Annals of Biomedical Engineering, Vol. 38, No. 6, June 2010 (*© 2010*) pp. 2068–2078
DOI: 10.1007/s10439-010-0048-9

Analysis of Hot Region Organization in Hub Proteins

Engin Cukuroglu, Attila Gursoy, and Ozlem Keskin

Center for Computational Biology and Bioinformatics and College of Engineering, Koc University, Rumeli Feneri Yolu, 34450 Sariyer Istanbul, Turkey

(*Received 14 January 2010; accepted 19 April 2010; published online 1 May 2010*)

Associate Editor Michael S. Detamore oversaw the review of this article.

Abstract—Protein interaction maps constructed from binary interactions reveal that some proteins are highly connected to others (acting as hub proteins), whereas some others have a few interactions (at the edges of the map). This paper addresses hub proteins from a structural point: interfaces. It investigates how hot spots are organized in hub proteins (hot regions). We annotate interfaces as the ones between two date-hubs (DD), two party hubs (PP), and two non-hubs (NN). We investigate the physico-chemical properties of these three types of interfaces focusing on the accessible surface area distribution, hot region organization, and amino acid composition differences. Results reveal that there are significant differences between DD and PP interfaces. More of the hot spots are organized into the hot regions in DD interfaces compared to PP ones. A high fraction of the interfaces are covered by hot regions in DD interfaces. There are more distinct hot regions in DDs. Since the same (or overlapping) DD interfaces should be used repeatedly, different hot regions can be used to bind to different partners. Further, these hot region characteristics can be used to predict whether a given hub interface is involved in a DD or a PP interface type with 80% accuracy.

Keywords—Hot region, Hot spot, Hub interface, Protein–protein interaction network, Date hub protein, Party hub protein, Protein–protein interfaces.

INTRODUCTION

Many biological processes are driven by the formation of protein–protein complexes.[18] Protein interaction maps constructed from binary interactions reveal that some proteins are highly connected to others (acting as hub proteins), whereas some others have a few interactions (at the edges of the map). 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 are enriched in repeats of tandem domains,[4] 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.[4,16] Flexibility[10] or disorder of the hubs can also cause them to bind to several proteins. Gerstein and coworkers stated that it is not the hubs, but the partners that are disordered.[17] On the other hand, Tsai *et al.*[26] 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.

Protein interactions can be found by experimentally. Yeast two hybrid method[11,29] is used for determining the transient interactions between proteins, and tandem affinity purification (TAP) with mass spectrometry[6] is another frequently used method to find assemblies of proteins interactions in complexes. Although data from these experiments are noisy, a recent study[31] indicates that the data have a sufficient quality for protein–protein interactions. By combining the interactions from these high-throughput experiments, a protein–protein interaction network (PPIN) can be generated. The topology of this network provides insights into 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

Address correspondence to Ozlem Keskin, Center for Computational Biology and Bioinformatics and College of Engineering, Koc University, Rumeli Feneri Yolu, 34450 Sariyer Istanbul, Turkey. Electronic mail: okeskin@ku.edu.tr

0090-6964/10/0600-2068/0 © 2010 Biomedical Engineering Society

profiles) and structures of complexes help to define two different hub types: party hubs and date hubs.[9,16] For example, Vidal and coworkers[9] 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.

In the study of Han *et al.*,[9] a protein–protein interaction network 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 is divided into many little networks representing the interactions of many biological processes all organized and combined by perturbed date hubs. Ekman *et al.*[4] 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 were 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.[23] Simply considering the geometrical constraints of a protein structure, we can state 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, it 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, our previous study pointed out that hub proteins have smaller, more planar, less tightly packed binding sites than non-hub proteins.[13] Kim *et al.*,[16] 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.[16] Previously, there was not a consensus whether hubs were slower-evolving than other proteins or not.[2,5,12,30] Kim *et al.*[16] 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 disordered than the proteome average, suggesting that their promiscuity is a result of disorder of their binding partners.[17]

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. Experimentally, a hot spot can be detected by alanine scanning mutagenesis.[3] If the binding-free energy change is more than 2 kcal/mol, the residue is flagged as a hot spot. Further, these hot spots are not randomly distributed in the interfaces but rather they are clustered. 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*.[14] 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.[24] This binding site organization rationalizes how a given protein molecule may bind to different protein partners.

