Hot Regions in Protein–Protein Interactions
&
Analysis of Hot Region Distribution in Hub Proteins

by

Engin Cukuroglu

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

Master of Science
in
Computational Science and Engineering

Koc University

August 2011

Koc University

Graduate School of Sciences and Engineering

This is to certify that I have examined this copy of a master's thesis by

Engin Cukuroglu

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:

_____

Ozlem Keskin, Ph. D. (Advisor)

_____

Attila Gursoy, Ph. D.

_____

Can Erkey, Ph. D.

Date: _____

Abstract

Protein interactions play a key role in many cellular processes, and proteins interact with each other in a highly specific manner. The interaction mechanism still remains a mystery. However, researchers work hard on identifying binding partners and binding regions of the proteins. In the cellular level, protein-protein interactions can be modeled as a network whose nodes are proteins and edges are interactions. Hub proteins are the mostly connected nodes in this network. Hence, hub proteins have a crucial role in the cell. Interfaces are the functional units of the proteins and any distortion in protein interfaces may lead to development of many diseases. Hot spots are the important residues at the interfaces which contribute more to binding energy. They are not uniformly distributed in the interface, but rather clustered (hot regions). Hot regions are important for binding affinity and specificity in protein–protein interactions, and drug targeting.

This study mainly focuses on answering the question "what are the roles of hot regions in the protein-protein interfaces". Towards this aim, the thesis concentrates on two major topics; (i) how a hot region can be identified in the interface, and (ii) analysis of hub proteins and their hot region distributions. When the protein–protein interactions are examined, hot regions of hub proteins are observed to vary with respect to interface properties. Also, in the hub protein classification (date and party), hot region properties have a significant role. A database, called HotRegion, is designed and implemented based on the role of hot region at the protein interactions.

This work shows how available structural information can help in examining the hot regions of these complexes. Also, HotRegion will help the researchers in detecting cooperativity of functionally important residues, mutagenesis targets and understand the stability and specificity of protein-protein interfaces.

ÖZET

Protein etkileşimleri birçok hücre işlemlerin gerçekleşmesinde önemli rol oynar ve proteinler son derece özel bir şekilde birbirleriyle etkileşirler. Etkileşim mekanızması hala sırrını korumaktadır. Ancak, araştırmacılar protein ortakları ve bağlayıcı bölgeleri tanımlamak için sıkı calışıyorlar. Hücresel düzeyde, protein–protein etkileşimleri düğümleri protein ve bağlantıları etkileşim olan bir ağ olarak modellenebilir. Merkez dügüm proteinleri ağdaki en çok bağlı düğümlerdir. Bu nedenle, Merkez düğüm proteinleri hücre içinde önemli bir role sahiptir. Arayüzler protein etkileşimlerinin fonksiyonel birimleridir ve protein arayüzlerindeki herhangi bir bozulma birçok hastalaığın gelişmesine neden olabilir. Sıcak noktalar, bağlanma enerjisine daha fazla katkıda bulunan önemli aminoasitlerdir. Bunlar arayüzeyde eşit dağıltılmamıştır daha ziyade kümelenmiştir (sıcak bölgeler). Sıcak bölgeler protein–protein etkileşimlerinde bağlanma eğilimi ve özgüllüğü, ve ilaç hedefleme için önemlidir.

Bu çalışma esas olarak protein–protein arayüzlerinde sıcak bölgelerin rolü ne sorusuna cevap vermeye yoğunlaşır. Bu amaç çerçevesinde, tez iki konu üzerinde yoğunlaşmaktadır; (i) sıcak bölgeler arayüzde nasıl tanımlanabilir, ve (ii) merkez düğüm ve bunların sıcak bölge analizleri. Protein–protein etkileşimleri incelendiğinde, Merkez düğümlerin sıcak bölgeleri arayüz özelliklerine göre değişiklik gösterir. Ayrıca, Merkez düğüm proteinleri sınıflandırmasında (date ve party), sıcak bölge özellikleri önemli rol oynar. HotRegion adında bir veritabanı sıcak bölgelerin protein etkileşimlerindeki rölüne dayalı olarak tasarlanmış ve kurulmuştur.

Bu çalışma, mevcut yapısal bilginin, sıcak bölgelerinin incelenmesinde nasıl yardımcı olabileceğini gösteriyor. Ayrıca, HotRegion araştırmacılara işlevsel olarak önemli aminoasitlerin işbirliğini, mutasyon hedeflerini, ve protein–protein arayüzlerinin sağlamlığının ve özgünlüğünün tespitinde yardımcı olabilir.

Acknowledgements

In the first place, I offer my sincerest gratitude to my supervisors, **Prof. Özlem Keskin** and **Prof. Attila Gürsoy** for the chance of being a member of their research group, supplying us a productive and comfortable research environment and their support during my work. Also, I would like to thank my thesis committee member **Prof. Can Erkey** for his critical reading and useful comments.

I offer my deep appreciation to **Prof. Ruth Nussinov** for her guidance and sharing her endless experience with us.

I thank everyone for the good times throughout the 7 years at Koc University. My special thanks go to Utku Boz, Başar Demir and Talha Akyol who are my colleques, my homemates and my friends. I would like to thank specially my past officemate Nurcan Tunçbağ for her assistance. I thank my current officemates (Gözde Kar, Billur Engin, Ece Özbabacan, Güray Kuzu) and all people at our partner office 227 and 110.

Finally, I thank my family for their patience and continuous support during every step of my education.

Table of Contents

List of Figures

List of Tables

Chapter 1
INTRODUCTION

Most of the biological functions in an organism are controlled by protein-protein interactions (PPI). To use an analogy, cells could be resembled to a chaos environment where proteins must find their best partners in nano seconds in order to fulfill their responsibilities. Hence, researchers pursuit to find the most challenging question "How it is possible that a protein recognize its partners". Deciphering the interaction pathways, methods, and mechanisms of the protein recognition are crucial for the disease research and drug discovery. With the improvement and diversification of the experimental methods, researchers have been determining the structural information of the proteins and protein complexes. Protein Data Bank (PDB) [2] is the depositor of the structural information of the proteins. The structural information of the protein is helpful to comprehend the protein recognition mechanisms and protein-protein interactions.

Proteins interact with other proteins through their interfaces in order to fulfill their functions. Although proteins are scattered in the cell, they can find their partners according to their interface properties. These properties like accessible surface area (ASA), residue preferences, hydrophobicity, residue energy distributions, cavity, residue conservations, and residue pairwise interactions are the identifier for the possible partner selection of the proteins.

According to energy distribution profiles of the residues, the residues in protein interfaces do not have equal contribution in binding, rather some residues, called "hot spots", play an exceptional role [3-5]. Also, these hot spot residues are not randomly distributed at the interface, rather clustered [6]. The combination of hot spot residues may

be considered as drug targets, and the existence of a network between hot spots addresses the question "what are the effects of hot spot clusters to the binding affinity and specificity". The residue targets for the drugs can be selected by using the hot spot information and their organization, after the analyzing the cooperativity of the hot spots.

Protein-protein interaction networks (PPIN) derived from experimental techniques enable the systematic analysis of the proteins. Pioneering studies on protein interaction networks and topological analysis of protein network provided insights into the different types of proteins such as hub proteins which are the mostly connected proteins in the network [7, 8]. Differentiation of hub proteins and non hub proteins are crucial to develop accurate drug targets for the protein networks because there is a positive correlation between lethality and connectivity of the protein [7].

This master thesis primarily focuses on structural properties of clustered hot spots in interfaces and their contribution to the binding specificity and affinity. While investigating hot spot clusters at the protein interfaces, PPIN analysis of *Saccharomyces cerevisiae* provides major key points in order to differentiate the multiple binding tendencies of varied hub proteins. As a result of this experience, a database of hot spot clusters of the known structural protein complexes are presented for researchers.

The outline of this thesis study is as follows:

In Chapter 2, a literature review of structural aspects of protein interactions is presented. This chapter includes characteristics of protein interfaces, hot spots, hot spot clusters, PPIN and hub proteins.

Chapter 3 includes the analysis of hub proteins in PPIN by using hot spot clusters in interfaces. Non-redundant interface dataset derivation from *Saccharomyces cerevisiae*'s protein-protein interaction network data is explained step by step and then, hot region construction at the interface is illustrated. For the hub protein classification, machine learning approaches are used and the feature selection is explained.

In Chapter 4, the database of hot spot clusters, HotRegion, is introduced which provides the interface properties of protein-protein complexes such as ASA, pair potentials, hot spot information and hot region information of the residues. Application of HotRegion is demonstrated by a case study of the colicin protein with two different partners at the same interface. Also, HotRegion database tutorial is presented in this section.

This thesis ends with a chapter which includes discussion of the results, future directions and conclusion of the study.

Chapter 2
LITERATURE REVIEW

In this chapter, the review of previous studies related to protein interfaces, hot spots, hot regions, protein interactions and hub proteins are presented.

2.1 Protein Interface

Proteins interact with proteins, peptides, DNA or RNA in order to form complexes. In **Figure 2.1**, some examples of these interactions are presented. These complex structures are the constituent of the many biological processes [9]. Actually, proteins use interfaces in order to build a complex and fulfill their functions. Interfaces are formed by residues whose properties determine binding specificity and affinity. As an interface example, **Figure 2.2** shows the interface between a multi-subunit E3 protein ubiquitin ligase (PDB Id: 1GQP). Protein interfaces have been studied for a long time and researchers deposited their findings to the protein interface databases to identify the general properties of them. Some of the available databases are PROTORP [10], InterPare [11], 3did [12-14], PIBASE [15] and PRINT [16, 17].

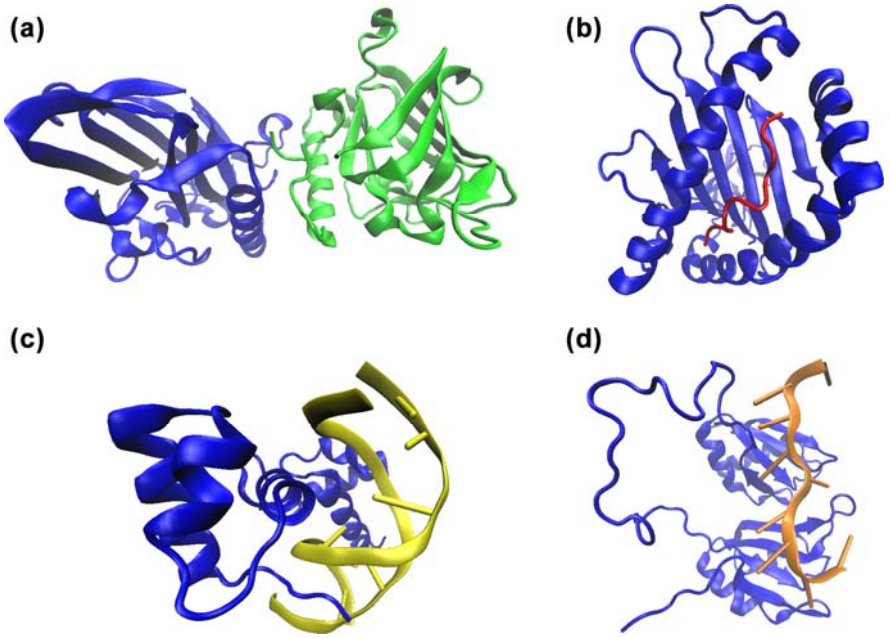


Figure 2.1 Protein interacts with different molecules. (a) Protein-protein interaction of 1GQP between chain A (blue) and chain B (green). (b) Protein-peptide interaction of 1HHH between chain A (blue, protein) and chain C (red, peptide). (c) Protein-DNA interaction of 3OSF between chain A (blue, protein) and chain EF (yellow, DNA). (d) Protein-RNA interaction of 2YH1 between chain A (blue, protein) and chain B (orange, RNA). (All figures visualized by using VMD [1].)

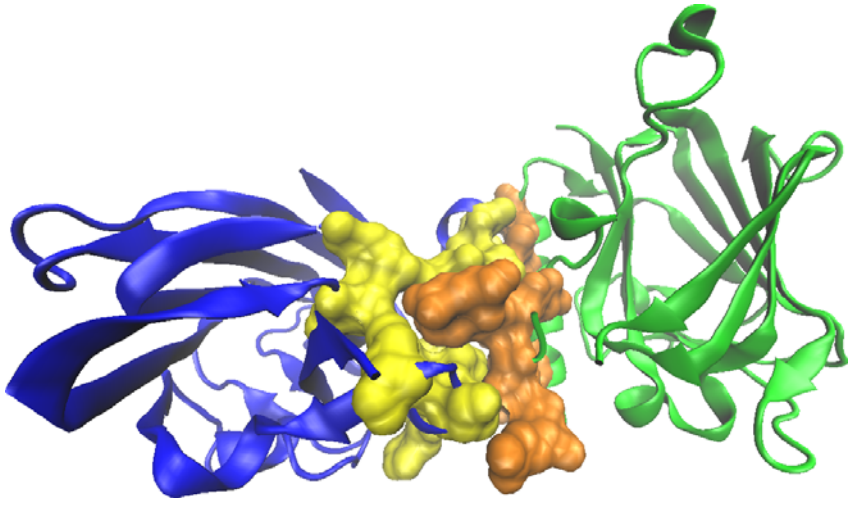


Figure 2.2 Protein interface of 1GQP between chain A (blue) and chain B (green). The interface residues of chain A are yellow and interface residues of chain B are orange.

2.2 Hot Spots: Critical Residues in Interface

One of the interesting features of interfaces is the degree of contribution of an amino acid to the binding free energy between two proteins. It is well known that not all residues contribute to the same extent in the binding, some are more important and these residues are called hot spots [3-5]. Experimentally, a hot spot can be detected by alanine scanning mutagenesis. If the binding free energy change is more than 2kcal/mol, the residue is flagged as a hot spot. Alanine Scanning Energetics Database (ASEdb) deposits hot spots from alanine scanning mutagenesis experiments [18]. Experimentally verified hot spots collected from literature are deposited to Binding Interface Database (BID) [19].

Amino acid composition of hot spots revealed that some residues are more favorable. Tyr, Arg and Trp are the mostly preferred residues for hot spots in the interfaces [5]. Also, Bogan and Thorn stated that hot spots are generally located at the center of the energetically less important residues which occlude bulk solvent (O-Ring hypothesis) [5]. Hot spot residues utilize occlusion solvent to generate highly energetic interactions [5, 20, 21]. According to the O-Ring hypothesis, less important residues for binding have an important role for shielding hot spot residues from contacting with bulk solvent, thus hot spot residues have small accessible surface area (ASA). Hot spots ASA are not increasing even though increasing interface size; they are buried in the interface [22].

