Similarity Networks of Protein Binding Sites

Ziding Zhang and Martin G. Grigorov*
Nestlé Research Center, Nestec Ltd, BioAnalytical Science, CH-1000 Lausanne 26, Switzerland

ABSTRACT **An increasing attention has been dedicated to the characterization of complex networks within the protein world. This work is reporting how we uncovered networked structures that reflected the structural similarities among protein binding sites. First, a 211 binding sites dataset has been compiled by removing the redundant proteins in the Protein Ligand Database (PLD) (http://www-mitchell.ch.cam.ac.uk/pld/). Using a clique detection algorithm we have performed all-against-all binding site comparisons among the 211 available ones. Within the set of nodes representing each binding site an edge was added whenever a pair of binding sites had a similarity higher than a threshold value. The generated similarity networks revealed that many nodes had few links and only few were highly connected, but due to the limited data available it was not possible to definitively prove a scale-free architecture. Within the same dataset, the binding site similarity networks were compared with the networks of sequence and fold similarity networks. In the protein world, indications were found that structure is better conserved than sequence, but on its own, sequence was better conserved than the subset of functional residues forming the binding site. Because a binding site is strongly linked with protein function, the identification of protein binding site similarity networks could accelerate the functional annotation of newly identified genes. In view of this we have discussed several potential applications of binding site similarity networks, such as the construction of novel binding site classification databases, as well as the implications for protein molecular design in general and computational chemogenomics in particular. Proteins 2006; 62:470–478.** © 2005 Wiley-Liss, Inc.

Key words: complex network; binding site; small world; scale free; sequence; fold; drug design

INTRODUCTION

Complex systems can be represented as networks of interactions occurring between their components. The study of such networks is gaining importance in many disciplines.[1] Any system of interconnected objects, such as the ensemble of Web sites linked by cross-references, the business and social relationships established between people, the electrical power grids, and the co-authorships and co-citations among scientists, can be regarded as networks.[1,2] Because the structure of a network directly affects its function, the global characterization of the topology is crucial. Studies of numerous and diverse real-world networks have provided interesting insights in the growth and dynamics of the related physical systems. Two common properties of these networks are that they are organized around a small-world[3] and scale-free[4] architecture. Recently, biological systems were also scrutinized, and researchers have demonstrated that molecular interaction networks involved in cellular, metabolic, and transcriptional regulatory processes exhibit some similar topological properties.[5–9]

Efforts have been made to apply network concepts to describe the protein molecular world, ranging from protein–protein interactions,[10] to interactions within families of protein domains,[11,12] to amino acids contacts within protein structures,[2,13] to conformational spaces of transition states in protein folding,[14] to protein family and fold occurrence and distribution in genomes,[15] to protein fold similarity networks in the protein structural universe.[16] These investigations have provided systematic and deeper understanding of the evolution and diversity of proteins.

With the progress of hundreds of genome projects, a huge number of genome sequences became available. However, a large fraction of gene products turned out to be difficult to annotate and remained a challenge for post-genomic bioinformatics. The most common way of inferring the biological function of a new gene is based on evaluating its sequence similarity with proteins of known function. A number of sequence comparison methods are currently available, and are widely applied. Recently, structural genomics initiatives embarked on high-throughput X-ray crystallography and NMR spectroscopy to obtain a comprehensive coverage of the protein structural world.[17] The tremendous increase of experimentally determined protein structures put at reach an even larger number of protein structures through the use of homology modelling and fold recognition. For example, Zhang and Skolnick[18] examined the "structural" completeness of the version of the Protein Data Bank (PDB) in 2004. The authors found that the protein-folding problem can, in principle, be solved for practically any sequence based on this version of the PDB library provided that efficient fold recognition algorithms can recover correct sequence alignments. The

*Correspondence to: Martin G. Grigorov, Nestlé Research Center, Nestec Ltd, BioAnalytical Science, CH-1000 Lausanne 26, Switzerland. E-mail: martin.grigorov@rdls.nestle.com

Received 21 April 2005; Revised 7 July 2005; Accepted 2 August 2005

Published online 18 November 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20752

true reason for the success of the structural annotation methods is that protein shape is more conserved than sequence and is closely related to protein function.[19] Structural comparisons were therefore able to identify functional relationships even when no clear sequence similarity was detectable. Protein structural analysis became an important source for understanding the functional role of new proteins, and is now often referred to as "structure-based functional annotation."

However, relatedness in sequence or fold does not necessarily imply a similar function. For example, functionally unrelated analogous proteins evolved from divergent ancestors to fold in similar structures. Moreover, it was found that the conservation of function between a pair of enzymes becomes questionable when sequence identity drops below 40%.[20] Alternatively, proteins with the same fold, like Tim barrels, can have multiple functions.[21] Protein function is very often encoded in a small number of residues located in the functional active site, which are dispersed around the primary sequence, but packed in a compact spatial region. Evidence is accumulating demonstrating that molecular recognition patterns are much more conserved within the binding pockets of proteins of similar function. Therefore, the detection and comparison of protein binding sites is emerging as an important topic in structural biology.

Several binding site comparison methods have been developed in the past decade.[22–28] A first group of methods is focused on analyzing the structural motifs among the binding sites. Artymiuk et al.[25] represented each side chain by pseudo-atoms and used a subgraph-isomorphism algorithm to identify the spatially conserved patterns. The TESS[29] method, based on the geometric hashing algorithm, has been used for the efficient comparison of a query protein to a template of a catalytic triad.[29] Russell and co-worker developed a different method that was able to detect side-chain geometric patterns common to two binding sites, and a statistical significance score was derived that quantified the degree of binding site similarity.[23] The second group of methods focused on the chemical nature of the binding sites to be compared. As an example, Kinoshita et al.[24] performed clique detection on the vertices of the triangulated solvent-accessible surface to address the similarity of binding sites. For the same purpose, Schmitt et al.[28] used generic pseudo-centers that efficiently encoded the physicochemical environments that are important in molecular interactions. Each amino acid residue of a protein was represented as a set of such centers. The clique detection algorithm was also used to retrieve cavities that are similar to a specific query cavity.[28] Very recently, a novel method, SiteEngine, based on an efficient hashing of the matching triangles of centers of physicochemical properties, has been developed.[30] Based on the available algorithms, several binding site databases were compiled, such as CaveBase,[28] eF-site,[24] and SURFACE,[26] to cite a few. These databases have been integrated into Web servers to allow for the identification of the potential binding sites in query protein structures.[23,24,26,31] A successful application of the available data was recently reported that allowed for the inference of a function for the non-annotated protein structures determined within the Structural Genomics Initiative.[24] However, efforts are required to improve the reliability of these databases, by improving the similarity measures for binding pockets as well as the criteria for quantifying statistical significance.

