Small-World Network Approach to Identify Key Residues in Protein–Protein Interaction

Antonio del Sol* **and Paul O'Meara**
Bioinformatics Research Project, Frontier Research Division, Fujirebio Inc., 51 Komiya-cho, Hachioji-shi, Tokyo 192-0031, Japan

ABSTRACT **We show that protein complexes can be represented as small-world networks, exhibiting a relatively small number of highly central amino-acid residues occurring frequently at protein–protein interfaces. We further base our analysis on a set of different biological examples of protein–protein interactions with experimentally validated hot spots, and show that 83% of these predicted highly central residues, which are conserved in sequence alignments and nonexposed to the solvent in the protein complex, correspond to or are in direct contact with an experimentally annotated hot spot. The remaining 17% show a general tendency to be close to an annotated hot spot. On the other hand, although there is no available experimental information on their contribution to the binding free energy, detailed analysis of their properties shows that they are good candidates for being hot spots. Thus, highly central residues have a clear tendency to be located in regions that include hot spots. We also show that some of the central residues in the protein complex interfaces are central in the monomeric structures before dimerization and that possible information relating to hot spots of binding free energy could be obtained from the unbound structures. Proteins 2005;58:672–682.**
© 2004 Wiley-Liss, Inc.

Key words: betweenness; hot spots; node centrality; protein–protein interface; protein conformational change

INTRODUCTION

The Small-World Network Approach

Small-world networks have recently been shown to successfully describe systems such as neural networks, social networks, disease spreading, scientific collaborations, and other self-organizing systems.[1–7] The name of this kind of network was given by Watts and Strogatz[1] by analogy with the small-world phenomenon (popularly known as six degrees of separation).

In order to quantify the structural characteristics of the small-world networks, Watts and Strogatz[1] used the characteristic path length L, and the clustering coefficient C. The characteristic path length is the average minimal distance between all pairs of vertices in the graph:

$$L = \frac{1}{N_p} \sum_{j>i} l_{ij}, \qquad (1)$$

Where N_p represents the number of pairs of vertices of the graph, and l_{ij} is the minimal path between vertices i and j. The clustering coefficient is the average over all vertices of the fraction of the number of connected pairs of neighbors for each vertex, i.e.,

$$C = \frac{1}{N_v} \sum_i \frac{n_i}{N_i(N_i - 1)/2}, \qquad (2)$$

Where N_v is the number of vertices, N_i is the number of neighbors of the vertex i, and n_i is the actual number of edges between the neighbors of N_i.

Regular graphs are characterized by high values of C and L, while random graphs are characterized by small values of C and L. However, small-world networks are shown to lie between these two extremes having small values of L and high values of C.[1,8]

The concept of small-world networks has been applied to biological networks by different authors.[9–11] Wuchty[12] studied the topology of protein domain networks in proteomes of different organisms, and showed that these networks exhibit a small-world character reflecting the complexity of the protein domain evolution. Vendruscolo et al.[13] showed that protein structures are good examples of small-world networks and used this fact to predict the key residues for the folding process. Wuchty[14] analyzed the conformational spaces spanned by *Escherichia coli tRNA*^{phe} suboptimal structures from the point of view of small-world networks. Cellular, metabolic, and transcriptional regulatory processes have also been shown to be examples of scale-free small-world networks.[15–17] Metabolic networks have been viewed as small-world networks where some specific metabolites play a central role in the network connectivity.[16] More recently, Greene et al.[18] showed that protein structures could be modeled as small-world networks exhibiting single-scale, and to certain extent, scale-free properties. They also showed that protein structure networks, including both long and short-range interactions, exhibit a small-world network character, while the

The Supplementary Materials referrerd to in this article can be found at http://www.interscience.wiley.com/jpages/0887-3585/suppmat/index.html

*Correspondence to: Antonio del Sol, Bioinformatics Research Project, Frontier Research Division, Fujirebio Inc., 51 Komiya-cho, Hachioji-shi, Tokyo 192-0031, Japan. E-mail: ao-mesa@fujirebio.co.jp

Received 3 June 2004; Accepted 9 September 2004

Published online 22 December 2004 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20348

protein structure networks including only long-range interactions are no longer small-world networks. Although within the context of this work, the authors do not analyze the structural characteristics of those residues important for the folding, function or structure of the protein, they stressed the robustness of the network approach for fulfilling this task.

The Protein–Protein Interaction Problem

Protein–protein interactions play a crucial role in many biological processes. Metabolic networks, signaling pathways, and cell adhesion, are some of the many examples of complex processes involving different types of protein–protein interactions. Although many studies have addressed the problem of characterizing binding sites in different examples of protein complexes,[19–21] the principles governing the interaction of proteins are not fully understood.

There have been several studies on protein–protein interfaces, aimed at unraveling different aspects of the protein–protein interaction mechanism, such as recognition, specificity or affinity of the interacting proteins, as well as the factors affecting the dimer interface stability. Early studies considered the change in accessibility of the monomer upon dimerization, and the complementarities between protein surfaces as important factors for protein–protein associations.[22,23] More recent analyses has been focused on the amino-acid conservation at protein interfaces, interface size, and shape, number and type of amino acids at the protein interface, or secondary structure analysis.[24–30]

Different physical approaches to the docking problem have also made progress in the prediction of the interacting proteins and the interacting surfaces.[31,32] These approaches have been mainly based on geometric complementarities of the interacting proteins, and on energetic calculations. Other characteristics such as hydrophobic patches, charge complementarities, or electrostatic and hydrogen bonding considerations have also been analyzed.[26,33] Fernandez and Scheraga,[21] showed that most of the backbone hydrogen bonds are wrapped intra-molecularly by non-polar groups except for a small number, which get stabilized after binding and contribute in determining the binding sites for proteins.

Alanine scanning mutagenesis[34] and phenylalanine substitution[35] of protein–protein interfaces has shown that the free energy contribution of individual amino acids in protein–protein binding is not uniformly distributed at the binding site.[34] There are hot spots of binding free energy comprising a small subset of residues particularly enriched in tryptophan, tyrosine, and arginine, which are surrounded by a shell of residues that very likely occlude bulk solvent from the hot spot.[34] Further studies on binding hot spots have supported and extended these conclusions by remarking on the role of polar residue hot spots.[24] Brinda et al.[36] proposed a graph–spectral-based method to detect side-chain clusters at the interface of a set of homodimeric proteins, which correlated well with the location of hot spots at the protein–protein interfaces. Recently, Buyong et al.[37] have remarked on the importance of the structural conservation of residues to discriminate between binding sites and exposed protein surfaces, showing the correspondence between energy hot spots and structurally conserved residues.

There have been other interesting approaches, aiming to understand different aspects of the protein–protein interaction mechanism, but the analysis of each of them would be beyond the scope of this paper.