Computational methods are widely used to predict the hot spot residues at the interface because extracting hot spot information from experimental studies are time consuming and expensive. Also, experimental studies are available for a very limited number of complexes. Research groups who worked to develop a reliable computational method in order to predict hot spots in the interface used different models. They are respectively energy based models [23, 24], learning based models [20, 21, 25-30], molecular dynamic based models [31-33] and graph based models [34, 35].

2.3 Hot Spot Clusters (Hot Regions)

Interfaces are formed by residues whose properties determine binding specificity and affinity. Correct orientations of the residues are critical for complex formation. Interactions between the residues in the binding sites are higher than the protein surface which shows that protein-protein interactions are highly depending on the cooperativity of the residues [36]. Cooperativity of the residues should have an important role on multi binding of the interface because a protein can bind different partners via the same interface, although ASA of an interface is limited [17, 37, 38]. These interfaces should have a mechanism that can identify the partner protein. The distribution of the residues across the interface and the residue–residue interactions may answer for question "How can the interfaces recognize their partners?" The residues tend to behave cooperatively during the interactions and they form modules in the interface [39]. Proteins utilize these modules in order to have specificity and affinity during interactions [6, 40-42] and also the combinations of these modules yield a powerful mechanism for binding multiple partners via unique interfaces [43, 44]. Previously, modules in interfaces are defined with various methods such as (i) the edge betweenness criteria in the residue-residue interaction network across the interface [41, 45], (ii) difference of energy profiles of residues in interfaces [40, 46, 47], and (iii) structurally conserved residues in interfaces [6, 43, 48]. In the edge betweenness approach, the authors used the topology of the network without considering residue energy profiles. The other two approaches used hot spot residues which are driven by energy profiles or structural conservation of residues.

According to the previous researches, hot spots are tightly packed and structurally conserved residues [6, 43, 48]. Keskin et al. showed that these hot spot residues are not randomly distributed along the protein-protein interfaces; rather clustered [38]. The assemblies of hot spots are located within densely packed regions. Within an assembly, the tightly packed hot spots form networks of interactions. These modular assembly regions

are called *hot regions* [43, 49]. An interface may contain none, single, or multiple hot regions. The tight, networked hot spot organization may imply that the contribution of the hot spots to the stability of the protein–protein complex within a hot region is cooperative [40]. This binding site organization rationalizes how a given protein molecule may bind to different protein partners.

2.4 Protein Interactions

Protein interactions can be found experimentally. Yeast two hybrid method [50, 51] which is used for determining the transient interactions between proteins, and tandem affinity purification (TAP) with mass spectrometry [52] which is used to find assemblies of proteins interactions in complexes. Although data from these experiments are noisy, a recent study [53] indicates that the data has a sufficient quality for protein protein interactions. By combining the interactions from these high throughput experiments, a PPIN can be generated. The topology of this network provides insights about the interactions. The PPIN of *Saccharomyces cerevisiae* has a power law connectivity distribution which means that some proteins are highly connected (hub proteins), although most proteins are not. High-throughput experiments (expression profiles) and structures of complexes help to define two different hub types; party hubs and date hubs [8, 37]. For example, Vidal and coworkers [8] used mRNA expression profiles of hubs and found that some hubs displayed similar mRNA expression patterns with their interacting partners indicating that their interactions are simultaneous and hence they were called party hubs. From a structural point of view, party hubs are found in static complexes where they interact with most of their partners at the same time. On the other hand, date hubs bind their interaction partners at different times and/or locations. Date hubs organize the proteome, connecting biological processes to each other, whereas party hubs take place inside processes. Thus, date hubs appear to be more important than party hubs for the

topology of the network because they cause more destruction of the network into small pieces when they are attacked [8].

In the study of Han et al. [8], a PPIN model was suggested for *S. cerevisiae* in which the date hubs are responsible for organizing biological modules whereas the party hubs have localized functions inside those modules. When an interactome is perturbed by deleting date hubs, it's divided into many little networks representing the interactions of many biological processes all organized and combined by perturbed date hubs. Ekman et al. [54] deduced that hub proteins of *S. cerevisiae* contain a higher fraction of multi-domain proteins and proteins with repeated domains (compared to the non-hubs). Having multiple interaction domains can explain their high connectivities. In their study, they also indicated that self-interaction and interacting with other proteins containing shared domains are observed more frequently in party hubs than date hubs. On the other hand, date hubs are shown to have long disordered regions explaining their flexible interactions.

Three dimensional structures of the protein complexes in interaction maps can help understanding the differences between hub proteins and others. Structural comparisons revealed that smaller hubs have fewer disordered residues and more charged residues on the surface than larger hubs [55]. Simply, considering the geometrical constraints of a protein structure, it can be stated that it is beyond the possibility of any protein surface to provide as many separate, isolated sites to bind to different proteins. This implies some binding sites can be specific to bind to a particular partner (most probably as in the case of party hubs) whereas the same or overlapping locations on the surface can be used to bind to several other proteins (presumably should be the mechanism for date hubs to interact with different proteins at different times). This suggests that there are binding sites that are repeatedly reused, although with different affinities and probably entailing differences in their specific interactions.

If some binding sites are uniquely used and some others are multiply used, then one expects to see some differences in the binding sites' physico-chemical and structural features. Indeed, Kar et al.'s study pointed out that hub proteins have smaller, more planar, less tightly packed binding sites than non-hub proteins [56]. Kim et al. [37] in a leading study identified the singlish and multi-interface hubs. Their analysis pointed out that the notion of hubs having a higher essentiality due to their network centrality was incomplete: It was rather the number of interaction interfaces that lead to higher essentiality [37]. Previously, there was not a consensus whether hubs were slower-evolving than other proteins or not [57-60]. Kim et al [37] by integrating structures into protein interaction networks stated that multi-interface hubs were more likely to be essential and more conserved, being members of large and stable complexes as opposed to singlish-interface hubs. In a proceeding study, they found that although singlish-interface hub proteins were more disordered, their interfaces were highly structured, as is the case for multi-interface hubs. Yet, they found that binding partners of single-interface hubs were more disorder than the proteome average, suggesting that their promiscuity is a result of disorder of their binding partners [61].

2.5 Hub Proteins

Protein interaction maps constructed from binary interactions reveal that some proteins, called hub proteins, are highly connected to others, whereas some others have a few interactions, called non-hub proteins. There are different views trying to explain what characteristics differentiate hubs from others and why and how a protein becomes a hub protein through evolution. One answer would be to have distinct binding sites on the surfaces of hub proteins. Hub proteins, given that they are larger, contain more domains and enriched in repeats of tandem domains [54], this could be true to an extent. Another answer would be that hub proteins bind to paralogs in the proteome. So actually the same binding site can be used to bind to several related proteins [37, 54]. Flexibility [62] or

disorder of the hubs can also contribute them to bind to several proteins. Gerstein and coworkers stated that it is not the hubs but the partners that are disordered [61]. On the other hand, Tsai et al. [63] recently suggested that a single structure cannot bind hundreds of different proteins, even if it is extremely flexible or disordered. They stated that the nodes in interaction maps are not a single protein but rather different forms of proteins (i.e., forms that result from post-translational modifications). Despite all these recent works, characteristics and interactions of hub proteins are not yet clearly understood.

Chapter 3
ANALYSIS OF HOT REGION ORGANIZATION IN HUB PROTEINS

3.1 Methodology

An interface is the contact region between two interacting proteins. Two residues are defined to be contacting if the distance between any two atoms of the two residues from different chains is less than the sum of their corresponding van der Waals radii plus 0.5 Å [16, 64]. An example of an interface is given in **Figure 3.1** displaying interface residues in ball-stick model.

In this study, interfaces are annotated as DD (interfaces between two date hubs), PP (between two party hubs), and NN (between two non-hub proteins) where D; P; N and X are for date hub, party hub, non-hub and any protein, respectively. **Figure 3.2** displays the different types of interfaces. Then, the hot regions are found in these interfaces. Various features such as change in accessible surface areas (ΔASAs) of hot regions and interfaces, ratio of hot region over interface areas and amino acid compositions are determined to understand the organization of hot regions and their relation to these interface types.

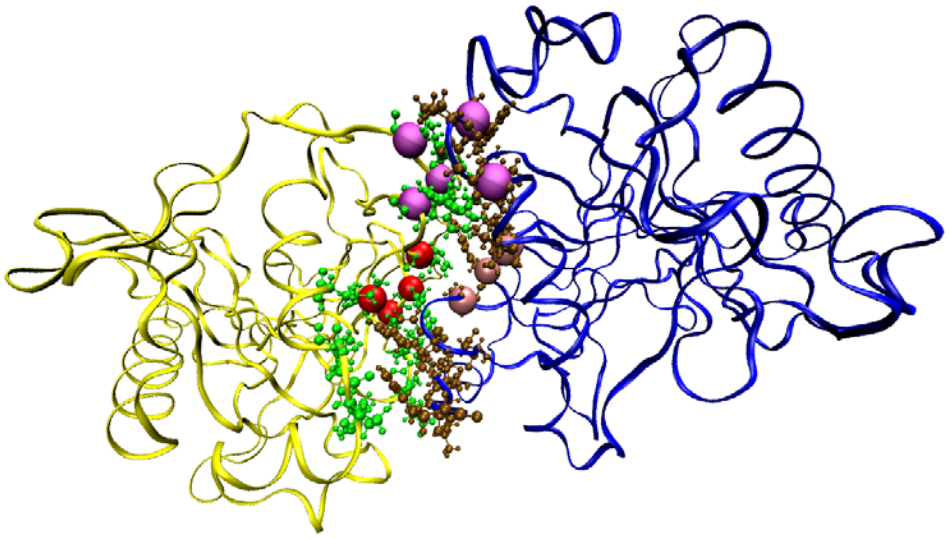


Figure 3.1 Interface representation of 1E9GBA. The yellow representation is the A chain and the blue representation is the B chain of the protein. The green ball-stick representation is the interface of chain A, and the ochre ball-stick representation is the interface of chain B. The red, magenta and pink ball representations are the different hot regions in the interface.

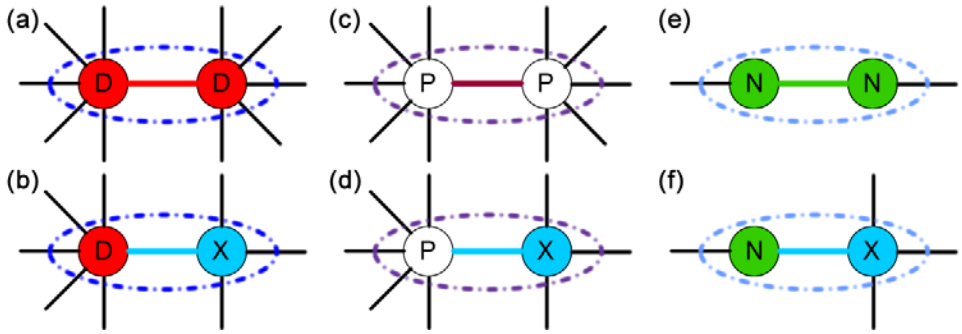


Figure 3.2 The nodes represent the protein; the edges represent the interfaces between the proteins. (a) Date hub—date hub interaction scheme in PPIN (DD). (b) Date hub—non labeled protein interaction scheme in PPIN (DX). (c) Party hub—party hub interaction scheme in PPIN (PP). (d) Party hub—non labeled protein interaction scheme in PPIN (PX). (e) Non hub—non hub interaction scheme in PPIN (NN). (f) Non hub—non labeled protein interaction scheme in PPIN (NX).

3.1.1 Interface Dataset

The interface dataset used in this study is generated from Ekman's PPIN. In Ekman's network, proteins are annotated as party, date or non-hubs [54] with ordered locus names (OLN) of the genes and their hub status. In order to determine and analyze hot regions in the binding sites of interfaces, the 3-dimensional structures of interfaces are necessary. Therefore, OLNs of the genes are cross referenced to the protein data bank (PDB) IDs using Uniprot. In some cases, different OLNs may map to the same 'PDB ID' despite the fact that they are labeled as different hub types in the Ekman's dataset. Such multiply labeled proteins are discarded from the dataset. The interfaces of complexes are fetched from interface dataset of Tuncbag et al.'s [17] resulting in 1199 PX, 602 DX and 1343 NX interfaces. In order to obtain non-biased statistics, the structurally redundant interfaces are removed and low resolution proteins (worse than 3.0 angstrom) resulting in 82 PXs, 83 DXs and 221 NXs. In PXs, 16 unique pdb ids generate 82 structurally non redundant interface data, 54 unique pdb ids generate 83 DXs and 133 unique pdb ids generate 221 NXs. A complete list of complexes is given in the Appendix A. This procedure is summarized in the flowchart shown in **Figure 3.3**.