As reported in the literature,[15,16] protein sequence or fold similarity networks within different genomes have been already derived and found to provide valuable insight on the evolution of proteins. However, to the best of our knowledge, a detailed and systematic study of the similarity networks of protein binding sites is still not available. In the present article, we attempted to uncover such similarity networks. In a first section we are deriving the similarity networks of protein binding sites within the PLD database[32] using a binding site comparison algorithm. In a second section we focused on the characterization of these similarity networks by comparing them to the similarity networks of protein sequences and protein structures. Finally, several potential applications of the obtained results are discussed.

METHODS

Protein Binding Site Database

In the present study, we employed the Protein Ligand Database (PLD)[32] to derive similarity networks of binding sites. We obtained the PLD (v1.3) from http://www-mitchell.ch.cam.ac.uk/pld/, which contained 485 protein–ligand complexes. After removing the redundant proteins in this database by keeping only one protein from pairs sharing more than 95% sequence identity, 211 protein–ligand complexes remained for further study. At this point three datasets were compiled: (1) The binding sites dataset. The binding site residues were selected as those residues having any atom closer than 4.5 Å from the ligand molecule; thus, we obtained the binding site PDB files for these 211 protein–complexes. (2) The PDB files for the 211 proteins. For those proteins with multiple chains, in some cases the binding site residues came from different chains. However, the combinatorial expansion (CE) algorithm[33] that we used to perform the pairwise structural comparison was not operational with multiple-chain proteins. Therefore, for those proteins with multiple chains, only the chain containing the maximal number of binding residues was taken into account. (3) We compiled a dataset of the 211 protein sequences in the Fasta format. In a way similar to protein structural comparisons, protein sequence matching algorithms were not functioning with multimer proteins. Again, for those proteins with more than one chain, we considered only the sequence of the chain bearing the maximal number of binding residues.

Comparison of Binding Site Similarity

To perform the binding site similarity comparison, the clique-detection method was applied, by taking the following four steps. First, the binding sites under investigation were converted into two input graphs. The geometric center of the side chain of each binding residue was

represented as a vertex in the graph. Because no side chain exists for glycine, the Cα atom was considered as a node in the case of this residue. Further, a vertex was colored in four (PLD-BSSN-I) or eight (PLD-BSSN-II) different colors according to the physicochemical properties of the parent residue [for PLD-BSSN-I: I (LVIMC), II (AGSTP), III (FYW) and IV (EDNQKRH), and for PLD-BSSN-II: I (LVIMC), II (AG), III (ST), IV (P), V (FYW), VI (EDNQ), VII (KR), VIII (H)].[34] The branching of the molecular graph used to encode a binding site was generated by connecting two vertices by an edge whenever the distance between them was less than 12.0 Å. In a second step, the similarity between any two colored graphs G1 and G2 was evaluated by deriving the so-called product graph P. Each node of P consisted of a pair of nodes with identical colors originating from the input graphs. Two nodes of P were connected by an edge if the respective matchings of G1 in G2 were compatible, that is, that the difference of the corresponding distances in G1 and G2 was less than 2.0 Å. The maximal complete subgraph (clique) of P was detected by the Bron-Kerbosh algorithm.[35] Finally, the binding site similarity was quantified by the Tanimoto coefficient, defined as:

$$\mathrm{Sim} = \frac{N_{\mathrm{sub}}}{N_1 + N_2 - N_{\mathrm{sub}}} \quad (1)$$

where N_1, N_2, and N_{sub} represented the number of nodes in **G1**, **G2**, and the maximal subgraph between **G1** and **G2**, respectively. Further details about the clique detection algorithm in the comparison of binding sites could be found in the works of Schmitt et al.[28] and Weskamp et al.[36]

Construction of Binding Site Similarity Networks

To construct the two binding similarity networks for the 211 binding sites available in PLD, a pairwise binding site comparison was performed by applying the procedure just discussed. The average Tanimoto similarity for the 22155 ($211 \times 210/2$) pairs was 0.20, with a standard deviation of 0.05 (PLD-BSSN-I). Therefore, we considered as similar in a statistically significant way the pairs of binding sites with Tanimoto similarity higher than 0.35, that is, with *Z*-scores greater than 3.0, to which corresponded a *P*-value of less than 0.005. Then, we repeated the pairwise comparison for the 211 binding sites for the PLD-BSSN-II network. In this case, the average similarity value was 0.165 ± 0.045, and we chose a score higher than 0.30 as the cutoff for the connectivity between two binding sites. Consequently, the two similarity networks for the 211 binding sites were constructed, by linking the relevant pairs with 214 edges in the case of the PLD-BSSN-I network and 176 edges in the case of the PLD-BSSN-II one.

Construction of Protein Sequence and Structural Similarity Networks

To better situate our findings concerning the similarity networks of protein binding sites, the sequence and structural similarity networks for the 211 proteins in the PLD were also constructed. The pairwise sequence similarity was evaluated by using FASTA,[37] and an edge between two proteins was added whenever the sequence identity between two sequences was larger than 40%. To evaluate pairwise structural similarities we applied the combinatorial extension (CE)[33] to perform an all-against-all comparison for the 211 proteins, where an edge between any two protein structures was set up if the CE *Z*-score was larger than 4.2, indicative of at least a fold level structural similarity.

