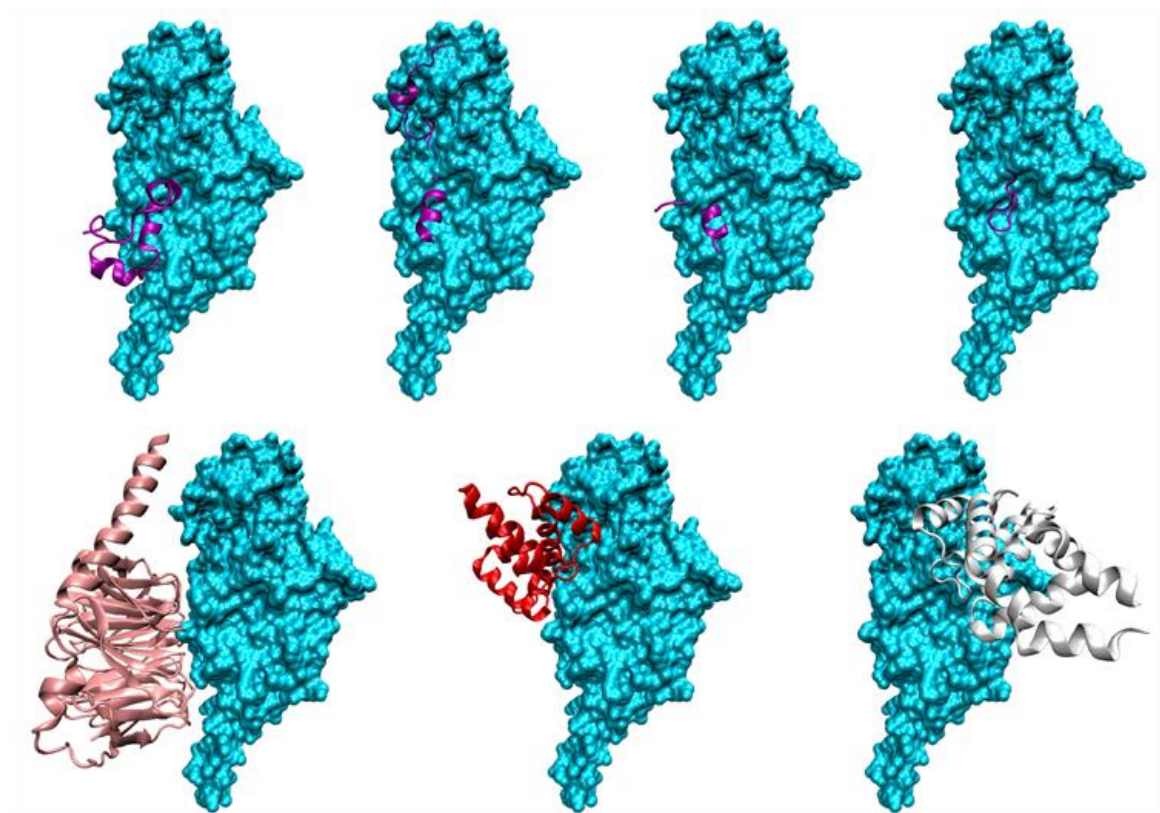


Figure 6.3 The available partners of guanine nucleotide binding protein. The first binding site on G protein is used by cGMP 3',5'-Cyclic Phosphodiesterase, Regulator of G-Protein Signaling (RGS) 14, KB752 peptide, KB-1753 phage display peptide, respectively (colored purple). The second region is used by G protein beta subunit, the third is used by RGS4, and the fourth one is used by RGS8.

In another example, interactions of G protein alpha subunit are shown in **Figure 6.3**. The partners cGMP 3',5'-Cyclic Phosphodiesterase, Regulator of G-Protein Signaling (RGS) 14, KB752 peptide and KB-1753 phage display peptide bind to the same region on guanine nucleotide binding protein. Hence, their interactions are not simultaneously possible. Distinct from this region, there are 3 more binding regions on guanine nucleotide

binding proteins where RGS4, RGS8 and G protein beta subunit bind, respectively. Although the binding regions of RGS4, G protein beta subunit and cGMP 3',5'-Cyclic Phosphodiesterase are not overlapping, the rest of the partner proteins interpenetrate each other. Hence, their simultaneous interactions are not possible. The four proteins using the same binding region do not share a common interaction hot spot. With the changing partners, the hot spots are also changing which may be the result of the small movements of G protein and different binding architectures of partner proteins.

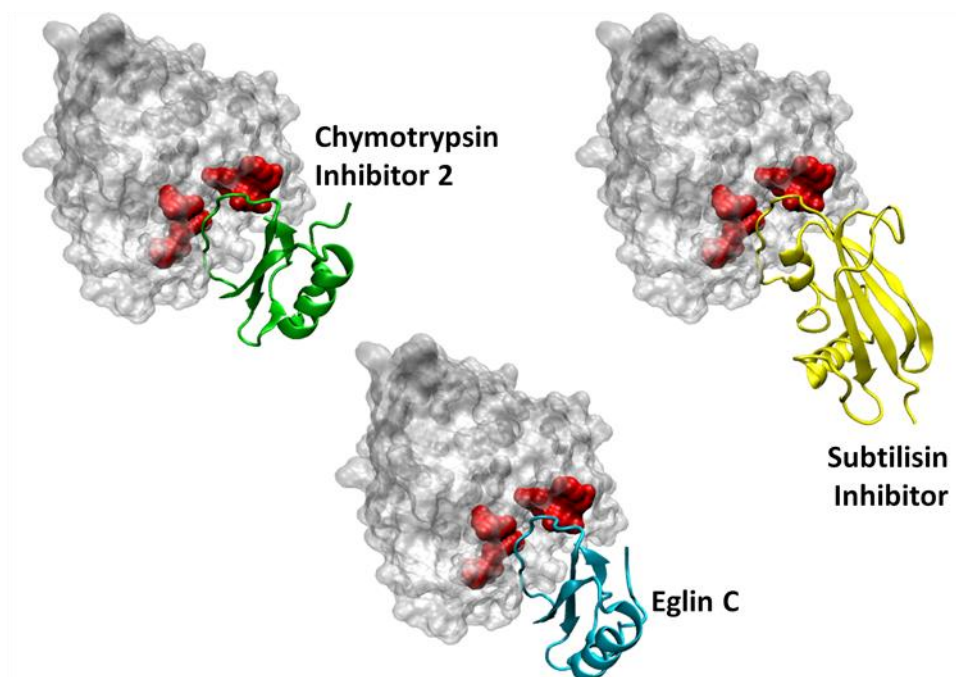


Figure 6.4 Subtilisin and its inhibitor molecules chymotrypsin inhibitor-2, subtilisin inhibitor and eglin C. Subtilisin is shown in surface representation colored white. The predicted hot regions are colored red.

The serine protease subtilisin BPN' is another multi-partner protein using the same region to interact with different proteins. *Streptomyces subtilisin inhibitor* (3sic:I), eglin C (1sib:I) and chymotrypsin inhibitor 2 (1y34:I) are the binding partners of subtilisin. Although their overall structures are dissimilar, their interface regions are structurally very similar. We checked the hot spots in the interface region and noticed that hot spots on subtilisin remain unchanged while interacting with different partners. These hot spots point out two common regions on subtilisin (residues 64, 96, 107, 126, 155, 189 and 220) like two clips holding the partners from up and down as shown in **Figure 6.4**.

Falcipain-2 is a cysteine protease and a promising drug target [232]. The falcipain-2 protein and its inhibitors are shown to illustrate the hot spot distribution in their interfaces in **Figure 6.5**. Computational hot spots are extracted using the HotPoint web server [209]. Most of the predicted hotspots on the falcipain-2 side are not changed despite the different partner proteins (cystatin and chagasin). Furthermore, the small ligand inhibitor E64 directly binds to the hot spots where other partners also bind. This is consistent with the statement that small molecules target hot spots.

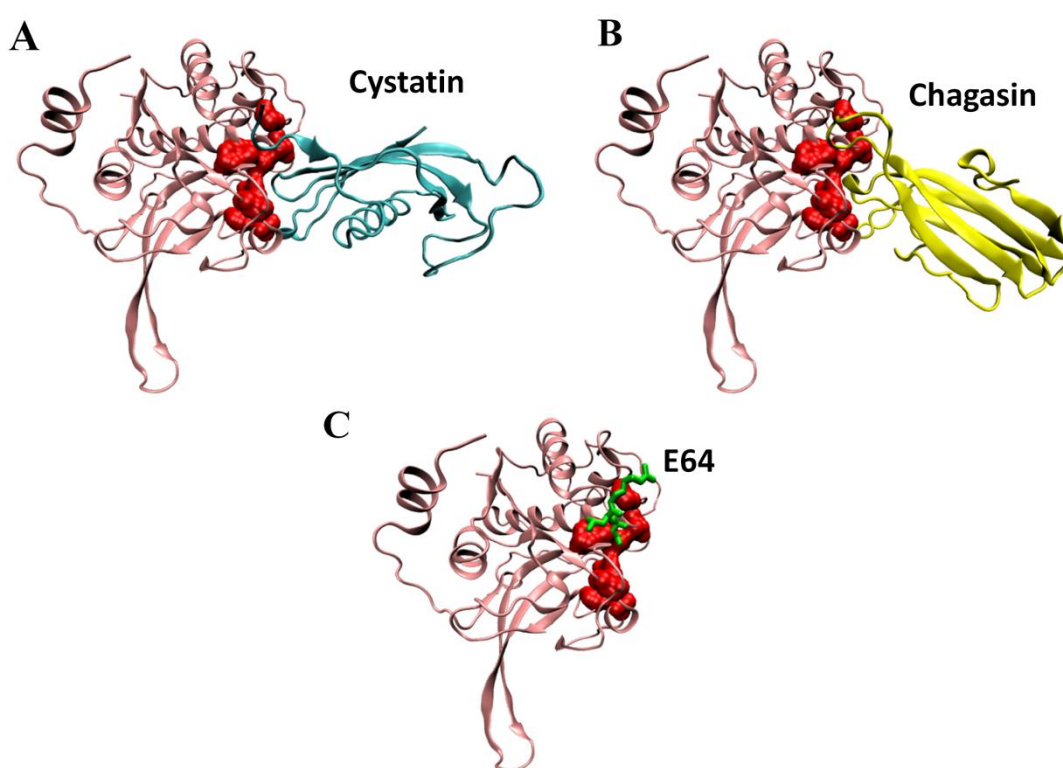


Figure 6.5 The cysteine protease Falcipain-2 and its inhibitors (A) Cystatin, (B) Chagasin and (C) the inhibitor molecule E64. Falcipain-2 is colored pink and the hotspots are shown in surface representation in red. Although the overall structures of the cystatin and chagasin are different their interface regions to falcipain-2 are structurally very similar. The intersection of their hot spots show that the small inhibitor molecule E64 also directly binds to these hot spots.

As a general trend, the enzyme-inhibitor complexes are more specific to their partners. While an enzyme is interacting with its partners, interface regions of the partner proteins is structurally very similar to each other. Hence, the hot spots remain unchanged with the changing partners and binding hot spots are important for the interaction specificity as in the subtilisin and falcipain-2 examples. On the other hand, despite the different overall

interface architecture, just a local similarity mediates the interaction with different partners as in actin example. In the G protein example, predicted hot spots change with small movements of the G protein while changing partner proteins. So we can conclude that if the partner proteins are using the similar interface to interact, the interaction hot spots remain unchanged; however, if partners associate using different architectures to the same region and the center protein make small movements, this leads to the change in the hot spot distribution. However, these statements are shown on limited number case studies; hence, this analysis needs a more rigorous and systematic large-scale examination.

6.2 Towards Inferring Time Dimensionality in Protein – Protein Interaction Networks by Integrating Structures

Here, to illustrate the time dimensionality concept we use the predictions obtained by PRISM [11, 13, 15] which is the previous version of the method in Chapter 5 without flexibility and energy calculations. However, the rationale is the same: the number of interface architectures in nature is limited; thus, if two surface regions of two single chain proteins are similar to two sides of a crystal (or NMR) complex, they can bind. Via PRISM, we are able to predict which interactions can and cannot co-exist at the same time. Here, the template set for predictions are composed of 158 non-obligate interfaces and 330 obligate interfaces which are also subsets of again 49512 interfaces clustered into 8205 clusters [10].