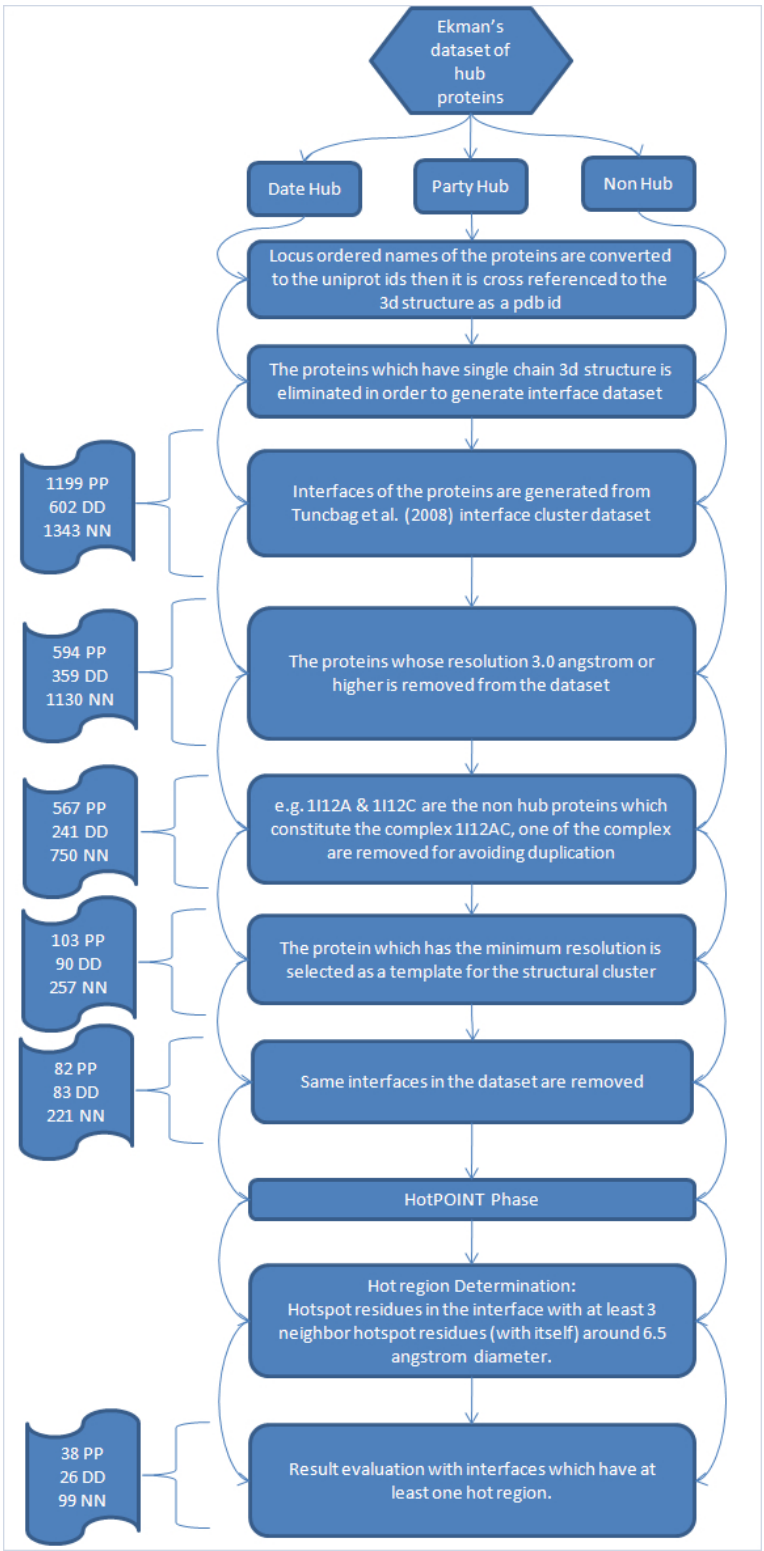


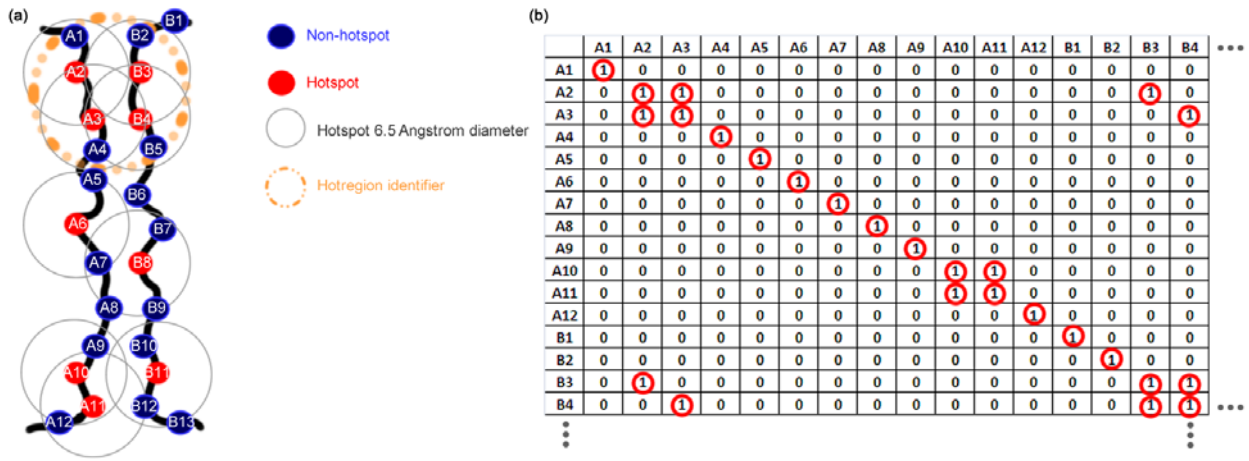
Figure 3.3 A flowchart of the methodology.

3.1.2 Hot Region Detection in the Interfaces

Interface properties (ASA values and hot spot status of residues) of the proteins are taken from the HotPOINT [21] server. HotPOINT is a server that predicts hotspot residues based on using ASA and knowledge-based pair energies. In addition to hotspot status of a residue in an interface, the server provides monomer and complex ASA values to calculate the ΔASA. The mean ΔASA on complexation (going from a monomeric state to a dimeric state) was calculated as the sum of the total ΔASA for both chains. There is not sufficient experimental hotspot data for hub proteins so computationally predicted hotspot data from HotPoint server is used in this study.

In order to define **hot regions**, a contact matrix is constructed by using the coordinates of the residues and hotspot status. It is an *nxn* matrix where *n* is the number of residues in the interface. Two residues are defined as contacting if the distance between their Cα atoms is smaller than 6.5 angstrom [6]. In the matrix, the ijth element is set to one if residues i and j are in contact and if both are hot spots. Otherwise, the element is zero (See **Figure 3.4**).
In a previous work, Reichman et al. defined residue modules as clusters of residues with at least 3 members [40]. Also, Shandar et al. labeled hot regions as the ones with at least three conserved residues [65]. Here, in a similar way, hot regions are defined as the group of hotspots which have at least 2 contacting hotspot neighbors in the interface (**Figure 3.4**). The contact matrix is used to find hot regions. **Figure 3.4** illustrates an example of hot regions in an interface. In order to find hot regions, first a column with at least three '1' entries are determined, this forms the initial cluster then for each element of the cluster corresponding column are merged to the existing cluster until no more additions are possible.

Some of the interfaces in the interface dataset did not yield any hot regions. The final interface dataset with hot regions includes 38 PPs, 26 DDs and 99 NNs.



	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	B1	B2	B3	B4
A1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
A2	0	1	1	0	0	0	0	0	0	0	0	0	0	0	1	0
A3	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	1
A4	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
A5	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
A6	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
A7	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
A8	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
A9	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
A10	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0
A11	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0
A12	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
B1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
B2	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
B3	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	1
B4	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	1

Figure 3.4 (a) Schematic representation of the hot region at the interface of the two proteins, (b) contact matrix of the interface. A2, A3, B3, and B4 columns have three '1' entries which means that the residues of A2-A3-B3, A2-A3-B4, B3-A2-B4, and B4-A3-B3 form a hot region. The hot regions which are obtained in this interface are also interconnected with each other in at least one hotspot. Therefore, their consensus builds only one hot region which includes A2-A3-B3-B4 residues.

3.1.3 Interface and Hot Region Features

This section summarizes various parameters used in assessing the organization of hot spots and also used in statistical analysis of DD, PP, and NN interfaces:

***Hot spot ratio*:** The ratio of the total number of hot spots in hot regions to the total number of hot spots in the interface. This parameter is an indicator of hot spot organization (the bigger the ratio, the more clustered hot spots in hot regions).

***Average hot region size*:** The average number of hot spots in hot regions. This parameter describes how big the hot regions are.

***Average number of hot regions*:** The average number of hot regions in the interface.

***Average hot region ΔASA to interface ΔASA ratio*:** The difference of accessible surface area upon complexation (ΔASA) is a widely used characteristic for estimating how much buried the interfaces become upon complexation. It is calculated as follows:

$HR_{\Delta ASA}$: Hot region ΔASA.

$I_{\Delta ASA}$: Interface ΔASA.

$HR_{ASA,A}$: Total monomer ASA values of the residues of chain A in the hot region.

$HR_{ASA,B}$: Total monomer ASA values of the residues of chain B in the hot region.

$HR_{ASA,AB}$: Total complex ASA values of the residues of in the hot region.

$I_{ASA,A}$: Total monomer ASA values of the residues of chain A in the interface.

$I_{ASA,B}$: Total monomer ASA values of the residues of chain B in the interface.

$I_{ASA,AB}$: Total complex ASA values of the residues of in the interface.

$$\frac{HR_{\Delta ASA}}{I_{\Delta ASA}} = \frac{HR_{ASA,A} + HR_{ASA,B} - HR_{ASA,AB}}{I_{ASA,A} + I_{ASA,B} - I_{ASA,AB}}$$

***Polar amino acid (aa) frequencies of interfaces*:** The ratio of polar amino acids to all amino acids in interfaces.

***Polar aa frequencies of hot spots*:** The ratio of the polar amino acids to non polar amino acids in hot spots.

***Polar aa frequencies of hot regions*:** The ratio of the polar amino acids to non polar amino acids in hot regions.

***Aa distribution in hot regions*:** Amino acid distribution of the hot spots in hot regions.

3.1.4 Automatic Classification of DD and PP Interfaces Based on Hot Regions

Machine learning (ML) methods are widely used for classification tasks. The differences in the organization hot spots in DD and PP interfaces can be used to automatically classify protein-protein interactions (for the ones with available complex structures) as hub/non hub interactions. 38 PPs, 26 DDs and 99 NNs which have hot regions in their interfaces are used in the training and prediction step by using 10 fold cross validation (In 10 fold cross validation method, the dataset is randomly divided into ten equal partitions. One of them is selected as the test set and the model is trained in the

remaining nine partitions. This procedure is repeated ten times). Support vector machine classifier (SVM) which is a well known ML classifier to demonstrate the success of classifying interfaces using hot region characteristics is used. SVM [66] is an algorithm which can classify the data by using features of the training data. Its output is robust to imperfect data. It classifies the data using a generated hyperplane. It maximizes the margin of the hyperplane using different kernel types such as, radial kernel, sigmodial kernel, linear kernel, Gaussian kernel and polynomial kernel. These kernels are utilized to find the best fit SVM model for the data which have different characteristic and pattern. In addition to SVM model, RBF network, nearest neighbor, decision tree, regression, naïve bayes and k-means clustering models are applied but SVM gives the best result. Therefore the results of SVM are provided in the following sections. The parameters used for classification and their significance between different types of protein protein interfaces (DD, PP, NN) are listed in **Table 3.1**. The p-values for candidate features are obtained by using ANOVA (analysis of variance) test. P-value is the probability of test statistics. If the p-values of the features are smaller than 0.05, they can be used as a feature for ML classification.

The assessment of the classification is done by the accuracy, precision and recall values of the ML methods. The definition and the meanings of the accuracy, precision and recall are:

TP: number of true positives

TN: number of true negatives

FP: number of false positives

FN: number of false negatives

$\text{accuracy} = \frac{TP+TN}{TP+FP+FN+TN}$ (the measure of closeness to the true value of the test)

$\text{precision} = \frac{TP}{TP+FP}$ (the measure of reproducibility of the test)

$\text{recall} = \frac{TP}{TP+FN}$ (the measure of completeness of the test)

Table 3-1 Statistical significance of the candidate features (p values, underlined number indicate the significant p values)

ANOVA significance test	PP-DD	PP-NN	DD-NN	(PP+DD) - NN
Hot spot ratio	$\underline{2.03 * 10^{-2}}$	$7.22 * 10^{-1}$	$1.18 * 10^{-1}$	$6.91 * 10^{-1}$
Average hot region size	$\underline{1.25 * 10^{-2}}$	$\underline{1.90 * 10^{-2}}$	$9.02 * 10^{-1}$	$7.56 * 10^{-2}$
Average number of hot regions	$9.02 * 10^{-2}$	$\underline{8.00 * 10^{-4}}$	$9.71 * 10^{-2}$	$\underline{5.00 * 10^{-4}}$
Average hot region ΔASA to interface ΔASA ratio	$\underline{8.00 * 10^{-4}}$	$\underline{4.00 * 10^{-4}}$	$1.59 * 10^{-1}$	$1.13 * 10^{-1}$
Polar amino acid (aa) frequencies of interface	$\underline{7.00 * 10^{-4}}$	$\underline{5.00 * 10^{-5}}$	$3.74 * 10^{-1}$	$\underline{1.98 * 10^{-2}}$
Polar aa frequencies of hot spots	$\underline{1.10 * 10^{-3}}$	$9.35 * 10^{-2}$	$\underline{4.10 * 10^{-3}}$	$5.95 * 10^{-1}$
Polar aa frequencies of hot regions	$\underline{2.68 * 10^{-2}}$	$5.47 * 10^{-1}$	$\underline{2.61 * 10^{-2}}$	$4.03 * 10^{-1}$

3.2 Results

A protein–protein interface consists of two binding sites of two proteins interacting with each other. Results presented in this section are based on the structural interface properties of the interface dataset that contains 26 DDs, 38 PPs and 99 NNs.

Figure 3.5 shows the ratio of hotspots clustered in the hot regions to the overall number of hotspots in the interfaces. The left hand side of the figure shows the distribution of the average fractions where diamonds, triangle and square shapes correspond to PP, DD and NN interfaces, respectively. The right hand side figure shows the histogram of the fractions for the three interface types. DD interfaces consist of a high fraction of their hot spots clustered in the hot regions (with an average of 0.75±0.21) as opposed to PP interfaces (an average of 0.62±0.21). It should be noted that standard deviations are quite high, but the two distributions are statistically significant different means. Details of the distributions are

provided as a box plot of the hot spot ratio given in **Figure 3.6**. The NN interfaces have an average of 0.69±0.17 (See **Table 3.2**). **Figure 3.7A** illustrates the histogram of the hot region sizes (average number of hot spots per hot region). The averages for DD, PP and NN interfaces are 6.99±3.92, 4.95±2.43, and 7.13±5.45, respectively. The results reveal that hot regions in DD interfaces are larger than that of PP interfaces. Part (**B**) of the figure shows the average number of hot regions in the three different types of interfaces. The averages are as follows for DD, PP and NN interfaces: 2.04, 1.58, and 2.75. Similarly, Part (C) displays the averages of the ratios of accessible surface areas of the hot regions to the overall interfaces. Overall, these two figures clearly show that DD interface hot spots are more organized in the hot regions. Hot spots are more clustered in DD interfaces compared to PP and NN interfaces. In other words, in PP interfaces one observes more isolated hot spots. On the other hand, hot regions in DD are the largest (both in terms of ASA and number of residues composed of) and they cover a high fraction of the total interface. These suggest that DD interfaces are mostly mediated by clustered hot spots (namely hot regions). The close contact among many hot spots may also indicate the cooperativity of these residues in DD interfaces. There are clear differences between the organization of hot spots and hot regions between the hub proteins and non-hub protein interfaces as well as significant differences between date and party hub interfaces.