RESULTS AND DISCUSSION

Connectivity within Networks of Binding Sites Similarities

In the present study, we have developed a binding site comparison method based on the well-known clique detection algorithm, which can be used for the fast evaluation of the similarity between two graphs. The binding site similarity was quantified by the Tanimoto coefficient, ranging between 0 and 1, with the higher values corresponding to a higher level of similarity. To construct the protein Binding Site Similarity Network for the PLD database (PLD-BSSN), an all-against-all binding site comparison was carried out. In this work we investigated how chemical diversity, provided by the amino acids located in the binding sites would reflect on the structure of the network. For this we divided the binding site residues into 4 (PLD-BSSN-I) and 8 classes (PLD-BSSN-II), based on the physicochemical properties of the 20 amino acids. Clearly, the PLD-BSSN-I was rather focused on the comparison of the binding site topology itself, while we expected the structure of the PLD-BSSN-II network to be influenced to a greater extent by the chemical nature of the binding sites. A typical example of a pair of proteins with low sequence identity, but high binding site similarity is shown in Figure 1. In this example, the binding site of a porcine metalloprotease (PDB entry: 1fbl, EC3.4.24.7), composed out of 16 residues, was complexing the HTA ligand (*n*-[3-(*n*′-hydroxycarboxamido)-2-(2-methylpropyl)-propanoyl]-*o*-tyrosine-*n*-methylamide). Although the sequence identity was only 28%, this porcine metalloprotease shared a relatively high similarity (Tanimoto coefficient of 0.52) with a human hydrolase (PDB entry: 1sln; EC3.4.24.17) with a binding site that comprised another 16 residues binding the INH ligand (*N*-(r-carboxyethyl)-alpha-(*s*)-(2-phenylethyl) glycyl-l-arginine-*n*-phenylamide). The size of the maximal subgraph of the two binding sites calculated with our algorithm was 11, thus giving a Tanimoto similarity of $11/(16 + 16 - 11) = 0.52$, which indicated a significant binding site similarity.

In the first network PLD-BSSN-I the vertices, representative of the 211 different binding sites, were connected by an edge if a pair of vertices had a similarity score larger than 0.35. As a result, 73 out of the 211 nonredundant protein binding sites in PLD were found to be singletons, as they had no other similar binding site. Thus, the PLD-BSSN-I network contained 138 vertices that had at least one neighbor, with an average number of nearly three edges per vertex. The graph representation of this network revealed a modular architecture that contained

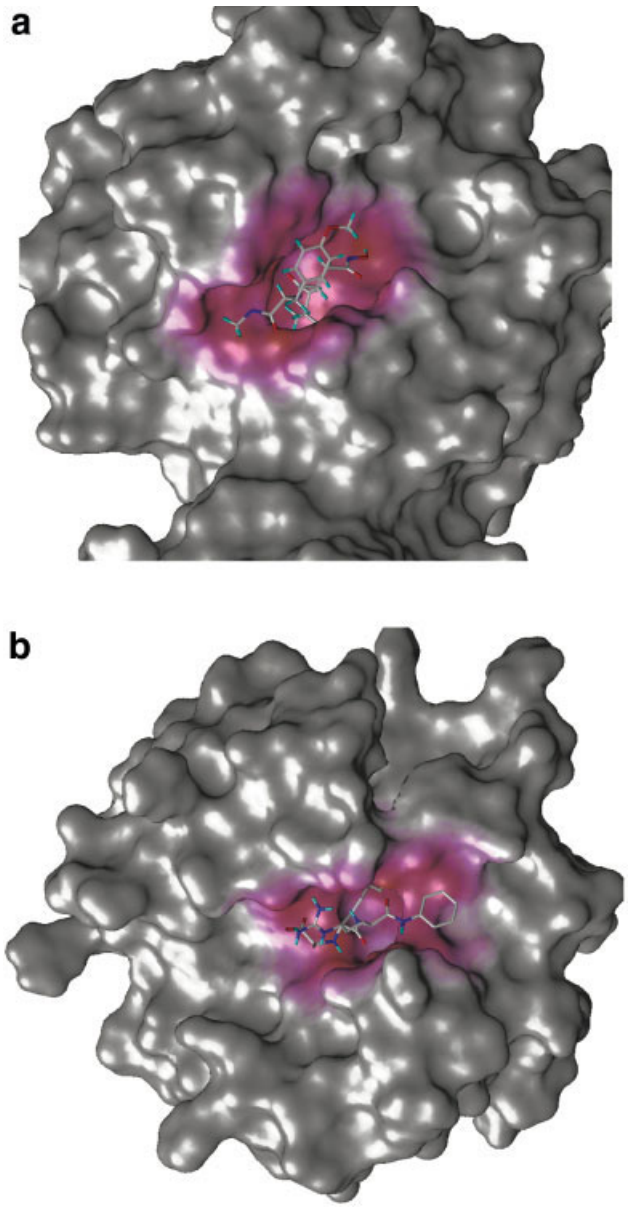

Fig. 1. Surface representation of two binding sites. The binding site of a porcine metalloprotease (PDB entry: 1fbl) (**a**) shared a relatively high similarity (Tanimoto coefficient of 0.52) with a human hydrolase (PDB entry: 1sln) (**b**), although their sequence identity was only 28%. The figure was generated by using the Sybyl molecular modeling software package (SYBYL 6.8, Tripos Inc., St. Louis, MO; 2000), with the respective ligands displayed in a sticks representation.

several highly connected clusters (Fig. 2). The same similarity threshold was applied to obtain the PLD-BSSN-II network that contained 142 connected vertices, with an average number of almost 2.5 edges per node. The chemical diversity introduced by the finer stratification of the amino acid types is reflected by the lower average connectivity in the PLD-BSSN-II network, indicating a lower average similarity.