A New Approach to Identify Hot Spots in Protein–Protein Complexes

The small-world networks are characterized by the presence of a small number of highly central vertices, which are important for the network topology.[13,18] These vertices have been associated with the central metabolites defining the core of metabolism,[16] key protein domains in proteome evolution,[12] and the key residues for the folding mechanism.[13] With respect to the latter, it was shown that protein structures can be modeled as small-world networks and that the small-world topology arises from the existence of a small number of amino acids (vertices) exhibiting a high measure of centrality. These central residues corresponded to the key residues, which play an essential role in the folding process.[38] In order to determine which residues are the most important in the generation of the small-world network, the "betweenness" B_k was used for each residue, which was defined as the number of times residue k was included in the shortest path between each pair of residues in the protein, normalized by the total number of pairs. Considering that there have been several indications that the overall physical principles behind protein folding and protein–protein association are similar,[39,40] we have aimed here to extend the small-world network approach to protein–protein interaction complexes, and try to identify the key residues contributing the most to the binding free energy.

Here we use a set of 48 examples of two-chain complexes to show that protein–protein complexes do indeed form small-world networks, with large values of the clustering coefficient and small values of the characteristic path length. As a general trend, we show that among all the amino-acid residues responsible for the small-world network character most of them are precisely at the interface. These highly central residues are most likely important to the protein complex interacting network stability, and therefore might be related to the most contributing residues to the binding free energy (hot spots). In order to test this hypothesis, we compiled a set of 18 protein–protein interfaces, including different biological cases of protein–protein interactions, with experimental information on hot spot residues. Our results show that 77% of the predicted residues coincide or are in direct contact with an experimentally annotated hot spot. We complemented our analysis with other properties of these highly central residues, such as, sequence conservation in homologous proteins, and relative accessible surface area upon dimerization. We kept among these residues only those completely or partially buried in the dimer structure and conserved (including conserved mutations) in sequence alignment (henceforth, the residues obtained with this criteria will be termed as "selected central residues"). The percentage of the selected central residues coinciding with or neighboring the experimentally observed hot spots resulted in 83%. The other residues predicted with no experimental information available (remaining 17%) are good candidates for hot spots, being conserved in sequence alignments, completely buried in the dimer structure and well correlated with the hot spot enrichment.

Interestingly, we found that in 11 out of 42 complexes (which contain at least one of their structures in the unbound form complexes) that at least one selected central residue, also exhibited statistically significant high betweenness in the corresponding unbound form (these residues will be termed "overlapping central residues"). In all the cases, where hot spot experimental information was available, each of these overlapping central residues corresponded to an annotated hot spot. In addition, different biological examples, showed the underlining topological changes occurring in the monomers to allow these central residues to remain central upon dimerization.

Finally, we would like to remark on the fact that by approaching protein–protein interaction complexes as small-world networks for identifying hot spots at protein interfaces, we consider the overall topology of the interaction network of protein complexes, rather than focusing on the individual interactions of each residue. On the other hand, the betweenness turned out to be a good measure of centrality of the interaction network, and it can be interpreted as a correction to the number of contacts of each residue. Although residues with high betweenness tend to have a high number of contacts, we show that the most central residues (highest betweenness) are not necessarily the ones with the highest number of contacts, and are the most correlated with the presence of hot spots.

MATERIALS AND METHODS

Selection of Data Sets

The initial data set of 48 dimer complexes were generated from a previous analysis on protein–protein interfaces.[41] Initially, 351 families of structurally nonredundant protein–protein interfaces were clustered using the CD-HIT program,[42] which produced a set of 197 nonredundant sequences. From these 197 sequences, 48 representative examples were selected, which exhibited different biological cases of protein–protein interface type. These include hormones, antigens, enzymes, hemoglobins and immunoglobulins.

A compiled data set of 42 protein structure complexes, which contained at least one structure in its unbound form, was obtained using the August 2003 release of SCOP[43] and by an exhaustive analysis of PDB.[44] The sets of noncomplexed structures were chosen if they had identical sequence to their bound form and if they had no insertions and deletions. If any of the sets contained more than two structures the most recently solved structures were used.

A list of 18 protein complexes with alanine scanning mutagenesis information (many containing at least one structure in its unbound form) was obtained from analyzing the mutated systems in ASEdb.[34] Some additional information was used from previous studies relating to phenylalanine substitutions.[35] Experimentally measured hot spots of binding free energy were defined as residues with a change in binding free energy greater than or equal to 1.0 $Kcal/mol$.

Analysis of Dataset Properties

The accessible surface areas (ASAs) of all the protein complexes were determined using the DSSP program.[45] The percentage relative accessibility was calculated as follows:

$$\mathrm{relASA(residue) = ASA(residue) / totalASA(residue)} \times 100, \quad (3)$$

A buried residue was defined as a residue with a percentage relative accessibility of $< 5\%$, a partially buried residue was defined as a residue with a percentage relative accessibility of $> 5\%$ and $< 20\%$, while an exposed residue was defined as a residue with a percentage relative accessibility of $> 20\%$. The total ASA for each residue type was obtained from literature.[46]

The conservation of residues in the protein complexes was analyzed based on sequence alignments generated by ClustalW,[47] using homologous protein sequences obtained from the Swiss-Prot data bank.[48]

In this present analysis, interacting residues were defined in such a way that two residues are in contact across an interface if at least a pair of atoms, one from each residue belonging to each chain, is at a distance smaller than the sum of their van der waii radii plus a threshold value of 0.5 Å. Along with interacting residues, neighboring residues were also taken into consideration in defining the protein interface. A neighboring residue was defined as a residue with a C_α atom at the most 6.0 Å away from the C_α atom of an already assigned interacting residue.

The propensity of residue type i to be among the selected central residues was calculated as:

$$P_i = (n_i/n_{Totpred}) \;/\; (N_{i(\mathrm{int})}/N_{Tot(\mathrm{int})}), \quad (4)$$

Where n_i is the number of selected central residues of type i at the interface, $n_{Totpred}$ is the total number of selected central residues predicted at the interface, $N_{i(int)}$ is the total number of residues of type i at the interface, and $NTotpred$ is the total number of interface residues.