Further, interface sizes of date hubs are observed to be larger (2066 $Å^2$) than party hubs (1823 $Å^2$) and smaller than non-hub proteins. Since party hubs interact with their partners through distinct sites, it is expectable to have smaller binding sites in party hubs. Physically, it would be impossible to locate large and many interfaces on a single protein surface. Non-hub proteins presumably interact with their partners through specific interactions, therefore one would expect to see larger binding sites which would be an indication of the strong interaction between the proteins. When the average sizes of the hot regions in these interfaces are investigated, it is observed that hot regions are much larger

in DD interfaces compared to PP interfaces. When the average change in accessible surface area of individual hot spots are investigated, in DD interfaces it is observed that hot spots are more exposed (change in accessible surface area is around 115 $Å^2$) compared to the ones in PP interfaces (change in accessible surface area of around 80$Å^2$). In NN interfaces this number is 135 $Å^2$. **Table 3.1** shows the p-values of the above parameters to discriminate PP, DD and NN interfaces. The underlined numbers are lower than 0.05 indicating that corresponding interface types are statistically significant from each other. This table clearly shows that PP and DD interfaces are the ones that show different characteristics. PP and NN can also be differentiated. On the other hand it is hard to discriminate DD from NN and hub from non-hub proteins in general.

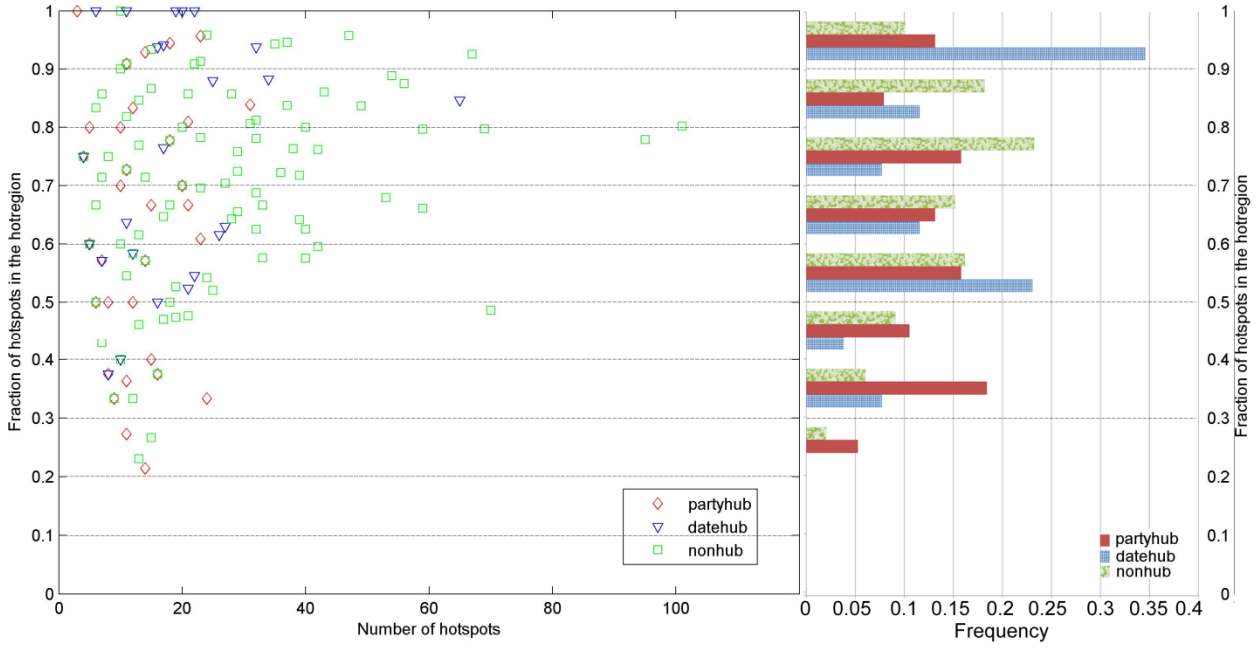


Figure 3.5 The distribution of the fraction of hot spots in the hot regions and their frequency according to their types. Date hub proteins have more tendencies to be involved in a hot region.

Table 3-2 Mean and standart deviation of the features.

	DD	PP	NN
Hot spot ratio	0.75 ± 0.21	0.62 ± 0.21	0.69 ± 0.17
Average hot region size	6.99 ± 3.92	4.95 ± 2.43	7.13 ± 5.45
Average number of hot regions	2.04 ± 1.37	1.58 ± 0.76	2.75 ± 2.04
Average hot region ΔASA to interface ΔASA ratio	0.36 ± 0.16	0.23 ± 0.13	0.32 ± 0.12
Hot region ASA(Å^2)	801.31 ± 649.99	397.31 ± 275.20	964.95 ± 829.32
Interface ASA(Å^2)	2066.54 ± 1235.19	1823.37 ± 871.88	2871.08 ± 1968.28
Number of residues in interfaces	43.19 ± 24.95	38.82 ± 18.67	61.35 ± 41.87

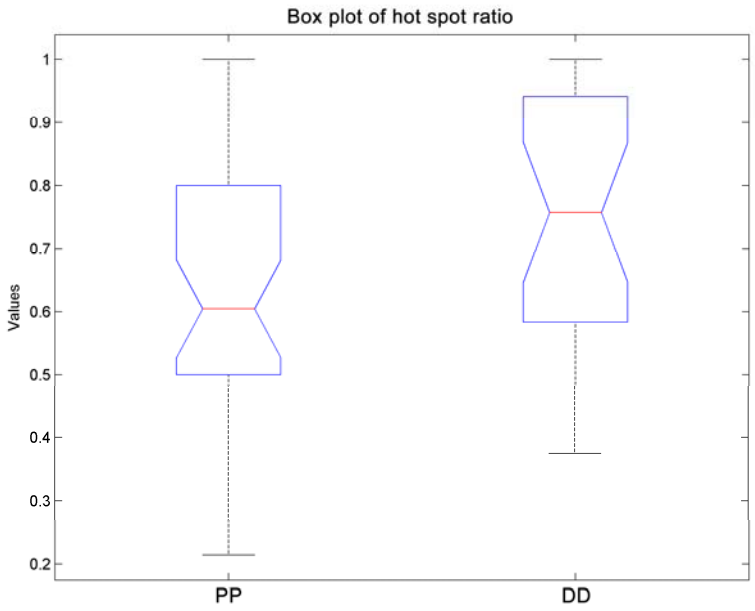


Figure 3.6 The notches are the confidence intervals in the box plot. If the notches do not overlap the two medians are significantly different. The notches of the box plot of the hot spot ratio do not overlap.

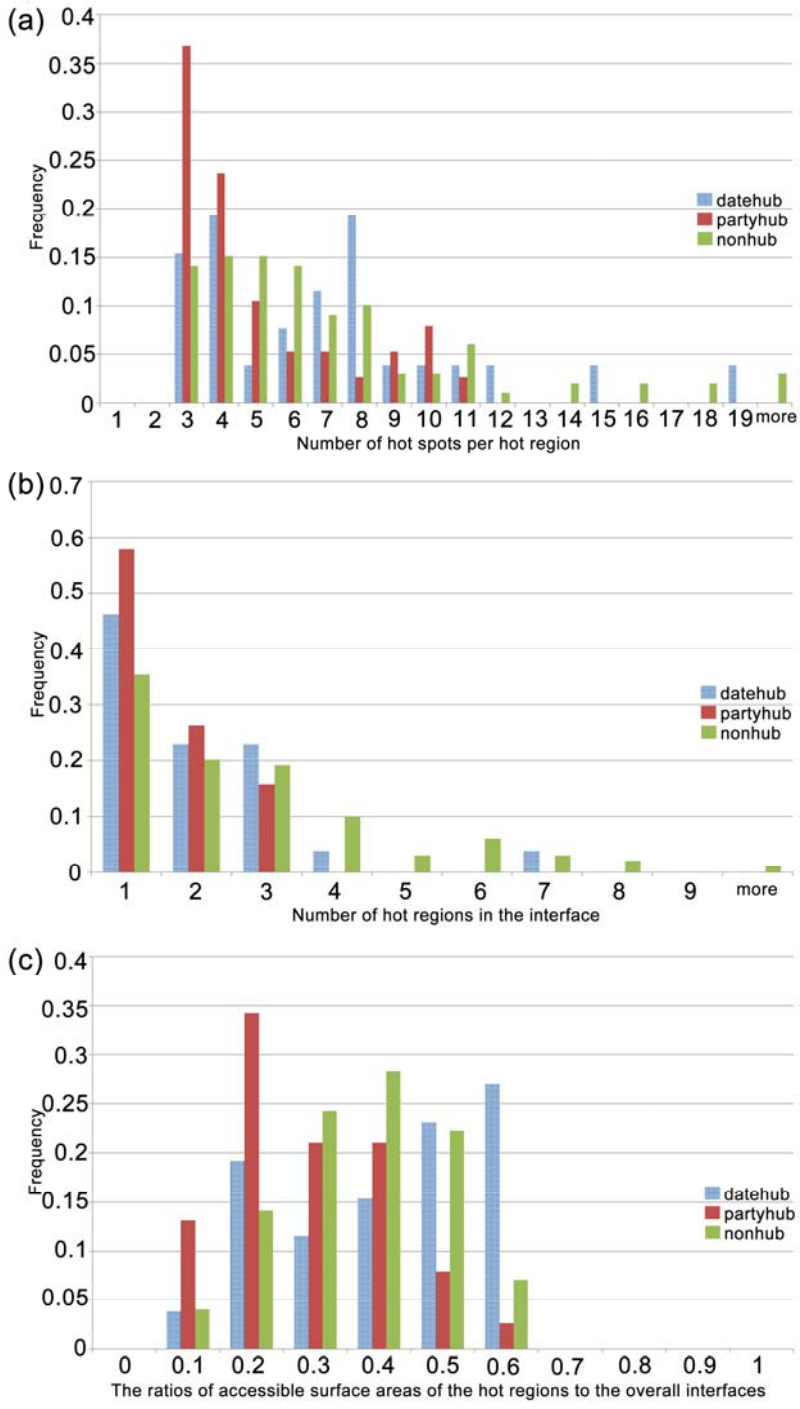


Figure 3.7 (a) The histogram of the hot region sizes. (b) The histogram shows the average number of hot regions in the interfaces. (c) The histogram displays the averages of the ratios of accessible surface areas of the hot regions to the overall interfaces.

3.2.1 Organization of Hot Regions in Hubs

Protein evolution is crucial in the sense that conserved functional domains of proteins generally correspond to specific binding surfaces which puts light to important biological processes in the cell. Studies so far have shown that rate of evolution of proteins are affected by dispensability of the protein for the cell, the level of transcription of the gene encoding the protein and the number of protein-protein interactions involved. There are two opposing ideas about the relationship between the evolutionary rate of proteins and the number of interactions they make. Fraser et al. [58] indicate that hubs of *S. cerevisiae* interactome evolve slowly with a suggested cause of them having larger regions responsible for interactions than that of non-hubs. Proteins with many interactors have smaller evolutionary rates since their structures are the key point in making so many interactions which limits the number of mutations acceptable and hence their evolution. In their study, they determined the evolutionary rates by comparing the orthologous sequences between *S. cerevisiae* and *C. elegans* and they analyzed the correlation between the evolutionary rate data and protein-protein interaction data. They also claimed that evolution rates for interacting pairs of proteins are very similar suggesting a co-evolution taking place. On the other hand, Jordan et al. [59] claimed that a simple dependence between evolution rate and high connectivity does not exist and the correlation is only due to slow evolution of a few proteins making many interactions. As a response to that, in another study Fraser et al. [58] showed a stronger correlation between evolutionary rate and connectivity than their previous study. This time, they compared yeast with closer species than *C. elegans* which are *S. pombe* and *C. albicans* to find the evolutionary rates and they used a more complete data of protein-protein interactions. They criticized Jordan et al.'s [59] conclusions for being based on less sufficient protein-protein interaction data than theirs. Later, when two different types of hubs (date and party) were determined, the discrepancy between different views could be explained to an extent. Usually the

evolutionary rate of date hubs was reported to be higher than party hubs, so party hubs were found to be more conserved.

By making an analogy between the hot spots and conserved residues [6, 67] (although these two terms are not fully correlated), here it is argued that date hub interfaces use a different strategy to locate their hot spots and thus communicate with their partners. There are more distinct hot regions in DD interfaces, maybe this might be due to the fact that DD interfaces should be re-used to bind to different partners, and different hot regions can be used to bind to different partners. Or, as another scenario, since hot regions are significantly larger in DD interfaces, some portions of the hot spots are used to bind to several partners whereas the other portions are used to bind to some others. As an example, protein G (a date hub) is illustrated in **Figure 3.8**. Protein G is represented as blue (dark) in all three figures. Three different proteins binding on the similar region of protein G are shown in yellow (parts A, B, C). Hot regions of protein G are shown in cyan whereas hot regions of the partner proteins are orange. This figure shows that different hot regions can be utilized to bind the different partners.

Previously, it is stated that hot regions can act as pre-organized binding sites even in unbound forms. Keeping in mind that a date hub usually interacts with a date hub and party hub interacts with a party hub [54], it makes sense that date hubs can reach the level of specificity as well as speed in recognizing each other with the hot regions on their binding sites. Therefore, similar organization of hot regions among date hubs can provide them advantage in their fast yet specific recognition.