Further, we focused our attention to the characterization of the global topological properties of the two similarity networks. Our first objective was to search for indications of these networks being small-world and scale-free. To find out if these networks were small-world, we considered random networks of equal size as references. Generally, a small-world network has a relatively short characteristic path length (L) and a high clustering coefficient (C). L is defined as the number of links in the shortest path between two vertices averaged over all pairs of vertices, while C is a measure of the local clustering within a network. As defined by Watts and Strogatz,[3] if a vertex v has k_v neighbors, then the maximum number of links between these neighbors is $[k_v\ (k_v - 1)]/2$. C_v gives the fraction of these possible links that actually exist, and C is then defined as the average C_v over all vertices v. In comparison with a random network of the same size, the parameters L and C in a small-world network satisfy to two criteria: (1) $C_{\text{small-world}}$ far exceeds C_{random}, and (2) $L_{\text{small-world}}$ slightly exceeds L_{random}. The values of these quantities that we calculated for the corresponding networks PLD-BSSN-I and PLD-BSSN-II are summarized in Table I. The results suggested that the PLD-BSSN-I network might be small-world, whereas the PLD-BSSN-II network clearly failed to satisfy to the conditions for a network to be considered as small-world.

Consequently, we tried to determine if the two similarity networks were scale-free. Scale-free networks typically have many nodes with few links, and only few highly connected ones. In contrast to a random network in which the connectivity distribution obeys a Poisson distribution, the probability $P(k)$ of nodes having k edges, decays as a power law $P(k) = k^{-\gamma}$ in scale-free networks. First, we analyzed the distribution patterns for the PLD-BSSN-I and PLS-BSSN-II binding site similarity networks. For this we plotted the connectivity distributions on a double logarithmic scale for the more reliable identification of a linear fit for the data, characteristic of a scale-free topology [cf. Fig. 3(b)]. The binding site similarity networks were approximately characterized by power laws, where $P(k) \approx k^{-142}$ ($R^2 = 0.87$) in the case of PLD-BSSN-I, and $P(k) \approx k^{-1.60}$ ($R^2 = 0.88$) for PLD-BSSN-II, respectively. We conducted a second analysis to verify if it was possible to model the data with an exponential distribution, as connectivity distributions of this type have been already observed in real-world networks. It was found that the data in the two binding site similarity networks can be equally well explained by exponential laws, where $P(k) \approx 10^{-0.207k}$ ($R^2 = 0.92$) in the case of PLD-BSSN-I and $P(k) \approx 10^{-0.1266k}$ ($R^2 = 0.85$) for PLD-BSSN-II. In particular, the data for the PLD-BSSN-I network deviated from a power law behavior in a very clear way, being closer to an exponential distribution.

It is well known that power law can be fitted reliably only when several orders of magnitude are considered. Although from our results it might be conjectured that the connectivity distributions of the PLD-BSSN-I and PLD-BSSN-II networks could follow such a distribution law, it should be emphasized that the data extends just one order of magnitude. Therefore, it was not possible to make a definitive conclusion about the scale-free architecture of the binding site similarity networks.

Fig. 2. The binding site similarity network PLD-BSSN-I. This figure was prepared using the Pajek software (http://vlado.fmf.uni-lj.si/pub/networks/pajek/).

TABLE I. Characteristic Path Lengths and Clustering Coefficents for the Networks[a]

	$L_{observed}$	L_{random}	$C_{observed}$	C_{random}
PLD-BSSN-I	3.97	3.22	0.153	0.022
PLD-BSSN-II	5.33	5.46	0.169	0.017

[a]We applied the formulae $L_{random} = \ln N/\ln k$ and $C_{random} = k/N$ for a random network with the same number of nodes (N) and average number of links (k).[2,42]

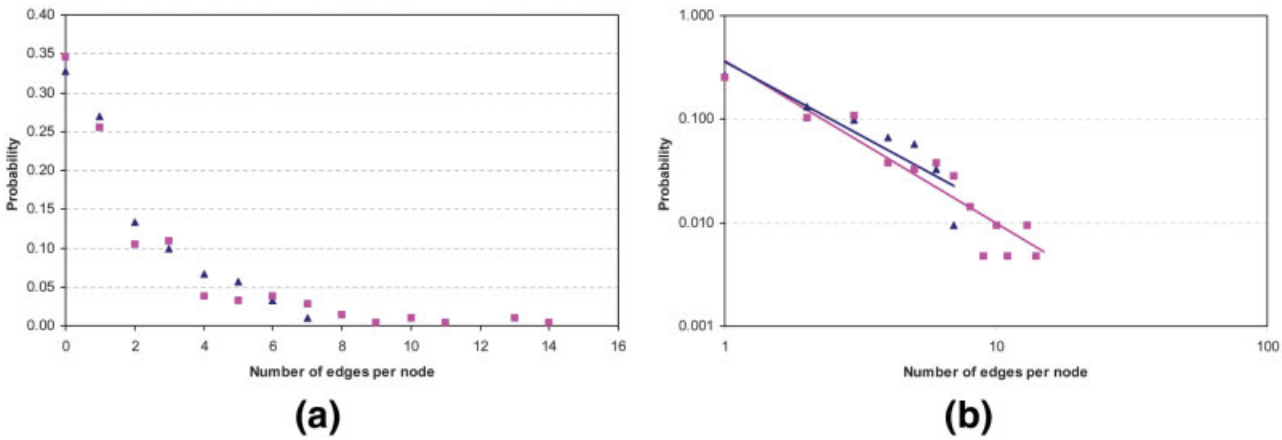

Fig. 3. Topological properties of the binding site similarity networks PLD-BSSN-I and PLD-BSSN-II. Blue and red data points belong to the PLD-BSSN-I and PLD-BSSN-II networks, respectively. (**a**) The distribution of node connectivity $P(k)$. (**b**) The log–log plot. The blue and red straight lines correspond to the power-law fits for PLD-BSSN-I and PLD-BSSN-II networks, respectively.

The current PLD dataset of some 211 binding sites is far from being complete, as it does not contain all of the protein structures for which the binding sites have been identified experimentally. To have a more reasonable characterization of the nature of binding site similarity networks, a larger dataset would be required.