This paper addresses hub proteins yet from another structural point: interfaces. It investigates how hot spots (hot regions) are organized in hub proteins. We annotate interfaces as the ones between two date-hubs (DD), two party hubs (PP) and two non-hubs (NN). We investigate the physico-chemical properties of these three types of interfaces focusing on the accessible surface area distribution, hot region organization, and amino acid composition differences. Results reveal that there are significant differences between DD and PP interfaces. More of the hot spots are organized into the hot regions in DD interfaces compared to PP ones.

A high fraction of the interfaces are covered by hot regions in DD interfaces. There are more distinct hot regions in DDs. Since the same (or overlapping) DD interfaces should be used repeatedly, different hot regions can be used to bind to different partners. Further, these hot region characteristics can be used to predict whether a given hub interface is involved in a DD or a PP interface type with accuracy of 80%.

MATERIALS AND METHODS

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.[15,25] An example of an interface is given in Fig. 1 displaying interface residues in a ball-stick model.

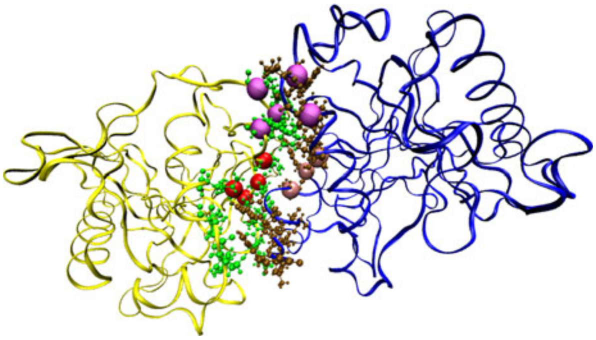


FIGURE 1. Interface representation of 1E9GAB. The yellow representation is the A chain and the blue representation is the B chain of the protein. The green ball-stick representation is the interface of chain A, and the ochre ball-stick representation is the interface of chain B. The *red*, *magenta*, and *pink* balls representation are the different hot regions in the interface.

In this study, we annotate interfaces 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 2 displays the different types of interfaces. Then, we find the hot regions in these interfaces. Various features such as the change in accessible surface areas (ΔASA) of hot regions and interfaces, the 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.

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[4] 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, we need the three-dimensional structures of interfaces. 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 the interface dataset of Tuncbag *et al.*'s[27] resulting in 1199 PX, 602 DX, and 1343 NX interfaces. In order to obtain non-biased statistics, we removed the structurally redundant interfaces and low resolution proteins (worse than 3.0 Å), resulting in 82 PXs, 83 DXs, and 221 NXs. In

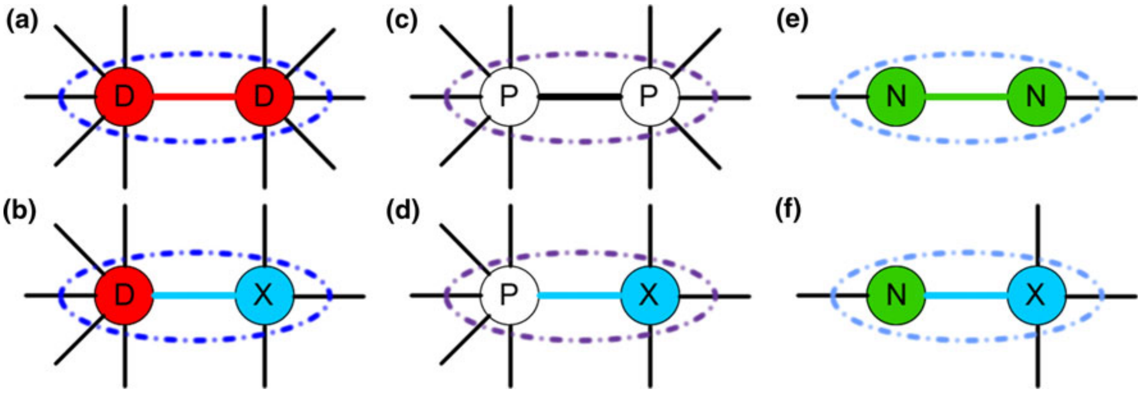


FIGURE 2. The nodes represent the protein; the edges represent the interfaces between the proteins. (a) Date hub—date hub interaction scheme in PPIN (DD). (b) Date hub—non-labeled protein interaction scheme in PPIN (DX). (c) Party hub—party hub interaction scheme in PPIN (PP). d) Party hub—non-labeled 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).