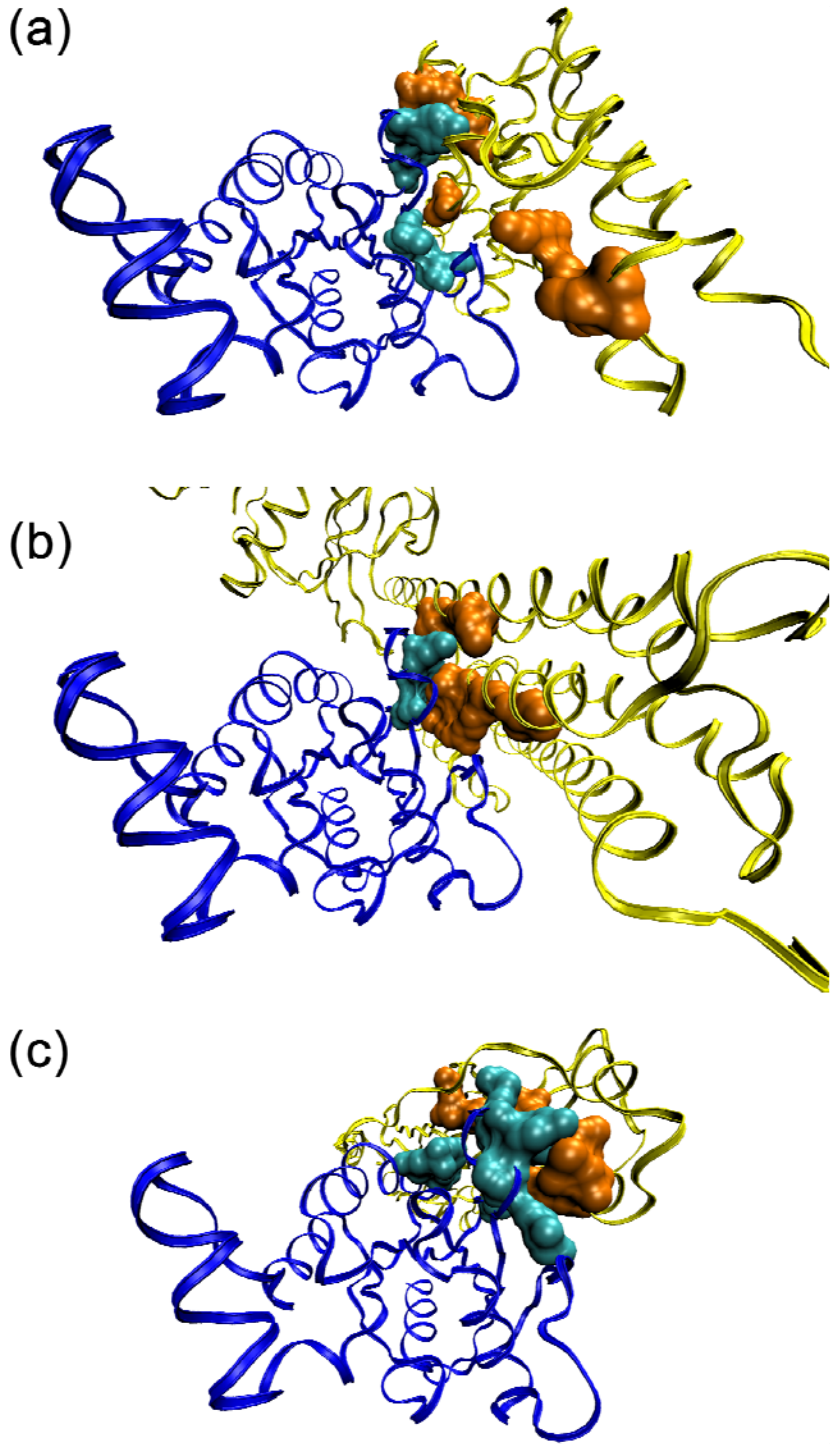


Figure 3.8 Protein G is represented as blue (dark) in all three figures. All complexes are taken from PDB [(a) 1GZS_CD, (b) 1KI1_CD, (c) 1DOA_AB]. Three different proteins binding on the similar region of protein G are shown in yellow. Hot regions of protein G are shown in cyan whereas hot regions of the partner proteins are orange. This figure shows that different hot regions can be utilized to bind the different partners.

3.2.2 Amino Acid Composition of Hot Regions in Hub Proteins

Amino acid composition of interfaces generally differs from the rest of the protein surfaces.[68] However, the differences are not pronounced significantly over all interfaces. If types of interfaces are considered such as homodimer interfaces, transient interfaces, or interfaces of disordered segments, the amino acid compositions can be more discriminative. Hydrophobic and polar interactions seem to be playing important role in protein interfaces. Therefore, amino acids are grouped into two categories: polar amino acids (R, N, D, E, Q, H, K, S, T, Y) and non-polar ones (A, C, G, I, L, M, F, P, W, V) to investigate if hot regions have a specific preference for hydrophobic or polar interactions. **Table 3.3** depicts the fraction of polar residues for all interface residues, for hot spot residues, and for hot regions.

The amino acid composition in interfaces, hotspots, and hot regions of DDs and PPs show differences. DD interfaces, which are likely more disordered, have lower polarity ratio than PPs. The ratio of polarity of hot spots is lower than that of interfaces; the ratio of polarity in hot regions is the lowest. The difference is significant particularly for DD type interfaces (0.18). Why the hot regions of DD type interfaces have more hydrophobic amino acids than that of PP or NN types? A recent study on disordered interfaces reports that, the interfaces that contain disordered regions (IUP interfaces) have higher ratio of hydrophobic amino acids compared to the ordered interfaces; also IUPs have more hydrophobic-hydrophobic interactions than ordered proteins [69-71]. These hydrophobic-hydrophobic interactions in the interface provide the recognition of the binding sites, re-use of the same interface in multiple biological processes and highly structured interface [69-71]. These findings suggest that DD type interfaces are likely to contain disordered regions and involved in transient interactions.

One would be curious to see if similar organization also exists in binding surfaces of monomeric parts of proteins, albeit not bound to their partners. The results show the same

conclusion does not hold for one sides of the protein interfaces. Date, party and non-hub protein binding sites cannot be differentiated by using the same features in only one side of the interfaces (i.e., hot spot ratio, average hot region size, average hot region ΔASA to interface ΔASA ratio, Polar aa frequencies of interfaces, Polar aa frequencies of hot spots, Polar aa frequencies of hot regions). The p-values in all cases are greater than 0.05.

Table 3-3 Mean and standart deviation of the features.

	DD	PP	NN
Polar amino acid frequencies of hot regions	0.18 ± 0.17	0.27 ± 0.14	0.25 ± 0.14
Polar amino acid frequencies of hot spots	0.25 ± 0.25	0.43 ± 0.18	0.37 ± 0.17
Polar amino acid frequencies of interface	0.50 ± 0.12	0.61 ± 0.12	0.52 ± 0.11

3.2.3 Automatic Classification of Hub Interfaces

The analysis shows that organization of hot regions and their hydrophobicity differ between DD, PP, and NN interfaces. One can use these properties to classify a given interface using machine learning techniques (widely used for classification). The performance of the classification task can indicate the significance of these properties as well. **Table 3.1** demonstrates the discriminative power of various features (hot region characteristics that are discussed already). The features that are statistically significant (ANOVA significance test) for discriminating a particular interface type marked (with p-values less that 0.5). These features can be used for classifying a given interface. The result using all parameters (explained in the methods) and Support Vector Machine (SVM) yields an accuracy of 80%, a precision of 0.80 and a recall of 0.80. This high accuracy supports that these characteristics are discriminative between DD and PP interfaces.

3.3 Concluding Remarks

PPINs indicate that some proteins are highly connected to others (acting as hub proteins), whereas some others have a few interactions. Structural properties of interacting proteins can make these networks less abstract and can indicate the structural and physical basis of interactions. For example, two proteins interact through their interfaces where each residue contributes differently to the binding. Some residues are more critical in binding known as hot spots. These hot spots are not distributed uniformly in the interfaces but rather cluster into highly packed hot regions.

In this chapter, it is concluded that there is a relationship between organization of hot spots (hot regions) and the status of hub proteins. Interfaces are annotated as the ones between two date-hubs (DD), two party-hubs (PP) and two non-hubs (NN). It is concluded that there are clear differences between the organization of hot spots and hot regions between the hub proteins and non-hub protein interfaces as well as significant differences between date and party hub interfaces. 1) More of the hot spots are organized into the hot regions in DD interfaces compared to PP ones. 2) A high fraction of the interfaces are covered by hot regions in DD interfaces. 3) The number of distinct hot regions in DDs is higher. As a result of this study, it is argued that date hub interfaces use a different strategy to locate their hot spots and thus communicate with their partners. There are more distinct hot regions in DD interfaces, maybe this might be due to the fact that DD interfaces should be re-used to bind to different partners, and different hot regions can be used to bind to different partners. Or, as another scenario, since hot regions are significantly larger in DD interfaces, some portions of the hot spots are used to bind to several partners whereas the other portions are used to bind to some others.

Further, these hot region characteristics (Hot spot ratio, average hot region size, average hot region ΔASA to interface ΔASA ratio, polar amino acid (aa) frequencies of interfaces, polar aa frequencies of hot spots, polar aa frequencies of hot regions) can be

used to predict whether an interface is formed between a DD or PP type of an interface with accuracy of 80%.

Chapter 4
HOTREGION: A DATABASE OF HOT SPOT CLUSTERS

In this chapter, we combine the residue network topology with the residue energy profile based clustering approaches. The residue clusters in interfaces are called 'hot regions' [6, 49]. Hot regions are useful to interpret the protein interface properties. We present the database 'HotRegion' in order to illustrate hot spot cooperativity information at protein-protein interfaces.

4.1 Design and Implementation of HotRegion

Hotspot residues in interfaces are predicted with HotPoint [72] using accessible surface area (ASA) and knowledge based pair energies of each residue [21]. In order to define hot regions, a network of hotspots is constructed. In the network, the nodes are the hotspot residues and the edges are linked between nodes if the two hotspot residues are in contact. Two hotspot residues are defined as contacting if the distance between their Cα atoms is smaller than 6.5 Å [6]. Afterwards, connected components of the network are found and if the nodes in a connected component are equal or greater than three, the connected component is labeled as a hot region and the hotspot residues in this connected component labeled as the members of this hot region **(Figure 4.1)**.

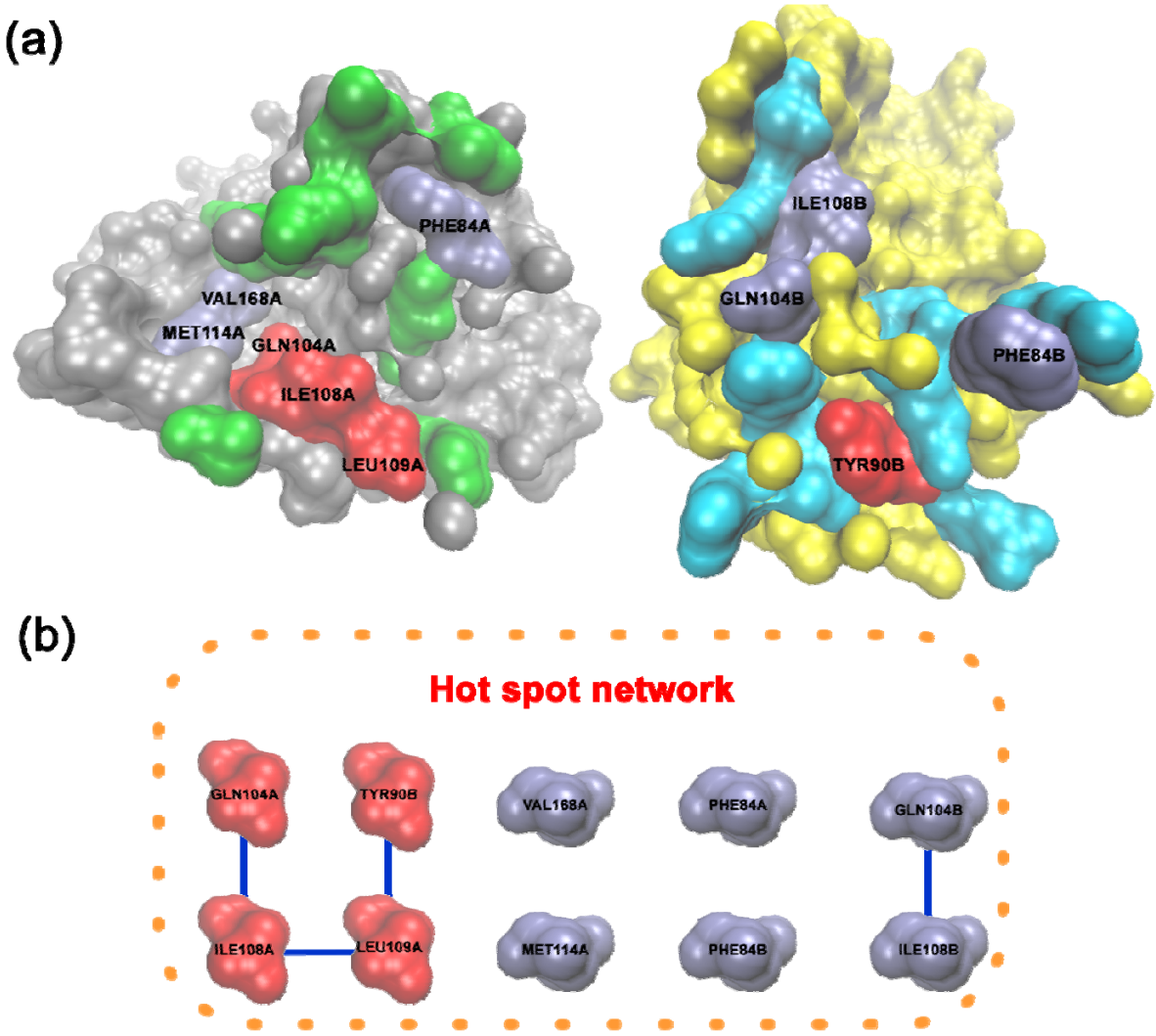


Figure 4.1 (a) Open form of interface 1GQPAB, the figure on the left is chain A, on the right is chain B. Red ones are the hot spot residues which construct hot region, ice blues are the hot spot residues which do not construct hot region. Greens are the chain A interface residues. Cyans are the chain B interface residues. (b) Hot region network and the connected components. Residues GLN104A, ILE108A, LEU109A and TYR90B are the members of the hot region.

4.1.1 Database Properties

The HotRegion database is available at http://prism.ccbb.ku.edu.tr/hotregion. HotRegion embraces three major components: a relational database management system for data storage and management, a web application to interface the database and a dynamically database update system. Data are stored in a relational MySQL database. The web application runs on an Apache web server hosted on a linux based system. PHP and JavaScript are used to implement the web application. The database can be updated dynamically.

4.1.2 Database Content

Currently, HotRegion contains all the PDB entries as of January 2011 (70695 PDB entries, 147892 protein-protein interfaces) and is using a dynamic update system which is based on the user's search queries. If a user searches hot region information of a complex (via PDB ID) which is not in the HotRegion database, the database can rapidly update itself and show the results. HotRegion has only protein-protein interface information. HotRegion database offers the researchers to find the hot regions of the protein complexes and provides structural properties of these complexes such as pair potentials of interface residues, ASA and relative ASA values of interface residues of both monomer and complex forms of proteins. Also, the visualization of the interface by using Jmol [73] and network of interactions of hot spot residues are presented in the results. An advanced search option is also available. Users can manipulate the HotRegion parameters by changing default values in advanced search section. Advanced searches are deposited in the database and users can retrieve their jobs by using email and job id from the 'Retrieve Job' section.

HotRegion needs atomic coordinates of the protein complexes in standard PDB format. If atoms are present in alternative locations, only the first location is considered. For NMR structures, the first model is used. HotRegion is specific to protein-protein interfaces; chains corresponding to DNA and RNA structures return no interface solutions.

If users do not supply enough information, the database asks for the missing information. The HotRegion database is free, open to all users and there are no login requirements.

