Localization of protein-binding sites within families of proteins

DMITRY KORKIN,[1] FRED P. DAVIS,[1,2] AND ANDREJ SALI[1]

[1]Departments of Biopharmaceutical Sciences and Pharmaceutical Chemistry, and California Institute for Quantitative Biomedical Research, and [2]Graduate Group in Biophysics, University of California at San Francisco, San Francisco, California 94143, USA

(RECEIVED May 5, 2005; FINAL REVISION May 27, 2005; ACCEPTED May 27, 2005)

Abstract

We address the question of whether or not the positions of protein-binding sites on homologous protein structures are conserved irrespective of the identities of their binding partners. First, for each domain family in the Structural Classification of Proteins (SCOP), protein-binding sites are extracted from our comprehensive database of structurally defined binary domain interactions (PIBASE). Second, the binding sites within each family are superposed using a structural alignment of its members. Finally, the degree of localization of binding sites within each family is quantified by comparing it with localization expected by chance. We found that 72% of the 1847 SCOP domain families in PIBASE have binding sites with localization values greater than expected by chance. Moreover, 554 (30%) of these families have localizations that are statistically significant (i.e., more than four standard deviations away from the mean expected by chance). In contrast, only 144 (8%) families have significantly low localization. The absence of a significant correlation of the binding site localization with the average sequence and structural conservations in a family suggests that localization can be helpful for describing the functional diversity of protein–protein interactions, complementing measures of sequence and structural conservation. Consideration of the binding site localization may also result in spatial restraints for the modeling of protein assembly structures.

Keywords: protein–protein interactions; protein interfaces; binding sites; evolution; protein family

Interactions between proteins play a key role in many cellular processes (Alberts and Miake-Lye 1992; Pawson and Nash 2003). An important step toward a mechanistic understanding of these processes is a structural description of the interactions within protein complexes (Park et al. 2001; Edwards et al. 2002; Sali et al. 2003). Studies of both proteins and their assemblies have generally used the domain as the principal unit of a protein (Jones et al. 2000; Spahn et al. 2001; Bashton and Chothia 2002; Janin et al. 2003; Ng et al. 2003; Zhang et al. 2003). Protein domains are structural units that carry out specific functions and are often identified as the basic evolutionary building blocks that are shuffled, duplicated, and fused during protein evolution (Wetlaufer 1973; Doolittle and Bork 1993; Kolkman and Stemmer 2001). Given the average length of 466 residues for a yeast protein and 173 residues for a domain in the CATH database, a protein is folded on average into approximately two domains (Chothia et al. 2003; Sali et al. 2003).

What features make protein-binding sites unique with respect to the rest of the surface? If we knew the answer to this question, we would be closer to solving a number of other problems, such as the identification and characterization of protein functional sites as well as the prediction of assembly structures. Many regularities

Reprint requests to: Andrej Sali, 1700 4th Street, Suite 503B, University of California at San Francisco, San Francisco, CA 94143-2552, USA; e-mail: sali@salilab.org; fax: (415) 514-4231.

Article published online ahead of print. Article and publication date are at http://www.proteinscience.org/cgi/doi/10.1110/ps.051571905.

 Copyright © 2005 The Protein Society

that are potentially useful for the recognition of protein-binding sites have been revealed in studies of the physicochemical properties of protein–protein interactions, including size, binding surface area, residue composition, polarity, and hydrophobicity (Tsai et al. 1996, 1997; Larsen et al. 1998; Lo Conte et al. 1999; Jones et al. 2000; Sheinerman et al. 2000; Chakrabarti and Janin 2002; Sheinerman and Honig 2002). Most protein-binding sites are flat (Jones and Thornton 1996). To form an interface, the component binding sites are almost always required to be geometrically complementary (Lawrence and Colman 1993). The sequence and structural conservation of both protein-binding sites and corresponding interfaces have been analyzed for specific families of proteins (Paavilainen et al. 2002; Campbell and Jackson 2003; Ofran and Rost 2003) as well as on a larger scale (Grishin and Phillips 1994; Tsai et al. 1997; Valdar and Thornton 2001; Aloy et al. 2003; Caffrey et al. 2004; Rajamani et al. 2004; Littler and Hubbard 2005). Some studies suggest that the interface residues are more highly conserved than the rest of the surface (Tsai et al. 1997; Valdar and Thornton 2001; Littler and Hubbard 2005), while others do not find a significant increase in conservation (Grishin and Phillips 1994; Caffrey et al. 2004).

Here we analyzed the degree of localization of protein-binding sites within families of related proteins. That is, we addressed the question of whether protein-binding sites on homologous proteins share similar relative positions on their surfaces, irrespective of the identities of their binding partners. We found that 72% of the domain families in our sample have binding sites with localization values greater than expected by chance.

In the next section, we quantify the localization of protein-binding sites within families of homologs and correlate it with common measures of sequence and structure divergence. We describe two specific families with highly and poorly localized protein-binding sites. Finally, we analyze the localization of binding sites in multidomain assemblies of known structure.

Results

Domain localization

The sample of binding sites was used to obtain the localization index and its value expected by chance for each SCOP family, as detailed in Materials and Methods. The localization of the protein-binding sites is on average higher than the localization of randomly generated binding sites (Fig. 1A). In particular, for 72% of the families, the protein-binding sites have greater localization index than expected by chance, with an average localization difference of ~0.1. In fact, 554 (30%)

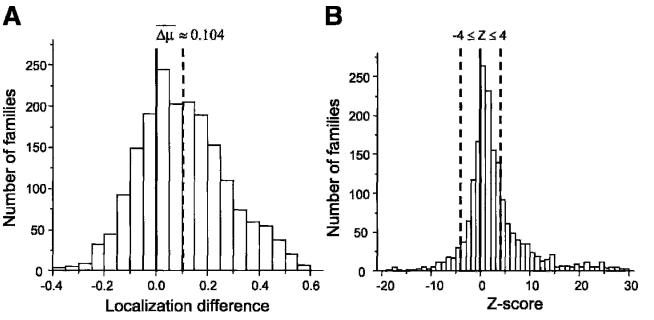

Figure 1. Distribution of the binding site localization. (*A*) Distribution of the difference between the localization of the existing binding sites and the localization of randomly generated binding sites. The average difference is ~0.1. (*B*) Distribution of the significance of localization (Z-score) values. There are 144 families with Z-scores < -4 and 554 families with Z-scores > 4 out of the 1847 SCOP families whose members participate in at least two nonredundant interactions.

families have a significantly greater binding site localization than expected by chance ($Z > 4.0$), and only 144 (8%) families have a significantly low binding site localization ($Z < -4.0$) (Fig. 1B). The distribution of the localization index versus domain family size revealed no correlation (data not shown).