6.2.1 p53 and its Binding Partners

p53 is a central protein, playing a key role in response to a broad range of stress signals such as DNA damage and oncogene activation; as such it has a large number of binding partners in the cell. p53 consists of five domains: the transactivation domain, the proline-rich domain, the DNA-binding domain, the tetramerization domain and the regulatory domain. A high resolution structure of full-length p53 is unavailable; however, structures of several individual domains are. **Figure 6.6** illustrates these domains with their corresponding structures. On its own, the transactivation domain is unfolded; it is folded when bound to the Mdm2 [233] or in a membrane environment [234]. The DNA-binding domain is the largest structured p53 domain with 191 residues. The tetramerization domain

is structured in the p53 tetramer, forming a single helix. The regulatory domain, to which ubiquitin attaches, is unfolded.

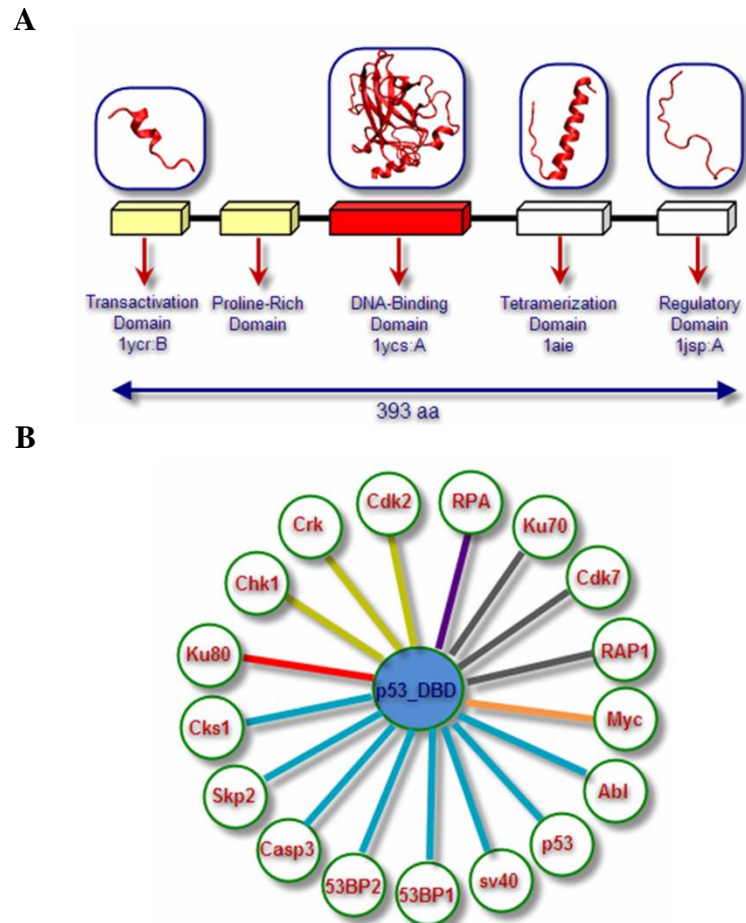


Figure 6.6 (A) The fragments of the p53 protein and the available crystal structures; (B) p53 DNA-binding domain interactions. Edges are colored according to the different binding sites; these contain both experimental and PRISM-predicted interactions.

To predict the p53 partners and their binding sites, we selected the DNA-binding domain (DBD); the helical structure of the tetramerization domain can match many helical template interfaces leading to a prediction bias. The structurally known and predicted interactions of the DBD clearly show a mechanistic multi-faced, multiple partner paradigm.

Figure 6.6B illustrates known and potential PRISM-predicted p53 DBD binding partners. In this small network, predicted interaction partners (Cdk2, Crk, Chk1, RPA, Ku70, Ku80, Abl, Casp3, RAP1A, Cdk7, Myc, Skp2, Cks1) and interaction partners known from available crystal structures of the p53 DBD complexes (53BP1, 53BP2, sv40,

p53 DBD) are drawn, where edges are colored according to their binding sites; protein partners binding p53 DBD at the same binding site are depicted by similar-color edges. **Figure 6.7** presents the detailed picture of this small network by combining the structures of the proteins and their binding regions. The p53 DNA binding domain is represented in ribbon and colored orange. The binding regions of the p53 interacting proteins are shown in ball representation; proteins interacting through the same p53 region are depicted using the same color. For example, Chk1, Crk, Cdk2 interact with p53 through the same region. They are colored red and their binding sites are colored yellow. Their interaction with p53 is represented by yellow edge.

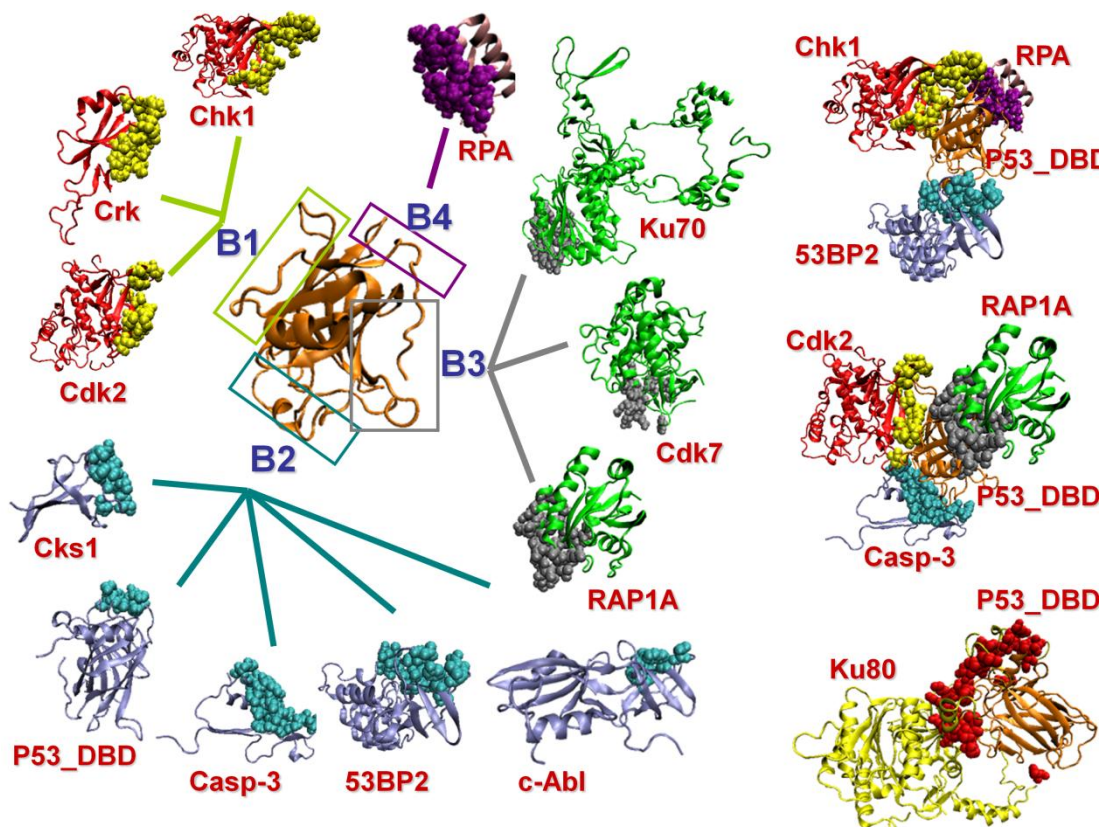


Figure 6.7 Predicted partners of the p53 DNA-binding domain (left panel), with representation of some in the complexed state (right panel).

The first binding site of p53 is the region (B1) where Cdk2, Crk and Chk1 bind. These proteins are PRISM-predicted p53 interaction partners. Cdk2 and Chk1 are members of the kinase family. Crk is a proto-onco protein and its SH2 domain is available in the target set (1ju5:A). The phosphorylation site of p53 is located at the transactivation- and regulatory

domains [235-237]. The catalytic site of the Cdk2 contains the residues Asp127, Lys129, Gln131, Asn132, Asp145 and Thr165. Also, the catalytic site of Chk1 contains Asp130, Asp132, Asn135 and Thr170. Here, we predict that Cdk2 and Chk1, two kinases, bind to DBD of p53 using a region different from their catalytic sites. The predicted binding site of Cdk2 overlaps partially with the region where Cdk2 binds to CycA. The template interface for these interactions is the obligate interface formed between two oligomerization domains of the arginine repressor (1b4bAC) which binds to DNA [238] with a transcription factor activity protein. Since Cdk2, Crk and Chk1 share the same binding site they are mutually exclusive and cannot interact simultaneously. When we compare their chemical contacts using MAPPIS (for details of MAPPIS see [239]), we observe seven structurally conserved contacts. These are illustrated in **Figure 6.8**. Thus, even though the partners are different, their putative interactions with p53 are conserved structurally and chemically [47]. NOXclass labels these interactions as biological and non-obligate, in agreement with the characteristics of mutually exclusive interactions.

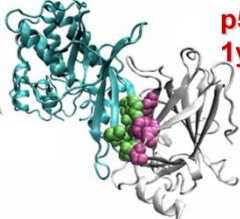
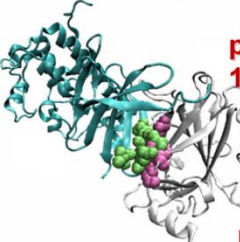

Structure		Conserved Contacts	
 <p>Cdk2 1jsu:A</p> <p>p53_DBD 1ycs:A</p>	CDK2		P53
	Val69 – Acc	↔	His233 – Dac
	Ile70 – Ali	↔	Thr231 – Ali
	Ile70 – Ali	↔	Leu114 – Ali
	Ile70 – Ali	↔	Pro142 – Ali
	His71 – Dac	↔	Thr140 – Dac
	Glu73 – Acc	↔	Ser116 – Don
Lys75 – Ali	↔	Leu114 – Ali	
 <p>Chk1 1nvq:A</p> <p>p53_DBD 1ycs:A</p>	CHK1		P53
	Leu114 – Acc	↔	His233 – Dac
	Ile115 – Ali	↔	Thr231 – Ali
	Ile115 – Ali	↔	Leu114 – Ali
	Ile115 – Ali	↔	Pro142 – Ali
	Lys105 – Don	↔	Thr140 – Dac
	Arg120 – Acc	↔	Ser116 – Don
Pro117 – Ali	↔	Leu114 – Ali	
 <p>Crk 1j5u:A</p> <p>p53_DBD 1ycs:A</p>	CRK		P53
	His73 – Acc	↔	His233 – Dac
	Arg74 – Ali	↔	Thr231 – Ali
	Arg74 – Ali	↔	Leu114 – Ali
	Arg74 – Ali	↔	Pro142 – Ali
	Arg75 – Don	↔	Thr140 – Dac
	Glu76 – Acc	↔	Ser116 – Don
Ile79 – Ali	↔	Leu114 – Ali	

Figure 6.8 Conserved contacts of Cdk2, Chk1 and Crk with p53 DBD predicted with MAPPIS.

Another set of interacting proteins competing for B2, the other region of p53 DBD, are the p53 binding protein 1 (53BP1), p53 binding protein 2 (53BP2), simian virus 40 large T antigen (sv40), p53 DBD, Casp-3, Cks1, and proto-oncogene tyrosine protein kinase c-Abl (2abl). These proteins have overlapping binding sites on the DBD. The structures of the complexes of 53BP1 (1gzh:B), 53BP2 (1ycs:B), p53 DBD (2geq:A) and sv40 (2h11:A) with the p53 DBD are available in PDB. The other partners are Prism- predicted interactions. The template interface of Casp3 – p53 interaction is the non-obligate interface between the acetylcholine receptor and its inhibitor (2br8BG). This interaction between Casp3 and p53 DBD is predicted by the human protein interaction prediction server (PIPs) [240], consistent with our results.

For two interactions to be simultaneously possible, it is insufficient that binding sites should not overlap; in addition, in the multimeric state there should not be residues overlapping between the partners. In our case here, the corresponding partners of B1 (where Cdk2, Crk, and Chk1 bind) and B2 (where 53BP1, 53BP2, sv40, p53 DBD, Casp3, Cks1, and c-Abl bind) do not interpenetrate each other in their trimeric states; consequently, simultaneous interactions of these two sets are possible. In the right column of **Figure 6.7**, some predicted multimeric co-interacting states are shown. For example, the first complex illustrates the simultaneous interaction of Chk1, RPA and 53BP2 with p53. The next complex presents the simultaneous interactions of Cdk2, RAP1A and Casp3 with p53. However, we should note that all the possible complexes we list above are derived from static structures and dynamics of the proteins are not considered. Proteins undergoing minor structural changes such as side chain rotations, can avoid some overlaps.

c-Abl is a proto-onco protein, necessary for normal growth and development. C-Abl regulates several cell cycle control genes. The interaction between c-Abl and p53 enhances p21 transcription [241]. Prism predicts the Abl – p53 interaction with their binding regions. The template for this interaction is the interface between DBD of p53 and binding protein of p53 (53BP2) (1ycsAB). Both 53BP2 and Abl contain SH3 domain and Abl matches well with the 53BP2 part of the template interface. The binding site of Abl on p53 overlaps with the binding region of 53BP1, 53BP2, sv40 etc. NOXclass classifies this interaction as biological (i.e. non-crystal).