The correlation of the propensities obtained in this analysis with experimental hot spots enrichments was preformed using the Pearson correlation coefficient.[49]

The C_α atom distances between the residues in the protein complexes and their corresponding residues in the structurally superimposed monomers were calculated using the sequence-independent analysis program LGA.[50]

Graph Representation of Protein Complexes

Each protein complex was modeled as an undirected graph, where each amino-acid residue is represented by a

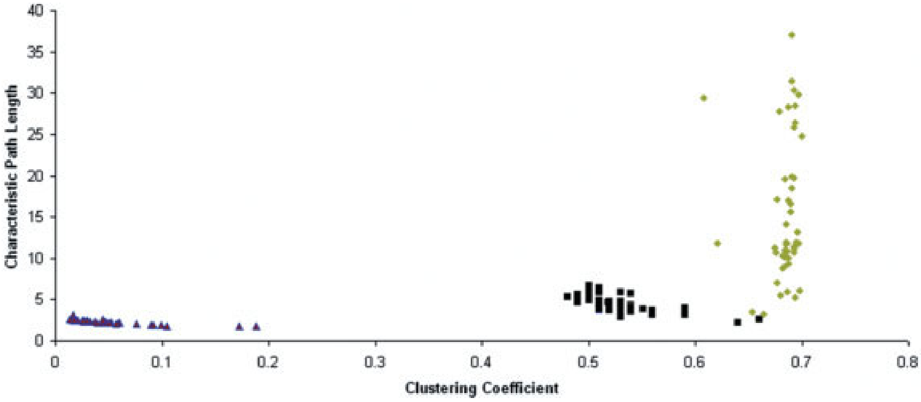

Fig. 1. Distribution of the values of clustering coefficients (C) and characteristic path lengths (L) of all the 48 protein complexes analyzed. The squares are small-world networks of the 48 protein complexes. The triangles are random networks with the same number of nodes and average connectivity and the diamonds are regular networks, again, with the same number of nodes and average connectivity. (Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.)

vertex, and an edge appeared between residues i and j if they are in contact, i.e., if the distance between at least one atom of residue i is at a distance less than or equal to 5.0 Å from an atom of residue j. The matrix elements of the adjacency matrix a_{ij} are equal to one if residues i and j are in contact, or zero otherwise.

The shortest path lengths between all pairs of residues are calculated using the Floyd's algorithm.[51] The clustering coefficient, characteristic path length, and betweenness are defined as in the introduction.

Statistical Analysis

To measure the significance of the betweenness of the residues, we considered their z-score values, defined as

$$Z_k = \frac{B_k - \bar{B}}{\sigma}, \tag{5}$$

Where B_k is the betweenness of residue k, B is the betweenness average value over all protein residues, and σ is the corresponding standard deviation.

Applying Chebyshev's inequality[49] to the distribution of B_k values, we considered the values of z_k greater than or equal to 3.0 for the betweenness to be significantly high. In other words, the relative frequency of those residues with z_k greater than or equal to 3.0 is less than 1/9. Due to the limitations in the number of examples, when dealing with the overlapping central residues, we used the criteria $z_k \geq 2.5$.

In order to measure how significantly close the selected central residues were to the annotated hot spots, we used the distance z-score:

$$Z_k = \frac{r - \bar{r}}{\sigma}, \tag{6}$$

Where r s the distance between each selected central residue to its closest annotated hot spot, r is the average value of the distances of all the interface residues respect to the same annotated hot spot, and σ is the distribution standard deviation. The expression (Equation 6) for the z-score values was also applied when dealing with the conformational changes experienced by the monomers upon dimerization. However, in this case, r stands for the distance of each interacting residue C_α atom in the unbound form and its corresponding structurally superimposed bound form.

RESULTS AND DISCUSSION

Protein Complexes and Small-World Networks

Our first goal was to prove that structures of protein complexes exhibit characteristics that resemble a small-world network. Initial analysis of a compiled dataset of 48 dimer complexes (see Methods), supported the fact that the structures of protein complexes are characterized by relatively high clustering coefficients C with an average value of (0.53 ± 0.035) and relatively small characteristic path lengths L with an average value of (4.46 ± 1.032).

Considering that for a random network with the same number of nodes N, and the same average number of links k, $L_{rand} \sim \ln N / \ln k = (2.27 \pm 0.25)$, $C_{rand} \sim k/N = (0.048 \pm 0.035)$, and for regular networks, $L_{reg} \sim N(N + k - 2) / [2k(N - 1)] = (15.42 \pm 8.44)$, $C_{reg} \sim 3(k - 2) / [4(k - 1)] = (0.68 \pm 0.016)$,[13] we find out that the protein complex structures are characterized by intermediate values of clustering coefficient and characteristic path length, which are typical values corresponding to small-world networks.[13,18] The distributions of the clustering coefficients and characteristic path lengths of all the protein complexes, the corresponding random graphs, and regular graphs are shown in Figure 1. Next we determined the most central residues, which make the most important contribution in generating the small-world networks. Our analysis revealed that of all the residues exhibiting a statistically significant high betweenness ($z - score \geq 3.0$) the majority were located at the dimer interface. In Figure 2, we show the percentage of statistically significant high betweenness residues occurring at the dimer interface. As illustrated from this figure, 38 protein dimer complexes have greater than or equal to 50 percent of their statistically significant high betweenness residues occurring at

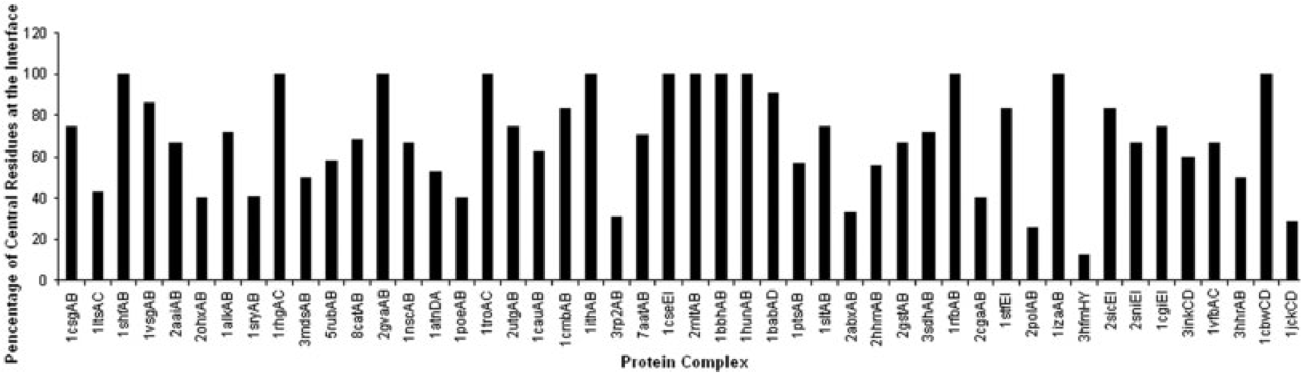

Fig. 2. Percentage of residues exhibiting high betweeness (*z-score* $\geq$ 3.0) at the dimer interface in the 48 complexes analyzed.

the interface. This initial result supports the idea that a vast majority of residues with a high measure of centrality, responsible for the small-world network occur at the protein–protein interface of dimer complexes and therefore might have a possible important role in the stability of the protein–protein complex.

Central Residues and Hot Spots

Building on this initial observation, efforts were focused to investigate whether these statistically significant high-betweenness interface residues were related to hot-spot residues, which significantly contribute to the binding free energy at the protein interface.