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 Supplementary Materials. This procedure is summarized in flowchart shown in Fig. 3.

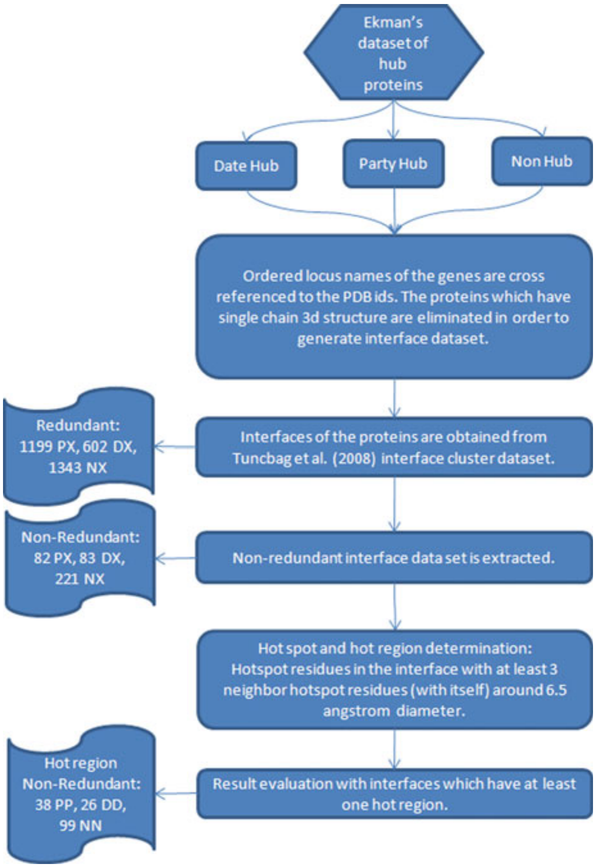

Ekman's dataset of hub proteins
Date Hub
Party Hub
Non Hub
Ordered locus names of the genes are cross referenced to the PDB ids. The proteins which have single chain 3d structure are eliminated in order to generate interface dataset.
Interfaces of the proteins are obtained from Tuncbag et al. (2008) interface cluster dataset.
Redundant: 1199 PX, 602 DX, 1343 NX
Non-redundant interface data set is extracted.
Non-Redundant: 82 PX, 83 DX, 221 NX
Hot spot and hot region determination: Hotspot residues in the interface with at least 3 neighbor hotspot residues (with itself) around 6.5 angstrom diameter.
Result evaluation with interfaces which have at least one hot region.
Hot region Non-Redundant: 38 PP, 26 DD, 99 NN

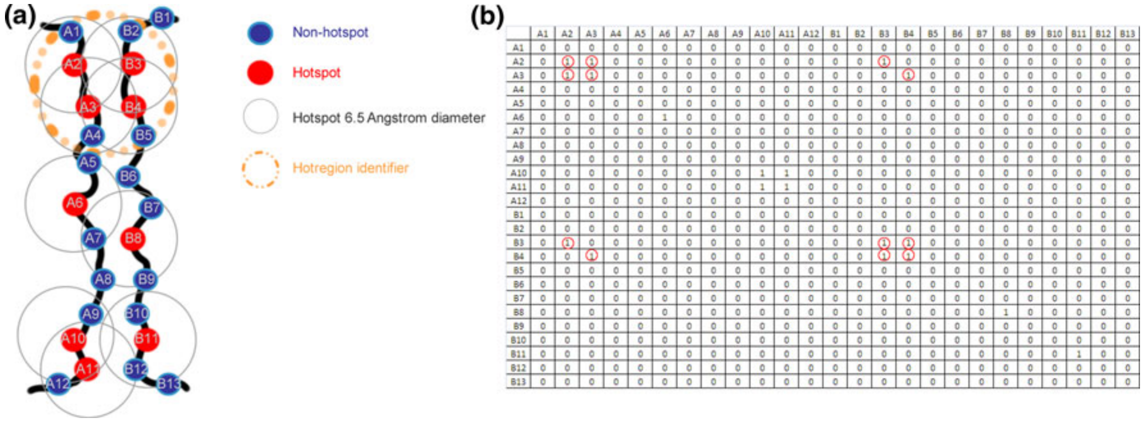
FIGURE 3. A flowchart of the methodology.