Ku80 is a repair protein which forms a heterodimer with Ku70; this heterodimer binds to broken DNA and repairs it; the Ku70-Ku80 heterodimer also has an important role in

growth regulation [242]. However, Ku70 and Ku80 also have functions independent from each other. Ku80 can move into the nucleus in its monomeric state independent from Ku70 using its own signals [243] or may transiently interact with a partner. The PRISM results lead us to propose that p53 DBD may be a potential partner of Ku80 in the nucleus. Deletion of Ku80 leads to an increase in p53-mediated DNA damage response [244]. The predicted interaction between Ku80 (1jeq:B) and p53 is shown in the lower right portion of Figure 4. This interaction is found to be 99.75% biologically relevant by NOXclass. The template interface for this putative interaction is the interface between the homodimer of DcoH protein (1dchAB). DcoH protein associates with specific DNA binding proteins. Ku80 covers just about the entire DBD surface, blocking the interaction of other proteins. Its binding region covers the B1 binding region (where Cdk2, Crk, Chk1 bind), the B2 (where 53BP1, 53BP2, sv40, p53 DBD, Casp3, Cks1, c-Abl bind) and the B3 (where Rap1A, Ku70, Cdk7 bind). Thus, while proteins interacting through B1, B2 and B3 can interact with p53 simultaneously, none of these proteins can bind p53 when Ku80 is bound. Replication protein-A (RPA) is a single-stranded DNA binding protein which has several functions in the cell and contains three subunits. RPA interacts with several transcription factors including p53. The 32 Kda subunit of RPA (1z1d:A) matches the template interface formed between the homodimer of human Flt3 ligand (1eteAB) [245].

6.2.2 Mdm2 and its Binding Partners

Mdm2 is a negative regulator of p53. p53 promotes the transcription of Mdm2; in turn, Mdm2 binds to p53 and stimulates the ubiquitination of the p53 carboxy terminus, marking it for degradation. This negative feedback loop leads to oscillation in the levels of p53 and Mdm2 in the cell. Over-expression of Mdm2 leads to attenuation in the p53 response to stress signals. While the Mdm2-p53 interaction has been well studied, Mdm2 also has p53 independent functions [246] and is a multi – interface cellular hub.

Mdm2 – pRb interaction disrupts the pRb – E2F association.

Like p53, mutated retinoblastoma protein (pRb) is observed in several human cancer types. Both p53 and pRb are inactivated in human tumor cells; the loss of their functions leads to tumor formation. Several viral groups target and inactivate these two tumor suppressors. Apart from the p53 dependent functions, Mdm2 physically and functionally interacts with

pRb. Mdm2 negatively regulates pRb, similar to p53, and inhibits its regulatory growth function. pRb interacts with the transactivation domain of the E2F family transcription factors, which are important regulators of DNA synthesis and cell cycle progression, and blocks E2F dependent transcription. The Mdm2 interaction with pRb disrupts the pRb-E2F binding leading to an increase in the E2F-dependent transcription [247, 248].

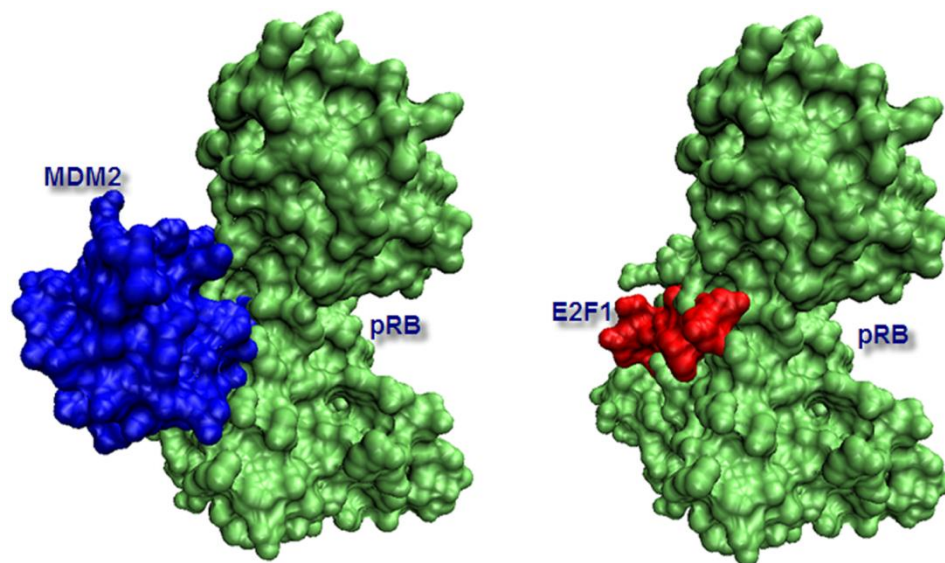


Figure 6.9 The Mdm2–pRb complex predicted by PRISM and the E2F1–pRb complex available in the PDB. Mdm2 associates with pRb through the same region where E2F1 interacts.

Using the interface of the homodimer of cytokine B10 (1o7zAB) as a template, PRISM predicts an Mdm2 – pRb interaction. The crystal structure of the pRb-E2F1 complex is available in the PDB (1n4m; the interface is labeled as 1n4mAC) [249]. The pRb region interacting with E2F1 matches an Mdm2 region, suggesting that Mdm2 binds to pRb at the same region as E2F1, thus blocking its interaction with pRb (shown in **Figure 6.9**). Consequently, the PRISM results suggest that the interactions of E2F1 and Mdm2 with pRb are mutually exclusive. Their binding sites on pRb share 16 residues, Glu533, Glu551, Glu554, His555, Ile536, Lys530, Lys537, Lys548, Lys652, Lys653, Leu649, Arg467, Arg656, Ser534, Thr645, and Val531.

p53 – Mdm2 pocket region

The transactivation domain of p53 interacts with the Swib domain of Mdm2. The interaction site is in a pocket region. When we focused on the PRISM-predicted putative

Mdm2 interaction partners in this region, we noticed that PCNA (Proliferating cell nuclear antigen), Casp3 (Caspase 3), Abl and TBP (TATA Box Binding Protein) all bind to Mdm2, blocking its pocket region. Thus, these proteins may compete with p53. **Figure 6.10** part A illustrates the interaction between Mdm2 and the transactivation domain of p53 (PDB code: 1ycr; 1ycrAB is the PRISM labeled interface). The predicted binding sites of PCNA, Casp3, Abl and TBP are shown in part B. This figure clearly illustrates that these putative exclusive interactions occur in the pocket region of Mdm2, just where p53 binds.

From experimental studies we know that Abl neutralizes the Mdm2-mediated degradation of p53. Abl binds to p53, enhances its transcriptional activity thus allowing p53 to overcome Mdm2-mediated degradation. Abl interacts with Mdm2 *in vivo* and *in vitro*. This interaction can occur in multiple Mdm2 sites [250]. Here, we predict that Abl binds to Mdm2 using the Swib domain of Mdm2. This region also corresponds to the pocket region where Mdm2 binds to p53. This scenario may be a mechanism to block the Mdm2-p53 interaction decreasing Mdm2-dependent degradation of p53.

TBP is a transcription factor which binds to specific, TATA box regions of DNA. TBP is predicted as a possible partner of Mdm2 protein by PRISM. The predicted binding region is illustrated in **Figure 6.10**. This putative interaction is also available in the IntAct database [33].

Mdm2 – Multi-interface Hub Protein

Mdm2 is a hub protein with multiple binding partners interacting at different binding sites. Prism points out two distinct binding sites in the Swib domain of Mdm2. The first is detailed above. The second predicted binding site is illustrated in **Figure 6.11**, where the left panel depicts the binding partners with their predicted binding sites. Among these, the interaction between PCAF and Mdm2 is verified in the literature. Mdm2 interacts with PCAF both *in vivo* and *in vitro* [251].

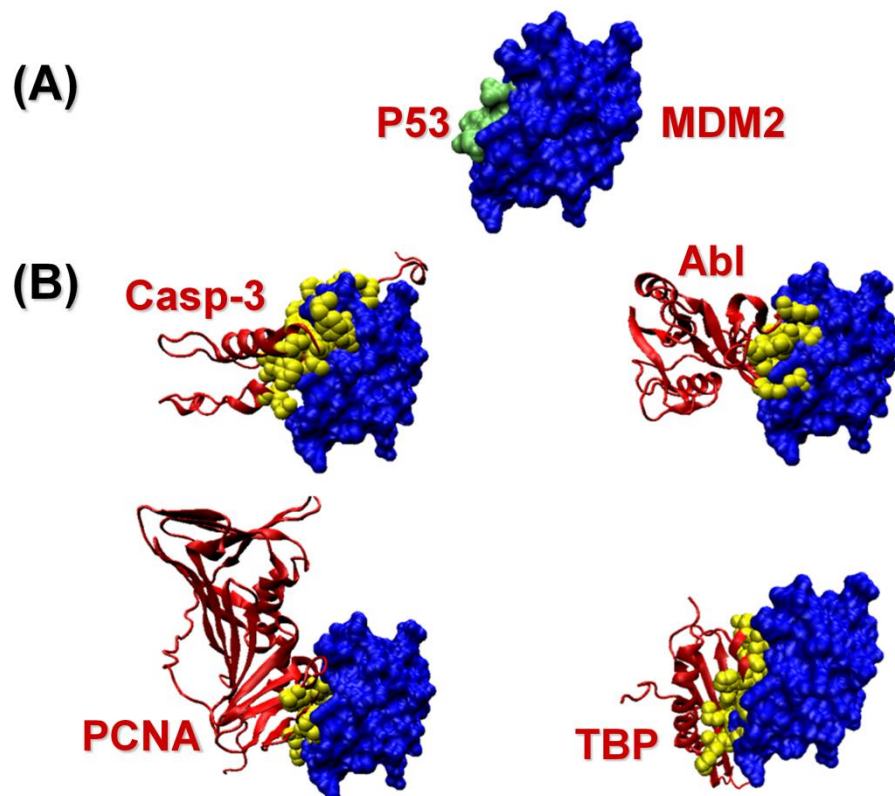


Figure 6.10 Predicted partners interacting at the pocket region of Mdm2. (A) Mdm2–p53 complex taken from PDB. (B) Possible partners predicted by PRISM.

Skp2 is also an E3 ligase like Mdm2. In several tumors, the expression levels of these two ligases are very high. However, inhibition of these ligases has more severe results in tumors [252]. Mdm2 displaces Skp2 and in this way the ubiquitination of the transcription factor E2F1 is inhibited [253]. The interaction between Skp2 and Mdm2 has also been validated experimentally. PRISM proposes a putative interaction between these two E3 ligases, Mdm2 and Skp2, using the template interface between the homodimer of human Flt3 ligand (1eteAB). The right panel shows some combination of interactions which can occur simultaneously; the Abl-Skp2-Mdm2, TBP-Ku80-Mdm2 and PCNA-PCAF-Mdm2 complexes are illustrated in their trimeric states. Here, none of the three proteins in each complex overlaps each other. If we find two proteins binding at different interfaces, microarray data can help in determining if they actually bind at the same time by looking at the correlation of their expression patterns. If their expression is correlated, most likely these two interactions can occur simultaneously.