Next, for each domain family we estimated the likelihood of a protein-binding site to occur on the cumulative binding map obtained from other known protein-binding sites in the same family. The jackknife procedure was used (see Materials and Methods). For 50% of the families, the binding site is more likely to be localized on the cumulative binding map (binding site coverage > 0.5) than outside of the map (Fig. 2A). On average, a cumulative binding map covers 53% of a new binding site.

The conservation of the location of a new binding site, given the other binding sites in the family, is even

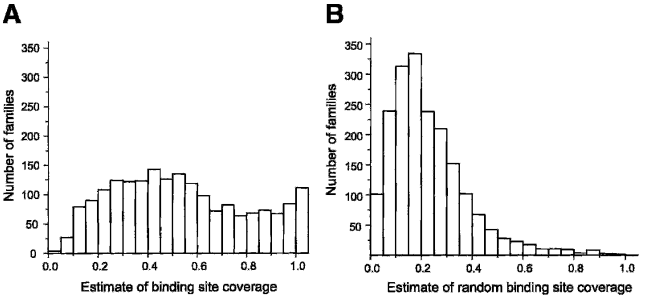

Figure 2. Binding site coverage. (*A*) Distribution of the estimated binding site coverage for each domain family. The estimate was obtained by a jackknife procedure (Materials and Methods). (*B*) Distribution of the binding site coverage obtained from a set of randomly generated binding sites. For 1752 of the 1847 SCOP families, the actual coverage of the protein-binding sites is higher than for the randomly located binding sites.

more clearly demonstrated by comparing the estimated coverage of known binding sites with the estimated coverage of binding sites generated by chance (Fig. 2B). In particular, for the randomly generated protein-binding sites, only 6% of the families have a binding site coverage higher than 0.5; on average, 23% of binding site residues obtained by chance are covered by an existing cumulative binding map.

How well does the conservation of location of protein-binding sites in a family correlate with the sequence and structural similarity among the family members? To answer this question, we compared localization with the following measures of sequence and structural conservation: root-mean-square deviation (RMSD) for C_α atoms, overall sequence identity, sequence identity of surface residues, and sequence identity of binding site residues. We found that binding site residues are no more conserved than the whole sequences of the family members (Fig. 3A), which is consistent with the previous results for a smaller set of proteins (Caffrey et al. 2004). In contrast to the high correlations among the overall sequence, binding site, and surface identities, the localization difference is not correlated with any of these three parameters (a representative plot is shown in Fig. 3B; R^2 is 0.19 for the correlation between the overall sequence identity and localization difference), although very similar proteins do tend to have similarly located binding sites and distantly related proteins do not. In addition, the localization difference is not correlated with C_α RMSD (Fig. 3C; $R^2 = 0.15$). We also observed no correlation between the localization difference of a family and the number of distinct families that interact with it; the distributions of localization differences for the families interacting with one, two, three, and four, as well as five and more different families are strikingly similar (Fig. 3D).

Assembly localization

In addition to studying the localization of protein-binding sites within individual domain families, we also obtained the localization index for complete multidomain assemblies using 20,639 assemblies stored in PIBASE, our comprehensive database of structurally defined binary domain interactions. The analysis included assemblies with 2–72 domains from 1 to 14 distinct families. The assembly localization index varies from −0.46 to 0.59 (Fig. 4A). PIBASE is mostly

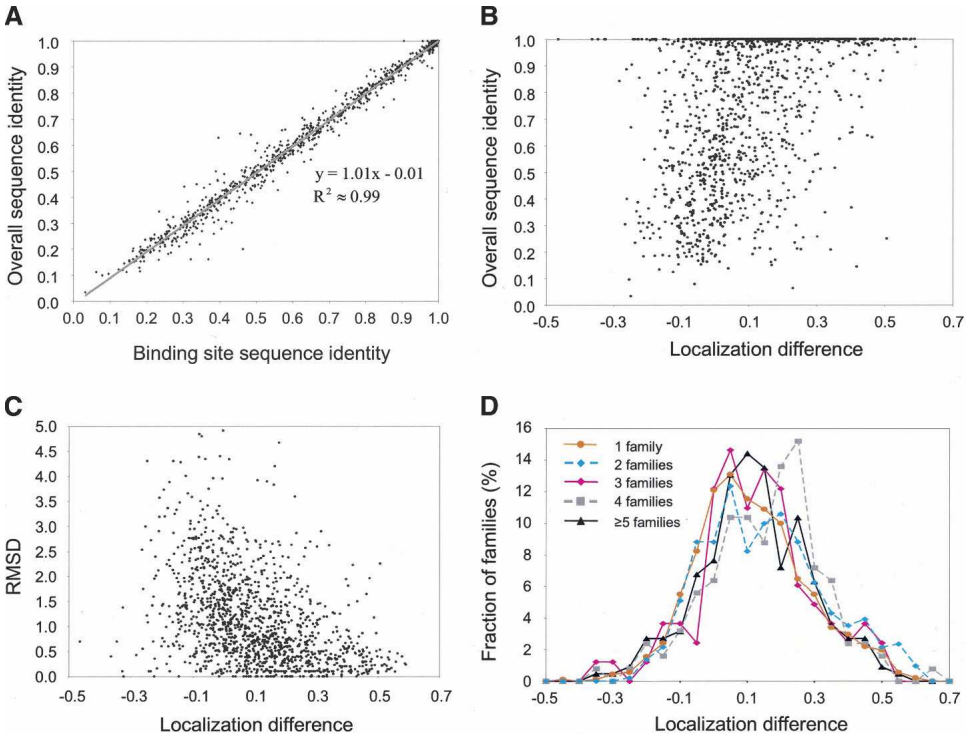

Figure 3. The relationship of the binding site localization to the structural and sequence divergence of family members. (*A*) Binding site residues are no more conserved than the overall sequence of the family members (best linear fit: $y = 1.01x - 0.01$, $R^2 = 0.99$). The distribution of the localization difference vs. overall sequence identity (*B*) and RMSD (*C*) revealed no apparent correlation (corresponding R^2 values for the best linear fit are 0.19 and 0.15, respectively). (*D*) The distributions of localization differences for families interacting with one, two, three, and four, as well as five and more distinct families, are similar to each other.