In order to carry out this analysis, a set of 18 protein complexes was compiled, from a variety of biological classifications, which contained experimentally measured hot spots of binding free energy at the protein–protein interface (see Table I). Five of the complexes illustrated in Table I are of enzyme-inhibitor type, five are of the antibody-antigen type, one is a hormone-receptor complex, two are cytokine complexes and the remaining six comprise of various other complexes as described in the table. Statistically significant high-betweenness residues and their correlation to experimentally annotated hot spots of binding free energy.

We calculated the betweenness of all interface residues for each of the 18 complexes. It was observed that a number of regions (two or three residues) exhibiting high betweenness values were surrounded by regions of small betweenness values. Due to this fact, our analysis was focused on identifying the most central residues in the high betweenness regions that could correspond to (or be in contact with) an annotated hot spot.

Interestingly, results showed that 77% of all statistically significant high-betweenness residues ($z - score \geq 3.0$) corresponded to either experimentally determined hot spots of binding free energy or to residues in direct contact with experimentally determined hot spots. Residues in direct contact with experimentally proven hot spots were considered due to the fact that in many cases only a small number of hot spots have been verified and also because we are identifying the regions, rather than individual residues, where hot spots are expected to be found.

The statistically significant high betweenness residues in the 1brsAD complex are shown in Figure 3. In this example, three of the four high-betweenness residues obtained from our analysis corresponded to a experimentally validated hot spot of binding free energy, whilst the remaining high-betweenness residue was a residue in contact with an experimentally validated hot spot.

Due to the importance of sequence conservation in identifying key residues involved in protein–protein interaction,[36,37] and because hot spot residues tend to be largely protected from solvent in the complex,[34] we only considered the high betweenness residues that were conserved (greater than or equal to 80% sequence identity or conserved mutation), and which were not completely exposed to solvent (< 20% relative accessibility) in the complex. This resulted in 83% of selected central residues that corresponded to either experimentally determined hot spots of binding free energy or to residues in direct contact with experimentally determined hot spots. Figure 4 shows the distribution of betweenness z-scores of the selected central residues for each protein complex. The position and type of these selected central residues are also shown for each of the protein complexes. Tryptophan, arginine, glutamine, and glutamic acid are the residues with the most significant z-score values.

In order to measure the significance of how close each selected central residue was to its nearest annotated hot spot, we considered the z-score of the distance between them respect to the distribution of distances from all the protein complex interfacing residues to that hot spot. Figure 5 shows that the most negative z-score values are associated with experimentally annotated hot spots or to residues in direct contact with an annotated hot spot. Clearly, there is a tendency for these selected central residues to be closer to an experimentally annotated hot spot than the other interface residues (negative z-score), therefore the regions of high betweenness represented by these selected central residues are well correlated with the presence of hot spots. The one exception to this case is the central tyrosine residue (46Y) in the Isomerase (1ypiAB) complex. However, it is worth noting that there are very few experimentally validated hot spots for this particular protein complex.

TABLE I. Statistically Significant High Betweenness Residues and their Correlation to Known Hot Spots

Protein	Type of complex	Species	PDB code	Predicted hot spots[a]	Exposed/buried in complex	Conservation of residue
Barnase-barstar complex	Enzyme inhibitor	*Bacillus amylolique faciens*	1brs	**27(A)K**	Partially exposed[b]	T.C[e]
			AD	**73(A)E**	Buried[c]	T.C
				38(D)W	Partially exposed	T.C
				39(D)D	Buried	T.C
β-Trypsin complex with pancreatic trypsin inhibitor	Enzyme/inhibitor	*Bos taurus*	2ptc	*41(E)F*	Buried	N.C[f]
			EI	**15(I)K**	Buried	N.C
				17(I)R	Partially exposed	N.C
				19(I)I	Exposed[d]	N.C
Colicin E9 dnase domain with Its cognate immunity protein Im9	Immune system protein	*Escherichia coli*	1bxi	**30(A)E**	Partially exposed	P.C[g]
			AB	**55(A)Y**	Exposed	C.M[h]
				54(B)R	Exposed	C.M
Immunoglobulin Fc and fragment B of protein A complex	Antibody/antigen	*Staphylococcus aureus* *Homo sapiens*	1fc2 CD	436(D)Y	Exposed	N.C
Ribonuclease inhibitor-angiogenin complex	Enzyme/inhibitor	*Homo sapiens*	1a4y	33(A)R	Exposed	T.C
			AB	63(A)R	Partially exposed	T.C
				150(A)Y	Exposed	T.C
				27(B)E	Exposed	P.C
				31(B)R	Partially exposed	N.C
				41(B)D	Partially exposed	T.C
				89(B)W	Exposed	N.C
				93(B)Q	Exposed	N.C
HIV-1 Gp120 core complexed with Cd4 and A neutralizing human antibody	Antibody/antigen	HIV-Type 1	1gc1	*29(C)K*	Buried	C.M
		Homo sapiens	CG	*43(C)F*	Buried	N.C
				46(C)K	Partially exposed	N.C
				279(G)D	Partially exposed	P.C
Interleukin 8	Cytokine	*Homo sapiens*	1i18	**25(A)L**	Buried	C.M
			AB	**25(B)L**	Buried	C.M
Calcium-free phospholipase A2	Hydrolase	*Crotalus atrox*	1pp2	*31(R)W*	Partially exposed	P.C
			RL	*31(L)W*	Partially exposed	P.C
Triose phosphate isomerase	Isomerase (intramolecular oxidoreductse)	*Saccharomyces cerevisiae*	1ypi	12(A)K	Partially exposed	C.M
			AB	*64(A)Q*	Buried	T.C
				77(A)E	Buried	T.C
				82(A)Q	Buried	N.C
				98(A)R	Buried	T.C
				12(B)K	Partially exposed	C.M
				46(B)Y	Buried	C.M
				77(B)E	Buried	T.C
				98(B)R	Buried	T.C
Ribonuclease inhibitor complexed with ribonuclease A	Enzyme/inhibitor	*Bos taurus*	1dfj	24(E)N	Partially exposed	T.C
		Sus scrofa	EI	28(E)Q	Exposed	P.C
				31(E)K	Exposed	N.C
				39(E)R	Partially exposed	T.C
				42(E)P	Partially exposed	T.C
				146(I)Y	Exposed	N.C
Interleukin 2 mutant	Cytokine	*Homo sapiens*	3ink	*43(D)K*	Exposed	T.C
			CD	45(D)Y	Exposed	T.C
				68(D)E	Partially exposed	T.C
IgG1 Fab fragment and lysozyme complex	Antibody/Antigen	*Mus musculus*	3hfm	**58(H)Y**	Exposed	C.M
		Gallus gallus	HY			ñ

TABLE I. (Continued)