Hot Region Detection in the Interfaces

Interface properties (ASA values and hot spot status of residues) of the proteins are taken from the HotPOINT[28] server. HotPOINT is a server that predicts hotspot residues based on using ASA and knowledge-based pair energies. In addition to the 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 the HotPoint server is used in this study.

In order to define hot regions, a contact matrix is constructed 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 Å.[14] 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 Fig. 4).

In a previous work, Reichman *et al.* defined residue modules as clusters of residues with at least three members.[24] Also, Ahmad *et al.* labeled hot regions as those with at least three conserved residues.[1] Here, in a similar way, we define hot regions as the group of hotspots which have at least two contacting hotspot neighbors in the interface (Fig. 4). The contact matrix is used to find hot regions. Figure 4 illustrates an example of hot regions in an interface. In order to find



	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	B1	B2	B3	B4	B5	B6	B7	B8	B9	B10	B11	B12	B13
A1	0	0	0	0	0	0	0	0	0	0	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	0	0	0	0	0	0	0	0	0
A3	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
A4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
A5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	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	0	0	0	0	0	0	0	0	0
A7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
A8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
A9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
A10	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
A11	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
A12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
B1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
B2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
B3	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0
B4	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0
B5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
B6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
B7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
B8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
B9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
B10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
B11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
B12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
B13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

FIGURE 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.

hot regions, first we find a column with at least three "1" entries, this forms the initial cluster then for each element of the cluster we merge corresponding column to the existing cluster until no more additions are possible.

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

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: This is 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 the 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 buried the interfaces become upon complexation. It is calculated as follows:

$HR_{\Delta ASA}$: Hot region ΔASA
$I_{\Delta ASA}$: Interface ΔASA
$HR_{ASA,\mathrm{A}}$: Total monomer ASA values of the residues of chain A in the hot region
$HR_{ASA,\mathrm{B}}$: Total monomer ASA values of the residues of chain B in the hot region
$HR_{ASA,\mathrm{AB}}$: Total complex ASA values of the residues of in the hot region
$I_{ASA,\mathrm{A}}$: Total monomer ASA values of the residues of chain A in the interface
$I_{ASA,\mathrm{B}}$: Total monomer ASA values of the residues of chain B in the interface
$I_{ASA,\mathrm{AB}}$: Total complex ASA values of the residues of in the interface

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

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.

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/nonhub interactions. 38 PPs, 26 DDs, and 99 NNs which have hot regions in their interfaces are used in the training and prediction step using 10-fold cross validation (In 10-fold cross validation method, the dataset is randomly divided into 10 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 10 times). We use support vector machine (SVM) classifier which is a well-known ML classifier to demonstrate the success of classifying interfaces using hot region characteristics. SVM[22] is an algorithm which can classify the data 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. SVM model with linear kernel gives the best classification of DD and PP interfaces on hot regions. In addition to the SVM model, the RBF network, nearest neighbor, decision tree, regression, naïve bayes, and k-means clustering models are applied, but SVM still gives the best result. Therefore, we provide the results of SVM in the following sections. The parameters used for classification and their significance between different types of protein–protein interfaces (DD, PP, and NN) are listed in Table 1. The p values for candidate features are obtained using the analysis of variance (ANOVA) 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

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

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

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

RESULTS

A protein–protein interface consists of two binding sites of two proteins interacting with each other. The 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 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 diamond, 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). We should note that standard deviations are quite high, but the two distributions have statistically significant different means. Details of the distributions are provided as a box plot of the hot spot ratio given in the Supplementary Materials. The NN interfaces have an average of 0.69 ± 0.17 (See Table 2). Figure 6a 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. Figure 6b 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

TABLE 1. Statistical significance of the candidate features (*p* values, underlined numbers indicate the significant *p* values).