4.2 Tutorial

4.2.1 Simple Search

Users retrieve the data of protein interfaces just by entering a PDB ID and two chain identifiers. Between the given monomers there must be an interface in order to get the hot region information. Also users have a control over the presentation of the results. Three properties of the interface (residue number, residue type, chain id) are always displayed in the result table and the output file, and the rest are displayed based on the preferences **(Figure 4.2)**.

4.2.2 Advanced Search

Users can retrieve the data based on their interface and hot region finding criteria. Users must enter email information in order to retrieve their jobs afterwards. They can supply a PDB file or enter a PDB code. After entering the chain information of the monomers which have interface between them, users can decide a valid interface extraction threshold which is summed with van der Waals radii of atoms. When the van der Waals threshold gets bigger, the number of interface residues will increase. Also users can change the hot spot neighbor criterion which is the Cα distance between the hot spots. When the hot region criterion gets bigger, the number of hot regions will decrease and hot regions start to merge in order to build larger hot regions.

4.2.3 Retrieve Job

The returning users can retrieve the results of previous jobs by using the job ids and their email addresses.

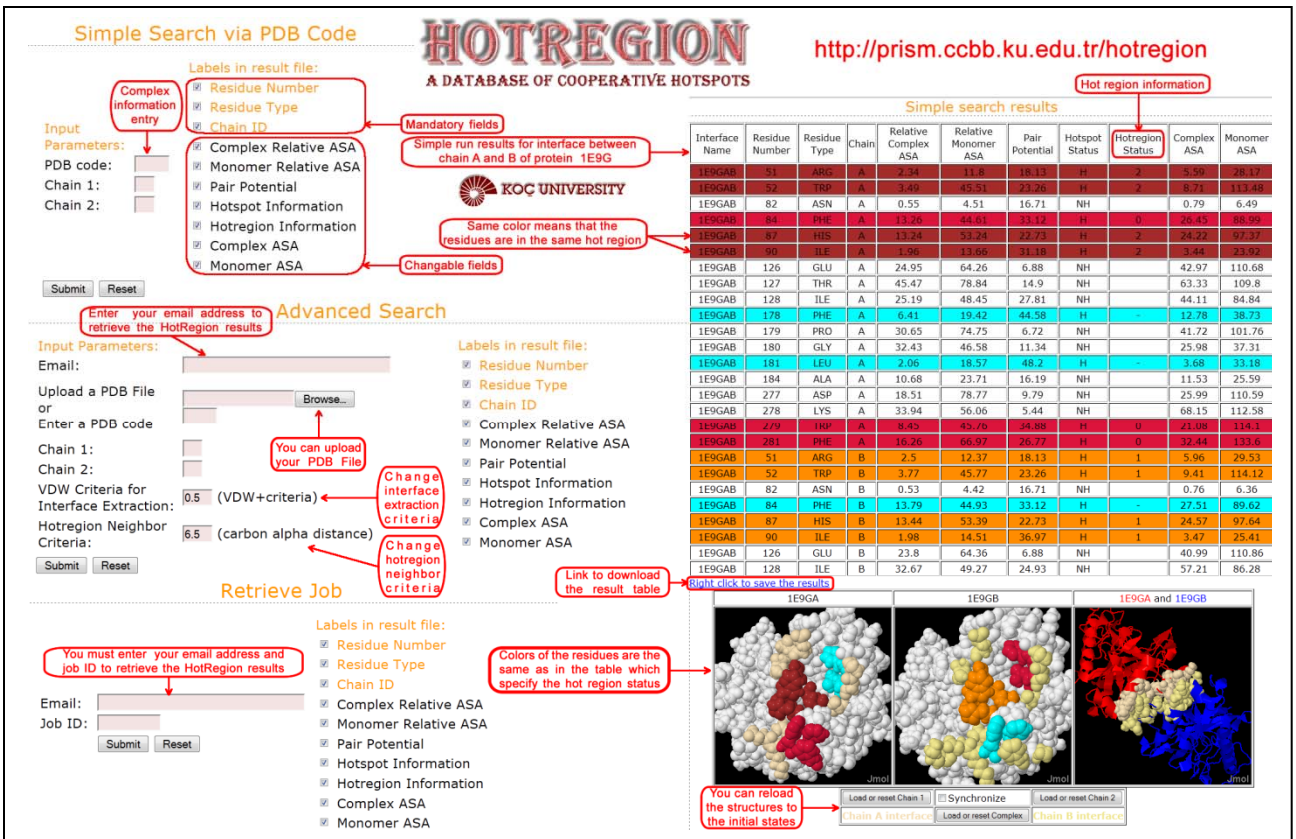


Figure 4.2 Properties of HotRegion Database in a quick view. On the left side of the figure, available search boxes and search requirements are presented, on the right side of the figure, an example of simple search results are presented. Also on the bottom-right, Jmol representation of the results are presented.

4.3 Case Study

Contribution to binding affinity of the proteins:

Colicins are plasmid-encoded, stress induced protein antibiotics that specifically target Escherichia coli cells. When it binds to a specific (cognate) partner, the nuclease can

protect the organism from endogenous and incoming colicin [74]. Kleanthous and coworkers showed that a limited number of mutations at the interface provide high-affinity binding to a noncognate partner [75]. According to this work, a noncognate complex between the colicin E9 endonuclease (E9 DNase) and immunity protein 2 (Im2) (PDB Id: 2WPT) has a weaker binding affinity than the cognate femtomolar E9 DNase – Im9 (PDB Id: 1EMV) interaction. When they substitute three Im2 residues with their Im9 counterparts (Im2 D33L/N34V/R38T) the binding energy is almost similar to the binding energy of cognate complex energy. HotRegion results for these complexes show that the predicted hot spots overlap with the experimental findings. The cognate complex has two hot regions but the noncognate complex has one hot region (Figure 4.3) (Table 4.1). The structural differences at the interface are based on the different side chain orientations. Possibly, cognate complex utilizes the two hot regions at the interface in order to increase the binding affinity of interaction. When the hot regions of both complexes are compared, it is observed that the only difference between the hot region residues at the cognate complex is L33 and V34 (they formed the extra hot region with T37 in cognate complex). When these residues used in the substitution experiment, they may probably form the extra hot region with T37 at noncognate complex in order to increase the binding affinity of the noncognate complex.

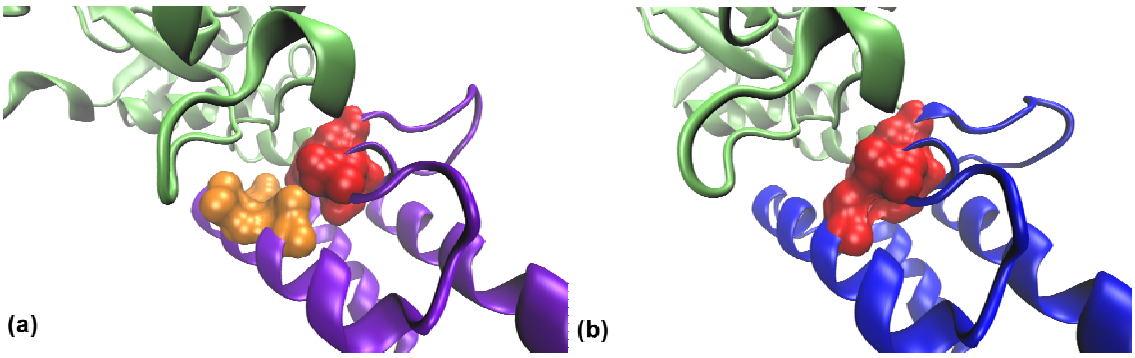


Figure 4.3 (a) Colicin E9 endonuclease (green) interacts with Im9 (purple) and the complex has two hot regions (red and orange). (b) Colicin E9 endonuclease (green) interacts with Im2 (blue) and the complex has one hot region (red).

Table 4-1 Hot region information search results from HotRegion Database for interfaces 1EMVAB and 2WPTAB.

interface	residue number	residue type	chain	hot region identifier
1EMVAB	33	LEU	A	1
1EMVAB	34	VAL	A	1
1EMVAB	37	VAL	A	1
1EMVAB	50	SER	A	0
1EMVAB	53	ILE	A	0
1EMVAB	54	TYR	A	0
2WPTAB	37	VAL	A	0
2WPTAB	50	SER	A	0
2WPTAB	53	ILE	A	0
2WPTAB	54	TYR	A	0

4.4 Concluding Remarks

A protein-protein interface consists of two binding sites of two proteins interacting with each other. For all different protein interactions, the binding energies of each complex are miscellaneous and the hot spot residues are distributed in a distinctive pattern. Extracting hot region information from not uniformly distributed binding energy of interfaces is important for analyzing the binding sites of the proteins. Some complexes are built upon more than one hot region, and size of the hot region is changing according to the binding site properties.

Previous research shows that such hot regions (hotspot clusters) are a signature for the protein-protein interfaces especially for hub proteins [49]. A hub protein binds different partner proteins by using different hot regions. These networked hotspot organization may imply that the contribution of the hotspots to the stability of the protein-protein complex within a hot region is cooperative. We hope the database will help in detecting cooperativity of functionally important residues, mutagenesis targets and understand the stability and specificity of protein-protein interfaces.

Chapter 5
CONCLUSION

As a consequence of improving experimental methods, structural data of proteins grows exponentially. To interpret tons of structural information is only possible with a systematic approach. Classification is one of the powerful sources to elucidate data. For that purpose, building PPINs of an organism and classifying proteins in the network according to their number of interactions to other proteins is a useful approach. Hub proteins which have multiple binding partners are extracted from PPIN and are used to draw a conclusion from PPIN. At the same time, structural properties of interacting proteins can make these networks less abstract and can indicate the structural and physical basis of interactions. Interfaces are the interaction components of the proteins and interface residues contribute differently to the binding. Hot spots are the key residues which can contribute the large part of the binding free energy. These hot spots are not distributed uniformly in the interfaces but rather clustered. The clustered hot spots are called hot regions. Evaluating hub proteins using hot region and interface properties showed that hub protein complexes can be classified as party-party hubs complexes and date-date hubs complexes. Date hubs which use single interface to interact many different partners have more hot regions than party hubs which use multiple interfaces to interact many different partners. It can be concluded that interfaces utilize combinations of these hot regions to bind multiple different partners.

We believe that the results provide insights for researchers working on characterization of protein interactions and multi partnered interfaces. Also, with its simple architecture and visualization tool, HotRegion would be useful both for experimentalist and computational scientist working on protein recognition, modeling of protein complexes and drug design.

In the future, multi binding partner interfaces in PDB can be derived and the cooperativity of the hot spots can be statistically determined. The hot region distribution across the interface of the multi binding partners can provide useful insights for protein interactions. Also, for the hot region detection, improvements and optimization in hot spot prediction method is crucial. For HotRegion, interface comparison tool which simplifies to evaluate hot region discrepancies across interface of different complexes can be added.

Once and for all, hot region definition is a useful method to evaluate protein – protein interactions and database of hot regions of all PDB entries is a rich source for studies about protein – protein interactions such as detection of the binding region patters, specificity and affinity of the binding sites, protein complex design, drug discovery etc.

Appendix A

Table 1. Non-redundant complexes

Date hub		
ordered locus name	uniprot name	interface
YFR034C	P07270	1A0AAB
YNL189W	Q02821	1BK5AB
YEL009C	P03069	1CE9AB
YEL009C	P03069	1CE9BD
YPL248C	P04386	1D66AB
YBR011C	P00817	1E9GAB
YNL189W	Q02821	1EE4AC
YNL189W	Q02821	1EE4AD
YDL185W	P17255	1EF0AB
YKR002W	P29468	1FA0AB
YPL153C	P22216	1G6GAB
YPL153C	P22216	1G6GAE
YEL009C	P03069	1GK6AB
YGL240W	P53068	1GQPAB
YPL240C	P02829	1HK7AB
YEL009C	P03069	1IJ2AB
YLR191W	P80667	1JQQAB
YLR191W	P80667	1JQQAD
YEL009C	P03069	1KQLAB
YEL009C	P03069	1LLMCD
YPL218W	P20606	1M2OCD
YBR011C	P00817	1M38AB
YML065W	P54784	1M4ZAB
YMR043W	P11746	1MNMAB
YMR043W	P11746	1MNMAD
YMR043W	P11746	1MNMBC
YLR191W	P80667	1N5ZAP
YGL153W	P53112	1N5ZBQ
YPR182W	P54999	1N9RAB
YPR182W	P54999	1N9REF
YER148W	P13393	1NGMAB
YER148W	P13393	1NGMFI
YER148W	P13393	1NH2AC
YER148W	P13393	1NH2AD
YEL009C	P03069	1NKNAB
YEL009C	P03069	1NKNAC
YJL041W	P14907	1O6OAD
YDR227W	P11978	1PL5AS
YLR044C	P06169	1PVDAB
YOL149W	Q12517	1Q67AB
YBR135W	P20486	1QB3AC
YBR135W	P20486	1QB3BC
YOL038W	P40303	1RYPCD
YOL038W	P40303	1RYPDK
YOL038W	P40303	1RYPDL
YDR228C	P39081	1SZ9AB
YDR228C	P39081	1SZ9AC
YDR228C	P39081	1SZ9BC
YER148W	P13393	1TBPAB
YDL140C	P04050	1TWFAB
YDL140C	P04050	1TWFAC
YDL140C	P04050	1TWFAE
YDL140C	P04050	1TWFAF
YDL140C	P04050	1TWFAH
YDL140C	P04050	1TWFAI
YDL140C	P04050	1TWFAK
YDL185W	P17255	1UM2AC

YNL189W	Q02821	1UN0AB
YNL189W	Q02821	1UN0AC
YNR052C	P39008	1UOCAB
YPL240C	P02829	1US7AB
YPL240C	P02829	1USUAB
YPL240C	P02829	1USVAC
YPL240C	P02829	1USVCE
YPL240C	P02829	1USVFG
YNL189W	Q02821	1WA5AB
YNL189W	Q02821	1WA5BC
YEL037C	P32628	1X3ZAB
YDR404C	P34087	1Y14BD
YDR404C	P34087	1Y14CD
YOL135C	Q08278	1YKHAB
YML065W	P54784	1ZBXAB
YDR283C	P15442	1ZXEAE
YDR283C	P15442	1ZXEBC
YDR283C	P15442	1ZY4AB
YBL016W	P16892	2B9HAC
YFL038C	P01123	2BCGGY
YLR347C	Q06142	2BKUAB
YLR347C	Q06142	2BKUBD
YLR347C	Q06142	2BPTAB
YPL240C	P02829	2BREAB
YNL189W	Q02821	2C1TAC
YDR477W	P06782	3HYHAB