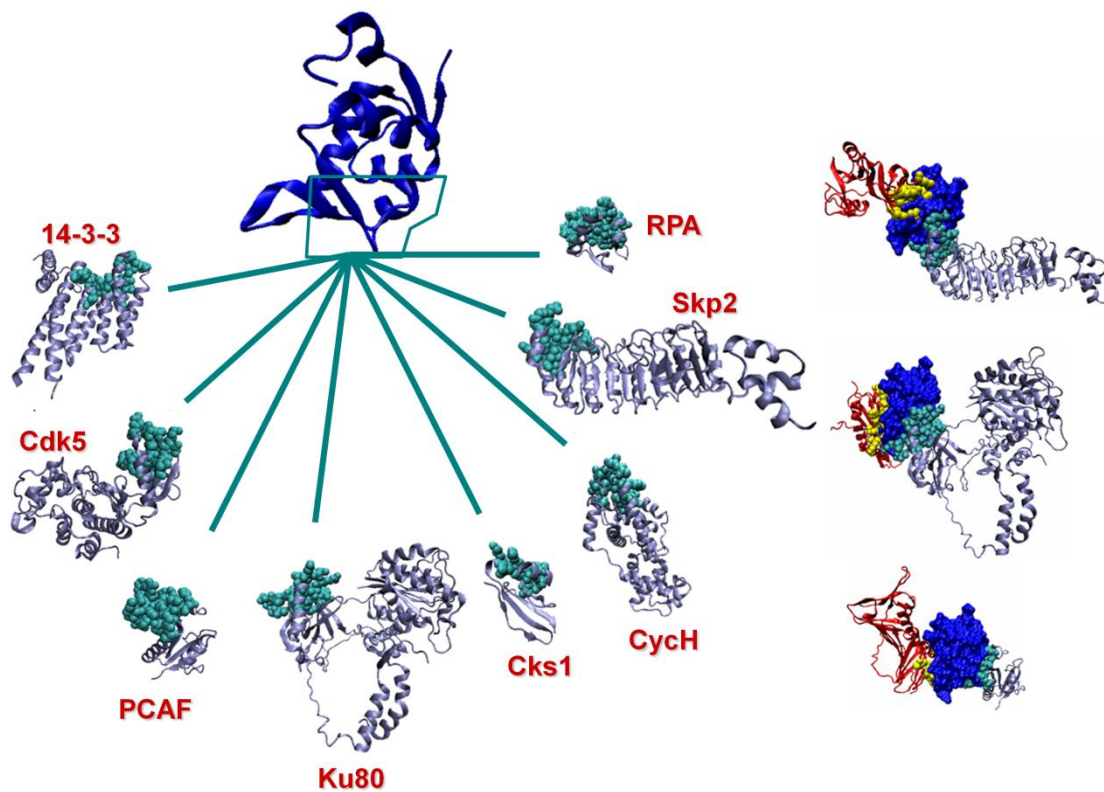


Figure 6.11 Predicted partners of the Swib domain of Mdm2 (left panel) and representation of some of them in the complex state (right panel).

6.3 Concluding Remarks

Here, we present a concept integrating time into protein interaction networks using three-dimensional protein structures and interfaces. The concept is illustrated by the multi-partner proteins available in PDB and two hub proteins, p53 and Mdm2. To figure out and characterize the interactions of the p53 and Mdm2, the PRISM server is used [11, 13, 15]. For the p53, we predict four distinct binding sites on the DNA binding domain. These sites are utilized to bind to at least 12 different proteins. Some of these interactions can occur at the same time while some others cannot. For Mdm2, we propose two binding regions. We believe that such a strategy should be immensely useful in the actual comprehension of the regulation of cellular processes beyond the common node-and-edge network picture.

Chapter 7

CONCLUSION

As implied in the previous chapters, this work is a collage of the studies related to multi-scale analysis of protein-protein interactions. The “multi-scale” term implies both molecular-level and proteome level investigations. For the molecular-level analysis, studies related to hot spots are presented. Hot spots are residues comprising only a small fraction of interfaces yet accounting for the majority of the binding energy. Hot spots are the targets of pharmaceutical agents and they are crucial for interaction specificity. In this dissertation, we introduced an efficient, intuitive model to predict computational hot spots in protein interfaces. This method uses two features for prediction, namely solvent accessibility and total knowledge-based pair potentials of the residues. It reaches an accuracy of 70% on an independent test set. When compared to other prediction methods, its performance exceeds all of them. Also, the case studies show the predictions. The hot spot analysis gives us the strength to incorporate the evolutionary data into protein modeling.

We believe that the results provide insights for researchers working on characterization of protein interaction sites. Such studies provide insights for function when clear evolutionary structural relationship between the sequences being compared exists and insights into what residues are most important in defining particular protein interface signatures. Also, with its simple architecture and visualization tool, HotPoint would be useful both for the experimentalists and computational scientist working on protein recognition, modeling of protein complexes and drug design.

In the future, the cooperative behavior of the hot spots can be examined using multiple point mutations. If we assume the protein interactions as therapeutic targets, the effect of a therapeutic agent on a cooperative hot region will be amplified when compared to an additive one. In Chapter 6, the hot spot distribution and conserved contacts are illustrated just for some case studies. In the future, further, hot spot organization for multi-partner proteins can be analyzed more systematically at large-scale.

For the proteome-level studies, we introduced our combinatorial approach to effectively predict the functional associations of proteins. This approach relies on the expectation that the number of protein-protein interface architectures in nature is limited; thus extrapolation of the known architecture space on target protein surfaces may help to identify protein interactions. This knowledge-based approach is made more physical by combining with docking strategies, flexible refinement of the solutions and energy calculations to rank them. We show how available structural information can help in modeling a pathway by using structural similarity *independent* from global fold homology and how such a knowledge-based approach can be enhanced by filters, flexibility and energy calculations. To show its functionality, we modeled the molecular interaction map of the p53 pathway. The nucleotide excision repair (NER) subsystem, cyclin-dependent kinase subsystem and some promising putative functional associations are illustrated in the p53 pathway.

Our strategy models the proteome based on local, sequence-order-independent homology to each side of an interface whose structure has been determined experimentally. As in any strategy which is motif-based there *must* be a similar motif in the template set. If there is no such motif, we cannot expect the method to find it. As in homology modeling, or threading of protein chains, a motif-based strategy is an advantage and a disadvantage: the advantage is that if such motif is available, the method is fast and reliable. At the same time it is a disadvantage, since the outcome depends on the presence of the motif in the template set. We expect that with the fast growth of the PDB the number of distinct interface motifs will grow making such fast strategies increasingly popular and useful for the modeling of protein interactions. The structural modeling of the interactions also let us to analyze the protein interactions from a different perspective where the time notion is implicitly integrated.

As a future work, for the modeling part, improvements and optimization in the template set is crucial. We know that for different types of interactions the characteristics are changing. For example, signaling proteins have smaller binding area, while enzyme-inhibitor interactions have higher affinity when compared to others and antibody-antigen interactions are promiscuous. So, target specific template dataset construction is a prerequisite for a future work. The classification of template dataset according to the interaction type, namely, enzyme-inhibitor complexes, antibody-antigen complexes, signaling proteins, cytokines will help to improve the prediction quality.

In overall, structural analysis of protein interactions both at molecular level and at proteome level where prediction and organization of hot spots in protein interfaces and modeling of protein complexes towards construction of structural protein interactions network at large scale are studied in this dissertation and substantial information is gained towards addressing the question *how do protein interactions take place?* These studies here will serve to functional and structural genomics, drug design and pathway analysis at the top level.

Appendix A

APPENDIX

A.1. Webservers, Softwares, Tools, Databases

A.1.1. NACCESS

Naccess is a program used for the calculation of the accessible surface areas of the molecules. It basely rolls a solvent probe on the desired molecule. The radius of the solvent can be chosen by the user, but the default value is 1.4 Å. The path gained by the center of the probe gives the accessible surface area. Naccess takes files in PDB format as input. Besides the accessible surface area, the output file of the Naccess gives also relative accessible area for each individual residue. Relative accessibility can be described as the percent accessibility of a residue relative to the accessibility of it in the tripeptide ALA-X-ALA. Generally if this value is larger than 5% then, this residue is identified as surface residue [15, 162]. In this work, we used Naccess with default values to calculate ASA [175].

A.1.2. MULTIPROT

Multiprot is fully automated software which identifies multiple structural alignments of a given set of protein structures. Structural alignment method is based on the Geometric Hashing Algorithm which detects common parts of the given structures in all possible ways. This is a sequence-order and directionality independent algorithm. Multiprot considers only C^α atoms. In the output file, the matched residue pairs, number of them and the RMSD value between these residues are present. The algorithm does not force all residues to participate in the alignment; on the contrary, it searches the best scored partial alignment for the given structures. In parameters file, by changing parameter user can change the alignment conditions. Its sequence order independent feature makes Multiprot appropriate for protein interfaces analysis. Multiprot is used both in clustering part and in cluster type separation part [208].

A.1.3. ClustalW

ClustalW is a multiple sequence alignment program for protein or DNA sequences. As input sequence information of the desired proteins or DNAs are given and in the output the multiple sequence alignment of these structures are produced by the program. It calculates the best match and shows the similarities, differences and identities. In global alignments overall sequences are aligned by using gaps. In local alignments, only particular regions are aligned to each other. ClustalW uses global alignment for multiple sequence alignment. It has some options like input file format, substitution matrix preference, etc. In the output, besides the multiple sequence alignment, pairwise alignments of the sequences and their scores are also provided. Phylogenetic trees are also produced by the multiple sequence alignment [254].

A.1.4. CytoScape (network visualization and analysis)

Cytoscape is molecular interaction network visualization software which also integrates biological information such as gene expression profiles, GO annotations etc. Additional features like network analyzer, functional enrichment generator, and additional file format support can be installed as plugins. Cytoscape user can visualize the protein – protein interaction network or other networks by loading .sif file which contains pairwise interaction information. Network visualization properties such as node shape, color, edge shape, color etc. can be defined by the user. It has also various filtering and selection tools. The more, the resulting graph can be organized several layouts such as hierarchical layout, spring embedded layout, circular layout etc [255]. Here, we used Cytoscape for visualization of functional interaction network of PDB. Cytoscape is downloadable through the web page <http://www.cytoscape.org/>.

A.1.5. VMD (molecule visualization)

VMD is a molecule visualization and analysis tool. Biological systems such as proteins, nucleic acids, lipid bilayer assemblies, etc. can be visualized by the help of VMD. VMD can read standard Protein Data Bank (PDB) files and display the contained structure. It has various molecular representation methods and an advanced coloring and rendering properties. VMD can be used also to animate and analyze the trajectory of molecular

dynamics (MD) simulations, and can interactively manipulate molecules being simulated on remote computers (Interactive MD) [256].

A.1.6. FiberDock

FiberDock [144, 257] is a flexible refinement program for docking. It considers docking solution candidates. The method models both side-chain and backbone flexibility and performs rigid body optimization on the ligand orientation. The movements of the backbone and side-chain are modeled according to the binding van der Waals forces between the receptor and ligand. The method uses both low and high frequency normal modes and therefore is able to model both global and local conformational changes, such as opening of binding sites and loop movement. After refining all the docking solution candidates, the refined models are re-scored according to an energy function. FiberDock is downloadable at <http://bioinfo3d.cs.tau.ac.il/FiberDock/>.

BIBLIOGRAPHY

1. Bogan AA, Thorn KS: **Anatomy of hot spots in protein interfaces.** *J Mol Biol* 1998, **280**(1):1-9.
2. Keskin O, Gursoy A, Ma B, Nussinov R: **Principles of protein-protein interactions: what are the preferred ways for proteins to interact?** *Chem Rev* 2008, **108**(4):1225-1244.
3. Giot L, Bader JS, Brouwer C, Chaudhuri A, Kuang B, Li Y, Hao YL, Ooi CE, Godwin B, Vitols E *et al*: **A protein interaction map of *Drosophila melanogaster*.** *Science* 2003, **302**(5651):1727-1736.
4. Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, Sakaki Y: **A comprehensive two-hybrid analysis to explore the yeast protein interactome.** *Proc Natl Acad Sci U S A* 2001, **98**(8):4569-4574.
5. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W *et al*: **Initial sequencing and analysis of the human genome.** *Nature* 2001, **409**(6822):860-921.
6. Stelzl U, Worm U, Lalowski M, Haenig C, Brembeck FH, Goehler H, Stroedicke M, Zenkner M, Schoenherr A, Koeppen S *et al*: **A human protein-protein interaction network: a resource for annotating the proteome.** *Cell* 2005, **122**(6):957-968.
7. Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR, Lockshon D, Narayan V, Srinivasan M, Pochart P *et al*: **A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*.** *Nature* 2000, **403**(6770):623-627.
8. Shoemaker BA, Panchenko AR: **Deciphering protein-protein interactions. Part I. Experimental techniques and databases.** *PLoS Comput Biol* 2007, **3**(3):e42.
9. Wodak SJ, Pu S, Vlasblom J, Seraphin B: **Challenges and rewards of interaction proteomics.** *Mol Cell Proteomics* 2009, **8**(1):3-18.