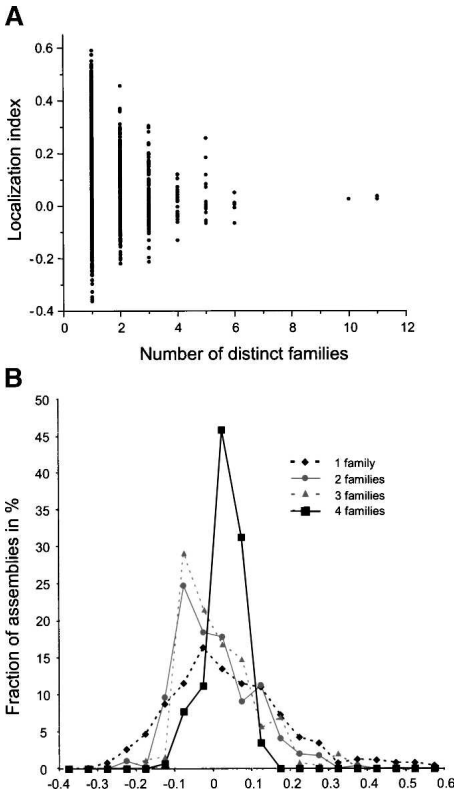

Figure 4. Assembly localization index. (*A*) Assembly localization index for the 20,639 known structures of multidomain assemblies. (*B*) Distribution of the assembly localization index values for the multidomain assemblies containing domains from one, two, three, and four distinct families. A non-zero index value is assigned to a multidomain assembly if each of the contributing domain families has a non-zero value of the binding site localization.

populated with assemblies containing domains from one to six distinct families with an average assembly localization index of 0.04. The distribution of the assembly localization index narrows with an increase in the number of distinct domain families, although the distributions have approximately the same mean values (Fig. 4B). The highest and lowest localization values occur for homo-oligomers, mainly dimers (Fig. 4). Sample assemblies with both a high assembly localization index and a large number of domains are listed in Table 1.

Here we present families from the two extremes of the localization distribution:

Example 1: Acyl-CoA dehydrogenase

We first surveyed the NM domains of acyl-CoA dehydrogenases (SCOP id e.6.1.1), a family of mitochondrial flavoproteins involved in the catabolism of fatty and amino acids (Thorpe and Kim 1995; Kim and Miura 2004). All acyl-CoA dehydrogenases form homotetramers, except for Very Long Chain acyl-CoA dehydrogenase, which forms a homodimer (Kim and Miura 2004). Each monomer consists of the NM and C-terminal domains (Kim and Miura 2004). There are 56 members in the NM family, participating in 186 nonredundant interactions. The localization of protein-binding sites is $\mu(F_D) = 0.71$, which is significantly higher than the localization value of $\overline{\mu}_R(F_D) = 0.40$ expected by chance ($S_R = 0.01$ and $Z = 27.3$). The cumulative binding map of the NM domains consists of the binding sites involved in interaction with the C-terminal domain and homo-oligomerization (Fig. 5A).

Comparison of the sequences and structures of the NM domains shows significant variation among surface (33%–100%), binding site (30%–100%), and overall sequence identities (33%–100%) as well as C_α RMSD (0.0–1.6 Å). The average values of surface (72%), binding site (72%), and overall sequence identities (70%) are high.

Example 2: C-type lectins

The family of C-type lectin domains (SCOP ID d.169.1.1) (Drickamer 1999; Dodd and Drickamer 2001; Nicholas and Hodgkin 2004; Vasta et al. 2004) contains 424 domains, 325 of which are interacting with other domains and participate in 665 nonredundant interactions in total. In contrast to alkaline phosphatases, the localization of the protein-binding sites in C-type lectins is significantly lower ($\mu(F_D) = 0.24$) than expected by chance ($\bar{\mu}_R(F_D) = 0.53$; $s_R = 0.07$; $Z = -10.6$). The cumulative binding map consists of contact residues with low interaction density that are spread across the domain surface (Fig. 5B). The poor localization is consistent with the known functional diversity of C-type lectins (Drickamer 1999; Vasta et al. 2004). Spe-

Table 1. *Examples of multidomain assemblies with high assembly localization index values*

PDB	Name	Localization index	Total number of domains	No. of distinct domain families
1kc6	Type II restriction enzyme Hincii	0.50	4	1
1mjg	Carbon-monoxide dehydrogenase	0.46	8	2
1hi9	D-Aminopeptidase Dppa	0.34	10	1
1iwp	Glycerol dehydratase-cyanocobalamin complex	0.31	6	3
1l9v	Nonstructural rotavirus protein	0.28	16	2

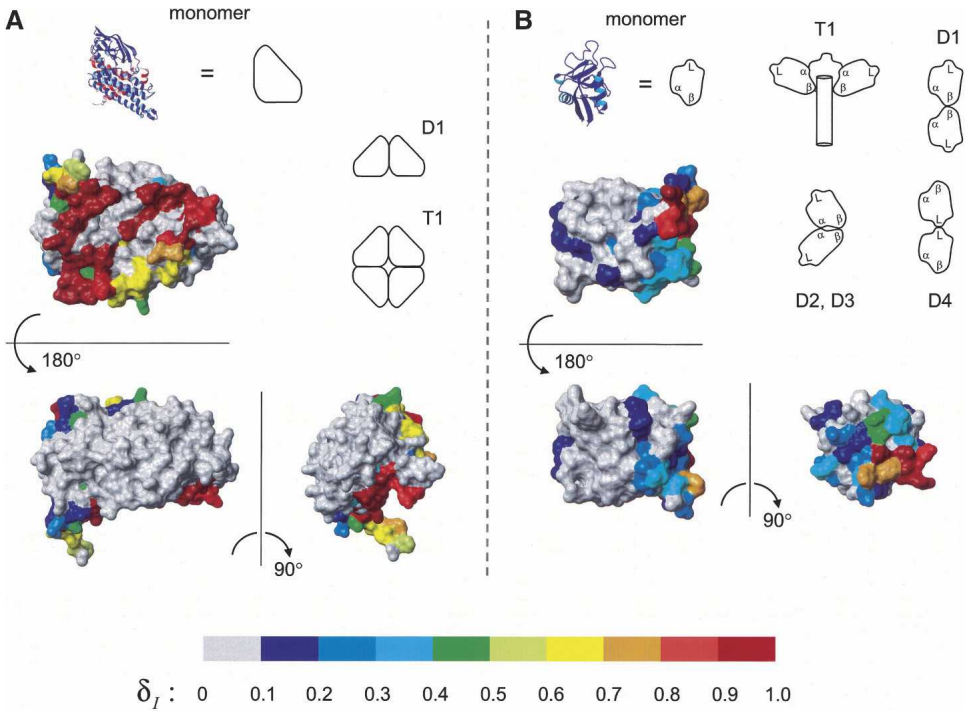

Figure 5. Examples of SCOP families with well and poorly localized protein-binding sites. (*A*) The NM domain of acyl-CoA dehydrogenases (SCOP ID e. 6.1.1). (*B*) C-type lectin domain (SCOP ID d.169.1.1). A family representative is shown in three projections as indicated, with the contact residues colored by interaction density δ_I. All known oligomer configurations are indicated schematically. Whereas most contact residues of the NM domains have high interaction density values and are located on one side of the representative domain, the contact residues of domains in the C-type lectin family have significantly lower interaction identity values and are scattered over the entire surface of the representative domain. There is one dimeric and one tetrameric configuration of the NM domains, while there are four dimeric configurations and one trimeric configuration of C-type lectins. Configurations D2 and D3 of C-type lectins, corresponding to *Polyandrocarpa* lectin and CD94, respectively, differ structurally at the interface and in their functions despite an appearance of overall structural similarity.

cifically, there are five different functional subfamilies, including carbohydrate-recognition domains, type II antifreeze proteins, oxidized LDL receptors, phospholipase receptors, and NK cell receptors. Moreover, there are five different oligomer configurations, including one trimeric and four different dimeric forms, that involve different surface regions as well as spatial orientations (Drickamer 1999).