In the PLD-BSSN networks that we derived, we have identified several highly connected hubs that were representative of the "generic" binding sites in the PLD database. In Table II we have listed the top 10 binding sites with the highest number of structural neighbors, referred to as the binding site "hubs". As reported by Park et al.,[12]

TABLE II. The 10 Most Highly Connected Binding Sites within the PLD-BSSN-I Network

	PDB Code	*Protein*	*Ligand*	k_v
1	1ive	Influenza A subtype n2 neuraminidase	4-(acetylamino)-3-aminobenzoic acid [BANA 108]	9
2	2pk4	Human plasminogen kringle 4	Aminocaproic acid	8
3	2msb	Mannose-binding protein a	Glycopeptide (oligomannose asparaginyl oligosaccharide)	8
4	1ptv	Protein tyrosine phosphatase 1b	Phosphotyrosine	8
5	2mcp	Immunoglobulin	Phosphocholine	7
6	1tet	Te33-Fab fragment of monoclonal antibody elicited against cholera toxin peptide 3 (CTP3)	Citrate	7
7	1slt	S-lectin	*N*-acetyllactosamine	7
8	1fig	Immunoglobulin g1 fab′ fragment	8-hydroxy-2-oxa-nicyclo [3:3.1] non-6-ene-3, 5-dicarboxylic acid	7
9	1bra	Trypsin	Benzamidine	7
10	1bcu	Alpha-thrombin	Hirugen and proflavin	7

immunoglobulins and serine proteases have been identified to be the most connected nodes in the networks of protein domain interactions. These "hub" proteins are quite versatile interaction partners, and they easily "combine" with many other protein domains to support diverse functional roles. Interestingly enough, we found out that the binding sites of immunoglobulins and serine proteases appeared as top-ranked hubs in our binding site similarity networks, a fact that further corroborates their functional polyvalence. Among the top 10 hubs in the PLD-BSSN-I network, six of the binding sites came from immunoglobulins or serine proteases. For instance, thrombin is involved in multiple functional roles: fibrinolysis, platelet aggregation, coagulation, peripheral blood cell activation, cell growth, and cellular migration. With seven partner nodes in PLD-BSSN-I, the binding site of thrombin was ranked as the top-10 hub. It might be hypothesized that the highly connected binding sites originated very early in evolution, but to prove this argument, further comparisons of the occurrence of these binding sites in different organisms would be required.

Relationships Among Protein Sequence, Structural, and Binding Site Similarity

To start with, we performed a detailed comparison of the binding site similarity networks based on the four-class (PLD-BSSN-I) and eight-class (PLD-BSSN-II) encodings. As illustrated in Figure 4(a), protein pairs always shared larger similarities based on the four-class encoding than that based on the eight-class encoding. According to the hierarchal classification of the 20 amino acids developed by Murphy et al.,[34] the four-class and eight-class encoding was consistent in reflecting the different levels of clustering of the side chains' physicochemical properties. As expected, a binding site representation of the detailed physicochemical properties of the amino acids (eight-class) led to a higher diversity and an increased dissimilarity of the binding sites [cf. Fig. 4(a)]. For instance, 214 pairs were found to have a significant similarity based on the four-class encoding, while only 176 pairs were recognized to share a similar binding site by considering the eight-class encoding. Therefore, in comparison to PLD-BSSN-II, PLD-BSSN-I has a larger number of highly connected nodes (hubs) and a higher average connectivity. For example, the average connectivity of PLD-BSSN-I was about 3.0, which is higher than that of PLD-BSSN-II where every node had 2.5 neighbors on average.

In the present study, we have also constructed the sequence and structural similarity networks for the 211 PLD proteins. Although the scale-free structure of the protein sequence and fold similarity networks have been demonstrated in previous works, we were not able to clearly identify these properties in the sequence and fold similarity networks that we derived in our work. As shown in Figure 5(a) and (b), the connectivity distributions that we observed significantly deviated from a power law behavior. A reason for this might be that some protein families were redundantly represented in the PLD database. Moreover, as in the case of the binding site similarity networks, the data spanned only a limited range on a log scale, therefore preventing a reliable characterization of the scale-free topology of the protein sequence and fold similarity networks to be made. In the similarity network of protein sequences, 99 proteins had only one neighbor, whereas 35 proteins were found to be totally isolated in the structural similarity network. In comparison with the similarity network of protein binding sites, a larger number of hubs were present in the protein sequence and structure similarity networks. For example, the maximal number of hubs in the PLD-BSSN-I binding site similarity network was 14, while this number was 19 and 25 for the sequence and structural similarity networks, respectively. The average connectivities for the protein sequence and structural similarity networks were 4.7 and 8.6, respectively, much higher than these of the binding site similarity networks reported above.

It is interesting to compare the similarity among the binding sites with the similarity of proteins at the sequence and structural levels, as this point was not addressed in previous studies. In principle, protein structure is better conserved than sequence, and protein pairs that share high sequence identity (e.g., >40%) usually are folding in the same way. However, protein pairs of low sequence identity often share significant structural similarity. We were able to verify these trends even within the very limited set of proteins included in PLD [cf. Fig. 4 (b)].