10. Tuncbag N, Gursoy A, Guney E, Nussinov R, Keskin O: **Architectures and functional coverage of protein-protein interfaces.** *J Mol Biol* 2008, **381**(3):785-802.
11. Shoemaker BA, Panchenko AR: **Deciphering protein-protein interactions. Part II. Computational methods to predict protein and domain interaction partners.** *PLoS Comput Biol* 2007, **3**(4):e43.
12. Tuncbag N, Kar G, Keskin O, Gursoy A, Nussinov R: **A survey of available tools and web servers for analysis of protein-protein interactions and interfaces.** *Brief Bioinform* 2009, **10**(3):217-232.
13. Aloy P, Bottcher B, Ceulemans H, Leutwein C, Mellwig C, Fischer S, Gavin AC, Bork P, Superti-Furga G, Serrano L *et al*: **Structure-based assembly of protein complexes in yeast.** *Science* 2004, **303**(5666):2026-2029.
14. Winter C, Henschel A, Kim WK, Schroeder M: **SCOPPI: a structural classification of protein-protein interfaces.** *Nucleic Acids Res* 2006, **34**(Database issue):D310-314.
15. Gong S, Park C, Choi H, Ko J, Jang I, Lee J, Bolser DM, Oh D, Kim DS, Bhak J: **A protein domain interaction interface database: InterPare.** *BMC Bioinformatics* 2005, **6**:207.
16. Stein A, Panjkovich A, Aloy P: **3did Update: domain-domain and peptide-mediated interactions of known 3D structure.** *Nucleic Acids Res* 2009, **37**(Database issue):D300-304.
17. Stein A, Russell RB, Aloy P: **3did: interacting protein domains of known three-dimensional structure.** *Nucleic Acids Res* 2005, **33**(Database issue):D413-417.
18. Davis FP, Sali A: **PIBASE: a comprehensive database of structurally defined protein interfaces.** *Bioinformatics* 2005, **21**(9):1901-1907.
19. Kundrotas PJ, Alexov E: **PROTCOM: searchable database of protein complexes enhanced with domain-domain structures.** *Nucleic Acids Res* 2007, **35**(Database issue):D575-579.
20. Keskin O, Tsai CJ, Wolfson H, Nussinov R: **A new, structurally nonredundant, diverse data set of protein-protein interfaces and its implications.** *Protein Sci* 2004, **13**(4):1043-1055.

21. Chakrabarti P, Janin J: **Dissecting protein-protein recognition sites.** *Proteins* 2002, **47**(3):334-343.
22. Jones S, Thornton JM: **Principles of protein-protein interactions.** *Proc Natl Acad Sci U S A* 1996, **93**(1):13-20.
23. Jones S, Thornton JM: **Analysis of protein-protein interaction sites using surface patches.** *J Mol Biol* 1997, **272**(1):121-132.
24. Li X, Keskin O, Ma B, Nussinov R, Liang J: **Protein-protein interactions: hot spots and structurally conserved residues often locate in complemented pockets that pre-organized in the unbound states: implications for docking.** *J Mol Biol* 2004, **344**(3):781-795.
25. Tsai CJ, Lin SL, Wolfson HJ, Nussinov R: **Studies of protein-protein interfaces: a statistical analysis of the hydrophobic effect.** *Protein Sci* 1997, **6**(1):53-64.
26. Tsai CJ, Xu D, Nussinov R: **Structural motifs at protein-protein interfaces: protein cores versus two-state and three-state model complexes.** *Protein Sci* 1997, **6**(9):1793-1805.
27. Landgraf C, Panni S, Montecchi-Palazzi L, Castagnoli L, Schneider-Mergener J, Volkmer-Engert R, Cesareni G: **Protein interaction networks by proteome peptide scanning.** *PLoS Biol* 2004, **2**(1):E14.
28. MacBeath G, Schreiber SL: **Printing proteins as microarrays for high-throughput function determination.** *Science* 2000, **289**(5485):1760-1763.
29. Bauer A, Kuster B: **Affinity purification-mass spectrometry. Powerful tools for the characterization of protein complexes.** *Eur J Biochem* 2003, **270**(4):570-578.
30. Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU, Eisenberg D: **The Database of Interacting Proteins: 2004 update.** *Nucleic Acids Res* 2004, **32**(Database issue):D449-451.
31. Chatr-aryamontri A, Ceol A, Palazzi LM, Nardelli G, Schneider MV, Castagnoli L, Cesareni G: **MINT: the Molecular INTeraction database.** *Nucleic Acids Res* 2007, **35**(Database issue):D572-574.
32. Bader GD, Betel D, Hogue CW: **BIND: the Biomolecular Interaction Network Database.** *Nucleic Acids Res* 2003, **31**(1):248-250.
33. Kerrien S, Alam-Faruque Y, Aranda B, Bancarz I, Bridge A, Derow C, Dimmer E, Feuermann M, Friedrichsen A, Huntley R *et al*: **IntAct--open source resource for**

- molecular interaction data.** *Nucleic Acids Res* 2007, **35**(Database issue):D561-565.
34. von Mering C, Krause R, Snel B, Cornell M, Oliver SG, Fields S, Bork P: **Comparative assessment of large-scale data sets of protein-protein interactions.** *Nature* 2002, **417**(6887):399-403.
35. Deane CM, Salwinski L, Xenarios I, Eisenberg D: **Protein interactions: two methods for assessment of the reliability of high throughput observations.** *Mol Cell Proteomics* 2002, **1**(5):349-356.
36. Zuiderweg ER: **Mapping protein-protein interactions in solution by NMR spectroscopy.** *Biochemistry* 2002, **41**(1):1-7.
37. Henderson R: **Realizing the potential of electron cryo-microscopy.** *Q Rev Biophys* 2004, **37**(1):3-13.
38. Lasker K, Phillips JL, Russel D, Velazquez-Muriel J, Schneidman-Duhovny D, Tjioe E, Webb B, Schlessinger A, Sali A: **Integrative structure modeling of macromolecular assemblies from proteomics data.** *Mol Cell Proteomics* 2010, **9**(8):1689-1702.
39. Robinson CV, Sali A, Baumeister W: **The molecular sociology of the cell.** *Nature* 2007, **450**(7172):973-982.
40. Argos P: **An investigation of protein subunit and domain interfaces.** *Protein Eng* 1988, **2**(2):101-113.
41. Chothia C, Janin J: **Principles of protein-protein recognition.** *Nature* 1975, **256**(5520):705-708.
42. Clackson T, Wells JA: **A hot spot of binding energy in a hormone-receptor interface.** *Science* 1995, **267**(5196):383-386.
43. Janin J, Chothia C: **The structure of protein-protein recognition sites.** *J Biol Chem* 1990, **265**(27):16027-16030.
44. 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**(5):1281-1294.
45. Keskin O, Ma B, Rogale K, Gunasekaran K, Nussinov R: **Protein-protein interactions: organization, cooperativity and mapping in a bottom-up Systems Biology approach.** *Phys Biol* 2005, **2**(2):S24-35.

46. Keskin O, Nussinov R: **Favorable scaffolds: proteins with different sequence, structure and function may associate in similar ways.** *Protein Eng Des Sel* 2005, **18**(1):11-24.
47. Keskin O, Nussinov R: **Similar binding sites and different partners: implications to shared proteins in cellular pathways.** *Structure* 2007, **15**(3):341-354.
48. Korn AP, Burnett RM: **Distribution and complementarity of hydrophathy in multisubunit proteins.** *Proteins* 1991, **9**(1):37-55.
49. Lijnzaad P, Berendsen HJ, Argos P: **Hydrophobic patches on the surfaces of protein structures.** *Proteins* 1996, **25**(3):389-397.
50. Lo Conte L, Chothia C, Janin J: **The atomic structure of protein-protein recognition sites.** *J Mol Biol* 1999, **285**(5):2177-2198.
51. Sheinerman FB, Norel R, Honig B: **Electrostatic aspects of protein-protein interactions.** *Curr Opin Struct Biol* 2000, **10**(2):153-159.
52. Tsai CJ, Lin SL, Wolfson HJ, Nussinov R: **A dataset of protein-protein interfaces generated with a sequence-order-independent comparison technique.** *J Mol Biol* 1996, **260**(4):604-620.
53. Xu W, Regnier FE: **Protein-protein interactions on weak-cation-exchange sorbent surfaces during chromatographic separations.** *J Chromatogr A* 1998, **828**(1-2):357-364.
54. Nooren IM, Thornton JM: **Diversity of protein-protein interactions.** *Embo J* 2003, **22**(14):3486-3492.
55. Neuvirth H, Raz R, Schreiber G: **ProMate: a structure based prediction program to identify the location of protein-protein binding sites.** *J Mol Biol* 2004, **338**(1):181-199.
56. Vakser IA: **Protein-protein interfaces are special.** *Structure* 2004, **12**(6):910-912.
57. Nooren IM, Thornton JM: **Structural characterisation and functional significance of transient protein-protein interactions.** *J Mol Biol* 2003, **325**(5):991-1018.
58. Mintseris J, Weng Z: **Structure, function, and evolution of transient and obligate protein-protein interactions.** *Proc Natl Acad Sci U S A* 2005, **102**(31):10930-10935.

59. Lawrence MC, Colman PM: **Shape complementarity at protein/protein interfaces.** *J Mol Biol* 1993, **234**(4):946-950.
60. Sheinerman FB, Honig B: **On the role of electrostatic interactions in the design of protein-protein interfaces.** *J Mol Biol* 2002, **318**(1):161-177.
61. Jones S, Thornton JM: **Prediction of protein-protein interaction sites using patch analysis.** *J Mol Biol* 1997, **272**(1):133-143.
62. Bahadur RP, Chakrabarti P, Rodier F, Janin J: **Dissecting subunit interfaces in homodimeric proteins.** *Proteins* 2003, **53**(3):708-719.
63. Larsen TA, Olson AJ, Goodsell DS: **Morphology of protein-protein interfaces.** *Structure* 1998, **6**(4):421-427.
64. Ofraan Y, Rost B: **Analysing six types of protein-protein interfaces.** *J Mol Biol* 2003, **325**(2):377-387.
65. Valdar WS, Thornton JM: **Protein-protein interfaces: analysis of amino acid conservation in homodimers.** *Proteins* 2001, **42**(1):108-124.
66. Caffrey DR, Somaroo S, Hughes JD, Mintseris J, Huang ES: **Are protein-protein interfaces more conserved in sequence than the rest of the protein surface?** *Protein Sci* 2004, **13**(1):190-202.
67. Cole C, Warwicker J: **Side-chain conformational entropy at protein-protein interfaces.** *Protein Sci* 2002, **11**(12):2860-2870.
68. Barabasi AL, Albert R: **Emergence of scaling in random networks.** *Science* 1999, **286**(5439):509-512.
69. Han JD, Bertin N, Hao T, Goldberg DS, Berriz GF, Zhang LV, Dupuy D, Walhout AJ, Cusick ME, Roth FP *et al*: **Evidence for dynamically organized modularity in the yeast protein-protein interaction network.** *Nature* 2004, **430**(6995):88-93.
70. Yook SH, Oltvai ZN, Barabasi AL: **Functional and topological characterization of protein interaction networks.** *Proteomics* 2004, **4**(4):928-942.
71. Jeong H, Mason SP, Barabasi AL, Oltvai ZN: **Lethality and centrality in protein networks.** *Nature* 2001, **411**(6833):41-42.
72. Gerdes SY, Scholle MD, Campbell JW, Balazsi G, Ravasz E, Daugherty MD, Somera AL, Kyrpides NC, Anderson I, Gelfand MS *et al*: **Experimental determination and system level analysis of essential genes in Escherichia coli MG1655.** *J Bacteriol* 2003, **185**(19):5673-5684.