Party hub		
ordered locus name	uniprot name	interface
YER009W	P33331	1GY7AB
YER009W	P33331	1GY7BD
YER009W	P33331	1GYBAE
YCR088W	P15891	1HQZ27
YCR088W	P15891	1HQZ35
YCR088W	P15891	1HQZ56
YLR127C	Q12440	1LDDAC
YLR127C	Q12440	1LDDAD
YPR181C	P15303	1M2OAB
YPR181C	P15303	1M2OAC
YIL109C	P40482	1M2VAB
YLR026C	Q01590	1MQSAB
YIL109C	P40482	1PD0AB
YBL041W	P23724	1RYP12
YGL011C	P21243	1RYPAB
YGL011C	P21243	1RYPAH
YGL011C	P21243	1RYPAI
YER094C	P25451	1RYPBJ
YER094C	P25451	1RYPCJ
YER012W	P22141	1RYPCK
YER012W	P22141	1RYPDK
YMR314W	P40302	1RYPFG
YMR314W	P40302	1RYPFM
YFR050C	P30657	1RYPFN
YOR362C	P21242	1RYPGH
YFR050C	P30657	1RYPGN
YFR050C	P30657	1RYPHN
YBL041W	P23724	1RYPI1
YER094C	P25451	1RYPIJ
YBL041W	P23724	1RYPJ1
YER012W	P22141	1RYPJK
YER094C	P25451	1RYPJZ
YER012W	P22141	1RYPKL
YER012W	P22141	1RYPKY
YER012W	P22141	1RYPKZ
YFR050C	P30657	1RYPMN
YOR157C	P25043	1RYPMW
YFR050C	P30657	1RYPNV
YFR050C	P30657	1RYPNW
YML092C	P23639	1RYPOP

YOR157C	P25043	1RYPOW
YGR135W	P23638	1RYPPQ
YML092C	P23639	1RYPPW
YML092C	P23639	1RYPPX
YGR135W	P23638	1RYPQX
YGR135W	P23638	1RYPQY
YPR103W	P30656	1RYPRZ
YBL041W	P23724	1RYPS1
YPR103W	P30656	1RYPSZ
YBL041W	P23724	1RYPT1
YOR157C	P25043	1RYPVW
YOR157C	P25043	1RYPWX
YPR103W	P30656	1RYPYZ
YBL041W	P23724	1RYPZ1
YOL094C	P40339	1SXJAB
YNL290W	P38629	1SXJAC
YBR087W	P38251	1SXJAE
YOL094C	P40339	1SXJBC
YBR088C	P15873	1SXJBG
YJR068W	P40348	1SXJCD
YBR088C	P15873	1SXJCF
YBR087W	P38251	1SXJDE
YBR088C	P15873	1SXJGH
YOR151C	P08518	1TWFAB
YIL021W	P16370	1TWFAC
YBR154C	P20434	1TWFAE
YPR187W	P20435	1TWFAF
YOR224C	P20436	1TWFAH
YGL070C	P27999	1TWFAI
YIL021W	P16370	1TWFBC
YGL070C	P27999	1TWFBI
YOR151C	P08518	1TWFBJ
YOR151C	P08518	1TWFBK
YOR151C	P08518	1TWFBL
YIL021W	P16370	1TWFCJ
YIL021W	P16370	1TWFCK
YIL021W	P16370	1TWFCL
YLR335W	P32499	1UN0AC
YFL039C	P60010	1YAGAG
YLL036C	P32523	2BAYAD
YER136W	P39958	2BCGGY
YLR335W	P32499	2C1TAC

Non hub		
ordered locus name	uniprot name	interface
YDR256C	P15202	1A4EAB
YDR256C	P15202	1A4EAC
YDR256C	P15202	1A4EAD
YIL160C	P27796	1AFWAB
YCL067C	Q6B2C0	1AKHAB
YPR074C	P23254	1AY0AB
YDL235C	Q07688	1C02AB
YLL050C	Q03048	1CFYAB
YBR035C	P38075	1CI0AB
YJR099W	P35127	1CMXAB
YJR099W	P35127	1CMXAC
YJR099W	P35127	1CMXBC
YPR060C	P32178	1CSMAB
YPR073C	P40347	1D1PAB
YPR073C	P40347	1D1QAB
YPL228W	O13297	1D8HAB
YPL228W	O13297	1D8HAC
YMR174C	P01094	1DPJAB
YEL021W	P03962	1DQWAB
YGR006W	P33411	1DVKAB
YJR139C	P31116	1EBFAB
YOL143C	P50861	1EJBAB
YPL020C	Q02724	1EUVAB
YBL045C	P07256	1EZVAB

Q0105	P00163	1EZVAC	YLR163C	P10507	1HR6GH
YEL024W	P08067	1EZVAE	YGL187C	P04037	1HR8BO
YJL166W	P08525	1EZVAG	YFL017C	P43577	1I12AB
YGR183C	P22289	1EZVAI	YFL017C	P43577	1I12AC
YOR065W	P07143	1EZVCD	YFL017C	P43577	1I21BM
YEL024W	P08067	1EZVCE	YFL017C	P43577	1I21BY
YDR529C	P00128	1EZVCF	YOR143C	P35202	1IG0AB
YJL166W	P08525	1EZVCG	YJR010W	P08536	1J70AB
YOR065W	P07143	1EZVDE	YGL087C	P53152	1JATAB
YDR529C	P00128	1EZVDF	YER057C	P40037	1JD1AB
YJL166W	P08525	1EZVDG	YDR419W	Q04049	1JIHAB
YFR033C	P00127	1EZVDH	YGR180C	P49723	1JK0AB
YGR183C	P22289	1EZVDI	YMR038C	P40202	1JK9AD
YEL024W	P08067	1EZVEG	YMR038C	P40202	1JK9BD
YEL024W	P08067	1EZVEI	YPR037C	Q12284	1JR8AB
YEL024W	P08067	1EZVEX	YPR037C	Q12284	1JRAAC
YEL024W	P08067	1EZVEY	YMR108W	P07342	1JSCAB
YDR529C	P00128	1EZVFG	YBR248C	P33734	1JVNAB
YFR033C	P00127	1EZVHG	YNL200C	P40165	1JZTAB
YKL069W	P36088	1F5MAB	YNL229C	P23202	1K0DBC
YLR351C	P49954	1F89AB	YNL229C	P23202	1K0DCD
YPR141C	P17119	1F9WAB	YML054C	P00175	1KBIAB
YPL154C	P07267	1FMXAB	YEL022W	P39993	1KU1AB
YDR177W	P21734	1FZYAB	YDR529C	P00128	1KYOAR
YML022W	P49435	1G2QAB	YPR191W	P07257	1KYOBM
YBR034C	P38074	1G6Q26	YDR529C	P00128	1KYOBR
YBR034C	P38074	1G6Q56	Q0105	P00163	1KYOCN
YPR074C	P23254	1GPUAB	YOR065W	P07143	1KYODO
YLR300W	P23776	1H4PAB	YEL024W	P08067	1KYOEN
YBR249C	P32449	1HFBAC	YBL045C	P07256	1KYOGL
YBR249C	P32449	1HFBBF	YPR191W	P07257	1KYOGM
YHR024C	P11914	1HR6AB	YJR048W	P00044	1KYOOW
YHR024C	P11914	1HR6AE	YBR213W	P15807	1KYQAB
YHR024C	P11914	1HR6AG	YBR213W	P15807	1KYQAC
YLR163C	P10507	1HR6BD	YOR176W	P16622	1LBQAB

YCR097W	P01366	1LE8AB
YOL049W	Q08220	1M0WAB
YHR124W	P38830	1M6UAB
YHR124W	P38830	1M7UAB
YCL067C	Q6B2C0	1MNMAD
YCL067C	Q6B2C0	1MNMBC
YAL012W	P31373	1N8PAB
YAL012W	P31373	1N8PAC
YAL012W	P31373	1N8PAD
YDR050C	P00942	1NEYAB
YNL168C	P53889	1NKQAB
YNL168C	P53889	1NKQAD
YNL168C	P53889	1NKQBE
YGL030W	P14120	1NMUBC
YGL030W	P14120	1NMUCD
YDR292C	P32916	1NRJAB
YOR357C	Q08826	1OCUAB
YKL186C	P34232	1OF5AB
YBR249C	P32449	1OF8AB
YDL235C	Q07688	1OXKAB
YDL235C	Q07688	1OXKAC
YDL235C	Q07688	1OXKAD
YDL235C	Q07688	1OXKAG
YIL147C	P39928	1OXKCF
YDL235C	Q07688	1OXKCK
YIL147C	P39928	1OXKEF
YJL153C	P11986	1P1HAC
YJL153C	P11986	1P1HAD
YJL153C	P11986	1P1HBD
YJL153C	P11986	1P1JAB
YML097C	P54787	1P3QQR
YML097C	P54787	1P3QQV
YML097C	P54787	1P3QRV
YPR062W	Q12178	1P6OAB
YMR092C	P46680	1PGUAB
YMR318C	Q04894	1PIWAB
YJL068C	P40363	1PV1AB
YJL068C	P40363	1PV1AC
YIL160C	P27796	1PXTAB
YKR018C	P36114	1PYOAE
YKR018C	P36114	1PYOBE
YPL015C	P53686	1Q17AB
YBR223C	P38319	1Q32AB
YMR216C	Q03656	1Q8YAB
YOR265W	P48606	1QSDAB
YPR193C	Q06592	1QSMAB
YPR193C	Q06592	1QSMAC
YMR038C	P40202	1QUPAB
YDR533C	Q04432	1QVVAB
YDR533C	Q04432	1QVZAB
YGL148W	P28777	1R52AB
YGL148W	P28777	1R52AC
YGL148W	P28777	1R52AD
YLR245C	Q06549	1R5TAB
YLR245C	Q06549	1R5TAC
YLR245C	Q06549	1R5TAD
YNL238W	P13134	1R64AB
Q0160	P03882	1R7MAB
YJR031C	P47102	1RE0AB
YDR435C	Q04081	1RJDAB
YDR435C	Q04081	1RJDAC
YMR020W	P50264	1RSGAB
YDR483W	P27809	1S4NAB
YJR048W	P00044	1S6VAD
YJR048W	P00044	1S6VCD
YIR029W	P25335	1SG3AB
YOR217W	P38630	1SXJAB
YOR217W	P38630	1SXJAC
YOR217W	P38630	1SXJAE
YOR217W	P38630	1SXJAG

YLR011W	Q07923	1T0IAB
YMR239C	Q02555	1T4OAB
YMR108W	P07342	1T9BAB
YMR108W	P07342	1T9DAD
YJR019C	P41903	1TBUAB
YJR019C	P41903	1TBUAC
YJR019C	P41903	1TBUAD
YDR044W	P11353	1TKLAB
YDR044W	P11353	1TLBAQ
YDR044W	P11353	1TLBSU
YOL005C	P38902	1TWFAK
YOL005C	P38902	1TWFBK
YHR143W-A	P40422	1TWFBL
YOL005C	P38902	1TWFCK
YHR143W-A	P40422	1TWFCL
YDR044W	P11353	1TXNAB
YDR440W	Q04089	1U2ZAC
YDR214W	Q12449	1USUAB
YDR214W	Q12449	1USVFG
YFL018C	P09624	1V59AB
YDR428C	Q04066	1VKHAB
YOR209C	P39683	1VLPAB
YOR209C	P39683	1VLPBC
YPR118W	Q06489	1W2WAB
YPR118W	Q06489	1W2WFJ
YGL238W	P33307	1WA5AC
YGL238W	P33307	1WA5BC
YJL020C	P47068	1WDXAD
YMR297W	P00729	1WPXAB
YPL096W	Q02890	1X3ZAB
YML079W	Q03629	1XE7AB
YML079W	Q03629	1XE7BC
YHR049W	P38777	1YCDAB
YNR071C	P53757	1YGAAB
YPL084W	P48582	1ZB1AB
YPL145C	P35844	1ZI7AB
YPL145C	P35844	1ZI7BC
YOR133W	P32324	1ZM9AB
YOR133W	P32324	1ZM9AF
YKL015W	P25502	1ZMECD
YMR058W	P38993	1ZPUBD
YMR058W	P38993	1ZPUCE
YMR058W	P38993	1ZPUEF
YNR051C	P53741	1ZX2AB
Q0115	Q9ZZW7	2AB5AB
YGR254W	P00924	2AL1AB
YJR048W	P00044	2B0ZAB
YNL053W	P38590	2B9IAC
YHR079C	P32361	2BE1AB
YHR042W	P16603	2BF4AB
YHR042W	P16603	2BN4AB
YER010C	P40011	2C5QAB
YER010C	P40011	2C5QBE
YDR217C	P14737	2FF4AE
YJR057W	P00572	3TMKDG
YJR057W	P00572	3TMKGH
YPR060C	P32178	4CSMAB

Table 2. Complexes which have at least one hot region and similar type (Date, party or non hub) binding partner

Date hub - Date hub interface	Party hub - Party hub interface	Non hub - Non hub interface
1A0AAB	1GY7AB	1A4EAB
1CE9AB	1GY7BD	1A4EAD
1D66AB	1HQZ27	1AFWAB
1E9GAB	1HQZ35	1AKHAB
1G6GAB	1HQZ56	1AY0AB
1GK6AB	1M2OAC	1C02AB
1GQPAB	1M2VAB	1CFYAB
1IJ2AB	1RYP12	1CI0AB
1KQLAB	1RYPAB	1CMXAC
1M38AB	1RYPAI	1CSMAB
1MNMAB	1RYPBJ	1D8HAB
1N5ZAP	1RYPCJ	1DPJAB
1N9RAB	1RYPCK	1DQWAB
1N9REF	1RYPFG	1EBFAB
1NKNAB	1RYPFN	1EJBAB
1PL5AS	1RYPI1	1EZVAB
1PVDAB	1RYPIJ	1EZVAC
1QB3AC	1RYPJ1	1EZVAG
1TBPAB	1RYPJK	1EZVCD
1UM2AC	1RYPJZ	1EZVCE
1UOCAB	1RYPKZ	1EZVCF
1Y14BD	1RYPMN	1EZVCG
1ZXEAE	1RYPMW	1EZVDH
1ZY4AB	1RYPNW	1EZVEI
2BKUBD	1RYPOP	1F5MAB
2BPTAB	1RYPOW	1F89AB
	1RYPPQ	1G2QAB
	1RYPPX	1G6Q56
	1RYPQY	1GPUAB
	1RYPWX	1HR6AB
	1RYPZ1	1HR6AE

	1SXJBC	1HR6GH
	1SXJCD	1HR8BO
	1SXJDE	1I12AC
	1SXJGH	1I21BY
	1TWFBC	1IG0AB
	1TWFBI	1J70AB
	2BAYAD	1JD1AB
		1JK9BD
		1JR8AB
		1JSCAB
		1JVNAB
		1JZTAB
		1K0DCD
		1KBIAB
		1KU1AB
		1KYOCN
		1KYOEN
		1KYQAB
		1LBQAB
		1M7UAB
		1N8PAB
		1N8PAD
		1NEYAB
		1NKQAB
		1NRJAB
		1OF8AB
		1OXKAB
		1OXKAD
		1OXKCF
		1OXKEF
		1P1HAC
		1P1JAB
		1P3QQR
		1P6OAB
		1PIWAB

		1PV1AB
		1PXTAB
		1Q17AB
		1QSMAB
		1QUPAB
		1R52AB
		1R52AC
		1R5TAB
		1R5TAC
		1R5TAD
		1RSGAB
		1S4NAB
		1T0IAB
		1T9BAB
		1TBUAB
		1TBUAC
		1TBUAD
		1TLBSU
		1TXNAB
		1U2ZAC
		1V59AB
		1VKHAB
		1W2WAB
		1W2WFJ
		1WDXAD
		1ZI7AB
		1ZMECD
		1ZX2AB
		2AL1AB
		2BE1AB
		2C5QAB
		2C5QBE
		3TMKGH

BIBLIOGRAPHY

[1] Humphrey, W., A. Dalke, and K. Schulten, *VMD: visual molecular dynamics.* J Mol Graph, 1996. **14**(1): p. 33-8, 27-8.