Protein	Type of complex	Species	PDB code	Predicted hot spots[a]	Exposed/buried in complex	Conservation of residue
Human growth hormone complexed with its receptor	Hormone/Receptor	*Homo sapiens*	3hhr	*21(A)H*	Buried	T.C
			AB	**178(A)R**	Partially exposed	P.C
				164(B)D	Partially exposed	C.M
				165(B)I	Exposed	C.M
Bovine chymotrypsin complexed to Bpti	Enzyme/inhibitor	*Bos taurus*	1cbw	*195(C)S*	Exposed	T.C
			CD	**15(D)K**	Buried	N.C
				17(D)R	Exposed	N.C
T-cell receptor β-chain complexed with Sec3 superantigen	Toxin/receptor	*Mus musculus*	1jck	**26(D)Y**	Exposed	N.C
		Staphylococcus aureus	CD	**60(D)N**	Partially buried	P.C
				176(D)F	Exposed	T.C
Crystal structure of the C2 fragment of streptococcal protein G in complex with the Fc domain of human IgG	Antibody/antigen	*Homo sapiens*	1fcc	**28(C)K**	Buried	N.C
		Streptococcus	AC			
X-ray structures of the antigen-binding domains from three variants of humanized anti-p185HER2 antibody 4D5	Antibody/antigen	*Homo sapiens*	1 fvc	*36(A)Y*	Buried	C.M
			AB	*38(A)Q*	Partially buried	T.C
				89(A)Q	Buried	P.C
				37(B)V	Buried	T.C
				39(B)Q	Partially buried	T.C
Cd2, N-terminal domain (1–99), truncated form	Immune system protein, receptor	*Rattus norvegicus*	1cdc	16(B)L	Buried	N.C
			BA	**31(B)R**	Buried	N.C
				32(B)W	Buried	T.C
				29(A)E	Buried	N.C
				31(A)R	Buried	N.C
				32(A)W	Buried	T.C
				33(A)E	Partially buried	N.C

[a] Predicted hot spots, shows the positions, the chain number (in brackets), and residue types of the statistically significant high betweenness residues. The entries in bold are residues corresponding to experimentally derived hot spots, entries in italic are residues in contact with experimentally derived hot spot residues and the plain text entries are residues with-no known experimental information.

[b] Partially exposed, residue with a percentage relative accessibility of $>5\%$ and $< 20\%$.

[c] Buried, residue with a percentage relative accessibility of $< 5\%$.

[d] Exposed, residue with a percentage relative accessibility of $> 20\%$.

[e] TC, totally conserved residue. Defined as having 100% conservation in a multiple sequence alignment.

[f] NC, non conserved residue. Defined as having $< 80\%$ conservation in a multiple sequence alignment

[g] PC, partially conserved residue. Defined as having $> 80\%$ and $< 100\%$ conservation in a multiple sequence alignment.

[h] CM, residue with a conserved mutation. Defined as having $> 80\%$ conservation (including conserved mutations) in a multiple sequence alignment.

In order to test whether our results corresponded well with experimental enrichment of hot spot information, the propensities of each residue type to be among the selected central residues at the complex interface were calculated and compared with the residue enrichment in hot spots compiled from the database of alanine-scanning mutagenesis.[34] Tryptophan, arginine, glutamine, and aspartic acid are among the residue types with the highest propensities, which also have high hot-spot enrichment. As shown in Figure 6, there is good correlation between all the observed propensities in this study with the experimentally determined amino-acid enrichment (correlation coefficient equal to 0.724). There are, however, some small disagreements between these propensities and the alanine-scanning data, such as in the cases of Tyr, Phe, Leu, and Met. This could possibly be because of limitations due to the size of the dataset incorporated in this study, but also because our analysis mainly identifies regions of residues exhibiting

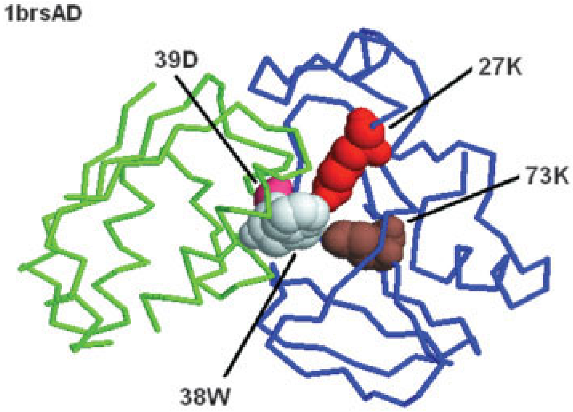

Fig. 3. Structure of the 1brsAD complex. Residues 27K, 73K, and 39D are statistically significant high-betweenness residues corresponding to known hot spots. 38W is a statistically significant high-betweenness residue in contact with known hot spots.

high betweenness, and, in some cases, the most central residues (with the highest value of betweenness) within these regions do not correspond precisely to experimentally validated hot spots, but to neighboring residues of hot spots, which might or might not be hot spots.

Interestingly, the remaining 17% of selected central residues residues with no known experimental hot-spot information are conserved in sequence alignment, completely buried at the interface complex, and well correlated with the experimental data on hot-spot enrichment and thus seem to be good candidates for being hot spots.

It is worth noting that the betweenness has been previously studied as a good measure of centrality,[13] and although there is a general tendency of statistically significant high-betweenness residues to exhibit a high number of contacts, betweenness gives a global topological measure of the relevance of a residue in the interacting network.[13] (See supplementary material, Figure 1.)

Overlapping Central Residues

Vendrusculo et al.[13] showed that it is possible to predict key residues for the folding process by identifying significant high-betweenness residues (central residues) in the transition state. In the native state this information is partially masked by the formation of the protein network, which includes both key and non-key residues for the folding mechanism. However, it is reasonable to think that the statistically significant central residues in the native state should play an essential role in the protein interaction network, and therefore to a certain extent be related to the stability of the protein structure. We show above that there is a general tendency for these central residues occurring at the protein–protein complex interfaces to be related to experimentally validated hot spots. Conversely, it would be interesting to see if some of the central residues in the protein complex interfaces were central in the monomeric structures before dimerization to ascertain if some information could be obtained from these unbound structures.

To pursue this point of interest, a data set of 42 protein structure complexes, with at least one existing structure in its unbound form was compiled. These structures were analyzed to see if there were some statistically significant high-betweenness (*z-score* ≥ 2.5) interface residues in the complexes, which also exhibited statistically significant high betweenness in the corresponding unbound forms. Eleven complexes were shown to contain at least one central residue in the complex interface, which was also central in its monomer structure. Only three of the 11 complexes contained experimental information on hot spots, and each of these overlapping central residues in the three complexes corresponded to an experimentally annotated hot spot.