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

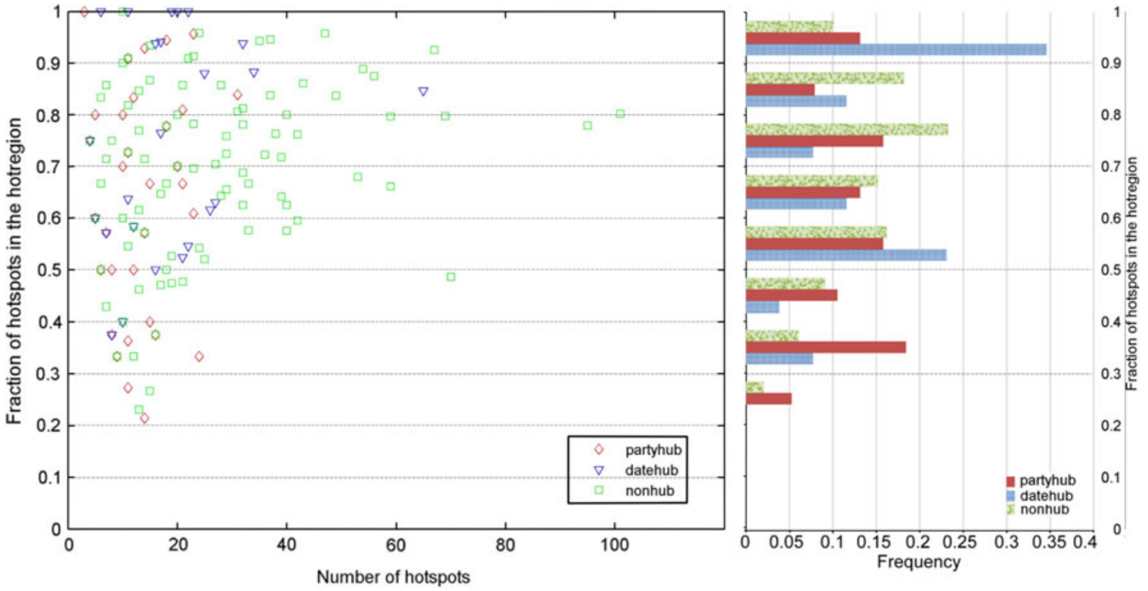


FIGURE 5. The distribution of the fraction of hot spots in the hot regions and their frequency according to their types. Date hub proteins have more tendencies to be involved in a hot region.

TABLE 2. Mean and standard 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 Δ*ASA* to interface Δ*ASA* 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