As in the case of the NM domains of acyl-CoA dehydrogenases, sequence and structural comparisons of C-type lectins reveal significant variation among surface (11%–100%), binding site (13%–100%), and overall sequence identities (16%–100%), with C_α RMSD varying from 0.0 to 2.4 Å. The average values are lower than in the NM domain family: 56% for surface, 56% for binding site, and 55% for the overall sequence identities.

Discussion

We aimed to investigate whether protein-binding sites on homologous proteins share similar relative positions on their surfaces. To achieve this, we performed a comprehensive analysis of binary interactions for domains in 1847 SCOP families (Figs. 6–8). For each domain family, we

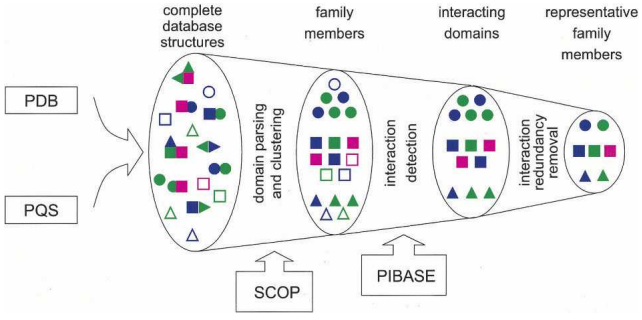

Figure 6. Obtaining the sample of nonredundant domain–domain interactions. Open and filled objects correspond to noninteracting and interacting domains, respectively. Filled objects of the same shape belong to the same SCOP family of interacting domains. Filled objects of the same shape and color are redundant with respect to each other.

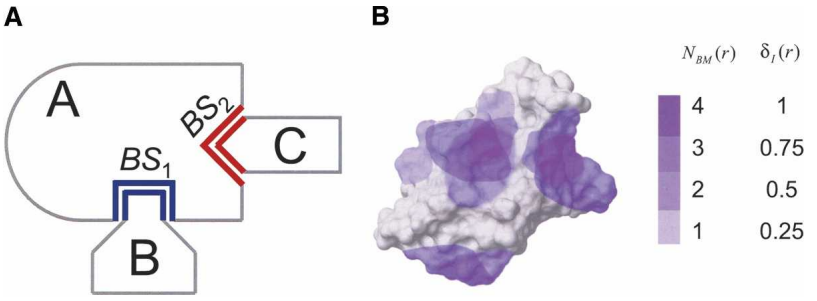

Figure 7. Protein interaction terms. (*A*) Definitions: Domain A participates in two interactions: one with domain B and another one with domain C providing two interfaces, A–B and A–C. As a result, domain A has two protein-binding sites, BS_1 and BS_2. The union of the two protein-binding sites, BS_1 and BS_2, gives a cumulative binding surface of domain A, which is noncontiguous in this case. (*B*) Cumulative binding map. Actual interaction residues of all members in a family, irrespective of the other partner domain, are mapped on the representative structure, following their superposition on the representative structure. The interaction density of a residue position *r*, $\delta_I(r)$, is indicated by a color scheme ranging from 0 (white) to 1 (dark violet) in steps of 0.25. The number of binding maps, $N_{BM}(r)$, per residue position r is proportional to $\delta_I(r)$ by construction.

extracted the protein-binding sites of its members from PIBASE (Pieper et al. 2004; Davis and Sali 2005) and superposed them via a structural alignment of member structures. Next, we described the localization index of protein-binding sites and qualified it by three additional measures, including localization difference relative to a random localization, statistical significance of localization given random localization, and coverage of a new binding site by the previously known binding sites in the same family (Figs. 1, 2). In addition, we correlated the localization difference with common measures of sequence and structure divergence (Fig. 3). We also described the localization index of multidomain assemblies (Fig. 4; Table 1). We illustrated localization by two sample families at the extremes of the localization distribution (Fig. 5; Table 2).

Of the 1847 domains analyzed, 554 and 144 exhibit significantly high and low localization of their protein-binding sites, respectively (Fig. 2). While only 28.1% of the 1847 families are enzymes, 71.8% of the families with higher localization than expected by chance are enzymes (Table 2), as illustrated by the NM domains of acyl-CoA dehydrogenases (Fig. 5A). This finding may be explained by the known preference of enzyme domains to interact with domains of the same family, often in a symmetric fashion, in order to avoid unwanted aggregation and perform cooperative binding functions (Goodsell and Olson 2000; Park et al. 2001). On the other hand, low protein-binding site localization may be associated with functional or structural diversity of the family members, as illustrated by C-type lectins (Fig. 5B). It is unlikely that poor localization is due to a larger number of family members; no correlation between family size and binding site localization was observed (data not shown).

Why does nature appear to favor localization of protein-binding sites on homologous proteins? We discuss the interplay of physics and evolution in creating localized binding sites that mediate binary protein interactions as well as higher-order complexes.

Evolution acts to preserve existing and generate novel biological functions. Preservation of a biological function mediated by a protein–protein interaction places constraints on sequence divergence, which tends to conserve the binding site location (Teichmann 2002). Generation of a novel function could be bootstrapped by reusing a previously existing binding site, taking advantage of its physical properties that have been evolved to favorably mediate protein–protein interactions (DeLano et al. 2000). Both scenarios result in a high binding site localization. Novel functions could also be mediated by binding through new surface regions of the proteins. This mechanism would lower the binding site localization of the protein family.

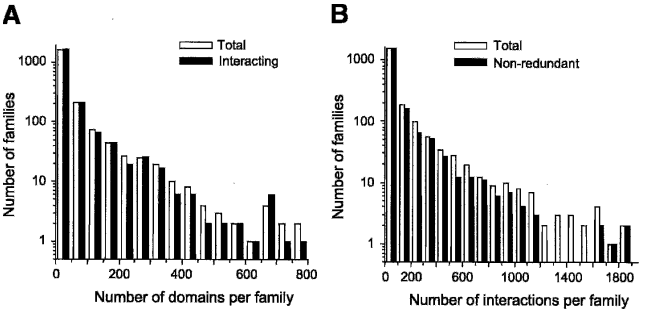

Figure 8. Interacting domains and interaction redundancy (logarithmic scale). (*A*) Distribution of the number of domains interacting with at least one other domain in a domain family (black) compared with the distribution of the total number of domains in the family (white). (*B*) Distribution of the nonredundant interactions in a domain family (black) compared with the distribution of the total number of interactions in the family (white).