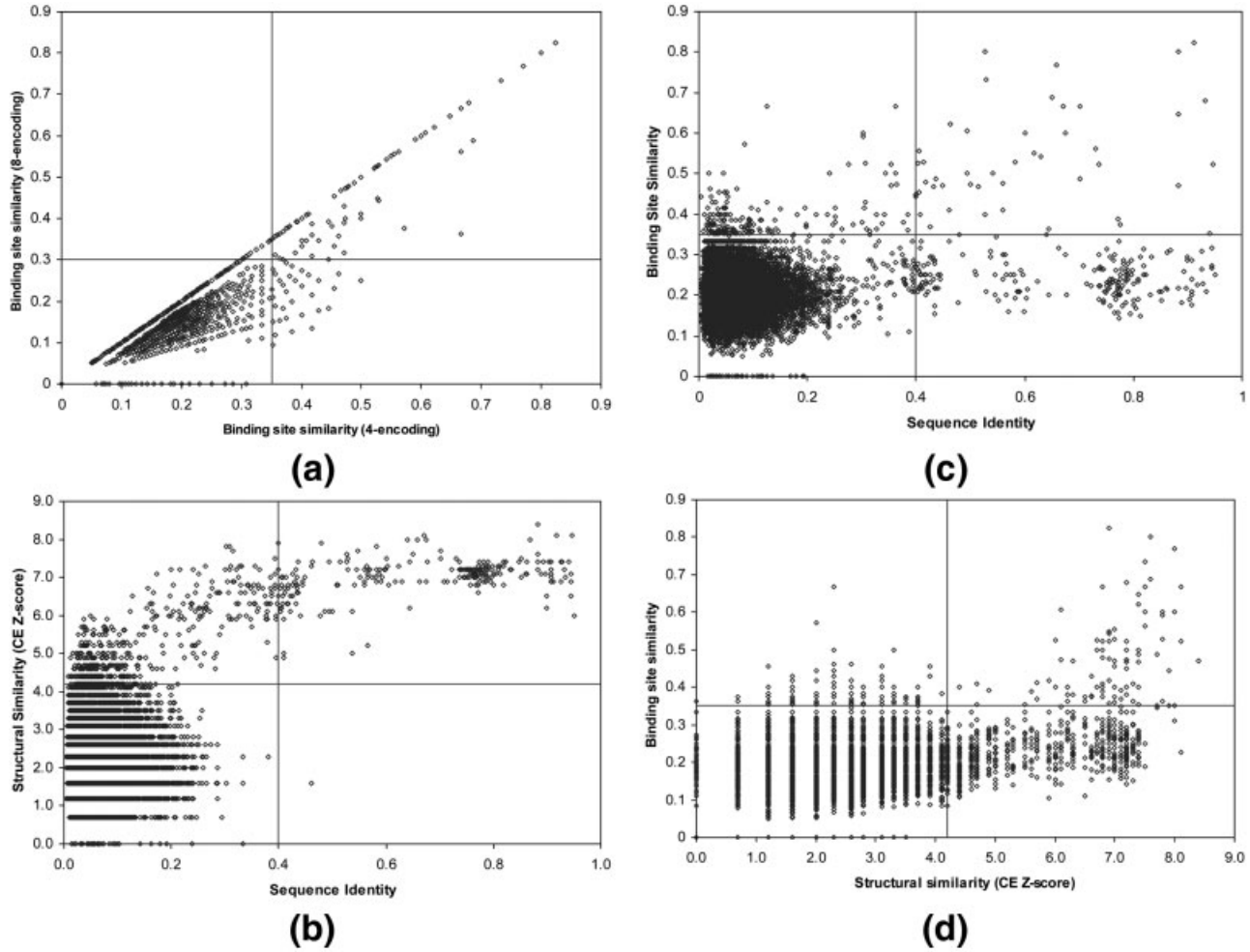

Fig. 4. The relationship between binding site, sequence, and structural similarity of proteins. (**a**) Comparison of binding similarities based on amino acid encoding in four and eight classes. (**b**) Comparison between sequence similarity and structural similarity. (**c**) Comparison between sequence similarity and binding site similarity (four-encoding). (**d**) Comparison between structural similarity and binding sites similarity (four-encoding). Significant sequence, structural, and binding site similarity (four-encoding and eight-encoding) are recognized in those pairs of proteins with a sequence identity larger than 40%, with CE *Z*-score greater than 4.2, and a Tanimoto similarities greater than 0.35 and 0.30, respectively. The lines showing the corresponding cutoff values are highlighted in each panel.

When the identity for two proteins was larger than 40%, they generally fold in similar structures, with CE *Z*-scores larger than 4.2. We found out that a significant sequence similarity did not necessary imply a significant similarity of the respective binding sites. Within the 240 protein pairs with sequence identities higher than 40%, only 46 (about 20%) shared a significant binding site similarity [Fig. 4(c)]. On the other hand, a low sequence similarity was not correlated with a lower binding site similarity. Of the 21965 pairs with low sequence identity (e.g., $<40\%$), only 168 pairs (0.8%) still had a significant binding site similarity (e.g., Tanimoto similarity higher than 0.35). In conclusion, the pairs with a sequence similarity higher than 40% were found to have a 25-fold higher chance to share a similar binding site. A similar result was also obtained by comparing the frequency of occurrence of binding site similarity with that of fold similarity among the 211 proteins in PLD. Among the 655 pairs of proteins that shared similar structures, 14% exhibited a significant binding site similarity [Fig. 4(d)]. On the contrary, only 0.6% of those pairs without significant structural similarity shared similar binding sites.

We have performed some investigations on those protein pairs with high sequence or fold similarities but without significant binding site similarity. It turned out that current binding site comparison algorithms experience difficulties to take in account protein flexibility, mostly due to the definition of the binding sites. For instance, porcine and avian citrate synthases (PDB entries 4cts and 2csc) shared 91% sequence identity, and the structural superimposition between them revealed that these two enzymes shared a similar binding site, because the superimposed ligands were closely located. However, the size of the ligands in the avian citrate synthase (D-malate and carboxymethyl coenzyme A) was larger than that of the ligand in the porcine enzyme (oxaloacetate). As a result, the sizes of the defined binding sites appearing within the current algorithm were quite different, which was reflected in a low binding site similarity.