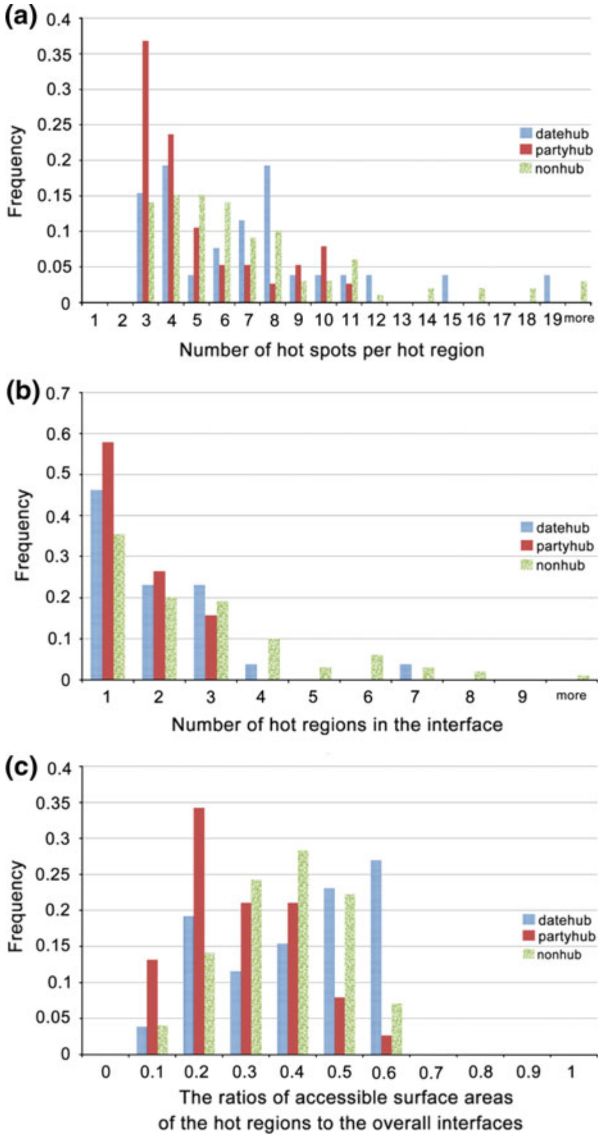


FIGURE 6. (a) The histogram of the hot region sizes. (b) The histogram shows the average number of hot regions in the interfaces. (c) The histogram displays the averages of the ratios of accessible surface areas of the hot regions to the overall interfaces.

2.75. Similarly, Fig. 6c 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 the number of residues they are composed of). 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 expected to have smaller binding sites in party hubs. Physically, it would be impossible to locate large and numerous 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 we look at the average sizes of the hot regions in these interfaces, we observe that hot regions are much larger in DD interfaces compared to PP interfaces. When we look at the average change in accessible surface area of individual hot spots, in DD interfaces we observe that hot spots are more exposed (change in accessible surface area is around 115 Å^2) compared to those in PP interfaces (change in accessible surface area of around 80 Å^2). In NN interfaces, this number is 135 Å^2. Table 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. Table 1 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.

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 on 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.*[5] indicate that hubs of *S. cerevisiae* interactome evolve slowly with a suggested cause of their 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.*[12] 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.*[5] 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[12] 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.

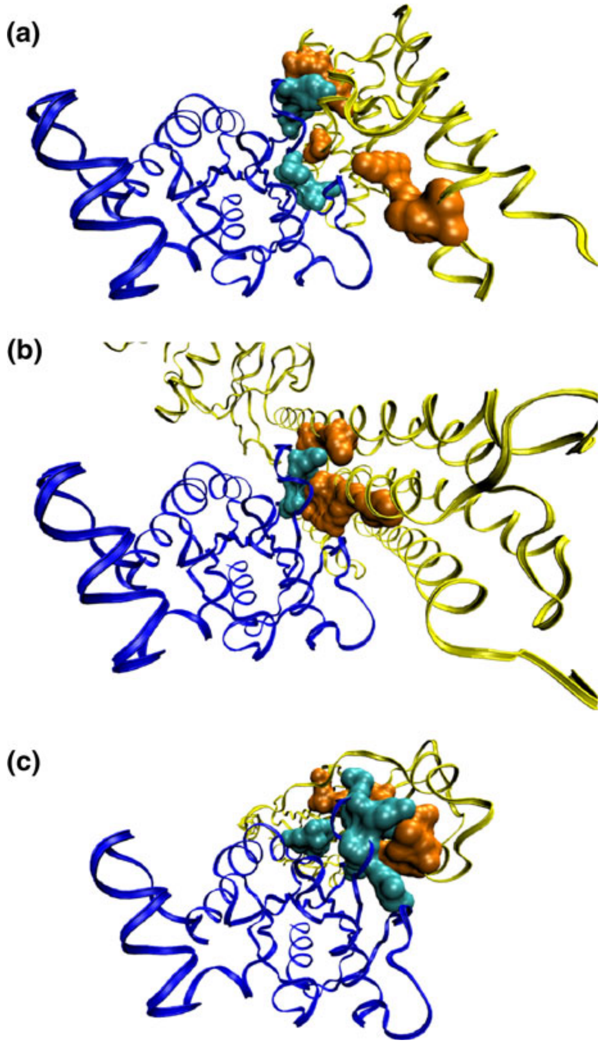


FIGURE 7. Protein G is represented as blue (*dark*) in all three figures. All complexes are taken from PDB [(a) 1GZS_CD, (b) 1KI1_CD, and (c) 1DOA_AB]. Three different proteins binding on the similar region of protein G are shown in *yellow*. Hot regions of protein G are shown in *cyan* whereas hot regions of the partner proteins are *orange*. This figure shows that different hot regions can be utilized to bind the different partners.

By making an analogy between the hot spots and conserved residues[14,20] (although these two terms are not fully correlated), here we argue 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, which 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, we illustrate protein G (a date hub) in Fig. 7. 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. Figure 7 shows that different hot regions can be utilized to bind the different partners.