[2] Berman, H.M., et al., *The Protein Data Bank.* Nucleic Acids Res, 2000. **28**(1): p. 235-42.

[3] Wells, J.A., *Systematic mutational analyses of protein-protein interfaces.* Methods Enzymol, 1991. **202**: p. 390-411.

[4] Clackson, T. and J.A. Wells, *A hot spot of binding energy in a hormone-receptor interface.* Science, 1995. **267**(5196): p. 383-6.

[5] Bogan, A.A. and K.S. Thorn, *Anatomy of hot spots in protein interfaces.* J Mol Biol, 1998. **280**(1): p. 1-9.

[6] Keskin, O., B. Ma, and R. Nussinov, *Hot regions in protein--protein interactions: the organization and contribution of structurally conserved hot spot residues.* J Mol Biol, 2005. **345**(5): p. 1281-94.

[7] Jeong, H., et al., *Lethality and centrality in protein networks.* Nature, 2001. **411**(6833): p. 41-2.

[8] Han, J.D., et al., *Evidence for dynamically organized modularity in the yeast protein-protein interaction network.* Nature, 2004. **430**(6995): p. 88-93.

[9] Kleanthous, C., *Protein–Protein Recognition, Frontiers in Molecular Biology*. 2000: Oxford University Press.

[10] Reynolds, C., D. Damerell, and S. Jones, *ProtorP: a protein-protein interaction analysis server.* Bioinformatics, 2009. **25**(3): p. 413-4.

[11] Gong, S., et al., *A protein domain interaction interface database: InterPare.* BMC Bioinformatics, 2005. **6**: p. 207.

[12] Stein, A., A. Ceol, and P. Aloy, *3did: identification and classification of domain-based interactions of known three-dimensional structure.* Nucleic Acids Res, 2011. **39**(Database issue): p. D718-23.

[13] Stein, A., A. Panjkovich, and P. Aloy, *3did Update: domain-domain and peptide-mediated interactions of known 3D structure.* Nucleic Acids Res, 2009. **37**(Database issue): p. D300-4.

[14] Stein, A., R.B. Russell, and P. Aloy, *3did: interacting protein domains of known three-dimensional structure.* Nucleic Acids Res, 2005. **33**(Database issue): p. D413-7.

[15] Davis, F.P. and A. Sali, *PIBASE: a comprehensive database of structurally defined protein interfaces.* Bioinformatics, 2005. **21**(9): p. 1901-7.

[16] Keskin, O., et al., *A new, structurally nonredundant, diverse data set of protein-protein interfaces and its implications.* Protein Sci, 2004. **13**(4): p. 1043-55.

[17] Tuncbag, N., et al., *Architectures and functional coverage of protein-protein interfaces.* J Mol Biol, 2008. **381**(3): p. 785-802.

[18] Thorn, K.S. and A.A. Bogan, *ASEdb: a database of alanine mutations and their effects on the free energy of binding in protein interactions.* Bioinformatics, 2001. **17**(3): p. 284-5.

[19] Fischer, T.B., et al., *The binding interface database (BID): a compilation of amino acid hot spots in protein interfaces.* Bioinformatics, 2003. **19**(11): p. 1453-4.
[20] Ofran, Y. and B. Rost, *Protein-protein interaction hotspots carved into sequences.* PLoS Comput Biol, 2007. **3**(7): p. e119.
[21] Tuncbag, N., A. Gursoy, and O. Keskin, *Identification of computational hot spots in protein interfaces: combining solvent accessibility and inter-residue potentials improves the accuracy.* Bioinformatics, 2009. **25**(12): p. 1513-20.
[22] Li, X., et al., *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): p. 781-95.
[23] Guerois, R., J.E. Nielsen, and L. Serrano, *Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations.* J Mol Biol, 2002. **320**(2): p. 369-87.
[24] Kortemme, T., D.E. Kim, and D. Baker, *Computational alanine scanning of protein-protein interfaces.* Sci STKE, 2004. **2004**(219): p. pl2.
[25] Darnell, S.J., D. Page, and J.C. Mitchell, *An automated decision-tree approach to predicting protein interaction hot spots.* Proteins, 2007. **68**(4): p. 813-23.
[26] Guney, E., et al., *HotSprint: database of computational hot spots in protein interfaces.* Nucleic Acids Res, 2008. **36**(Database issue): p. D662-6.
[27] Lise, S., et al., *Prediction of hot spot residues at protein-protein interfaces by combining machine learning and energy-based methods.* BMC Bioinformatics, 2009. **10**: p. 365.
[28] Gao, Y., R. Wang, and L. Lai, *Structure-based method for analyzing protein-protein interfaces.* J Mol Model, 2004. **10**(1): p. 44-54.
[29] Assi, S.A., et al., *PCRPi: Presaging Critical Residues in Protein interfaces, a new computational tool to chart hot spots in protein interfaces.* Nucleic Acids Res, 2010. **38**(6): p. e86.
[30] Cho, K.I., D. Kim, and D. Lee, *A feature-based approach to modeling protein-protein interaction hot spots.* Nucleic Acids Res, 2009. **37**(8): p. 2672-87.
[31] Huo, S., I. Massova, and P.A. Kollman, *Computational alanine scanning of the 1:1 human growth hormone-receptor complex.* J Comput Chem, 2002. **23**(1): p. 15-27.
[32] Gonzalez-Ruiz, D. and H. Gohlke, *Targeting protein-protein interactions with small molecules: challenges and perspectives for computational binding epitope detection and ligand finding.* Curr Med Chem, 2006. **13**(22): p. 2607-25.
[33] Rajamani, D., et al., *Anchor residues in protein-protein interactions.* Proc Natl Acad Sci U S A, 2004. **101**(31): p. 11287-92.
[34] Brinda, K.V., N. Kannan, and S. Vishveshwara, *Analysis of homodimeric protein interfaces by graph-spectral methods.* Protein Eng, 2002. **15**(4): p. 265-77.
[35] del Sol, A. and P. O'Meara, *Small-world network approach to identify key residues in protein-protein interaction.* Proteins, 2005. **58**(3): p. 672-82.
[36] Illingworth, C.J., et al., *Connectivity and binding-site recognition: applications relevant to drug design.* J Comput Chem, 2010. **31**(15): p. 2677-88.
[37] Kim, P.M., et al., *Relating three-dimensional structures to protein networks provides evolutionary insights.* Science, 2006. **314**(5807): p. 1938-41.

[38] Keskin, O. and R. Nussinov, *Similar binding sites and different partners: implications to shared proteins in cellular pathways.* Structure, 2007. **15**(3): p. 341-54.
[39] Liu, T., S.T. Whitten, and V.J. Hilser, *Functional residues serve a dominant role in mediating the cooperativity of the protein ensemble.* Proc Natl Acad Sci U S A, 2007. **104**(11): p. 4347-52.
[40] Reichmann, D., et al., *The modular architecture of protein-protein binding interfaces.* Proc Natl Acad Sci U S A, 2005. **102**(1): p. 57-62.
[41] Del Sol, A., et al., *Modular architecture of protein structures and allosteric communications: potential implications for signaling proteins and regulatory linkages.* Genome Biol, 2007. **8**(5): p. R92.
[42] Tyagi, M., et al., *Exploring functional roles of multibinding protein interfaces.* Protein Sci, 2009. **18**(8): p. 1674-83.
[43] Keskin, O., et al., *Protein-protein interactions: organization, cooperativity and mapping in a bottom-up Systems Biology approach.* Phys Biol, 2005. **2**(2): p. S24-35.
[44] 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): p. e1000821.
[45] Carbonell, P., R. Nussinov, and A. del Sol, *Energetic determinants of protein binding specificity: insights into protein interaction networks.* Proteomics, 2009. **9**(7): p. 1744-53.
[46] Moza, B., et al., *Long-range cooperative binding effects in a T cell receptor variable domain.* Proc Natl Acad Sci U S A, 2006. **103**(26): p. 9867-72.
[47] Humphris, E.L. and T. Kortemme, *Design of multi-specificity in protein interfaces.* PLoS Comput Biol, 2007. **3**(8): p. e164.
[48] Halperin, I., H. Wolfson, and R. Nussinov, *Protein-protein interactions; coupling of structurally conserved residues and of hot spots across interfaces. Implications for docking.* Structure, 2004. **12**(6): p. 1027-38.
[49] Cukuroglu, E., A. Gursoy, and O. Keskin, *Analysis of hot region organization in hub proteins.* Ann Biomed Eng, 2010. **38**(6): p. 2068-78.
[50] Ito, T., et al., *A comprehensive two-hybrid analysis to explore the yeast protein interactome.* Proc Natl Acad Sci U S A, 2001. **98**(8): p. 4569-74.
[51] Uetz, P., et al., *A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae.* Nature, 2000. **403**(6770): p. 623-7.
[52] Gavin, A.C., et al., *Functional organization of the yeast proteome by systematic analysis of protein complexes.* Nature, 2002. **415**(6868): p. 141-7.
[53] Yu, H., et al., *High-quality binary protein interaction map of the yeast interactome network.* Science, 2008. **322**(5898): p. 104-10.
[54] Ekman, D., et al., *What properties characterize the hub proteins of the protein-protein interaction network of Saccharomyces cerevisiae?* Genome Biol, 2006. **7**(6): p. R45.
[55] Patil, A. and H. Nakamura, *Disordered domains and high surface charge confer hubs with the ability to interact with multiple proteins in interaction networks.* FEBS Lett, 2006. **580**(8): p. 2041-5.
[56] Kar, G., A. Gursoy, and O. Keskin, *Human cancer protein-protein interaction network: a structural perspective.* PLoS Comput Biol, 2009. **5**(12): p. e1000601.

[57] Bloom, J.D. and C. Adami, *Apparent dependence of protein evolutionary rate on number of interactions is linked to biases in protein-protein interactions data sets.* BMC Evol Biol, 2003. **3**: p. 21.
[58] Fraser, H.B., D.P. Wall, and A.E. Hirsh, *A simple dependence between protein evolution rate and the number of protein-protein interactions.* BMC Evol Biol, 2003. **3**: p. 11.
[59] Jordan, I.K., Y.I. Wolf, and E.V. Koonin, *No simple dependence between protein evolution rate and the number of protein-protein interactions: only the most prolific interactors tend to evolve slowly.* BMC Evol Biol, 2003. **3**: p. 1.
[60] Wuchty, S., *Evolution and topology in the yeast protein interaction network.* Genome Res, 2004. **14**(7): p. 1310-4.
[61] Kim, P.M., et al., *The role of disorder in interaction networks: a structural analysis.* Mol Syst Biol, 2008. **4**: p. 179.
[62] Higurashi, M., T. Ishida, and K. Kinoshita, *Identification of transient hub proteins and the possible structural basis for their multiple interactions.* Protein Sci, 2008. **17**(1): p. 72-8.
[63] Tsai, C.J., B. Ma, and R. Nussinov, *Protein-protein interaction networks: how can a hub protein bind so many different partners?* Trends Biochem Sci, 2009. **34**(12): p. 594-600.
[64] Tsai, C.J., et al., *A dataset of protein-protein interfaces generated with a sequence-order-independent comparison technique.* J Mol Biol, 1996. **260**(4): p. 604-20.
[65] Ahmad, S., et al., *Protein-DNA interactions: structural, thermodynamic and clustering patterns of conserved residues in DNA-binding proteins.* Nucleic Acids Res, 2008. **36**(18): p. 5922-32.
[66] Noble, W.S., *What is a support vector machine?* Nat Biotechnol, 2006. **24**(12): p. 1565-7.
[67] Ma, B., et al., *Protein-protein interactions: structurally conserved residues distinguish between binding sites and exposed protein surfaces.* Proc Natl Acad Sci U S A, 2003. **100**(10): p. 5772-7.
[68] Lo Conte, L., C. Chothia, and J. Janin, *The atomic structure of protein-protein recognition sites.* J Mol Biol, 1999. **285**(5): p. 2177-98.
[69] Gsponer, J. and M.M. Babu, *The rules of disorder or why disorder rules.* Prog Biophys Mol Biol, 2009. **99**(2-3): p. 94-103.
[70] Gunasekaran, K., C.J. Tsai, and R. Nussinov, *Analysis of ordered and disordered protein complexes reveals structural features discriminating between stable and unstable monomers.* J Mol Biol, 2004. **341**(5): p. 1327-41.
[71] Meszaros, B., et al., *Molecular principles of the interactions of disordered proteins.* J Mol Biol, 2007. **372**(2): p. 549-61.
[72] Tuncbag, N., O. Keskin, and A. Gursoy, *HotPoint: hot spot prediction server for protein interfaces.* Nucleic Acids Res, 2010. **38**(Web Server issue): p. W402-6.
[73] Herraez, A., *Biomolecules in the computer: Jmol to the rescue.* Biochem Mol Biol Educ, 2006. **34**(4): p. 255-61.
[74] Kleanthous, C. and D. Walker, *Immunity proteins: enzyme inhibitors that avoid the active site.* Trends Biochem Sci, 2001. **26**(10): p. 624-31.
[75] Meenan, N.A., et al., *The structural and energetic basis for high selectivity in a high-affinity protein-protein interaction.* Proc Natl Acad Sci U S A, 2010. **107**(22): p. 10080-5.