Table 2. *Domain families with most (rows 1–5) and least (rows 6–10) localized protein binding sites*

SCOP ID	Family name	No. of nonredundant interactions	Localization difference
b.30.5.4	Aldose 1-epimerase homolog	57	0.59
c.76.1.1	Alkaline phosphatases	97	0.54
c.39.1.1	5,6-dimethylbenzimidazole phosphoribosyltransferase (CobT)	58	0.56
d.2.1.3	Phage T4 lysozymes	653	0.41
b.30.5.1	Hyaluronate lyase-like, central domain	40	0.40
b.62.1.1	Cyclophilin	119	− 0.26
a.7.3.1	Succinate dehydrogenase/fumarate reductase flavoprotein C-terminal domain	41	− 0.26
b.40.2.2	Superantigen toxins, N-terminal domain	104	− 0.27
d.169.1.1	C-type lectin domain	195	− 0.30
d.87.1.1	FAD/NAD-linked reductases, dimerization (C-terminal) domain	138	− 0.36

A family was included in this ranking if its members participated in at least 40 nonredundant interactions.

Most proteins in modern cells appear to be symmetrical oligomeric complexes (Goodsell and Olson 2000). Symmetry has been hypothesized to be favorable in the evolution of oligomeric proteins, as it facilitates the stability of association and provides fewer kinetic barriers to folding, compared to asymmetric complexes (Blundell and Srinivasan 1996; Wolynes 1996). The symmetry of oligomeric proteins often implies symmetric interfaces formed by similarly located binding sites on all subunits, increasing the binding site localization. In addition, this use of similarly located binding sites can help to avoid unwanted aggregation, creating oligomers of defined copy number by imposing point group symmetry (Goodsell and Olson 2000).

Sequence conservation appears to be insufficient for an accurate prediction of the protein-binding site location because sequence conservation in binding sites is not significantly different from the overall and surface sequence conservations (Fig. 3A). Moreover, the absence of a significant correlation of the binding site localization with the average sequence and structural conservations in a family (Fig. 6B,C) suggests that localization may be a helpful measure for annotating functions of interacting domains, complementing the sequence and structural similarities of binding sites. For example, the binding interface, surface, and overall sequence residues in the NM domains of acyl-CoA dehydrogenases do not vary significantly, and the protein-binding sites are well localized on the same surface region (Fig. 5A), which is consistent with the conservation of the oligomer configuration and the biological function of the family members. In contrast, the family of C-type lectins, containing homologs with sequence and structure variabilities comparable to the NM domains, has extremely poor localization of the protein-binding sites (Fig. 5B). The poor localization is consistent with considerable variation in the functions of C-type lectins. The lack of correlation of the localization with sequence and structure divergence suggests the difficulty of ab initio prediction of binding site localization when only few or no interaction data are available.

When analyzing the relationship between the localization of a family and the number of distinct families interacting with it, one would intuitively expect a decrease in the average localization with an increase in the number of interacting families. However, a shift in the distributions of localization differences for the families interacting with one, two, three, and four, as well as five and more families was not observed (Fig. 6D). This suggests that a protein may employ the same binding site for interactions with domains from different families. In the future, we will explore this relationship in light of the evolutionary age of a domain family.

Knowledge of domain–domain interactions would provide helpful spatial restraints for modeling the arrangement of structurally defined domains into a multidomain assembly (Wodak and Mendez 2004). It is possible to build a structural model of a binary complex by comparative modeling when the structure of a complex of homologous domains is available. If the whole complex structure is not known, but structures of the homologs of the individual domains comprising the complex are available, then building a model of the complex raises the following question: Does homology of the pair of interacting domains imply similarity of interaction? This question can be addressed by studying

the sequence conservation, structure conservation, and localization of binding sites, as well as the sequence and structure conservation of the interfaces within a family of protein homologs (Camacho and Vajda 2002; Aloy et al. 2003; Lu et al. 2003). Knowledge of binding site localization can also provide guidance in the modeling of protein assembly structures by highlighting the likely locations of protein-binding sites.

The distribution of the localization differences for the domain families contributing to an assembly narrows with an increase in the number of domain families, with an average assembly localization difference remaining constant at a fairly low value of 0.04 (Fig. 4). The assembly localization index could be used to select a benchmark set of multidomain assembly structures for an assessment of protein docking methods. Assemblies with high localization indices contain domains with highly localized binding sites that may restrict the possible configurations of the interacting domains. As a result, these assembly structures are expected to be easier to model.

In summary, we determined the localization of protein-binding sites within families of homologous proteins. We found that 30% of SCOP domain families have binding sites with localization values significantly greater than expected by chance, whereas only 8% of the families have significantly low localization ($|Z| > 4.0$). The localization is expected to be a helpful criterion for investigating associations between primary, tertiary, and quaternary structures as well as their functions from an evolutionary point of view (Fig. 5). In future work, we will use localization to facilitate derivation of spatial restraints on the relative orientation of pairs of interacting domains that will be used in the modeling of quaternary structure.

Materials and methods

Our goal was to determine whether protein-binding sites on homologous proteins share similar relative positions on their surfaces. To achieve this, we first identified a nonredundant set of protein-binding sites for each SCOP family of homologous structures (Murzin et al. 1995) represented in PIBASE (Fig. 6) (Pieper et al. 2004; Davis and Sali 2005). Second, we mapped the protein-binding sites onto a family representative using structure-based alignments between each of the family members and the representative. Third, we defined a measure of the binding site localization and further characterized it by (1) the difference between the binding site localization and localization expected by chance, (2) statistical significance of the localization, and (3) the binding site coverage that determines the likelihood of a protein-binding site to overlap with the previously obtained set of binding sites.

Definitions

An "interacting domain" has at least one residue in contact with a residue of another domain. Residue r_1 of one domain is in contact with residue r_2 of another domain if it has at least one atom within 5.5 Å of an atom of r_2. A domain–domain "interaction" is defined by a triple (D_1, D_2, O), where D_1 and D_2 are the two interacting domains, and O is their relative orientation. Given an interaction (D_1, D_2, O), the protein "binding site" of D_1 is the set of all residues of D_1 that are in contact with any residue of D_2. The protein-binding site of its partner, D_2, is defined similarly. The "cumulative binding surface" of a domain is the set of all residues of all protein-binding sites from all known interactions of the domain. Given an interaction (D_1, D_2, O), the "interface" is the triple (B_1, B_2, O_B), where B_1 is the protein-binding site of D_1, B_2 is the protein-binding site of D_2, and O_B is the relative orientation of the binding sites (Fig. 7A).