Previously, we made a proposition 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,[4] 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.

Amino Acid Composition of Hot Regions

Amino acid composition of interfaces generally differs from the rest of the protein surfaces.[19] 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, we group amino acids 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 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 do 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 a higher ratio of hydrophobic amino acids compared to the ordered interfaces; also IUPs have more hydrophobic–hydrophobic interactions than ordered proteins.[7,8,21] 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.[7,8,21] These findings suggest that DD-type interfaces are likely to contain disordered regions and involved in transient interactions.

TABLE 3. Fraction of polar residues.

	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

One would be curious to see if a 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 side of the protein interfaces. Date, party, and non-hub protein binding sites cannot be differentiated 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, and polar aa frequencies of hot regions). The p values in all cases are greater than 0.05.

Automatic Classification of Hub Interfaces

Our analysis shows that organization of hot regions and their hydrophobicity differ among DD, PP, and NN interfaces. One can use these properties to classify a given interface using ML techniques (widely used for classification). The performance of the classification task can indicate the significance of these properties as well. Table 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 are marked (with $p < 0.5$). These features can be used for classifying a given interface. The result of using all parameters (explained in the "Materials and Methods" section) and 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.

CONCLUSION

Protein–protein interaction networks 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 paper, we conclude that there is a relationship between organization of hot spots (hot regions) and the status of hub proteins. We annotate interfaces as the ones between two date-hubs (DD), two party-hubs (PP), and two non-hubs (NN). We conclude 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, we argue 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, which 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 80% accuracy.

ELECTRONIC SUPPLEMENTARY MATERIAL

The online version of this article (doi:10.1007/s10439-010-0048-9) contains supplementary material, which is available to authorized users.

ACKNOWLEDGMENTS

We thank Nurcan Tuncbag for providing Protein G complexes. This project has been supported by TUBITAK (Research Grant No 109T343 and 109E207).

REFERENCES

[1] Ahmad, S., O. Keskin, A. Sarai, and R. Nussinov. Protein–DNA interactions: structural, thermodynamic and clustering patterns of conserved residues in DNA-binding proteins. *Nucleic Acids Res.* 36(18):5922–5932, 2008.

[2] 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.* 3:21, 2003.

[3] Clackson, T., and J. A. Wells. A hot spot of binding energy in a hormone–receptor interface. *Science* 267(5196):383–386, 1995.

[4] Ekman, D., S. Light, A. K. Bjorklund, and A. Elofsson. What properties characterize the hub proteins of the protein–protein interaction network of *Saccharomyces cerevisiae*? *Genome Biol.* 7(6):R45, 2006.

[5] 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.* 3:11, 2003.

[6] Gavin, A. C., M. Bosche, R. Krause, P. Grandi, M. Marzioch, A. Bauer, J. Schultz, J. M. Rick, A. M. Michon, C. M. Cruciat, M. Remor, C. Hofert, M. Schelder, M. Brajenovic, H. Ruffner, A. Merino, K. Klein, M. Hudak, D. Dickson, T. Rudi, V. Gnau, A. Bauch, S. Bastuck, B. Huhse, C. Leutwein, M. A. Heurtier, R. R. Copley, A. Edelmann, E. Querfurth, V. Rybin, G. Drewes, M. Raida, T. Bouwmeester, P. Bork, B. Seraphin, B. Kuster, G. Neubauer, and G. Superti-Furga. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* 415(6868):141–147, 2002.

[7] Gsponer, J., and M. M. Babu. The rules of disorder or why disorder rules. *Prog. Biophys. Mol. Biol.* 99(2–3):94–103, 2009.

[8] 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.* 341(5):1327–1341, 2004.

[9] Han, J. D., N. Bertin, T. Hao, D. S. Goldberg, G. F. Berriz, L. V. Zhang, D. Dupuy, A. J. Walhout, M. E. Cusick, F. P. Roth, and M. Vidal. Evidence for dynamically organized modularity in the yeast protein–protein interaction network. *Nature* 430(6995):88–93, 2004.

[10] Higurashi, M., T. Ishida, and K. Kinoshita. Identification of transient hub proteins and the possible structural basis for their multiple interactions. *Protein Sci.* 17(1):72–78, 2008.