73. Kim WK, Henschel A, Winter C, Schroeder M: **The many faces of protein-protein interactions: A compendium of interface geometry.** *PLoS Comput Biol* 2006, **2**(9):e124.
74. Aloy P, Ceulemans H, Stark A, Russell RB: **The relationship between sequence and interaction divergence in proteins.** *J Mol Biol* 2003, **332**(5):989-998.
75. Kim PM, Lu LJ, Xia Y, Gerstein MB: **Relating three-dimensional structures to protein networks provides evolutionary insights.** *Science* 2006, **314**(5807):1938-1941.
76. Bullock AN, Debreczeni JE, Edwards AM, Sundstrom M, Knapp S: **Crystal structure of the SOCS2-elongin C-elongin B complex defines a prototypical SOCS box ubiquitin ligase.** *Proc Natl Acad Sci U S A* 2006, **103**(20):7637-7642.
77. Kamura T, Maenaka K, Kotoshiba S, Matsumoto M, Kohda D, Conaway RC, Conaway JW, Nakayama KI: **VHL-box and SOCS-box domains determine binding specificity for Cul2-Rbx1 and Cul5-Rbx2 modules of ubiquitin ligases.** *Genes Dev* 2004, **18**(24):3055-3065.
78. Beckett D: **Functional switches in transcription regulation; molecular mimicry and plasticity in protein-protein interactions.** *Biochemistry* 2004, **43**(25):7983-7991.
79. Valdar WS, Thornton JM: **Conservation helps to identify biologically relevant crystal contacts.** *J Mol Biol* 2001, **313**(2):399-416.
80. Elcock AH, McCammon JA: **Identification of protein oligomerization states by analysis of interface conservation.** *Proc Natl Acad Sci U S A* 2001, **98**(6):2990-2994.
81. Mintseris J, Weng Z: **Atomic contact vectors in protein-protein recognition.** *Proteins* 2003, **53**(3):629-639.
82. Henrick K, Thornton JM: **PQS: a protein quaternary structure file server.** *Trends Biochem Sci* 1998, **23**(9):358-361.
83. Shoemaker BA, Panchenko AR, Bryant SH: **Finding biologically relevant protein domain interactions: conserved binding mode analysis.** *Protein Sci* 2006, **15**(2):352-361.
84. Carugo O, Argos P: **Protein-protein crystal-packing contacts.** *Protein Sci* 1997, **6**(10):2261-2263.

85. Zhu H, Domingues FS, Sommer I, Lengauer T: **NOXclass: prediction of protein-protein interaction types.** *BMC Bioinformatics* 2006, **7**:27.
86. Liu Q, Li J: **Propensity vectors of low-ASA residue pairs in the distinction of protein interactions.** *Proteins*, **78**(3):589-602.
87. Zor T, De Guzman RN, Dyson HJ, Wright PE: **Solution structure of the KIX domain of CBP bound to the transactivation domain of c-Myb.** *J Mol Biol* 2004, **337**(3):521-534.
88. Klemm JD, Pabo CO: **Oct-1 POU domain-DNA interactions: cooperative binding of isolated subdomains and effects of covalent linkage.** *Genes Dev* 1996, **10**(1):27-36.
89. Moza B, Buonpane RA, Zhu P, Herfst CA, Rahman AK, McCormick JK, Kranz DM, Sundberg EJ: **Long-range cooperative binding effects in a T cell receptor variable domain.** *Proc Natl Acad Sci U S A* 2006, **103**(26):9867-9872.
90. Monod J, Changeux JP, Jacob F: **Allosteric proteins and cellular control systems.** *J Mol Biol* 1963, **6**:306-329.
91. Kuriyan J, Eisenberg D: **The origin of protein interactions and allostery in colocalization.** *Nature* 2007, **450**(7172):983-990.
92. Gunasekaran K, Ma B, Nussinov R: **Is allostery an intrinsic property of all dynamic proteins?** *Proteins* 2004, **57**(3):433-443.
93. Lockless SW, Ranganathan R: **Evolutionarily conserved pathways of energetic connectivity in protein families.** *Science* 1999, **286**(5438):295-299.
94. Chi CN, Elfstrom L, Shi Y, Snall T, Engstrom A, Jemth P: **Reassessing a sparse energetic network within a single protein domain.** *Proc Natl Acad Sci U S A* 2008, **105**(12):4679-4684.
95. Del Sol A, Arauzo-Bravo MJ, Amoros D, Nussinov R: **Modular architecture of protein structures and allosteric communications: potential implications for signaling proteins and regulatory linkages.** *Genome Biol* 2007, **8**(5):R92.
96. Thorn KS, Bogan AA: **ASEdb: a database of alanine mutations and their effects on the free energy of binding in protein interactions.** *Bioinformatics* 2001, **17**(3):284-285.
97. Fischer TB, Arunachalam KV, Bailey D, Mangual V, Bakhru S, Russo R, Huang D, Paczkowski M, Lalchandani V, Ramachandra C *et al*: **The binding interface**

- database (BID): a compilation of amino acid hot spots in protein interfaces.** *Bioinformatics* 2003, **19**(11):1453-1454.
98. Ofraan Y, Rost B: **Protein-Protein Interaction Hotspots Carved into Sequences.** *PLoS Comput Biol* 2007, **3**(7):e119.
99. Tuncbag N, Gursoy A, Keskin O: **Identification of computational hot spots in protein interfaces: combining solvent accessibility and inter-residue potentials improves the accuracy.** *Bioinformatics* 2009, **25**(12):1513-1520.
100. Burgoyne NJ, Jackson RM: **Predicting protein interaction sites: binding hotspots in protein-protein and protein-ligand interfaces.** *Bioinformatics* 2006, **22**(11):1335-1342.
101. Guharoy M, Chakrabarti P: **Conservation and relative importance of residues across protein-protein interfaces.** *Proc Natl Acad Sci U S A* 2005, **102**(43):15447-15452.
102. Moreira IS, Fernandes PA, Ramos MJ: **Hot spots--a review of the protein-protein interface determinant amino-acid residues.** *Proteins* 2007, **68**(4):803-812.
103. Volkov AN, Bashir Q, Worrall JA, Ubbink M: **Binding hot spot in the weak protein complex of physiological redox partners yeast cytochrome C and cytochrome C peroxidase.** *J Mol Biol* 2009, **385**(3):1003-1013.
104. Ma B, Elkayam T, Wolfson H, Nussinov R: **Protein-protein interactions: structurally conserved residues distinguish between binding sites and exposed protein surfaces.** *Proc Natl Acad Sci U S A* 2003, **100**(10):5772-5777.
105. Smith GR, Sternberg MJ, Bates PA: **The relationship between the flexibility of proteins and their conformational states on forming protein-protein complexes with an application to protein-protein docking.** *J Mol Biol* 2005, **347**(5):1077-1101.
106. Rajamani D, Thiel S, Vajda S, Camacho CJ: **Anchor residues in protein-protein interactions.** *Proc Natl Acad Sci U S A* 2004, **101**(31):11287-11292.
107. Yogurtcu ON, Erdemli B, Turkay M, Nussinov R, Keskin O: **Restricted mobility of interface residues.** *Biophysical J* 2008.
108. DeLano WL: **Unraveling hot spots in binding interfaces: progress and challenges.** *Curr Opin Struct Biol* 2002, **12**(1):14-20.

109. Guerois R, Nielsen JE, Serrano L: **Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations.** *J Mol Biol* 2002, **320**(2):369-387.
110. Kortemme T, Kim DE, Baker D: **Computational alanine scanning of protein-protein interfaces.** *Sci STKE* 2004, **2004**(219):pl2.
111. Darnell SJ, Page D, Mitchell JC: **An automated decision-tree approach to predicting protein interaction hot spots.** *Proteins* 2007, **68**(4):813-823.
112. Guney E, Tuncbag N, Keskin O, Gursoy A: **HotSprint: database of computational hot spots in protein interfaces.** *Nucleic Acids Res* 2008, **36**(Database issue):D662-666.
113. Lise S, Archambeau C, Pontil M, Jones DT: **Prediction of hot spot residues at protein-protein interfaces by combining machine learning and energy-based methods.** *BMC Bioinformatics* 2009, **10**:365.
114. Gao Y, Wang R, Lai L: **Structure-based method for analyzing protein-protein interfaces.** *J Mol Model* 2004, **10**(1):44-54.
115. Assi SA, Tanaka T, Rabbitts TH, Fernandez-Fuentes N: **PCRPI: Presaging Critical Residues in Protein interfaces, a new computational tool to chart hot spots in protein interfaces.** *Nucleic Acids Res* 2010, **38**(6):e86.
116. Cho KI, Kim D, Lee D: **A feature-based approach to modeling protein-protein interaction hot spots.** *Nucleic Acids Res* 2009, **37**(8):2672-2687.
117. Gonzalez-Ruiz D, Gohlke H: **Targeting protein-protein interactions with small molecules: challenges and perspectives for computational binding epitope detection and ligand finding.** *Curr Med Chem* 2006, **13**(22):2607-2625.
118. Huo S, Massova I, Kollman PA: **Computational alanine scanning of the 1:1 human growth hormone-receptor complex.** *J Comput Chem* 2002, **23**(1):15-27.
119. Brinda KV, Kannan N, Vishveshwara S: **Analysis of homodimeric protein interfaces by graph-spectral methods.** *Protein Eng* 2002, **15**(4):265-277.
120. del Sol A, O'Meara P: **Small-world network approach to identify key residues in protein-protein interaction.** *Proteins* 2005, **58**(3):672-682.
121. Kortemme T, Baker D: **A simple physical model for binding energy hot spots in protein-protein complexes.** *Proc Natl Acad Sci U S A* 2002, **99**(22):14116-14121.

122. Grosdidier S, Fernandez-Recio J: **Identification of hot-spot residues in protein-protein interactions by computational docking.** *BMC Bioinformatics* 2008, **9**:447.
123. Haliloglu T, Seyrek E, Erman B: **Prediction of binding sites in receptor-ligand complexes with the Gaussian Network Model.** *Phys Rev Lett* 2008, **100**(22):228102.
124. Reichmann D, Rahat O, Albeck S, Meged R, Dym O, Schreiber G: **The modular architecture of protein-protein binding interfaces.** *Proc Natl Acad Sci U S A* 2005, **102**(1):57-62.
125. Carbonell P, Nussinov R, del Sol A: **Energetic determinants of protein binding specificity: insights into protein interaction networks.** *Proteomics* 2009, **9**(7):1744-1753.
126. Cukuroglu E, Gursoy A, Keskin O: **Analysis of hot region organization in hub proteins.** *Ann Biomed Eng*, **38**(6):2068-2078.
127. Andrusier N, Mashiach E, Nussinov R, Wolfson HJ: **Principles of flexible protein-protein docking.** *Proteins* 2008, **73**(2):271-289.
128. Gray JJ: **High-resolution protein-protein docking.** *Curr Opin Struct Biol* 2006, **16**(2):183-193.
129. Halperin I, Ma B, Wolfson H, Nussinov R: **Principles of docking: An overview of search algorithms and a guide to scoring functions.** *Proteins* 2002, **47**(4):409-443.
130. de Vries SJ, van Dijk M, Bonvin AM: **The HADDOCK web server for data-driven biomolecular docking.** *Nat Protoc* 2010, **5**(5):883-897.
131. Lesk VI, Sternberg MJ: **3D-Garden: a system for modelling protein-protein complexes based on conformational refinement of ensembles generated with the marching cubes algorithm.** *Bioinformatics* 2008, **24**(9):1137-1144.
132. Cheng TM, Blundell TL, Fernandez-Recio J: **pyDock: electrostatics and desolvation for effective scoring of rigid-body protein-protein docking.** *Proteins* 2007, **68**(2):503-515.
133. Moont G, Gabb HA, Sternberg MJ: **Use of pair potentials across protein interfaces in screening predicted docked complexes.** *Proteins* 1999, **35**(3):364-373.

134. Janin J, Henrick K, Moult J, Eyck LT, Sternberg MJ, Vajda S, Vakser I, Wodak SJ: **CAPRI: a Critical Assessment of PRedicted Interactions**. *Proteins* 2003, **52**(1):2-9.
135. Wodak SJ, Mendez R: **Prediction of protein-protein interactions: the CAPRI experiment, its evaluation and implications**. *Curr Opin Struct Biol* 2004, **14**(2):242-249.
136. Mendez R, Leplae R, Lensink MF, Wodak SJ: **Assessment of CAPRI predictions in rounds 3-5 shows progress in docking procedures**. *Proteins* 2005, **60**(2):150-169.
137. Hwang H, Pierce B, Mintseris J, Janin J, Weng Z: **Protein-protein docking benchmark version 3.0**. *Proteins* 2008, **73**(3):705-709.
138. Hwang H, Vreven T, Janin J, Weng Z: **Protein-protein docking benchmark version 4.0**. *Proteins* 2010.
139. Janin J: **Protein-protein docking tested in blind predictions: the CAPRI experiment**. *Mol Biosyst* 2010.
140. Kastritis PL, Bonvin AM: **Are scoring functions in protein-protein docking ready to predict interactomes? Clues from a novel binding affinity benchmark**. *J Proteome Res*, **9**(5):2216-2225.
141. Mosca R, Pons C, Fernandez-Recio J, Aloy P: **Pushing structural information into the yeast interactome by high-throughput protein docking experiments**. *PLoS Comput Biol* 2009, **5**(8):e1000490.
142. Andrusier N, Nussinov R, Wolfson HJ: **FireDock: fast interaction refinement in molecular docking**. *Proteins* 2007, **69**(1):139-159.
143. Li L, Chen R, Weng Z: **RDOCK: refinement of rigid-body protein docking predictions**. *Proteins* 2003, **53**(3):693-707.
144. Mashiach E, Nussinov R, Wolfson HJ: **FiberDock: Flexible induced-fit backbone refinement in molecular docking**. *Proteins* 2010, **78**(6):1503-1519.
145. Korkin D, Davis FP, Alber F, Luong T, Shen MY, Lucic V, Kennedy MB, Sali A: **Structural modeling of protein interactions by analogy: application to PSD-95**. *PLoS Comput Biol* 2006, **2**(11):e153.
146. Huang B, Schroeder M: **Using protein binding site prediction to improve protein docking**. *Gene* 2008, **422**(1-2):14-21.