Domain–domain interaction sample

Before our analysis of the localization of protein-binding sites, the protein structures needed to be divided into families of homologous domains. Several schemes define protein domain boundaries as well as classify them, based on sequence and/or structure information (Murzin et al. 1995; Orengo et al. 1997; Holm and Sander 1998; Mulder et al. 2003). For the present study, the Structural Classification of Proteins (SCOP) was chosen (Murzin et al. 1995). SCOP identifies and classifies protein structure domains based on evolutionary and structural relationships. It defines a four-level hierarchy (classes, folds, superfamilies, and families), of which we used the most detailed level, the family. Domains that belong to the same SCOP family usually share at least 30% sequence identity or the same biological function. SCOP families that are inappropriate for our analysis, including low-resolution protein structures, peptides, and designed, small, and coiled-coil proteins (classes g–k in the SCOP classification) were removed. In addition, we removed all domain structures with only backbone coordinates or with less than 30 residues.

The initial set of binary domain interfaces for each SCOP domain family was obtained from PIBASE (Pieper et al. 2004; Davis and Sali 2005), our comprehensive relational database of all structurally characterized interfaces between pairs of protein domains. The domain–domain interfaces in PIBASE are extracted from protein structures in the Protein Data Bank (PDB) (Westbrook et al. 2002) and Protein Quaternary Structure (PQS) server (Henrick and Thornton 1998) using domain definitions from the SCOP and CATH domain classification systems (Murzin et al. 1995; Orengo et al. 1997). There are 121,169 binary domain interactions of known structure between 85,366 domains from 1910 SCOP families (Fig. 8A).

To obtain the final sample of pairwise domain interactions for our analysis, we removed redundant interactions from the initial set of the 121,169 interactions. No interaction in the final nonredundant set has more than a specified level of similarity to any other interaction in the nonredundant set. This similarity filter is triggered when both the sequences of the constituent domains as well as the structures of the two compared interfaces are too similar. The domain similarity filter is triggered at 90% sequence identity. The interface similarity filter is triggered when two interfaces associated with the same PDB code (i.e., the original PDB entry and any PQS derivative structures) are clustered in PIBASE using the number of interface residues, buried surface area, and residue-type contacts (Davis and

Sali 2005); interface similarity across PDB entries is not considered. Also, we retained only those SCOP families with at least two interactions. The final nonredundant sample contains 79,354 interactions between 74,204 domains from 1847 SCOP families (Fig. 8B).

Structural superposition of protein-binding sites on a representative structure

To study the localization of binding sites for each domain family, we needed to obtain their structural superposition. First, each of the domain family members was superposed on an arbitrarily chosen family representative. The choice of the representative domain is inconsequential for a statistical study of the localization because the members of a SCOP family generally share high structural similarity. The structure-based alignment was performed using DaliLite, which uses a Monte Carlo procedure to find the best alignment by optimizing a similarity score defined in terms of equivalent intramolecular distances (Holm and Park 2000). Then, the binding sites of each family member were mapped onto the representative structure, using the DaliLite alignment. A "binding map" is a mapping of a single binding site of one of the family members onto the representative structure. A "cumulative binding map" is the union of binding maps of all members of the domain family.

Binding site localization

For each residue position r of the representative structure, the interaction density function, $\delta_I(r)$, is defined as the normalized number of members in the same family whose binding site residues align with residue position r:

$$\delta_I(r) = \frac{N_{BS}(r)}{\max\limits_r (N_{BS}(r))},$$

where $N_{BS}(r)$ is the number of binding sites at the residue position r, and $\max\limits_r(N_{BS}(r))$ is the maximal number of overlapping binding maps in this family (Fig. 7B). Next, the "localization index" of binding sites in a family, F, of interacting domains is calculated as an average interaction density for all contact residues:

$$\mu(F) = \frac{\sum\limits_{r \in CR} \delta_I(r)}{N_{CR}},$$

where CR is the set of size N_{CR} of all contact residues that are mapped onto the representative structure. When all interactions occur via the same region of the domain surface, $\mu(F)$ reaches its maximum value of 1. In contrast, when each interaction occurs via a different nonoverlapping binding site, $\mu(F)$ approaches its minimum (greater than zero).

Characterization of localization

We assess a given localization value $\mu(F)$ with the aid of a distribution of localization values expected by chance. The "localization difference" estimates the degree to which the protein-binding sites in a family are localized as opposed to dispersed, compared to what is expected by chance, and is defined as:

$$\Delta\mu(F) = \mu(F) - \bar{\mu}_R(F),$$

where $\mu(F)$ is the localization index of known protein-binding sites in a family F and $\bar{\mu}_R(F)$ is the mean of the localization indices for 50 randomly distributed sets of protein-binding sites on the surface of a family representative. The "significance of localization" determines how atypical the real distribution of the binding sites on the protein surface is, compared to the random distribution. As a measure of the statistical significance of localization, we used the Z-score defined by the sample of localization indices for each family F:

$$Z(F) = \frac{\Delta\mu(F)}{S_R(F)},$$

where $s_R(F)$ is the standard deviation of the localization index for 50 randomly distributed sets of protein-binding sites on the surface of a family representative. Each random distribution of protein-binding sites was generated using the following protocol. For each family, we first calculated three parameters: N_F, the average number of contiguous fragments per each protein-binding site; N_{BS}, the total number of binding sites for the family; and N_A, the average number of exposed atoms per each binding site. Then, we created a set of random binding sites that has the same N_F, N_{BS}, and N_A as the set of the actual binding sites in the family. More specifically, we define ($N_{BS}N_F$) contiguous fragments on the surface of the representative, each of N_A atoms, by the MODELLER's subroutine MAKE_REGION (Sali and Blundell 1993; Marti-Renom et al. 2000); this routine constructs a contiguous patch of exposed atoms of the specified size by first picking the seed atom randomly among the exposed atoms, and then iteratively adding the exposed atom that is closest to the gravity center of the currently selected patch atoms.