The conformational changes occurring in these overlapping central residues upon dimerization were analyzed to ascertain the topological changes occurring in the structure of the monomers, which allow these central residues in the monomers to become central in the protein–protein interacting networks. (See supplementary material, Figures 2 and 3.) Figure 7 shows an illustrative example of this analysis. In this figure we show the 1dfjEI endonuclease/inhibitor complex with its corresponding unbound 2bnh inhibitor protein. In this example, the conformational change (calculated as the residue C_α atom distance between the unbound and superimposed bound structures) of the overlapping central residue, which also corresponds to an experimentally annotated hot spot of binding free energy, was less than the average conformational change of all the interacting residues. However, the overall C_α RMSD value for all the interacting residues was relatively high revealing a large conformational change at the interface. As shown, the ribonuclease inhibitor is shaped like a horseshoe, with a large cavity at its center. Ribonuclease inhibitors can bind different RNAses despite their differences in shape and in this example, the large conformational change in the interface residues is most probably due to the fact that the ribonuclease inhibitor contains many leucine rich repeats (LLRs), which can move slightly allowing a change in shape of the inner cavity to accommodate the different types of RNAses onto which it can bind.[52] Other examples of the overlapping central residues are illustrated in Figures 4 and 5 of the supplementary material.

In this analysis, the majority of the statistically significant high-betweenness residues at the protein–protein interface are newly formed after dimerization. However, in the minority of cases we see that residues that are central in the protein complex interfaces are also central in the monomer structures and in the cases where experimental information on hot spots of binding free energy was available, these overlapping central residues corresponded to a hot spot. It is worth noting, that only 18 protein complexes containing binding free energy information were used in this present study and within this dataset not all of the protein complexes had monomer structures available. However, the fact that all of the overlapping central residues for which experi-

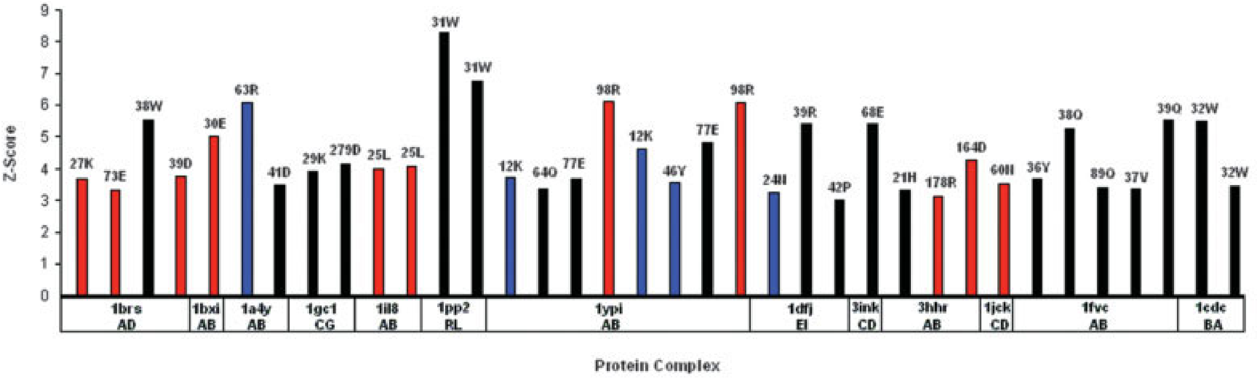

Fig. 4. Betweenness z-score distribution of the selected central residues. The red bars are selected central residues corresponding to experimentally validated hot spots, the black bars represent selected central residues in contact with experimentally validated hot spots and the blue bars are selected central residues with no known experimental information on hot spots. The residue type and numeration in sequence are shown over each z-score value. The residues 25L in the 1i18AB complex, 31W in the 1pp2RL complex, 12K, 77E and 98R in the 1ypiAB complex and 32W in the 1 cdcBA complex are shown to be repeated in the figure as these residues were predicted for each chain in their respective homodimer complex.

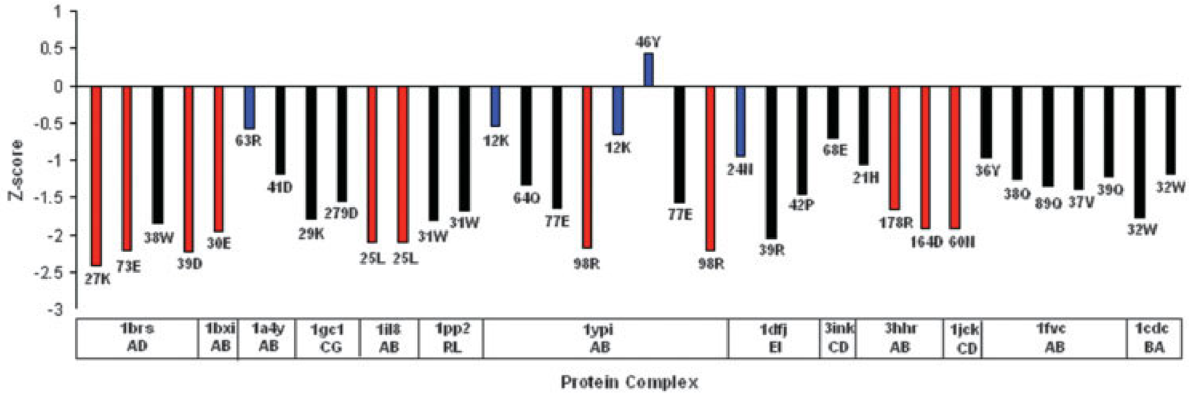

Fig. 5. Distance Z-score distribution of the selected central residues to their nearest hot spots with respect to all interface residues. The red bars represent selected central residues corresponding to experimentally validated hot spots, the black bars represent selected central residues in contact with experimentally validated hot spots and the blue bars are residues with no known experimental information. The residue type and numeration in sequence are shown over each z-score value.

mental information on binding free energy was available corresponded to hot spots is interesting and in the future our studies will be directed to explore if more information based on the monomers could predict hot spots of binding free energy.

Conclusion

Here we show that protein complexes form small-world networks, and that a high percentage of highly connected residues correspond to or are in contact with experimentally validated hot spots.

Initially, 48 protein complexes including a good balance of interface type, showed clustering coefficients and characteristic path lengths corresponding to small-world networks, when compared to random and regular graphs with the same number of nodes and average connectivity.

Next, we used the concept of betweenness to extract information about the most central residues, which make the most important contribution in generating the small-world character of the structures of the protein complexes. Our study illustrated that in 38 of these complexes greater than or equal to 50 percent of these statistically significant high-betweenness residues occur at the protein–protein interfaces.