[11] Ito, T., T. Chiba, R. Ozawa, M. Yoshida, M. Hattori, and Y. Sakaki. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl Acad. Sci. USA* 98(8):4569–4574, 2001.

[12] 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.* 3:1, 2003.

[13] Kar, G., A. Gursoy, and O. Keskin. Human cancer protein–protein interaction network: a structural perspective. *PLoS Comput. Biol.* 5(12):e1000601, 2009.

[14] 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.* 345(5):1281–1294, 2005.

[15] Keskin, O., C. J. Tsai, H. Wolfson, and R. Nussinov. A new, structurally nonredundant, diverse data set of

protein–protein interfaces and its implications. *Protein Sci.* 13(4):1043–1055, 2004.

[16] Kim, P. M., L. J. Lu, Y. Xia, and M. B. Gerstein. Relating three-dimensional structures to protein networks provides evolutionary insights. *Science* 314(5807):1938–1941, 2006.

[17] Kim, P. M., A. Sboner, Y. Xia, and M. Gerstein. The role of disorder in interaction networks: a structural analysis. *Mol. Syst. Biol.* 4:179, 2008.

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

[19] Lo Conte, L., C. Chothia, and J. Janin. The atomic structure of protein–protein recognition sites. *J. Mol. Biol.* 285(5):2177–2198, 1999.

[20] Ma, B., T. Elkayam, H. Wolfson, and R. Nussinov. Protein–protein interactions: structurally conserved residues distinguish between binding sites and exposed protein surfaces. *Proc. Natl Acad. Sci. USA* 100(10):5772–5777, 2003.

[21] Meszaros, B., P. Tompa, I. Simon, and Z. Dosztanyi. Molecular principles of the interactions of disordered proteins. *J. Mol. Biol.* 372(2):549–561, 2007.

[22] Noble, W. S. What is a support vector machine? *Nat. Biotechnol.* 24(12):1565–1567, 2006.

[23] 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.* 580(8):2041–2045, 2006.

[24] Reichmann, D., O. Rahat, S. Albeck, R. Meged, O. Dym, and G. Schreiber. The modular architecture of protein–protein binding interfaces. *Proc. Natl Acad. Sci. USA* 102(1):57–62, 2005.

[25] Tsai, C. J., S. L. Lin, H. J. Wolfson, and R. Nussinov. A Dataset of Protein–Protein Interfaces Generated with a Sequence-Order-Independent Comparison Technique. *J. Mol. Biol.* 260(4):604–620, 1996.

[26] 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.* 34(12):594–600, 2009.

[27] Tuncbag, N., A. Gursoy, E. Guney, R. Nussinov, and O. Keskin. Architectures and functional coverage of protein–protein interfaces. *J. Mol. Biol.* 381(3):785–802, 2008.

[28] 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* 25(12):1513–1520, 2009.

[29] Uetz, P., L. Giot, G. Cagney, T. A. Mansfield, R. S. Judson, J. R. Knight, D. Lockshon, V. Narayan, M. Srinivasan, P. Pochart, A. Qureshi-Emili, Y. Li, B. Godwin, D. Conover, T. Kalbfleisch, G. Vijayadamodar, M. Yang, M. Johnston, S. Fields, and J. M. Rothberg. A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*. *Nature* 403(6770):623–627, 2000.

[30] Wuchty, S. Evolution and topology in the yeast protein interaction network. *Genome Res.* 14(7):1310–1314, 2004.

[31] Yu, H., P. Braun, M. A. Yildirim, I. Lemmens, K. Venkatesan, J. Sahalie, T. Hirozane-Kishikawa, F. Gebreab, N. Li, N. Simonis, T. Hao, J. F. Rual, A. Dricot, A. Vazquez, R. R. Murray, C. Simon, L. Tardivo, S. Tam, N. Svrzikapa, C. Fan, A. S. de Smet, A. Motyl, M. E. Hudson, J. Park, X. Xin, M. E. Cusick, T. Moore, C. Boone, M. Snyder, F. P. Roth, A. L. Barabasi, J. Tavernier, D. E. Hill, and M. Vidal. High-quality binary protein interaction map of the yeast interactome network. *Science* 322(5898):104–110, 2008.