The "binding site coverage" determines the likelihood of a protein-binding site for a new family member to overlap with the cumulative binding map previously obtained for that family. Given a protein-binding site B and the cumulative binding map BM, the binding site coverage is:

$$\alpha_F(B) = \frac{N(B \cap BM)}{N(B) + N(BM) - N(B \cap BM)},$$

where $N(B)$ is the number of residues in binding site B, $N(BM)$ is the number of distinct residues in binding map BM, and $N(B \cap BM)$ is the number of residue positions of that overlap with the residue positions of BM. To estimate the binding site coverage for each family, we used a jackknife procedure (Quenouille 1949). Given a family of n domains, the cumulative binding map is obtained for n subsets, each containing $(n-1)$ domains, and overlapped with the protein-binding site of the remaining n-th member. The binding site coverage, $\alpha(F)$, is then estimated as the average of the binding site coverage values for each of the subsets of size $(n-1)$. Due to the high computational load of this calculation, we limited the jackknife calculation to 10 subsets for $n > 10$.

Next, we defined the localization index for an assembly of domains as the average localization difference of all domains participating in the assembly:

$$\mu(A) = \begin{cases} \dfrac{\sum_{D \in A} \Delta\mu(F_D)}{N_D}, & \mu(F_D) > 0 \\ 0, & \mu(F_D) = 0 \end{cases},$$

where N_D is the number of domains comprising assembly A, and the sum is taken over all domains.

Finally, to compare binding site localization with sequence and structure conservation, we calculated average values of the following four parameters for the pairs of all domains participating in nonredundant interactions in each SCOP family: RMSD for C_α atoms, overall sequence identity, sequence identity of surface residues, and sequence identity of binding site residues. We employed the same protocol that was used to calculate the localization, including the same set of family members, same family representative, and the same all-to-one pairwise structural alignment scheme using DaliLite. To calculate the surface identity, we first calculated the surface exposure of all residues using MODELLER (the fractional residue solvent accessibility cutoff is 20%) and then filtered out buried residues from the DaliLite alignment. To calculate the binding site identity, we used the family cumulative binding map, previously calculated for the localization analysis, to filter out from the DaliLite alignment those residues whose localization index is less than 0.1.

Acknowledgments

We thank the members of the Sali lab for their valuable comments and especially Dr. Rachel Karchin for her help in preparing this manuscript. F.P.D. acknowledges a Howard Hughes Medical Institute predoctoral fellowship. We are also grateful for the support of the NSF EIA-032645, Human Frontier Science Program, The Sandler Family Supporting Foundation, SUN, IBM, and Intel.

References

Alberts, B. and Miake-Lye, R. 1992. Unscrambling the puzzle of biological machines: The importance of the details. *Cell* **68:** 415–420.

Aloy, P., Ceulemans, H., Stark, A., and Russell, R.B. 2003. The relationship between sequence and interaction divergence in proteins. *J. Mol. Biol.* **332:** 989–998.

Bashton, M. and Chothia, C. 2002. The geometry of domain combination in proteins. *J. Mol. Biol.* **315:** 927–939.

Blundell, T.L. and Srinivasan, N. 1996. Symmetry, stability, and dynamics of multidomain and multicomponent protein systems. *Proc. Natl. Acad. Sci.* **93:** 14243–14248.

Caffrey, D.R., Somaroo, S., Hughes, J.D., Mintseris, J., and Huang, E.S. 2004. Are protein–protein interfaces more conserved in sequence than the rest of the protein surface? *Protein Sci.* **13:** 190–202.

Camacho, C.J. and Vajda, S. 2002. Protein–protein association kinetics and protein docking. *Curr. Opin. Struct. Biol.* **12:** 36–40.

Campbell, S.J. and Jackson, R.M. 2003. Diversity in the SH2 domain family phosphotyrosyl peptide binding site. *Protein Eng.* **16:** 217–227.

Chakrabarti, P. and Janin, J. 2002. Dissecting protein–protein recognition sites. *Proteins* **47:** 334–343.

Chothia, C., Gough, J., Vogel, C., and Teichmann, S.A. 2003. Evolution of the protein repertoire. *Science* **300:** 1701–1703.

Davis, F.P. and Sali, A. 2005. PIBASE: A comprehensive database of structurally defined protein interfaces. *Bioinformatics* **21:** 1901–1907.

DeLano, W.L., Ultsch, M.H., de Vos, A.M., and Wells, J.A. 2000. Convergent solutions to binding at a protein–protein interface. *Science* **287:** 1279–1283.

Dodd, R.B. and Drickamer, K. 2001. Lectin-like proteins in model organisms: Implications for evolution of carbohydrate-binding activity. *Glycobiology* **11:** 71R–79R.

Doolittle, R.F. and Bork, P. 1993. Evolutionarily mobile modules in proteins. *Sci. Am.* **269:** 50–56.

Drickamer, K. 1999. C-type lectin-like domains. *Curr. Opin. Struct. Biol.* **9:** 585–590.

Edwards, A.M., Kus, B., Jansen, R., Greenbaum, D., Greenblatt, J., and Gerstein, M. 2002. Bridging structural biology and genomics: Assessing protein interaction data with known complexes. *Trends Genet.* **18:** 529–536.

Goodsell, D.S. and Olson, A.J. 2000. Structural symmetry and protein function. *Annu. Rev. Biophys. Biomol. Struct.* **29:** 105–153.

Grishin, N.V. and Phillips, M.A. 1994. The subunit interfaces of oligomeric enzymes are conserved to a similar extent to the overall protein sequences. *Protein Sci.* **3:** 2455–2458.

Henrick, K. and Thornton, J.M. 1998. PQS: A protein quaternary structure file server. *Trends Biochem. Sci.* **23:** 358–361.

Holm, L. and Park, J. 2000. DaliLite workbench for protein structure comparison. *Bioinformatics* **16:** 566–567.

Holm, L. and Sander, C. 1998. Touring protein fold space with Dali/FSSP. *Nucleic Acids Res.* **26:** 316–319.

Janin, J., Henrick, K., Moult, J., Eyck, L.T., Sternberg, M.J., Vajda, S., Vakser, I., and Wodak, S.J. 2003. CAPRI: A Critical Assessment of PRedicted Interactions. *Proteins* **52:** 2–9.

Jones, S. and Thornton, J.M. 1996. Principles of protein–protein interactions. *Proc. Natl. Acad. Sci.* **93:** 13–20.

Jones, S., Marin, A., and Thornton, J.M. 2000. Protein domain interfaces: Characterization and comparison with oligomeric protein interfaces. *Protein Eng.* **13:** 77–82.

Kim, J.J. and Miura, R. 2004. Acyl-CoA dehydrogenases and acyl-CoA oxidases. Structural basis for mechanistic similarities and differences. *Eur. J. BioChem.* **271:** 483–493.

Kolkman, J.A. and Stemmer, W.P. 2001. Directed evolution of proteins by exon shuffling. *Nat. Biotechnol.* **19:** 423–428.

Larsen, T.A., Olson, A.J., and Goodsell, D.S. 1998. Morphology of protein–protein interfaces. *Structure* **6:** 421–427.