In order to see if these statistically significant high-betweenness residues occurring at protein–protein interfaces play an important role in the structures of the protein complexes, a set of 18 protein complexes was compiled, which contained experimental information on the most-contributing residues for the binding free energy (hot spots). The analysis of the residue betweenness in each of these complexes, revealed that there were regions at the protein–protein interfaces comprised of statistically significant high betweenness residues, surrounded by low betweenness residues. Further analysis showed that 77% of the statistically significant high betweenness residues corresponded either to an experimentally validated hot spot, or to a residue in direct contact with an annotated hot spot. The set of statistically significant high-betweenness residues, which were conserved in sequence alignments and not completely exposed to the solvent, comprised of

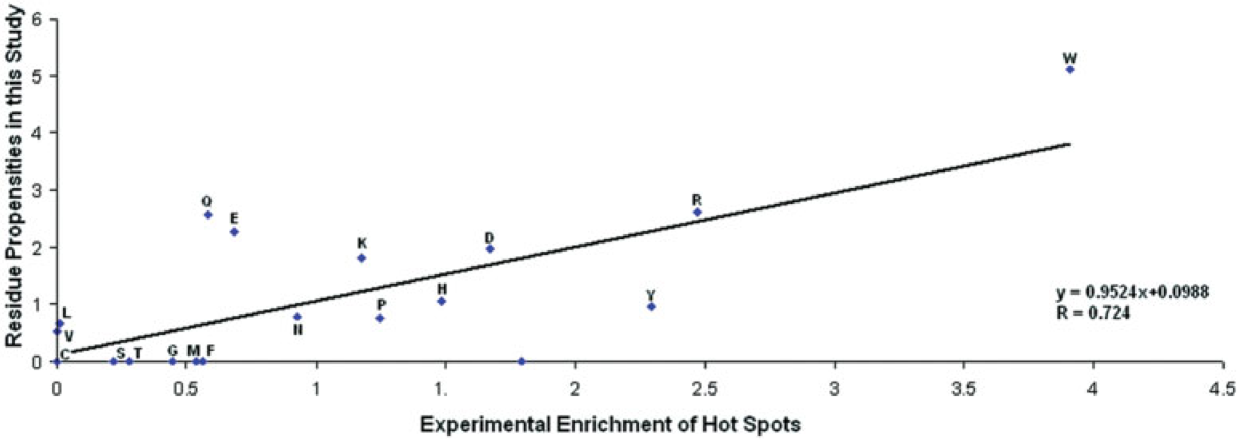

Fig. 6. Correlation of the propensities of the selected central residues with experimental enrichment of hot-spot information. The correlation coefficient value is 0.724. (Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.)

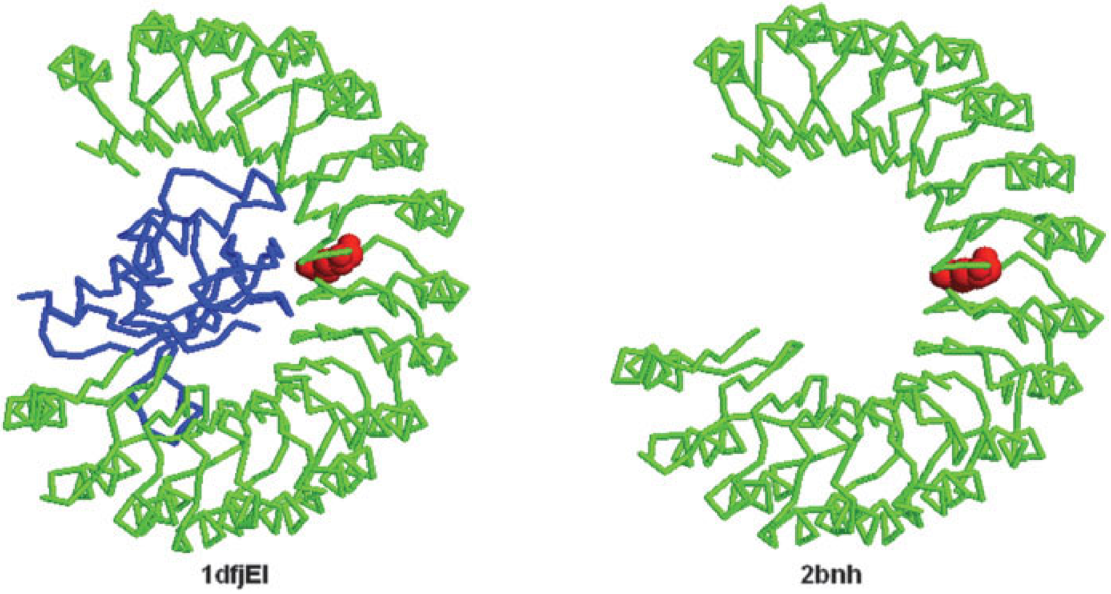

Fig. 7. Structures of the 1dfjEI complex and its unbound inhibitor, 2bnh. The overlapping central residue, which also corresponds to a hot spot of binding free energy, is shown in red.

83% of residues (selected central residues), which either corresponded to an annotated hot spot, or to a residue in contact with a hot spot. Additional analysis (see Results and Discussion) showed that there was a clear general tendency for these selected central residues to be close to an annotated hot spot, showing that high-betweenness residues tend to be located in regions where experimentally validated hot spots are present.

A good correlation was observed between the propensities for each residue type to be among the selected central residues at protein interfaces with the experimental enrichment of hot-spot information. As a result of this observation, and including conservation and solvent accessibility information, it is reasonable to say that, in general, the remaining 17% of selected residues with no hot-spot experimental information could be considered good candidates as hot spots. We show that betweenness, rather than the residue number of contacts, gives valuable information for identifying free energy hot spots.

Interestingly, 11 complexes analyzed were shown to contain at least one selected central residue at its interface, which also exhibited statistically significant high betweenness in the unbound form (overlapping central residues). In all the cases where experimental information on hot spots was available, these overlapping residues corresponded to a hot spot. The example of the inhibitor 2bnh shows that a large conformational change in its binding site, facilitates a central residue in its structure to remain central in the interacting network of the 1dfj endonuclease/inhibitor complex, however, the overlapping central residue in this example, which is a hot spot of binding free energy, shows less conformational change than the average of all the interacting residues (Fig. 7). This analysis suggests that some information on the important residues for protein–protein association could be extracted from the topology of the monomer structures. In the future, we plan to continue further studies relating to this matter.

ACKNOWLEDGMENTS

We would like to acknowledge interesting discussions in issues related to small-world view of protein structures with Dr. Alfonso Valencia (CNB), and thank Professor Ruth Nussinov (NCI, Tel Aviv University) for helpful discussions on protein–protein interactions.