147. Chothia C: **Proteins. One thousand families for the molecular biologist.** *Nature* 1992, **357**(6379):543-544.
148. Park SY, Beel BD, Simon MI, Bilwes AM, Crane BR: **In different organisms, the mode of interaction between two signaling proteins is not necessarily conserved.** *Proc Natl Acad Sci U S A* 2004, **101**(32):11646-11651.
149. Aloy P, Russell RB: **Interrogating protein interaction networks through structural biology.** *Proc Natl Acad Sci U S A* 2002, **99**(9):5896-5901.
150. Bahar I, Jernigan RL: **Coordination geometry of nonbonded residues in globular proteins.** *Fold Des* 1996, **1**(5):357-370.
151. Keskin O, Bahar I, Badretdinov AY, Ptitsyn OB, Jernigan RL: **Empirical solvent-mediated potentials hold for both intra-molecular and inter-molecular inter-residue interactions.** *Protein Sci* 1998, **7**(12):2578-2586.
152. Aloy P, Russell RB: **InterPreTS: protein interaction prediction through tertiary structure.** *Bioinformatics* 2003, **19**(1):161-162.
153. Lu L, Lu H, Skolnick J: **MULTIPROSPECTOR: an algorithm for the prediction of protein-protein interactions by multimeric threading.** *Proteins* 2002, **49**(3):350-364.
154. Lu H, Lu L, Skolnick J: **Development of unified statistical potentials describing protein-protein interactions.** *Biophys J* 2003, **84**(3):1895-1901.
155. Lu L, Arakaki AK, Lu H, Skolnick J: **Multimeric threading-based prediction of protein-protein interactions on a genomic scale: application to the *Saccharomyces cerevisiae* proteome.** *Genome Res* 2003, **13**(6A):1146-1154.
156. Chen H, Skolnick J: **M-TASSER: an algorithm for protein quaternary structure prediction.** *Biophys J* 2008, **94**(3):918-928.
157. Launay G, Simonson T: **Homology modelling of protein-protein complexes: a simple method and its possibilities and limitations.** *BMC Bioinformatics* 2008, **9**:427.
158. Kundrotas PJ, Lensink MF, Alexov E: **Homology-based modeling of 3D structures of protein-protein complexes using alignments of modified sequence profiles.** *Int J Biol Macromol* 2008, **43**(2):198-208.
159. Kundrotas PJ, Zhu Z, Vakser IA: **GWIDD: Genome-wide protein docking database.** *Nucleic Acids Res*, **38**(Database issue):D513-517.

160. Davis FP, Braberg H, Shen MY, Pieper U, Sali A, Madhusudhan MS: **Protein complex compositions predicted by structural similarity.** *Nucleic Acids Res* 2006, **34**(10):2943-2952.
161. Davis FP, Barkan DT, Eswar N, McKerrow JH, Sali A: **Host pathogen protein interactions predicted by comparative modeling.** *Protein Sci* 2007, **16**(12):2585-2596.
162. Aytuna AS, Gursoy A, Keskin O: **Prediction of protein-protein interactions by combining structure and sequence conservation in protein interfaces.** *Bioinformatics* 2005, **21**(12):2850-2855.
163. Ogmen U, Keskin O, Aytuna AS, Nussinov R, Gursoy A: **PRISM: protein interactions by structural matching.** *Nucleic Acids Res* 2005, **33**(Web Server issue):W331-336.
164. Kar G, Gursoy A, Keskin O: **Human cancer protein-protein interaction network: a structural perspective.** *PLoS Comput Biol* 2009, **5**(12):e1000601.
165. Tuncbag N, Kar G, Gursoy A, Keskin O, Nussinov R: **Towards inferring time dimensionality in protein-protein interaction networks by integrating structures: the p53 example.** *Mol Biosyst* 2009, **5**(12):1770-1778.
166. Gunther S, May P, Hoppe A, Frommel C, Preissner R: **Docking without docking: ISEARCH--prediction of interactions using known interfaces.** *Proteins* 2007, **69**(4):839-844.
167. Jung SH, Hyun B, Jang WH, Hur HY, Han DS: **Protein complex prediction based on simultaneous protein interaction network.** *Bioinformatics* 2010, **26**(3):385-391.
168. Martin J: **Beauty is in the eye of the beholder: proteins can recognize binding sites of homologous proteins in more than one way.** *PLoS Comput Biol* 2010, **6**(6):e1000821.
169. Sinha R, Kundrotas PJ, Vakser IA: **Docking by structural similarity at protein-protein interfaces.** *Proteins.*
170. Kundrotas PJ, Vakser IA: **Accuracy of protein-protein binding sites in high-throughput template-based modeling.** *PLoS Comput Biol*, **6**(4):e1000727.

171. Shulman-Peleg A, Mintz S, Nussinov R, Wolfson H.J.: **Protein-Protein Interfaces: Recognition of Similar Spatial and Chemical Organizations.** *Algorithms in Bioinformatics* 2004:194-205.
172. Zhu H, Sommer I, Lengauer T, Domingues FS: **Alignment of non-covalent interactions at protein-protein interfaces.** *PLoS One* 2008, **3**(4):e1926.
173. Gao M, Skolnick J: **iAlign: a method for the structural comparison of protein-protein interfaces.** *Bioinformatics* 2010.
174. Wang G, Dunbrack RL, Jr.: **PISCES: a protein sequence culling server.** *Bioinformatics* 2003, **19**(12):1589-1591.
175. Hubbard SJ, Thornton JM: **NACCESS.** In. Department of Biochemistry and Molecular Biology, University College, London; 1993.
176. Miller S, Lesk AM, Janin J, Chothia C: **The accessible surface area and stability of oligomeric proteins.** *Nature* 1987, **328**(6133):834-836.
177. 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**:S71-77.
178. Sander C, Schneider R: **Database of homology-derived protein structures and the structural meaning of sequence alignment.** *Proteins* 1991, **9**(1):56-68.
179. Dill KA, Bromberg S, Yue K, Fiebig KM, Yee DP, Thomas PD, Chan HS: **Principles of protein folding--a perspective from simple exact models.** *Protein Sci* 1995, **4**(4):561-602.
180. Jernigan RL, Bahar I: **Structure-derived potentials and protein simulations.** *Curr Opin Struct Biol* 1996, **6**(2):195-209.
181. Ponder JW, Case DA: **Force fields for protein simulations.** *Adv Protein Chem* 2003, **66**:27-85.
182. Godzik A, Skolnick J: **Sequence-structure matching in globular proteins: application to supersecondary and tertiary structure determination.** *Proc Natl Acad Sci U S A* 1992, **89**(24):12098-12102.
183. Miyazawa S, Jernigan R.L.: **Estimation of effective interresidue contact energies from protein crystal structures: quasichemical approximation.** *Macromolecules* 1985, **18**:534-552.

184. Rost B, Sander C: **Prediction of protein secondary structure at better than 70% accuracy.** *J Mol Biol* 1993, **232**(2):584-599.
185. Sippl MJ: **Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins.** *J Mol Biol* 1990, **213**(4):859-883.
186. Tanaka S, Scheraga HA: **Medium- and long-range interaction parameters between amino acids for predicting three-dimensional structures of proteins.** *Macromolecules* 1976, **9**(6):945-950.
187. Thomas PD, Dill KA: **Statistical potentials extracted from protein structures: how accurate are they?** *J Mol Biol* 1996, **257**(2):457-469.
188. Grishin NV, Phillips MA: **The subunit interfaces of oligomeric enzymes are conserved to a similar extent to the overall protein sequences.** *Protein Sci* 1994, **3**(12):2455-2458.
189. Bordner AJ, Abagyan R: **Statistical analysis and prediction of protein-protein interfaces.** *Proteins* 2005, **60**(3):353-366.
190. Witten IH, Frank E: **Data Mining: Practical machine learning tools and techniques.** San Francisco: Morgan Kaufmann Press; 2005.
191. Livnah O, Stura EA, Johnson DL, Middleton SA, Mulcahy LS, Wrighton NC, Dower WJ, Jolliffe LK, Wilson IA: **Functional mimicry of a protein hormone by a peptide agonist: the EPO receptor complex at 2.8 Å.** *Science* 1996, **273**(5274):464-471.
192. Humphrey W, Dalke A, Schulten K: **VMD - Visual Molecular Dynamics.** *J Molec Graphics* 1996, **14**:33-38.
193. Bjorck L, Blomberg J: **Streptococcal protein G: a sensitive tool for detection of antibodies to human immunodeficiency virus proteins in Western blot analysis.** *Eur J Clin Microbiol* 1987, **6**(4):428-429.
194. Herraiz A: **Biomolecules in the computer: Jmol to the rescue.** *Biochemistry and Molecular Biology Education* 2006, **34**:255-261.
195. Thanos CD, DeLano WL, Wells JA: **Hot-spot mimicry of a cytokine receptor by a small molecule.** *Proceedings of the National Academy of Sciences of the United States of America* 2006, **103**(42):15422-15427.

196. Anashkina A, Kuznetsov E, Esipova N, Tumanyan V: **Comprehensive statistical analysis of residues interaction specificity at protein-protein interfaces.** *Proteins* 2007, **67**(4):1060-1077.
197. West DB: **Introduction to Graph Theory.** 1996.
198. Gomory RE, Hu, T.C.: **Multi-terminal network flows.** *J Soc Indust Appl Math* 1961, **9**(4):551-570.
199. Flake GW, Tarjan, R.E., Tsioutsoulis, K.: **Graph clustering and minimum cut trees.** *Internet Mathematics* 2004, **1**(4):385-408.
200. Goldberg AV, Tsioutsoulis: **Cut tree algorithms: an experimental study.** *Journal of Algorithms* 2001, **38**:51-83.
201. Mitrofanova A, Farach-Colton M, Mishra B: **Efficient and robust prediction algorithms for protein complexes using Gomory-Hu trees.** *Pac Symp Biocomput* 2009:215-226.
202. Buckle AM, Schreiber G, Fersht AR: **Protein-protein recognition: crystal structural analysis of a barnase-barstar complex at 2.0-A resolution.** *Biochemistry* 1994, **33**(30):8878-8889.
203. Schreiber G, Fersht AR: **Energetics of protein-protein interactions: analysis of the barnase-barstar interface by single mutations and double mutant cycles.** *J Mol Biol* 1995, **248**(2):478-486.
204. Loewenthal R, Sancho J, Fersht AR: **Fluorescence spectrum of barnase: contributions of three tryptophan residues and a histidine-related pH dependence.** *Biochemistry* 1991, **30**(27):6775-6779.
205. Serrano L, Sancho J, Hirshberg M, Fersht AR: **Alpha-helix stability in proteins. I. Empirical correlations concerning substitution of side-chains at the N and C-caps and the replacement of alanine by glycine or serine at solvent-exposed surfaces.** *J Mol Biol* 1992, **227**(2):544-559.
206. Huang M, Syed R, Stura EA, Stone MJ, Stefanko RS, Ruf W, Edgington TS, Wilson IA: **The mechanism of an inhibitory antibody on TF-initiated blood coagulation revealed by the crystal structures of human tissue factor, Fab 5G9 and TF.G9 complex.** *J Mol Biol* 1998, **275**(5):873-894.