Lawrence, M.C. and Colman, P.M. 1993. Shape complementarity at protein/protein interfaces. *J. Mol. Biol.* **234:** 946–950.

Littler, S.J. and Hubbard, S.J. 2005. Conservation of orientation and sequence in protein domain–domain interactions. *J. Mol. Biol.* **345:** 1265–1279.

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

Lu, L., Arakaki, A.K., Lu, H., and Skolnick, J. 2003. Multimeric threading-based prediction of protein–protein interactions on a genomic scale: Application to the *Saccharomyces cerevisiae* proteome. *Genome Res.* **13:** 1146–1154.

Marti-Renom, M.A., Stuart, A.C., Fiser, A., Sanchez, R., Melo, F., and Sali, A. 2000. Comparative protein structure modeling of genes and genomes. *Annu. Rev. Biophys. Biomol. Struct.* **29:** 291–325.

Mulder, N.J., Apweiler, R., Attwood, T.K., Bairoch, A., Barrell, D., Bateman, A., Binns, D., Biswas, M., Bradley, P., Bork, P., et al. 2003. The InterPro Database, 2003 brings increased coverage and new features. *Nucleic Acids Res.* **31:** 315–318.

Murzin, A.G., Brenner, S.E., Hubbard, T., and Chothia, C. 1995. SCOP: A structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247:** 536–540.

Ng, S.K., Zhang, Z., Tan, S.H., and Lin, K. 2003. InterDom: A database of putative interacting protein domains for validating predicted protein interactions and complexes. *Nucleic Acids Res.* **31:** 251–254.

Nicholas, H.R. and Hodgkin, J. 2004. Responses to infection and possible recognition strategies in the innate immune system of *Caenorhabditis elegans*. *Mol. Immunol* **41:** 479–493.

Ofran, Y. and Rost, B. 2003. Analysing six types of protein–protein interfaces. *J. Mol. Biol.* **325:** 377–387.

Orengo, C.A., Michie, A.D., Jones, S., Jones, D.T., Swindells, M.B., and Thornton, J.M. 1997. CATH—A hierarchic classification of protein domain structures. *Structure* **5:** 1093–1108.

Paavilainen, V.O., Merckel, M.C., Falck, S., Ojala, P.J., Pohl, E., Wilmanns, M., and Lappalainen, P. 2002. Structural conservation between the actin monomer-binding sites of twinfilin and actin-depolymerizing factor (ADF)/cofilin. *J. Biol. Chem.* **277:** 43089–43095.

Park, J., Lappe, M., and Teichmann, S.A. 2001. Mapping protein family interactions: Intramolecular and intermolecular protein family interaction repertoires in the PDB and yeast. *J. Mol. Biol.* **307:** 929–938.

Pawson, T. and Nash, P. 2003. Assembly of cell regulatory systems through protein interaction domains. *Science* **300:** 445–452.

Pieper, U., Eswar, N., Braberg, H., Madhusudhan, M.S., Davis, F.P., Stuart, A.C., Mirkovic, N., Rossi, A., Marti-Renom, M.A., Fiser, A., et al. 2004. MODBASE, a database of annotated comparative protein structure models, and associated resources. *Nucleic Acids Res*. **32:** D217–D222.

Quenouille, M. 1949. Approximate tests of correlation in time series. *J. R. Stat. Soc. B* **11:** 18–84.

Rajamani, D., Thiel, S., Vajda, S., and Camacho, C.J. 2004. Anchor residues in protein–protein interactions. *Proc. Natl. Acad. Sci.* **101:** 11287–11292.

Sali, A. and Blundell, T.L. 1993. Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* **234:** 779–815.

Sali, A., Glaeser, R., Earnest, T., and Baumeister, W. 2003. From words to literature in structural proteomics. *Nature* **422:** 216–225.

Sheinerman, F.B. and Honig, B. 2002. On the role of electrostatic interactions in the design of protein–protein interfaces. *J. Mol. Biol.* **318:** 161–177.

Sheinerman, F.B., Norel, R., and Honig, B. 2000. Electrostatic aspects of protein–protein interactions. *Curr. Opin. Struct. Biol.* **10:** 153–159.

Spahn, C.M., Beckmann, R., Eswar, N., Penczek, P.A., Sali, A., Blobel, G., and Frank, J. 2001. Structure of the 80S ribosome from *Saccharomyces cerevisiae*—tRNA-ribosome and subunit–subunit interactions. *Cell* **107:** 373–386.

Teichmann, S.A. 2002. The constraints protein–protein interactions place on sequence divergence. *J. Mol. Biol.* **324:** 399–407.

Thorpe, C. and Kim, J.J. 1995. Structure and mechanism of action of the acyl-CoA dehydrogenases. *FASEB J.* **9:** 718–725.

Tsai, C.J., Lin, S.L., Wolfson, H.J., and Nussinov, R. 1996. Protein–protein interfaces: Architectures and interactions in protein–protein interfaces and in protein cores. Their similarities and differences. *Crit. Rev. Biochem. Mol. Biol.* **31:** 127–152.

———. 1997. Studies of protein–protein interfaces: A statistical analysis of the hydrophobic effect. *Protein Sci.* **6:** 53–64.

Valdar, W.S. and Thornton, J.M. 2001. Protein–protein interfaces: Analysis of amino acid conservation in homodimers. *Proteins* **42:** 108–124.

Vasta, G.R., Ahmed, H., and Odom, E.W. 2004. Structural and functional diversity of lectin repertoires in invertebrates, protochordates and ectothermic vertebrates. *Curr. Opin. Struct. Biol.* **14:** 617–630.

Westbrook, J., Feng, Z., Jain, S., Bhat, T.N., Thanki, N., Ravichandran, V., Gilliland, G.L., Bluhm, W., Weissig, H., Greer, D.S., et al. 2002. The Protein Data Bank: Unifying the archive. *Nucleic Acids Res*. **30:** 245–248.

Wetlaufer, D.B. 1973. Nucleation, rapid folding, and globular intrachain regions in proteins. *Proc. Natl. Acad. Sci.* **70:** 697–701.

Wodak, S.J. and Mendez, R. 2004. Prediction of protein–protein interactions: The CAPRI experiment, its evaluation and implications. *Curr. Opin. Struct. Biol.* **14:** 242–249.

Wolynes, P.G. 1996. Symmetry and the energy landscapes of biomolecules. *Proc. Natl. Acad. Sci.* **93:** 14249–14255.

Zhang, W., Chipman, P.R., Corver, J., Johnson, P.R., Zhang, Y., Mukhopadhyay, S., Baker, T.S., Strauss, J.H., Rossmann, M.G., and Kuhn, R.J. 2003. Visualization of membrane protein domains by cryo-electron microscopy of dengue virus. *Nat. Struct. Biol.* **10:** 907–912.