REFERENCES

1. Watts DJ, Strogatz SH. Collective dynamics of small-world networks. Nature (London) 1998;393:440–442.
2. Woolhouse M, Donaldson A. Managing foot-and-mouth. Nature (London) 2001;410:515–516.
3. Newman MEJ. The structure of scientific collaboration networks. Proc Natl Acad Sci USA 2001;98:404–409.
4. Alon U, Surette MG, Barkai N, Leibler S. Robustness in bacterial chemotaxis. Nature (London) 1999;397:168–171.
5. Albert R, Jeong H, Barabasi A-L. Internet: diameter of the world-wide web. Nature (London) 1999;401:130–131.
6. Barabasi A-L, Albert R. Emergence of scaling in random networks. Science 1999;286:509–512.
7. Strogatz, SH. Exploring complex networks. Nature 2001;410:268–276.
8. Watts DJ. Small worlds: The dynamics of networks between order and randomness. Princeton: Princeton University Press; 1999.
9. Fell D, Wagner A. The small world of metabolism. Nat Biotechnol 2000;18:1121–1122.
10. Jeong H, Tombor B, Albert R, Oltvai Z, Barabasi A-L. The large-scale organization of metabolic networks. Nature 2000;407: 651–654.
11. Girvan M, Newman MEJ. Community structure in social and biological networks. Proc Natl Acad Sci USA 2002;99:7821–7826.
12. Wuchty S. Scale-free behavior in protein domain networks. Mol Biol Evol 2001;18:1694–1702.
13. Vendruscolo M, Dokholyan NV, Paci E, Karplus M. Small-world view of the amino acids that play a key role in protein folding. Phys Rev E 2002;65:061910-1–061910-4.
14. Wuchty S. Small world in RNA structures. Nucleic Acids Res 2003;31:1108–1117.
15. Ravasz E, Somera AL, Mongru DA, Oltvai ZN, Barabasi A-L. Hierarchical organization of modularity in metabolic networks. Science 2002;297:1551–1555.
16. Wagner A, Fell DA. The small world inside large metabolic networks. Proc R Soc Lond B 2001;268:1803–1810.
17. Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D, Alon U. Network motifs: simple building blocks of complex networks. Science 2002;298:824–827.
18. Greene LH, Higman VA. Uncovering network systems within protein structures. J Mol Biol 2003;334:781–791.
19. Schmitt S, Kuhn D, Klebe G. A new method to detect related function among proteins independent of sequence and fold homology. J Mol Biol 2002;323:387–406.
20. Ofran Y, Rost B. Analysing six types of protein-protein interfaces. J Mol Biol 2003;325:377–387.
21. Fernandez A, Scheraga HA. Insufficiently dehydrated hydrogen bonds as determinants of protein interactions. Proc Natl Acad Sci USA 2003;100:113–118.
22. Chonthia C, Janin J. Principles of protein-protein recognition. Nature 1975;256:705–708.
23. Janin J, Miller S, Chonthia C. Surface, subunit interfaces and interior of oligomeric proteins. J Mol Biol 1988;204:155–164.
24. Hu Z, MaB, Wolfson H, Nussinov R. Conservation of polar residues as hot spots at protein interfaces. Proteins 2000;39:331–342.
25. Valdar WS, Thornton JM. Protein-protein interfaces: analysis of amino acid conservation in homodimers. Proteins 2001;42:108–124.
26. Jones S., Thornton JM, Principles of protein-protein interactions. Proc Natl Acad Sci USA 1996;93:13–20.
27. Tsai CJ, Lin SL, Wolfson HJ, Nussinov R. Protein-protein interfaces: architectures and interactions in protein-protein interfaces and in protein cores. Their similarities and differences. Crit Rev Biochem Mol Biol 1996;31:127–152.
28. Jones S, Marin A, Thornton JM. Protein domain interfaces: characterization and comparison with oligomeric protein interfaces. Protein Eng 2000;13:77–82.
29. Nooren IMA, Thornton JM. Diversity of protein-protein interactions. EMBO J. 2003;22: 3486–3492.
30. Bahadur RP, Chakrabarti P, Rodier F, Janin J. A dissection of specific and non-specific protein-protein interfaces. J Mol Biol 2004;336:943–955.
31. Sternberg MJ, Gabb HA, Jackson RM. Predictive docking of protein-protein and protein-DNA complexes. Curr Opin Struct Biol 1998;8:250–256.
32. Lengauer T, Rarey M. Computational methods for biomolecular docking. Curr Opin Struct Biol 1996;6:402–406.
33. Palma PN, Krippahl L, Wampler JE, Moura JJ. BIGGER: a new (soft) docking algorithm for predicting protein interactions. Proteins 2000;39:372–384.
34. Bogan AA, Thorn KS. Anatomy of hot spots in protein interfaces. J Mol Biol 1998;280:1–9.
35. Mainfroid V, Mande SC, Hol WG, Martial JA, Goraj K. Stabilization of human triosephosphate isomerase by improvement of the stability of individual alpha-helices in dimeric as well as monomeric forms of the protein. Biochemistry 1996;35:4110–4117.
36. Brinda KV, Kannan N, Vishveshwara S. Analysis of homodimeric protein interfaces by graph-spectral methods. Protein Eng 2002;15: 265–277.
37. Ma B, Elkayam T, Wolfson H, Nussinov R. Protein-protein interactions: structurally conserved residues distinguish between binding sites and exposed protein surfaces. Proc Natl Acad Sci USA 2003;100:5772–5777.
38. Vendruscolo M, Paci E, Dobson CM, Karplus M. Three key residues form a critical contact network in a protein folding transition state. Nature 2001;409:641–645.
39. Walls PH, Sternberg MJ. New algorithm to model protein-protein recognition based on surface complementarity. Applications to antibody-antigen docking. J Mol Biol 1992;228:277–297.
40. Young L, Jernigan RL, Covell DG. A role for surface hydrophobicity in protein-protein recognition. Protein Sci 1994;3:717–729.
41. Tsai CJ, Lin SL, Wolfson HJ, Nussinov RJ. A dataset of protein-protein interfaces generated with a sequence-order-independent comparison technique. J Mol Biol 1996;260:604–620.
42. Li W, Jaroszewski L, Godzik A. Clustering of highly homologous sequences to reduce the size of large protein databases. Bioinformatics 2001;17:282–283.
43. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 1995;247:536–540.
44. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res 2000;28:235–242.
45. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983;22:2577–2637.
46. Creighton TE. Proteins structures and molecular properties. New York: Freeman and Company; 1993.
47. Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 1994;22:4673–4680.
48. Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I, Pilbout S, Schneider M. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res 2003;31: 365–370.
49. Rice JA. Mathematical statistics and data analysis. Belmont, CA: Wadsworth Pub Co; 1994.
50. Zemla A. LGA: A method for finding 3D similarities in protein structures. Nucleic Acids Res 2003;31:3370–3374.
51. Rosen KH. Discrete mathematics and its applications. Singapore: McGraw-Hill; 1999.
52. Kobe B, Deisenhofer J. A structural basis of the interactions between leucine-rich repeats and protein ligands. Nature 1995;374: 183–186.