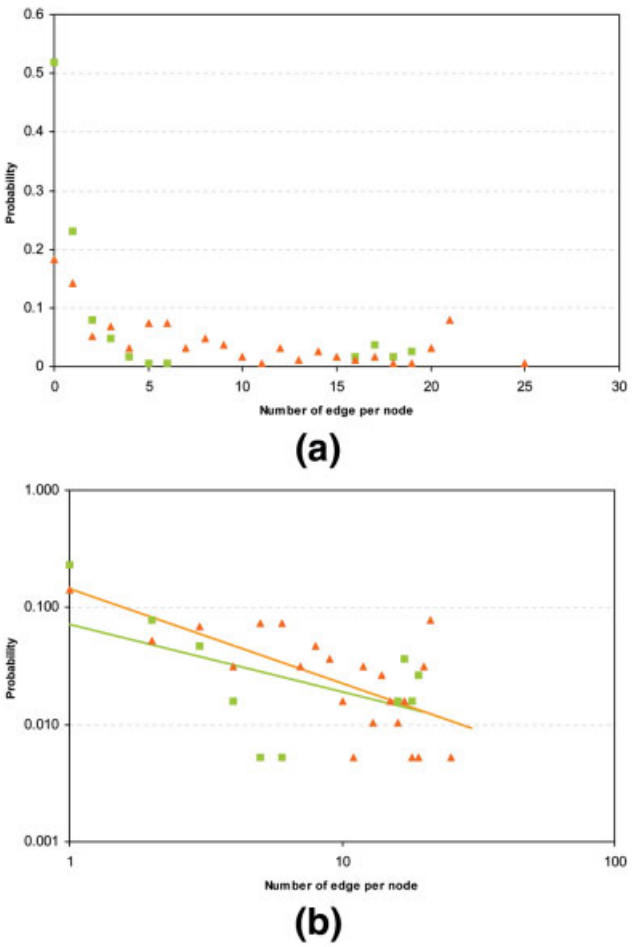

Fig. 5. Topological properties of sequence and structural similarity networks constructed for the 211 PLD proteins. Data points encoded with green triangles and yellow rectangles represent the sequence similarity network and the structural similarity network, respectively. (**a**) The distribution of node connectivity $P(k)$. (**b**) The log–log plot. The green and yellow straight lines correspond to the power-law fits for the sequence and structural similarity networks, respectively.

Even so, the current observation within the PLD database has demonstrated that the relationship between the similarity of protein binding sites and the similarity of protein sequences and structures is not straightforward. A further investigation should certainly take in account the flexibility of binding sites upon ligand docking. However, we have obtained already here indications that the similarity of protein binding sites seems less conserved than the relatedness among protein sequences or structures.

Based on the above detailed comparisons, it can be conjectured that in a arbitrary sample of protein molecules, here represented by the PLD database, the protein similarity should be the highest at the structural level, followed by sequence similarity, while protein molecules would turn to be mostly dissimilar with respect to the structure of their binding sites. However, to prove this argument improved binding site comparison algorithms should be used, that take into account protein flexibility upon ligand binding, and that are to be applied to much larger protein datasets.

Applications

The possibility to determine pairwise binding site similarities is opening avenues for potential new applications. The characterized similarity networks of protein binding sites imply that these binding sites could be clustered in a hierarchical way, so that an entirely new classification scheme could be proposed for the implementation of protein binding site databases. In turn, such modular databases, enriched with the already available protein sequence family and structural information, will provide new tools to understand and predict protein function. The detection of a binding site similarity might be used for the functional annotation of those known or predicted gene products for which function cannot be inferred by the classical comparative sequence- or structure-based methods. Indeed, the comparison of binding sites is increasingly used to accelerate the in silico annotation of functionally unknown genes.[38]

A second application could be identified in the area of de novo enzyme molecular design. The catalytic function of enzymes is the result of natural evolution, and biochemists have tried for more than one century to understand the underlying chemical mechanisms.[39] The ultimate test of our understanding of enzyme catalysis would be the design and production of enzyme chimeras from scratch. To achieve this goal, an important step consists in manipulating the catalytic residues of an already known binding site. Using computer-based rational design, for instance, Dwyer et al.[40] turned the inverting ribose-binding protein into a triose phosphate isomerase. Undoubtedly, the availability of an extensive similarity network of protein binding sites will be useful to assess the de novo designability of novel enzymatic functions.

A third important application concerns structure-based drug design activities. Indeed, provided that an extensive similarity network of protein binding sites is at hand, the neighbors of a target binding site will provide valuable information for the most relevant potential ligands to be screened on the orphan receptor. The position of the target binding site in the hierarchical similarity network might be indicative of the level to which specificity could be obtained. Indeed, in the case when the binding site of the protein target turns out to be a "hub," with a large number of related sites, the design of a highly specific bioactive molecule will hardly succeed, as other targets with similar binding sites might bind the same molecule. However, an advantage in such situations is that the designed bioactive molecule might have multiple functionalities. In any case, the analysis of the similarity network of protein binding sites in the early phases of a project has the potential to become a routine tool for structure-based drug design. Such a tool would be especially informative for emerging disciplines such as chemogenomics,[41] which has the ambitious task of identifying all possible drugs for all possible targets.

CONCLUSIONS

We have uncovered novel network structures in the world of proteins that consisted in a hierarchically organized structural relationship among protein binding sites. The generated similarity networks of protein binding sites

have been found to have a few highly connected hubs as well as many nodes with few links. Considering the small data set (PLD database) used in the present study, we were not able to definitively prove that these networks have scale-free and small-world properties. Due to complicated evolution of protein binding sites, the relationship between protein binding similarity and protein sequence or structural similarity is not straightforward. Generally, those pairs with higher sequence or structural similarity tend to exhibit a higher binding site similarity. This allows for integration of the binding site comparisons into the pipeline of computational functional annotation that until now was exclusively based on the assessment of sequence similarity, and to a lesser extent on structural homology. The binding site similarity networks uncover new perspectives in the construction of hierarchical classification databases, in the improvement of de novo design of enzymes, and in the enhancement of the efficiency of structure-based drug design activities.

ACKNOWLEDGMENTS

We thank Dr. Nils Weskamp (Dept. of mathematics and computer science, University of Marburg, Germany) for his generous help and stimulating discussion on clique detection algorithm. We also thank the referees whose constructive comments were very helpful in improving the quality of this work.