207. Nussinov R, Wolfson HJ: **Efficient detection of three-dimensional structural motifs in biological macromolecules by computer vision techniques.** *Proc Natl Acad Sci U S A* 1991, **88**(23):10495-10499.
208. Shatsky M, Nussinov R, Wolfson HJ: **A method for simultaneous alignment of multiple protein structures.** *Proteins* 2004, **56**(1):143-156.
209. Tuncbag N, Keskin O, Gursoy A: **HotPoint: hot spot prediction server for protein interfaces.** *Nucleic Acids Res* 2010, **38** Suppl:W402-406.
210. Kohn KW: **Molecular interaction map of the mammalian cell cycle control and DNA repair systems.** *Mol Biol Cell* 1999, **10**(8):2703-2734.
211. Chen R, Li L, Weng Z: **ZDOCK: an initial-stage protein-docking algorithm.** *Proteins* 2003, **52**(1):80-87.
212. Schneidman-Duhovny D, Inbar Y, Nussinov R, Wolfson HJ: **PatchDock and SymmDock: servers for rigid and symmetric docking.** *Nucleic Acids Res* 2005, **33**(Web Server issue):W363-367.
213. Croteau DL, Peng Y, Van Houten B: **DNA repair gets physical: mapping an XPA-binding site on ERCC1.** *DNA Repair (Amst)* 2008, **7**(5):819-826.
214. Gillet LC, Scharer OD: **Molecular mechanisms of mammalian global genome nucleotide excision repair.** *Chem Rev* 2006, **106**(2):253-276.
215. de Laat WL, Jaspers NG, Hoeijmakers JH: **Molecular mechanism of nucleotide excision repair.** *Genes Dev* 1999, **13**(7):768-785.
216. Saijo M, Kuraoka I, Masutani C, Hanaoka F, Tanaka K: **Sequential binding of DNA repair proteins RPA and ERCC1 to XPA in vitro.** *Nucleic Acids Res* 1996, **24**(23):4719-4724.
217. Ikegami T, Kuraoka I, Saijo M, Kodo N, Kyogoku Y, Morikawa K, Tanaka K, Shirakawa M: **Solution structure of the DNA- and RPA-binding domain of the human repair factor XPA.** *Nat Struct Biol* 1998, **5**(8):701-706.
218. Shen Z, Cloud KG, Chen DJ, Park MS: **Specific interactions between the human RAD51 and RAD52 proteins.** *J Biol Chem* 1996, **271**(1):148-152.
219. Kurumizaka H, Aihara H, Kagawa W, Shibata T, Yokoyama S: **Human Rad51 amino acid residues required for Rad52 binding.** *J Mol Biol* 1999, **291**(3):537-548.

220. Motycka TA, Bessho T, Post SM, Sung P, Tomkinson AE: **Physical and functional interaction between the XPF/ERCC1 endonuclease and hRad52.** *J Biol Chem* 2004, **279**(14):13634-13639.
221. Ranatunga W, Jackson D, Flowers IR, 2nd, Borgstahl GE: **Human RAD52 protein has extreme thermal stability.** *Biochemistry* 2001, **40**(29):8557-8562.
222. Plate I, Albertsen L, Lisby M, Hallwyl SC, Feng Q, Seong C, Rothstein R, Sung P, Mortensen UH: **Rad52 multimerization is important for its nuclear localization in *Saccharomyces cerevisiae*.** *DNA Repair (Amst)* 2008, **7**(1):57-66.
223. Smith ML, Chen IT, Zhan Q, Bae I, Chen CY, Gilmer TM, Kastan MB, O'Connor PM, Fornace AJ, Jr.: **Interaction of the p53-regulated protein Gadd45 with proliferating cell nuclear antigen.** *Science* 1994, **266**(5189):1376-1380.
224. Hall PA, Kearsley JM, Coates PJ, Norman DG, Warbrick E, Cox LS: **Characterisation of the interaction between PCNA and Gadd45.** *Oncogene* 1995, **10**(12):2427-2433.
225. Zhan Q, Antinore MJ, Wang XW, Carrier F, Smith ML, Harris CC, Fornace AJ, Jr.: **Association with Cdc2 and inhibition of Cdc2/Cyclin B1 kinase activity by the p53-regulated protein Gadd45.** *Oncogene* 1999, **18**(18):2892-2900.
226. Jin S, Antinore MJ, Lung FD, Dong X, Zhao H, Fan F, Colchagie AB, Blanck P, Roller PP, Fornace AJ, Jr. *et al*: **The GADD45 inhibition of Cdc2 kinase correlates with GADD45-mediated growth suppression.** *J Biol Chem* 2000, **275**(22):16602-16608.
227. Kontopidis G, Wu SY, Zheleva DI, Taylor P, McInnes C, Lane DP, Fischer PM, Walkinshaw MD: **Structural and biochemical studies of human proliferating cell nuclear antigen complexes provide a rationale for cyclin association and inhibitor design.** *Proc Natl Acad Sci U S A* 2005, **102**(6):1871-1876.
228. Shishodia S, Aggarwal BB: **Nuclear factor-kappaB: a friend or a foe in cancer?** *Biochem Pharmacol* 2004, **68**(6):1071-1080.
229. Benyamini H, Leonov H, Rotem S, Katz C, Arkin IT, Friedler A: **A model for the interaction between NF-kappa-B and ASPP2 suggests an I-kappa-B-like binding mechanism.** *Proteins* 2009, **77**(3):602-611.
230. Jensen LJ, Kuhn M, Stark M, Chaffron S, Creevey C, Muller J, Doerks T, Julien P, Roth A, Simonovic M *et al*: **STRING 8--a global view on proteins and their**

- functional interactions in 630 organisms.** *Nucleic Acids Res* 2009, **37**(Database issue):D412-416.
231. Carrano AC, Eytan E, Hershko A, Pagano M: **SKP2 is required for ubiquitin-mediated degradation of the CDK inhibitor p27.** *Nat Cell Biol* 1999, **1**(4):193-199.
232. Wang SX, Pandey KC, Somoza JR, Sijwali PS, Kortemme T, Brinen LS, Fletterick RJ, Rosenthal PJ, McKerrow JH: **Structural basis for unique mechanisms of folding and hemoglobin binding by a malarial protease.** *Proc Natl Acad Sci U S A* 2006, **103**(31):11503-11508.
233. Kussie PH, Gorina S, Marechal V, Elenbaas B, Moreau J, Levine AJ, Pavletich NP: **Structure of the MDM2 oncoprotein bound to the p53 tumor suppressor transactivation domain.** *Science* 1996, **274**(5289):948-953.
234. Rosal R, Pincus MR, Brandt-Rauf PW, Fine RL, Michl J, Wang H: **NMR solution structure of a peptide from the mdm-2 binding domain of the p53 protein that is selectively cytotoxic to cancer cells.** *Biochemistry* 2004, **43**(7):1854-1861.
235. Meek DW, Eckhart W: **Phosphorylation of p53 in normal and simian virus 40-transformed NIH 3T3 cells.** *Mol Cell Biol* 1988, **8**(1):461-465.
236. Samad A, Anderson CW, Carroll RB: **Mapping of phosphomonoester and apparent phosphodiester bonds of the oncogene product p53 from simian virus 40-transformed 3T3 cells.** *Proc Natl Acad Sci U S A* 1986, **83**(4):897-901.
237. Wang Y, Eckhart W: **Phosphorylation sites in the amino-terminal region of mouse p53.** *Proc Natl Acad Sci U S A* 1992, **89**(10):4231-4235.
238. Ni J, Sakanyan V, Charlier D, Glansdorff N, Van Duyne GD: **Structure of the arginine repressor from *Bacillus stearothermophilus*.** *Nat Struct Biol* 1999, **6**(5):427-432.
239. Shulman-Peleg A, Shatsky M, Nussinov R, Wolfson HJ: **MultiBind and MAPPIS: webservers for multiple alignment of protein 3D-binding sites and their interactions.** *Nucleic Acids Res* 2008, **36**(Web Server issue):W260-264.
240. McDowall MD, Scott MS, Barton GJ: **PIPs: human protein-protein interaction prediction database.** *Nucleic Acids Res* 2009, **37**(Database issue):D651-656.

241. Jing Y, Wang M, Tang W, Qi T, Gu C, Hao S, Zeng X: **c-Abl tyrosine kinase activates p21 transcription via interaction with p53.** *J Biochem* 2007, **141**(5):621-626.
242. Walker JR, Corpina RA, Goldberg J: **Structure of the Ku heterodimer bound to DNA and its implications for double-strand break repair.** *Nature* 2001, **412**(6847):607-614.
243. Koike M, Ikuta T, Miyasaka T, Shiomi T: **Ku80 can translocate to the nucleus independent of the translocation of Ku70 using its own nuclear localization signal.** *Oncogene* 1999, **18**(52):7495-7505.
244. Holcomb VB, Rodier F, Choi Y, Busuttill RA, Vogel H, Vijg J, Campisi J, Hasty P: **Ku80 deletion suppresses spontaneous tumors and induces a p53-mediated DNA damage response.** *Cancer Res* 2008, **68**(22):9497-9502.
245. Savvides SN, Boone T, Andrew Karplus P: **Flt3 ligand structure and unexpected commonalities of helical bundles and cystine knots.** *Nat Struct Biol* 2000, **7**(6):486-491.
246. Daujat S, Neel H, Piette J: **MDM2: life without p53.** *Trends Genet* 2001, **17**(8):459-464.
247. Martin K, Trouche D, Hagemeyer C, Sorensen TS, La Thangue NB, Kouzarides T: **Stimulation of E2F1/DP1 transcriptional activity by MDM2 oncoprotein.** *Nature* 1995, **375**(6533):691-694.
248. Xiao ZX, Chen J, Levine AJ, Modjtahedi N, Xing J, Sellers WR, Livingston DM: **Interaction between the retinoblastoma protein and the oncoprotein MDM2.** *Nature* 1995, **375**(6533):694-698.
249. Lee C, Chang JH, Lee HS, Cho Y: **Structural basis for the recognition of the E2F transactivation domain by the retinoblastoma tumor suppressor.** *Genes Dev* 2002, **16**(24):3199-3212.
250. Goldberg Z, Vogt Sionov R, Berger M, Zwang Y, Perets R, Van Etten RA, Oren M, Taya Y, Haupt Y: **Tyrosine phosphorylation of Mdm2 by c-Abl: implications for p53 regulation.** *Embo J* 2002, **21**(14):3715-3727.
251. Jin Y, Zeng SX, Dai MS, Yang XJ, Lu H: **MDM2 inhibits PCAF (p300/CREB-binding protein-associated factor)-mediated p53 acetylation.** *J Biol Chem* 2002, **277**(34):30838-30843.

-
252. Garber K: **Missing the target: ubiquitin ligase drugs stall.** *J Natl Cancer Inst* 2005, **97**(3):166-167.
253. Zhang Z, Wang H, Li M, Rayburn ER, Agrawal S, Zhang R: **Stabilization of E2F1 protein by MDM2 through the E2F1 ubiquitination pathway.** *Oncogene* 2005, **24**(48):7238-7247.
254. Thompson JD, Higgins DG, Gibson TJ: **CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.** *Nucleic Acids Res* 1994, **22**(22):4673-4680.
255. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T: **Cytoscape: a software environment for integrated models of biomolecular interaction networks.** *Genome Res* 2003, **13**(11):2498-2504.
256. Humphrey W, Dalke A, Schulten K: **VMD - Visual Molecular Dynamics.** *Journal of Molecular Graphics* 1996, **14**:33-38.
257. Mashiach E, Nussinov R, Wolfson HJ: **FiberDock: a web server for flexible induced-fit backbone refinement in molecular docking.** *Nucleic Acids Res* 2010, **38** Suppl:W457-461.

VITA

Nurcan Tunçbağ was born in Istanbul, Turkey, on December 9, 1982. She received the B.Sc. Degree in Chemical Engineering from Istanbul Technical University (ITU) in 2005 and M.Sc. Degree in Computational Science and Engineering from Koc University in 2007. She received the Ph.D. degree from Koc University in Computational Science and Engineering in 2010. From September 2005 to September 2010 she worked as teaching and research assistant at Koç University. She was a TUBITAK scholar during her M.Sc. and PhD studies.

Her research focuses on structural analysis and characterization of protein interactions using computational methods. She has published articles in prestigious journals such as *Bioinformatics*, *Nucleic Acids Research*, *Protein: Structure, Function, and Bioinformatics*, *Journal of Molecular Biology*, *Molecular Biosystems*.

She will continue her academic career as a Postdoctoral Associate in Biological Engineering Department at Massachusetts Institute of Technology (MIT), where she will focus on revealing how the networks of interactions among proteins and genome are altered in cells during disease.