REFERENCES

1. Strogatz SH. Exploring complex networks. Nature 2001;410:268–276.
2. Greene LH, Higman VA. Uncovering network systems within protein structures. J Mol Biol 2003;334:781–791.
3. Watts DJ, Strogatz SH. Collective dynamics of "small-world" networks. Nature 1998;393:440–442.
4. Barabasi AL, Albert R. Emergence of scaling in random networks. Science 1999;286:509–512.
5. Maslov S, Sneppen K. Specificity and stability in topology of protein networks. Science 2002;296:910–913.
6. Ravasz E, Somera AL, Mongru DA, Oltvai ZN, Barabasi AL. Hierarchical organization of modularity in metabolic networks. Science 2002;297:1551–1555.
7. Shen-Orr SS, Milo R, Mangan S, Alon U. Network motifs in the transcriptional regulation network of Escherichia coli. Nat Genet 2002;31:64–68.
8. Wuchty S. Evolution and topology in the yeast protein interaction network. Genome Res 2004;14:1310–1314.
9. Grigorov MG. Global properties of biological networks. Drug Discovery Today 2005;10:365–372.
10. Bork P, Jensen LJ, von Mering C, Ramani AK, Lee I, Marcotte EM. Protein interaction networks from yeast to human. Curr Opin Struct Biol 2004;14:292–299.
11. Wuchty S. Scale-free behavior in protein domain networks. Mol Biol Evol 2001;18:1694–1702.
12. Park J, Lappe M, Teichmann SA. Mapping protein family interactions: intramolecular and intermolecular protein family interaction repertoires in the PDB and yeast. J Mol Biol 2001;307:929–938.
13. Amitai G, Shemesh A, Sitbon E, Shklar M, Netanely D, Venger I, Pietrokovski S. Network analysis of protein structures identifies functional residues. J Mol Biol 2004;344:1135–1146.
14. Rao F, Caflisch A. The protein folding network. J Mol Biol 2004;342:299–306.
15. Qian J, Luscombe NM, Gerstein M. Protein family and fold occurrence in genomes: power-law behaviour and evolutionary model. J Mol Biol 2001;313:673–681.
16. Dokholyan NV, Shakhnovich B, Shakhnovich EI. Expanding protein universe and its origin from the biological Big Bang. Proc Natl Acad Sci USA 2002;99:14132–14136.
17. 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.
18. Zhang Y, Skolnick J. The protein structure prediction problem could be solved using the current PDB library. Proc Natl Acad Sci USA 2005;102:1029–1034.
19. Orengo CA, Todd AE, Thornton JM. From protein structure to function. Curr Opin Struct Biol 1999;9:374–382.
20. Wilson CA, Kreychman J, Gerstein, M. Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores. J Mol Biol 2000;297:233–249.
21. Nagano N, Orengo CA, Thornton JM. One fold with many functions: the evolutionary relationships between TIM barrel families based on their sequences, structures and functions. J Mol Biol 2002;321:741–765.
22. Morris RJ, Najmanovich RJ, Kahraman A, Thornton JM. Real spherical harmonic expansion coefficients as 3D shape descriptors for protein binding pocket and ligand comparisons. Bioinformatics 2005;21:2347–2355.
23. Stark A, Sunyaev S, Russell RB. A model for statistical significance of local similarities in structure. J Mol Biol 2003;326:1307–1316.
24. Kinoshita K, Nakamura H. Identification of protein biochemical functions by similarity search using the molecular surface database eF-site. Protein Sci 2003;12:1589–1595.
25. Artymiuk PJ, Poirrette AR, Grindley HM, Rice DW, Willett P. A graph-theoretic approach to the identification of three-dimensional patterns of amino acid side-chains in protein structures. J Mol Biol 1994;243:327–344.
26. Ferre F, Ausiello G, Zanzoni, A, Helmer-Citterich M. SURFACE: a database of protein surface regions for functional annotation. Nucleic Acids Res 2004;32:D240–D244.
27. Binkowski TA, Adamian L, Liang J. Inferring functional relationships of proteins from local sequence and spatial surface patterns. J Mol Biol 2003;332:505–526.
28. 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.
29. Wallace AC, Borkakoti N, Thornton JM. TESS: a geometric hashing algorithm for deriving 3D coordinate templates for searching structural databases. Application to enzyme active sites. Protein Sci 1997;6:2308–2323.
30. Shulman-Peleg A, Nussinov R, Wolfson HJ. Recognition of functional sites in protein structures. J Mol Biol 2004;339:607–633.
31. Jambon M, Imberty A, Deleage G, Geourjon C. A new bioinformatic approach to detect common 3D sites in protein structures. Proteins 2003;52:137–145.
32. Puvanendrampillai D, Mitchell JB. L/D Protein Ligand Database (PLD): additional understanding of the nature and specificity of protein-ligand complexes. Bioinformatics 2003;19:1856–1857.
33. Shindyalov IN, Bourne PE. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. Protein Eng 1998;11:739–747.
34. Murphy LR, Wallqvist A, Levy RM. Simplified amino acid alphabets for protein fold recognition and implications for folding. Protein Eng 2000;13:149–152.
35. Bron C, Kerbosch J. Algorithm 457: Finding all cliques of an undirected graph. Commun ACM 1973;16:575–577.
36. Weskamp N, Kuhn D, Hullermeier E, Klebe G. Efficient similarity search in protein structure databases by k-clique hashing. Bioinformatics 2004;20:1522–1526.
37. Pearson WR, Lipman DJ. Improved tools for biological sequence comparison. Proc Natl Acad Sci USA 1988;85:2444–2448.
38. Laskowski RA, Watson JD, Thornton JM. From protein structure to biochemical function? J Struct Funct Genomics 2003;4:167–177.
39. Sterner R, Schmid FX. Biochemistry. De novo design of an enzyme. Science 2004;304:1916–1917.
40. Dwyer MA, Looger LL, Hellinga HW. Computational design of a biologically active enzyme. Science 2004;304:1967–1971.
41. Mestres J. Computational chemogenomics approaches to systematic knowledge-based drug discovery. Curr Opin Drug Discov Dev 2004;7:304–313.
42. 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.